Bayesian Methods in Finance

Eric Jacquier and Nicholas Polson[*]

Forthcoming in "The Handbook of Bayesian Econometrics"
John Geweke, Gary Koop, Herman Van Dijk editors

September 2010

Abstract

This chapter surveys Bayesian Econometric methods in finance. Bayesian methods provide a natural framework for addressing central issues in finance. In particular, they allow investors to assess return predictability, estimation and model risk, formulated predictive densities for variances, covariances and betas. This can be done through decision theoretic problems, such as option pricing or optimal portfolio allocation. Bayesian predictive distributions are straightforward to calculate and summarize the investor's future views for return distribution and expected utility computation. Nonlinear functionals, such as market efficiency measures and Sharpe ratios, are easily dealt with from a Bayesian perspective. A central theme in this chapter is the use of simulation-based estimation and prediction via Markov Chain Monte Carlo (MCMC) and particle filtering (PF) algorithms. We provide detailed applications of these methods to the central issues in finance.

[*] Nick Polson is at the Booth School of Business, University of Chicago. Eric Jacquier is the corresponding author at HEC Montreal and MIT Sloan School of Management, email: jacquier@mit.edu. Eric Jacquier acknowledges financial support from the HEC professorship in derivatives products. The chapter has benefited from comments from the handbook editors, discussions with Shirley Miller and Maïa Cullen, and editing by Linda Tracey of Inkslingers. I am especially grateful to John Geweke and Herman Van Dijk for their innumerable insights and support. There is a rich body of Bayesian work in finance, and, as this chapter has no encyclopedic ambition, a number of excellent papers are not discussed, with no implication on their quality; apologies where they are due. A general acknowledgement is owed to Arnold Zellner whose foundational work and mentoring are at the source of most everything discussed in this chapter.

Contents

List of Tables

List of Figures

1 Introduction

This chapter discusses the use of Bayesian methods in finance. A first fundamental aspect of modern finance is that it asks questions of predictability. Discussions on the efficiency of markets center on the degree of predictability, if any, of financial series. This maps directly onto the Bayesian use of the predictive density. Further, as one must compare multiple competing models of predictability, the Bayesian perspective on testing, namely odds ratios for model comparison or averaging, is well suited. Second, the quantities of interest in many finance applications, e.g., the period of a cycle, hedge ratios, option prices, correlations between portfolios, and Sharpe ratios, are non-linear functions of the base parameters, used to write the likelihood function. For example, the period of a cycle in autocorrelation is a function of the AR parameters. Bayesian methods, especially when posteriors are simulated, easily deliver the exact posterior density of such non-linear function of the parameters. Third, in recent years, models of time-varying volatility have become increasingly complex, and Bayesian methods in conjunction with Markov Chain Monte Carlo techniques have produced highly effective estimation and prediction algorithms for these models. Finally, rationally-based financial decision-making contains a vast normative aspect. Following Bayes' rule, an agent updates her beliefs on the predictive distribution of asset returns, possibly via the optimal averaging of competing models, none of which she holds to be true. She then devises a portfolio to maximize the expected utility of her predictive wealth.

With this in mind, section 2 addresses the classic portfolio optimization introduced by Markowitz (1952) and Merton (1969). We first show how estimation error is addressed in the Bayesian framework. Even within the restrictive framework of one single model, e.g., one-period as in Markowitz or independently identically distributed log-normal as in Merton, expected-utility is now random since it is a function of parameters themselves random for the Bayesian.[1] Decision theory shows that conditioning on point estimates, no matter how *good*, does not yield an optimal strategy. One needs to integrate out the parameters using their posterior density, which yields the predictive density of future returns. This point is made early, for example, in Zellner and Chetty (1962), and Brown (1978). While the Bayesian methodology is optimal given the information available to the agent, diffuse priors alone offer limited improvement over the classical approach. Therefore, section 2.2 discusses a remedy, empirical Bayes. This refers to the use proper priors calibrated from the sample

[1] This is even before accounting for model uncertainty. Especially in Bayesian econometrics, there is no sense in which a given model is seen as true. Models are, hopefully, convenient windows through which to view the data, and make needed inference, prediction, or decision. See Poirier (2010) in this handbook.

itself, rather than subjective priors reflecting actual views. We discuss the relation between these priors and the James and Stein (1961) shrinkage estimators, and how to calibrate both the mean vector and the covariance matrix as in Frost and Savarino (1986). Perhaps because the numerical effect did not appear too large at the time, for the one-period framework, the ideas in Brown (1978) and others, did not get immediate widespread use. We show in section 2.3 that the impact of uncertainty in the mean increases dramatically with the investment horizon, we discuss Jacquier (2008) and Barberis (2000). Section 2.4 discusses how the Bayesian econometrician can use priors to incorporate beliefs in asset pricing models, as in Black and Litterman (1991) and Pastor (2000). We complete the section with a discussion of further issues.

Section 3 discusses the predictability of the mean of asset returns, central to finance, as it relates to the efficiency of financial markets. Predictability can be analyzed in a rather classic statistical time series approach or from the viewpoint of its economic relevance. We start with the time series approach. Here the benefits of Bayesian analysis reside in the use of posterior odds, that allow the ranking of multiple models. The initial literature on predictability typically analyzed the ability of one or more variables to predict stock returns with classical statistical tools such as t-statistics, R-squares, or root mean-squared errors. The standard classical framework with one null hypothesis nested in an alternative does not allow the ranking of multiple, possibly non-nested models of predictability. In contrast, model odds ratios are perfectly fitted for that task, and we discuss Avramov (2002) and Cremers (2002) who contribute to the understanding of predictability by using Bayesian model comparison and, more importantly, averaging in this context. Predictability is often assessed by a measure that is a non-linear function of the basic parameters used to write the likelihood function. Classical estimation techniques are ill-equipped for this situation because of the approximations inherent in the asymptotic theory. We show such an example with the analysis of cyclicality in stock returns, and with Lamoureux and Zhou (1996)'s Bayesian approach n the long-horizon return predictability,

Section 3.2 discusses the economic relevance of predictability, namely its impact on optimal allocation. A classic first paper that initiated this way of thinking is Kandel and Stambaugh (1996). They analyze predictability through the classic regression of stock returns on the dividend yield, specifically its impact on asset allocation, when parameter uncertainty is properly accounted for in the predictive density. Recognizing the stochastic nature of the predictor, they formulate the predictive regression as a bivariate VAR. Stambaugh (1999) provides a thorough analysis of this predictive regression. We conclude with a discussion of Barberis (2001) who pays special attention to the multi-period case.

Section 4 discusses some major contributions of Bayesian econometrics to the literature on empirical asset pricing. First we show how McCulloch and Rossi (1990, 1991) implement a Bayesian test of Ross's (1976) arbitrage pricing theory (APT) from statistical and economic perspectives. Second, a major issue in tests of latent factor models is the needed preliminary estimation of the factor scores and loadings. In contrast, for the CAPM and the index models, one usually only worries about the estimations of the loadings (betas). Classical approaches typically assume asymptotics in the time series or in the cross-section of assets. Geweke and Zhou (1995) show that neither assumption is necessary. They jointly estimate the scores and the factors with a simple MCMC algorithm. Third, within the CAPM world, it has long been established that tests of asset pricing models were akin to testing whether the index at hand was ex-post efficient. This has led to a rich Bayesian literature tackling tests of asset pricing models this way, which we discuss. Another perspective on the efficiency of markets is whether managed portfolios can beat passive indexes. We discuss Baks et al. (2001) and others who study mutual fund performance from a Bayesian perspective. We contrast with the approach in Jones and Shanken (2005) who study the funds jointly.

Section 5 discusses volatility and covariance modeling. It starts with a review of Bayesian GARCH, and continues with Stochastic volatility (SV) modeling. MCMC algorithms have resulted in a tremendous growth in the use of SV models in financial econometrics, because they make possible the estimation of complex non-linear latent variable models for which the Kalman filter is not optimal. For example, the MCMC algorithms in Jacquier, Polson and Rossi (1994, 2004) obtain the posterior densities of both the latent variables, here the volatilities, and the parameters. While MCMC methods can be applied to the maximization of the likelihood function, the Bayesian approach does not require this complicated step since it only needs to draw from the posterior distribution. We review a Bayesian MCMC algorithm for classic SV models with both leverage effects and fat tails. We show first how to design and diagnose the algorithm, then how to conduct model comparison by looking at the predictions of the model and by computing odds ratios. We show simple ways to obtain the odds ratios that only rely on the readily available posterior draws, thus bypassing the specific integration needed for the computation of the marginal likelihood. We then mention extensions to the model. We complete this section by discussing Bayesian strategies for estimating covariance matrices. We discuss two different approaches to the matrix, first where the individual volatilities and the correlation matrix are modeled separately, second, where factor model allow to constraint the covariance matrix. We discuss Bayesian estimation of the factor loadings , the β's. Cosemans et al. (2009) model jointly the cross-sectional and time-series dynamics of betas and show that it results in improved portfolio performance.

Jostova and Philipov (2004) implement a MCMC algorithm for latent betas.

Section 6 reviews the area of empirical option pricing. We first discuss simulation based methods to compute the option price on the basis of draws of the uncertain volatility. Early method only reflect the uncertainty in volatility. We then discuss the explicit addition of a pricing error to the model entertained, so that the likelihood incorporates model error. We discuss Jacquier and Jarrow (2000) who model the Black-Scholes pricing error as a function of observable variables. While their likelihood exclusively follows from the pricing error, they incorporate the historical returns via a prior on σ. In the same spirit, Eraker (2004) implements a vastly more general model where volatility is stochastic and can jump. Jones (2003) links implied and historical volatilities by a linear relationship, allowing for errors. This in fact incorporates both the historical and risk-neutral process in the likelihood function

Section 7 discusses a promising recent development in finance, filtering with parameter learning. Filtering techniques have gained recognition over the past years, (see Kohn et al. in this handbook for classic filtering algorithms). MCMC methods allow inference for complex models with parameters θ and latent variables h by breaking their joint posterior distribution into conditionals, $(h|\theta)$ and $(\theta|h)$. This produces the joint distribution of the smoothed estimates of the latent variables, for example $(h_t|y_1, \ldots y_T)$. However, one often wants the distribution of the filtered values of the latent variable, $(h_t|y_1, \ldots, y_t) \;\; \forall t \in [1, T]$. A feasible solution is to repeat a MCMC smoother for all desired subsamples $[1, t]$. It is, however, not attractive computationally. Particle filters deliver the desired filtered densities of the latent variables, but until recently, conditioned on a parameter value $(h_t|y_1, \ldots, y_t, \theta)$, which was not very interesting. In contrast, the more recent algorithms which we discuss allow for parameter learning. That is, at any time $t \in [1, T]$, the algorithm produces the density of both latent variables and parameters using only the data until time t, $(h_t|y_1, \ldots, y_t, \theta)$ and $(\theta_t|y_1, \ldots, y_t)$. We discuss implementations from Jacquier and Miller (2010).

2 Optimal Portfolio Design

2.1 The basic optimal portfolio setup with parameter uncertainty

Before introducing parameter uncertainty, we briefly review some key results of one-period optimal portfolio theory. See Markowitz (1952), Merton (1972), Roll (1977), Brandt (2009), or classic graduate finance textbooks for derivations. Markowitz's (1952) one-period framework assumes N jointly normal asset returns $\mathbf{R}$ with known mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. A portfolio with weights w in the N assets, has mean $\mu = w'\boldsymbol{\mu}$ and variance $\sigma^2 = w'\Sigma w$. This yields the well-known efficient frontier in the mean versus variance space, customarily plotted versus standard deviation. In brief, with short sales allowed, the locus of expected returns versus standard deviation of optimal portfolios that minimize variance subject to a desired expected return, is a hyperbola. Its vertex is the global minimum variance portfolio (MVP) whose vector of weights is $\Sigma^{-1}i/i'\Sigma^{-1}i$, where i is a vector of ones. Note that the weights sum to one due to the denominator. Without a risk-free asset, investors select a portfolio on this frontier, so as to maximize their expected utility, or certainty equivalent (CE) which represents the trade-off between mean and variance. For investors with constant relative risk aversion, the CE is $\mu - \frac{\gamma}{2}\sigma^2$. The weights maximizing this CE are equal to $\frac{1}{\gamma}\Sigma^{-1}(\boldsymbol{\mu} - \mu_0 i)$, where μ_0 is also a function of Σ^{-1} and $\boldsymbol{\mu}$.

The combinations of a risk-free asset with a risky asset on this frontier occur on a straight line, known as the capital allocation line (CAL), in this mean versus standard deviation space. The slope of this line is the Sharpe ratio, the ratio of the expected return in excess of the risk-free rate over the standard deviation, which investors seek to maximize. The resulting tangency portfolio of the N risky assets is located where the CAL is tangent to the frontier of risky assets. Its vector of weights is:

$$\frac{\Sigma^{-1}(\boldsymbol{\mu} - R_f i)}{i'\Sigma^{-1}(\boldsymbol{\mu} - R_f i)}. \tag{1}$$

Investors allocate their wealth between this tangency portfolio and the risk free rate according to their risk aversion. The optimal allocation, the weight in the risky portfolio which maximizes this certainty equivalent is:

$$w^\star = \frac{\mu - R_f}{\gamma\sigma^2}, \tag{2}$$

where μ, σ are the mean and standard deviation of the tangency portfolio found in (1), and $1 - w^\star$ is allocated to the risk-free rate.

Parameters are unknown in actual implementations. Early practice was to substitute point estimates of the parameters $\boldsymbol{\mu}$ and Σ into the standard optimal portfolio formulas, or into an optimizer. However, decision theory shows that conditioning the problem on point estimates, as good as they may be, of model parameters leads to suboptimal portfolios. As pointed out by Zellner and Chetty (1965), accounting properly for estimation error requires the computation of the predictive density. The predictive density, an essentially Bayesian concept, is the joint density of future data, conditional only on the model used and the data already observed $\mathbf{R}$. In our case, the joint predictive density of the N asset returns for time T+1 is:

$$p(\mathbf{R}_{\mathrm{T+1}}|\mathbf{R}) = \int \mathrm{p}(\mathbf{R}_{\mathrm{T+1}}|\mathbf{R}, \boldsymbol{\mu}, \Sigma)\ \mathrm{p}(\boldsymbol{\mu}, \Sigma|\mathrm{R})\ \mathrm{d}\boldsymbol{\mu}\ \mathrm{d}\Sigma. \tag{3}$$

Note how the posterior density of the parameters is used to integrate them out of the density of the future returns $p(R_{T+1}|\mathbf{R}, \boldsymbol{\mu}, \Sigma)$. Similarly, the predictive density of the return on a portfolio with weights w, follows by integrating its mean $\mu = w'\boldsymbol{\mu}$ and variance $\sigma^2 = w'\Sigma w$, out of the conditional density of its return.

Klein and Bawa (1976) demonstrate that computing, and then optimizing, expected utility around the predictive density is the optimal strategy. The intuition is clear in the Bayesian framework: the Sharpe ratio and the expected utility, (or CE), $\mu - \frac{\gamma}{2}\sigma^2$ are random due to parameter uncertainty. How can one maximize a random function and hope to find a fixed answer? Also, going forward and substituting point estimates of μ, Σ in the CE or Sharpe ratio clearly omits an uncertainty that should be accounted for, especially by risk averse investors. In this spirit, Brown (1976, 1978) and Bawa et al. (1979) incorporate parameter uncertainty into the optimal portfolio problem. They (mostly) use improper priors $p(\boldsymbol{\mu}, \Sigma) \propto |\Sigma|^{-(N+1)/2}$ to compute the predictive density of the parameters, and maximize expected utility for that predictive density.

The multivariate predictive density of returns is shown to be a student-t with mean $\hat{\boldsymbol{\mu}}$, degrees of freedom $T - N$, and covariance matrix $k\widehat{\Sigma}$, where the variance inflation factor k is $(1 + \frac{1}{T})\frac{T+1}{T-N-2}$. This modifies optimal allocation, especially when N is sizable relative to T. Relative to the portfolio based on point estimates, Bayesian optimal portfolios take smaller positions on the assets with higher risk, for example those with high $\widehat{\mu}$. If Σ is known, k reduces to $1 + \frac{1}{T}$, and the correction is far less dramatic. Consider, for example, the risky versus risk-free asset allocation. With an improper prior, the posterior density of μ, is $N(\hat{\mu}, \frac{\sigma^2}{T})$, where $\hat{\mu}$ is the sample mean. The predictive density of the future return is $N(\hat{\mu}, \sigma^2(1 + \frac{1}{T}))$. Intuitively, the future variance faced by the investor is the sum of the return's variance given the mean and the variance of the uncertain mean. Computing the Merton allocation with respect to this predictive density of returns lowers the allocation on the tangency portfolio in (2) by the factor $1 + \frac{1}{T}$. However, it does not affect the weights of the risky assets in the tangency portfolios in (1).

Initially, these corrections did not appear important for the one-period model, when N was deemed small enough relative to T. The practice of substituting point estimates of $\boldsymbol{\mu}$ and Σ in the theoretical solutions remained common with both practitioners and academic researchers. However, practitioners eventually recognized that this *plug-in* approach was sensitive to estimation error (see for example Michaud (1989)). Consider an investor who minimizes variance subject to a given desired mean return, and uses a point estimate of the mean vector. The highest individual point estimates are such because the corresponding mean may be high, and the sample of data used may lead to a positive estimation error. The next sample, corresponding to the investment period, will likely lead to lower point estimates for these means. This investor will then be over-invested in these estimation errors.

Jobson and Korkie (1980), and a following literature, discuss the sampling variability of the optimal portfolio weights due to the sampling variability of the vector of means and covariance matrix. The major problem with this approach is that, for an investor at decision time, the weights are not a random variable, they are a decision variable. Therefore, the statistical exercise of characterizing their sampling variability offers little insight to the investor, who needs to optimize the predictive utility on the basis of the sample at hand, as decision theory stipulates. A second problem is that the relationship between the portfolio weights and the mean and variance is non linear. It can be shown that large variations in weights can result in small variations in portfolio mean and variance. Therefore, the frequentist degree of uncertainty in the weights is a poor indicator of the uncertainty in the future returns of optimized portfolios.

Another approach proposed computes the *"resampled"* frontier. One simulates returns data from the sampling distribution of the estimators of $\boldsymbol{\mu}$ and Σ, and then computes a frontier for each simulated data set. The resampled frontier is an average of these simulated frontiers. This frequentist simulation is very different from the Bayesian decision theory approach, that computes one frontier on the basis of the predictive density of returns obtained from the data at hand. The two methods are qualitatively similar in that both penalize the mean estimates of the more variable assets. The extreme mean estimates in the actual sample are averaged out in the simulated data sets, leading to smaller weights on these assets in the resampled frontier. Bayesian optimization based on the predictive density leads to smaller weights on these same assets, to the extent that they have a large variance. Harvey et al. (2008) compare the two approaches and conclude that the Bayesian method dominates.

Stambaugh (1997) generalizes the problem to the case where a subset of the assets has a shorter history, as when new assets are introduced. Consider N_1 assets with returns $\mathbf{R}_1$ on $[1, T]$, and N_2 assets with returns $\mathbf{R}_2$ on [s,T]. Earlier methods either used only the truncated common sample [s,T], foregoing the information in [1,s-1], or estimated separately subsets of μ, Σ, using the relevant subsamples. In the second case, μ_1, Σ_{11} were based upon [1,T], while $\mu_2, \Sigma_{22}, \Sigma_{1,2}$, were based on [s,T]. The second approach can produce singular estimates of Σ

and still does not use all the information in the likelihood. Stambaugh rewrites the joint density $p(\mathbf{R}_1, \mathbf{R}_2|\boldsymbol{\mu}, \Sigma)$ as $p(\mathbf{R}_2|\mathbf{R}_1)p(\mathbf{R}_1)$, and parameterizes it in terms of the regression of R_2 on R_1.

Using this full-sample likelihood function has two effects; first, μ_1 and Σ_{11} benefit from the added precision of the full sample, second, and less obvious, $\mu_2, \Sigma_{12}, \Sigma_{22}$, also benefit from the longer sample because the covariance between R_1 and R_2 is fully exploited. For example, with diffuse priors, the posterior mean of μ_2 is not just the unconditional $\widehat{\mu}_2$, it also uses information from the discrepancy in $\widehat{\mu}_1$ between [1,s-1] and [s,T]. Similar results follow for Σ_{12} and Σ_{22}. The key intuition is that inference on the shorter assets differs from the truncated method if the two samples [1,s-1] and [s,T] produce different inference on the longer assets. Stambaugh then derives the predictive density of returns and implements the method on a portfolio of longer developed market indices and a shorter index of emerging markets. For these data, the two methods produce drastically different tangency portfolios and optimal allocations in the tangency portfolio. Posterior analysis shows that the two methods produce very different inference on the mean of the emerging market, but basically identical covariance matrices.

2.2 Shrinkage and empirical Bayes for the portfolio problem

The mean vector of financial assets is particularly difficult to estimate precisely, even with calendar spans of data as long as many decades. This is due to the magnitude of the standard deviation relative to the mean for typical financial returns. Further, due to the low autocorrelation of financial returns, sampling them at a higher frequency does not reduce the uncertainty in the mean, because mean and variance time-aggregate at the same rate for i.i.d returns. For example, the posterior distribution of the annualized mean obtained from $252T$ daily returns is not tighter than the posterior distribution of the annual mean obtained from T annual returns. Econometricians and investors have to live with this fact. In contrast, one can reduce the uncertainty on variance by increasing the sampling frequency. This is why, in the Merton world with constant variance and continuous time trading, the agent can be assumed to know the variance. This makes uncertainty in the mean the first order effect to address in portfolio optimization.

The optimization process tends to put higher (lower) weights on the assets with higher (lower) mean. Due to parameter uncertainty, the extreme point estimates in the mean vector for the estimation period, are likely to be closer to the central estimates next period, the investment period. An optimizer which merely uses point estimates takes positions too extreme, and experience poor performance during the investment period. The phenomenon is more serious for the more risk tolerant investors who load up more on the extreme mean returns. Jobson and Korkie (1981) use 25 years of monthly returns and show that the realized Sharpe ratios of a portfolio that optimizes on the basis of point estimates of $\boldsymbol{\mu}, \Sigma$ is 0.08, versus 0.34 for a portfolio using the true quantities. The substitution approach is clearly costly. Frost and Savarino (1986) show that Bayesian optimization based on diffuse priors indeed improve over this classical substitution approach, but the amount of uncertainty in the mean is still too high to make the Markowitz framework appealing over passive strategies such as value or equal weighting. For example, the estimates and resulting portfolio weights still vary too much from period to period. We now discuss how portfolio performance can be improved with informative priors.

James and Stein (1961) prove the inadmissibility of the maximum likelihood estimator (MLE) of a multivariate mean, of dimension larger than 2, by showing that it is dominated by a shrinkage, therefore biased estimator. Their shrinkage estimator is:

$$\hat{\mu}_{JS} = (1-w)\hat{\mu} + w\mu_0 i, \tag{4}$$

where w is a data-based scalar weight, μ_0 is the scalar central value towards which shrinkage occurs, and i is a vector of ones. w, formula omitted here, is inversely proportional to a quadratic form in $(\hat{\mu} - \mu_0 i)$ and Σ^{-1}. This estimator shrinks the original mean estimate to a common value μ_0.

Shrinkage is a natural approach to reduce the effect of parameter uncertainty in the mean. They counter the tendency of the optimizer to load on extreme value by bringing them closer to the center, replacing the MLE of the mean vector with a linear combination of that estimate and a chosen central mean. This reduces the cross-sectional dispersion of the vector of means. This effect is achieved in the Bayesian framework with the use of a prior. With normal conjugate priors, the posterior mean is a linear combination of the prior mean and the MLE. The weights are the respective precisions of these two components. Therefore, a given shrinkage estimation is consistent with some Bayesian prior. Note that individual prior means need not be equal, and the shrinkage does not have to occur toward a vector of equal values such as $\mu_0 i$ in (4). However classic shrinkage corresponds to *"empirical Bayes"*, where the prior parameters are based on the data, a convenience to reduce parameter uncertainty, rather than a representation of the econometrician's subjective prior views.

An important question is whether there is a *better* central value toward which to shrink the initial estimate. Initial work proposed shrinking toward the grand mean. Jorion (1986) makes the important point that, under basic assumptions, μ_0 should be the mean of the global minimum variance portfolio (MVP). The intuition for this is clear. First, the MVP is robust to uncertainty in the mean because we do not require the mean vector to identify it. Second, the mean of the MVP is subject to the least uncertainty, since it has, by definition, the smallest variance of all portfolios. Shrinking toward a mean with the smallest possible uncertainty is precisely the desired objective. Jorion writes an empirical Bayes estimator where the prior on μ is $N(\eta, \lambda\Sigma)$. The hyperparameter λ calibrates the tightness of the prior, η is the prior mean. Jorion puts a diffuse prior on η. Integrating out μ, η, one finds that the mean of the predictive density of the returns r is:

$$E(r|\mathbf{R}) = (1 - \mathrm{w})\hat{\mu} + \mathrm{w}\mu_0 \mathrm{i},$$

where μ_0 is the mean of the MVP and $w = \lambda/(\lambda + T)$, showing that λ has the intuition of a notional sample size for the prior. As λ increases relative to T, the cross-sectional dispersion of $E(r|\mathbf{R})$ vanishes and the means shrink toward the global minimum variance portfolio. Jorion (1985) implements optimal portfolio selection on international assets with this empirical Bayes method. He shows that it dominates those approaches based upon the basic sample estimates. The portfolio weights are more stable through time as they do not take large bets on point estimates.

Dumas and Jacquillat (1990) implement a similar empirical approach on currency portfolios. They use logarithmic utility and lognormal returns while Jorion was assuming normal returns. They argue that shrinking to the MVP introduces a country-specific behavior undesirable for them because they want to model the asset allocation of a universal investor. This country-specific behavior arises if one admits deviations from the purchasing power parity. Instead, they engineer an empirical Bayes prior which produces a shrinkage toward an equal weighted portfolio of currencies.

Frost and Savarino (1986) assume normal returns and exponential utility, and also shrink the covariance matrix by empirical Bayes. They formulate a conjugate Normal Inverse-Wishart prior for (μ, Σ), centered on equal means, variances, and covariances. The prior on the mean is $p(\mu) \sim N(\mu_0 i, \frac{1}{\tau}\Sigma)$, where μ_0 is the MLE of the grand mean assuming equal means for the N returns, and τ is a notional sample size representing the strength of prior belief. The prior on Σ is an inverse Wishart which prior mean is a covariance matrix Ω with equal variances δ and correlations ρ. A parameter ν with the interpretation of a notional sample size models the prior strength of belief. Frost and Savarino estimate all prior parameters, including ν and τ by maximum likelihood, a strong use of empirical Bayes.

To do this, they write the likelihood of the data, modeled by this prior, and maximize it.

The posterior mean vector is the weighted average seen above. The covariance matrix of the predictive density of returns is a weighted average of three quantities, the prior mean Ω, the sample covariance $\widehat{\Sigma}$, and an outer-product of the vector of discrepancies between prior and sample means $(\hat{\mu} - \mu_0)$. This latter term is typical of posterior covariance matrices when proper priors are used on means or regression coefficients. In term of optimization, this amounts to a shrinkage toward the equal weighted portfolio since no asset has preferred characteristics in the prior. With an investment universe of 25 randomly selected securities, they compare the realized returns of optimized portfolios based on the classical point estimates, the Bayesian predictive densities with diffuse priors, and their empirical Bayes priors. Their results show that while the use of the predictive density with diffuse priors does improve on the classical method, the empirical Bayes estimator leads to a vast additional improvement over the diffuse prior.

2.3 Parameter uncertainty and long-run asset allocation

We will now see that parameter uncertainty compounds over time, becoming very important in the long run. Namely, we discuss inference on the compound excess return of the market index over the risk free rate, and its impact on long-run asset allocation.

Merton (1969) derives the optimal asset allocation between one risky and one riskless asset in continuous time, generalizing the one-period result in (2). Consider an i.i.d. log-normal risky asset, where $\log(1 + R_t) \sim N(\mu, \sigma^2)$, its compound return over H periods is:

$$\exp\left(\mu H + \sigma \sum_{i=1}^{H} \epsilon_{t+i}\right), \quad \epsilon_t \sim N(0, 1),$$

This H-period compound return is lognormal $(\mu H, \sigma^2 H)$, its expectation is therefore:

$$\exp\left(H\mu + \frac{1}{2} H\sigma^2\right) \tag{5}$$

Consider a risk-free return r_0, and a power utility of final wealth per dollar invested, $U(V_H) = \frac{1}{1-\gamma} \exp[(1-\gamma)\log(1 + R_H)]$, where γ is the constant relative risk aversion. One of Merton's key assumptions is continuous rebalancing, it guarantees that the portfolio of the two assets is log-normal, see Dumas and Jacquillat (1990) for a discussion of the approximation to log-normality. Then, by Ito's lemma, the multi-period compound return for a constantly rebalanced allocation w is shown to be:

$$\log(V_H|\alpha,\sigma) \sim N\left[(r_0(1-w)+w\alpha-0.5w^2\sigma^2)H, w^2\sigma^2 H\right], \tag{6}$$

where $\alpha = \mu + 0.5\sigma^2$. The expected utility is:

$$E[U(V_H)] = \frac{1}{1-\gamma}\exp\left[(1-\gamma)H(r_0+w(\alpha-r_0)-0.5w^2\sigma^2+0.5(1-\gamma)w^2\sigma^2)\right]. \tag{7}$$

The maximization of (7) over w gives the well-known Merton allocation:

$$w^* = \frac{\alpha - r_0}{\gamma\sigma^2} \tag{8}$$

The allocation in (8) offers an added insight over its one-period counterpart in (2), even though they appear similar. Merton's i.i.d. log-normal framework is a multi-period problem. Yet, the horizon H present in the expected utility (7) drops out of the optimal solution in (8). This is the well known irrelevance of the horizon in the optimal asset allocation. when returns are i.i.d.

In contrast, most of the subsequent intertemporal portfolio literature entertains the predictability of risky asset returns, with predominantly negative autocorrelations. Then, the variance grows with the horizon at a slower rate than the mean. One, therefore, optimally allocates more to the risky asset in the long than the short run. Additionally, in a dynamic strategy, the investor can reallocate her optimal weight within the investment horizon, reaping a further benefit from the long-run horizon. See Brandt (2006) for a survey of intertemporal portfolio strategies.

There is an ongoing debate in the finance literature, between those who consider that there is strong evidence of predictability and those who are unconvinced. However, one fact that is not up for debate is that mean returns are estimated with large errors. It is, therefore, curious that most of the finance literature has spent much more energy on predictability assuming known parameters, rather than the opposite. We now incorporate uncertainty in the mean into the optimal allocation problem.

For both classical and Bayesian frameworks, the sample mean, $\hat{\mu}$, computed from T years of data is a key sample statistic. For long-term forecasts, practitioners choose a point estimate by compounding the sample geometric return $G = \frac{1}{T}\log(\frac{P_T}{P_1})$. This amounts to estimating $E(V_H)$ by $e^{\widehat{\mu}H}$. Academics, however, tend to substitute $\widehat{\mu}, \widehat{\sigma}$ in the theoretical expectation (5), possibly because of the maximum likelihood justification, where the estimator of a function is approximated by the function of the estimator. The difference in these point estimates becomes very large in the long run. Using Siegel's (1994) geometric and arithmetic averages of 7% and 8.5%, the two approaches grow \$1 to \$160 versus \$454 over 75 years.

Even in the classical framework, a solution that does not invoke asymptotic approximation can be found. Jacquier et al. (2005) assume that σ is known and show that, for this problem, the uncertainty in σ is secondary to the uncertainty in μ. They derive a minimum mean squared error classical estimator of $E(V_h)$:

$$M = e^{H(\widehat{\mu} - \frac{\sigma^2}{2}(1-3\frac{H}{T})}.$$

The penalty for estimation error in μ increases with the horizon. The MLE estimator obtains as $T/H \to \infty$. Even with 100 years of data, as in the most mature market, one is never close to asymptotic assumptions for the purpose of long-term forecasts. Panel (b) in Figure 1 plots the compounding factor in M versus H as the dashed line, for realistic values of μ and σ, a sample of $T = 50$ years, horizons H from 1 to 40 years. It decreases linearly with $\frac{H}{T}$. The penalty in Figure 1 is so severe that one may wonder if it is even reasonable. For a very long horizon, it implies negative estimates of the compounding excess return, which does not make economic sense.

Figure 1 here

Let us see how the rational Bayesian investor incorporates uncertainty in the mean into her long horizon asset allocation. To do this, we repeat the asset allocation, with estimation error as in Bawa et al. (1976), but for the long run. The density of V_H in (6) is now conditional on μ, which must be integrated out to produce the predictive density of V_H. Then, the expected utility can be computed. Jacquier (2008) does this for a normal conjugate prior on μ. Consider for simplicity a diffuse prior, so that the posterior on μ is $N(\hat{\mu}, \frac{\sigma^2}{T})$. Because the integrations over the parameter and over the distribution of returns can be exchanged, one can also view this as integrating μ out of the conditional expected utility in (7). The expected (predictive) utility becomes:

$$E[U(V_H)] = \frac{1}{1-\gamma} \exp\left[(1-\gamma)H[r_0 + w(\hat{\alpha} - r_0) - 0.5w^2\sigma^2 + 0.5(1-\gamma)w^2\sigma^2(1 + \frac{H}{T})]\right]. \quad (9)$$

Recall that $\alpha = \mu + 0.5\sigma^2$, it is replaced by its posterior mean $\hat{\alpha}$, and there is a new term in H/T at the end. Maximizing the expected utility in (9), Jacquier (2008) finds the optimal asset allocation:

$$w^* = \frac{\widehat{\alpha} - r_0}{\sigma^2 \left[\gamma(1 + \frac{H}{T}) - \frac{H}{T}\right]}. \tag{10}$$

It is a function of the horizon H relative to the sample size T. It is in the spirit of Bawa et al., but the numerical effect is very large for long horizons. Panel (a) in Figure 1 compares this Bayesian optimal allocation. with Merton's. As the horizon H increases, The Bayesian allocation decreases drastically, even for a moderate risk aversion of $\gamma = 5$.

This allocation is consistent with an implicit point estimate of $E(V_H)$, optimal for the Bayesian investor given her risk aversion, the sample size T, and the horizon H. Equation (10) can be used to find this implicit point estimate of α, denoted $\alpha^\star$. This is, in essence, a change of measure to find the estimation risk adjusted estimate of α. In the new measure where expected returns are risk-adjusted, Merton's optimal allocation (8) applies, that is, $w^* = \frac{\alpha^\star - r_0}{\gamma\sigma^2}$. Equating this and (10) solves for $\alpha^\star$:

$$\alpha^\star - r_0 = \frac{\widehat{\alpha} - r_0}{1 + \frac{H}{T}(1 - \frac{1}{\gamma 1})}$$

This point estimate incorporates the Bayesian investor's expected utility and her optimal long-term allocation under estimation risk. Panel (b) in Figure 1 displays $\alpha^\star$. It strongly penalizes uncertainty for long horizons, even more than the classical MMSE for $\gamma = 8$. The finance literature points at typical risk aversions in this range, possibly even greater. Unlike M, $\alpha^\star$ never implies negative estimates of the risk premium. This is because it is the implication of the investor's optimal asset allocation in (10), which can at worst be 0.

In summary, estimation error in the mean has a far greater effect on long than on short term optimal allocation. A Bayesian optimal allocation shows that the investor should have drastically smaller optimal weights for the long run than for the short run. By a change of measure, the Bayesian optimal allocation provides us with an estimate of long-term expected returns consistent with the investor's risk aversion. This estimate implies a severe downward penalty relative to the posterior mean of μ for long-term forecasts. For the more risk averse investors, this penalty is larger than that implied by conventional statistical estimators. The set-up described above assumes i.i.d. log-normal returns, known variance, and a continuous time rebalancing, which provides us with analytical results. Barberis (2000) implements this optimal allocation with discrete-time rebalancing and unknown variance. The predictive density of returns, then, does not have an analytical form and must be simulated. Barberis finds results extremely close to those pictured in panel (a) of Figure 1.

2.4 Economically motivated priors

We now consider subjective Bayesian prior specifications of the mean vector, derived from economic considerations. These are not empirical Bayes priors because their hyperparameters are not based on the sample at hand. They still effect some shrinkage in the sense that the cross-sectional dispersion of the resulting posterior means is generally much smaller than for the sample means.

Consider incorporating in the prior views on the mean vector, the implications of an equilibrium model of expected returns such as the CAPM. Portfolio weights consistent with the CAPM are the relative capitalization weights, as in the market portfolio. In the absence of additional information on specific returns, these are good weights to start from. Alternatively, an extreme of the passive investment framework is simply to replace expected returns by β's, since the CAPM states that expected excess returns are proportional to the asset beta. The uncertainty in the mean would be the uncertainty in betas, still an improvement as inference on β's is more precise than inference on the mean. This, however, is not convenient for an investor who wants to incorporate private information on some of the assets, arising from proprietary analysis. This investor views the CAPM prediction for expected returns as prior information. She may have an econometric model to predict abnormal expected returns in excess of the CAPM, the so-called Jensen α. That will be the source of the likelihood function. In realistic settings it will be more sophisticated than the mere computation of average returns for a given sample period.

Black and Litterman (1991) (BL) is in this spirit. It specifically accounts for the fact that even active managers do not have private information on every asset in their investment universe. They notice that portfolio managers often modify only a few elements of the vector of means, for which they have private information. BL show that this practice has a large and undesirable impact on the entire vector of weights, rather, they combine investor views and market equilibrium, in the spirit of shrinkage. The originality of BL is that they do not write the prior expected returns from the asset pricing model. Instead, they reverse engineer it from the observed weights of the different assets, assuming that these weights arise from the market's optimization of these expected returns. This amounts to inverting the well-known formula for the optimal weights; $w^* = \frac{1}{\gamma}\Sigma^{-1}\mu$, where γ is the representative investor's risk aversion. The remainder $(1 - w)$ is invested in the risk-free rate. The mean μ consistent with the observed capitalization weights w^* is BL's prior mean for the mean vector. They use a prior covariance matrix proportional to the data, $\frac{1}{T_0}\lambda\Sigma$, where T_0 is a notional prior sample size. BL are primarily concerned with uncertainty in the mean vector, in their implementation they use the sample covariance matrix.

BL combine this economic-based prior view with the investor's private views on μ. These private views are formulated as a set of linear combinations of the mean vectors, where normal error terms allow to model the precision of the view. Essentially, they can be written as a multivariate normal distribution on μ. Assumedly these views come from some data analysis, but they do not need to. The posterior views, in BL's sense, result from combining the prior, asset pricing based view, with the private view. BL's formulation is conjugate, and the posterior mean is a simple weighted average of the economic and private means. This is a Bayesian learning scheme but curiously absent a formal likelihood function.

Zhou (2008) proposes to formally incorporate the data into the BL framework. He allows for three sources of information, the views from the economic model, the private views that may not use data, and the data. To do this, Zhou considers the final BL estimate of the mean as a prior. He then multiplies this prior by the likelihood to obtain a posterior mean. Zhou's motivation is that the data provide valuable information in the likely case that the economic model is inaccurate. This motivation runs a bit against the mainstream view, which, as we have seen, worries that it is the noisy data that lead to erratic estimation of the mean and, therefore, need to be restrained by shrinkage or an economic based prior. Yet, Zhou's implementation shows how to easily blend the above three fundamental sources of information in the Bayesian framework. Also, in quantitative portfolio optimization, the private views most certainly come from the estimation of some structure on the data, e.g., a predictive regression. Then, the problem simplifies as the prior can model the degree of belief in the departure from the equilibrium asset pricing model.

Pastor (2000) demonstrates a natural approach to incorporating an asset-pricing model in the portfolio allocation by using the prior to reflect the strength of belief in the model. This encompasses the two extremes of an investor with complete disregard for the model, and one with total confidence in the model. Reasonable investors have beliefs somewhere in between. Pastor determines the predictive density of returns and the resulting maximum Sharpe portfolio. Pastor's methodology can incorporate modern multi-factor asset-pricing models, where the efficient frontier is spanned by a set of benchmark portfolios that mimic the realizations of factors. In these models, expected returns in excess of the risk-free rate are typically of the form $E(R_i) = \beta_i' E(F)$, where β_i is the vector of k factor loadings of asset i, and $E(F)$ is the vector of k factor expected excess returns. In the case of the CAPM, the single benchmark portfolio is the capitalization weighted market portfolio.

Pastor considers a one-period setup with normal returns and a risk-free asset where the investor maximizes the Sharpe ratio. The investment universe includes N ordinary assets and K benchmark, factor-mimicking, portfolios. These $N+K$ assets have mean vector E and covariance matrix V. The likelihood function, consistent with the existing literature, comes from the multivariate regression of the N assets excess returns on the K portfolio excess returns. Let R be the $T \times N$ matrix of asset returns, $X = [i_T, F]$ includes a vector of ones and the $T \times K$ matrix of benchmark excess returns. Then, the multivariate regression of the assets on the benchmark returns is:

$$R = XB + U, \quad vec(U) \sim N(0, \Sigma \otimes I),$$

where $B' = [\alpha', B_2']$ includes the vector of abnormal returns, α and the factor loadings B_2. This can also be written using Zellner's (1962) seemingly unrelated regressions framework. It helps to write one observation of the multivariate regression above, for asset i at time t:

$$R_{it} = \alpha_i + \beta_i' F_t + u_{it}.$$

α represents the deviation from the asset pricing model, Jensen's alpha for the CAPM. If the asset pricing model holds, α is zero. On the other hand, if the model is useless, α is unconstrained and the above regression delivers inference on $E(R)$.

The benchmark returns F are assumed to be i.i.d. $N(E_F, V_F)$. The likelihood function $p(R, F|B, \Sigma, E_F, V_F)$ is multivariate normal and decomposed as $p(R|F, .)p(F|.)$. As a technical aside, Pastor uses the results in Stambaugh (1997) to allow a longer history for the benchmark portfolios than for the N assets. The priors on E_F, V_F are left diffuse. The prior on B, Σ is normal for $B|\Sigma$ and inverted Wishart for Σ, where $E(\Sigma) = s^2 I$, it is left largely non informative for B_2 and Σ. The key prior for modeling the degree of belief in the model is that on α. Recall that α and B_2 are subsets of B. The prior on $\alpha|\Sigma$ is modeled as $N(0, \sigma_\alpha^2 \frac{\Sigma}{s^2})$. The zero mean centers the prior on the asset pricing model; σ_α reflects the investor's degree of skepticism toward the model. The prior mean could also be centered on a non-zero value if the investor practiced some fundamental analysis, σ_α would then reflect the precision of the analyst's views.

The predictive density of returns $p(R_{T+1}|R, F)$ is simulated by drawing in sequence from $p(F_{T+1}|F)$ a multivariate student-t, the posterior $(B, \Sigma|R, F)$, and the conditional density $p(R_{T+1}|B, \Sigma, F_{T+1})$. This integrates out the parameters and future benchmark returns from the conditional density of future returns. The draw of B, Σ can be done with a very effective Gibbs sampling. A large sample of draws allows the computation of the predictive mean and variance up to an arbitrary degree of precision. These two moments are then used to compute the maximum Sharpe ratio portfolio as per (1).

Intuitively, the predictive mean of returns is $\tilde{\alpha} + \tilde{\beta}\hat{E}_F$, where $\tilde{\alpha}$ is the posterior mean of α. This posterior mean is a linear combination of the prior mean, here zero, and the sample estimate. The posterior mean of β is almost always very close to the sample estimate as the prior on B_2 is diffuse. Therefore, the predictive mean is shrunk to the asset pricing model prediction $\tilde{\beta}E(F)$, away from the sample estimate $\hat{\alpha} + \tilde{\beta}E(F)$. As σ_α increases, the predictive mean of returns tilts toward $\hat{\alpha}$, showing that the investor pays more attention to the sample estimate $\hat{\alpha}$ of mispricings.

An interesting use of the analysis is to document the effect of variations in σ_α on the optimal portfolio. Pastor implements the model to study the home bias effect. He uses one benchmark asset, the US market, and one other asset, an international index. He finds that an investor would require very strong belief in the global validity of the domestic CAPM, $\sigma_\alpha < 1\%$ annually, to justify holding as little in the foreign stock as is commonly observed. A second application sheds light on the Fama-French factors, especially the book-to-market benchmark portfolio. Pastor finds that even an investor with very strong prior beliefs in the CAPM would hold a sizable fraction of her portfolio in the book-to-market portfolio.

As a portfolio paper, Pastor (2000) shows how a rational Bayesian investor naturally incorporate her degree of belief in asset-pricing models into her optimal investment strategy. Possibly more strikingly, the paper recasts empirical asset pricing in terms of the usefulness and impact of asset pricing models on investing decisions. For a given sample and prior belief in the CAPM, we can see how optimal choice differs from that implied by the asset-pricing model.[2] This is in contrast with the earlier classical literature which argues whether an asset-pricing model, such as the CAPM, can or cannot be rejected by the data, with no obviously useful implication for the investor.

2.5 Other aspects of parameter and model uncertainty

While some of the work discussed so far does integrates out Σ to obtain the predictive density, e.g., Frost and Savarino (1986), it is mostly done assuming normal returns. This yields a multivariate Student-t predictive density of returns. The assumption of log-normal returns, coupled with a power utility, is often preferred in finance, especially for multi-period problems. The log-normal distribution is preserved under time aggregation, as seen in section

[2] As such, the paper could have been discussed in the asset pricing section below. It is not the first paper to use this approach, we discuss Kandel and Stambaugh (1996) later.

2.3. However, the integration of the variance results in a log-Student t predictive density. Geweke (2001) shows that expected utility then ceases to exist. The common practice has been, for asset allocation, to arbitrarily constrain the weight in the risky asset to be below one. Alternatively, a truncation of the log-Student t distribution eliminates the problem.

Pastor (2000) assumes normal returns and ignores the fat-tailness inherent in the predictive density. This allows him to use an analytical formula for the optimum Sharpe ratio. In addition, one may want to model skewness and kurtosis in the conditional distribution of returns. Such generalizations can quickly render the maximization of expected utility intractable. One can use simulation methods, or one can expand the utility function beyond the second moment of the distribution of returns. Harvey et al. (2010) allow for skewness and co-skewness in returns, with a skew-normal distribution, and formulate a utility function linear in skewness as well as mean and variance. This may be important because, while individual stocks may not exhibit strong skewness, portfolios of these stocks can have skewed returns. Therefore, portfolio weights should also be judged on their ability to increase skewness, since rational investors like skewness.

De Miguel, Garlappi and Uppal(2007) run a horse-race of 13 portfolio optimization models against a basic equal-weighted portfolio, named *one-over-N*. They use several combinations of the classic domestic and international portfolios encountered in the literature. Their sample period goes as far as 1963, they roll windows of ten years of monthly data. The 13 models include several of those mentioned so far. De Miguel et al. conclude that these models would have generated lower Sharpe ratios and certainty equivalent returns than the naive one-over-N rule over the past several decades. Their approach is based on the comparison of realized Sharpe ratios. Tu and Zhou (2008) revisit the results, using priors that incorporate, not only parameter uncertainty, but also the economic objective, here Sharpe ratio maximization. Their prior on the parameters is derived from a prior weight vector. Tu and Zhou document utility improvement over other prior specifications and also over the the 1-over-N portfolio.

Finally, the presence of multiple competing models brings up the issue of model uncertainty. Posterior odds can be used for model comparison, but also for optimal model averaging. The optimal model is a (normalized) posterior odds weighted average of the competing models. It would be interesting to revisit studies, such as De Miguel et al. by incorporating the optimal combination model.

3 Predictability of Returns

A large finance literature has studied the predictability, or lack thereof, of stock returns. The ability to predict future returns is at the core of the debate on market efficiency. See Campbell, Lo and Mackinlay (1996) chapter 2, (CLM) for a review of key classical results and techniques. Past returns, as well as firm and economy characteristics, have been used as predictors in time-series regressions or cross-sectional regressions where the returns lag the right hand variables. The horizon may be short, a month or less, or long, up to business cycle horizons. First, predictability can be studied through its statistical significance. We will show examples that highlight some pitfalls of classical analysis when the econometrician is concerned with non-linear functions of the base parameters used in the likelihood function. Second, predictability can be studied through its core economic implications, for example its impact on optimal portfolio design. This is conveniently done in a Bayesian framework.

3.1 Statistical analysis of predictability

3.1.1 Long-run predictability

In the study of long-run predictability, Bayesian methods can lead to conclusions very different from classical methods. Consider the popular permanent-transitory component:

$$\begin{aligned} p_t &= q_t + z_t \\ q_t &= \mu + q_{t-1} + u_t, \; u_t \sim N(0, \sigma_u^2) \\ z_t &= \phi z_{t-1} + \epsilon_t, \; \epsilon_t \sim N(0, \sigma_\epsilon^2), \end{aligned} \tag{11}$$

where the log-price p_t is the sum of a random walk and a stationary AR(1). It generates the long-term negative autocorrelations observed in asset returns (but fails to generate the intermediate term positive autocorrelations, see CLM chapter 2). A common approach has been to estimate directly the autocorrelation of long-term returns by the regressions:

$$r_{t,k} = \alpha_k + \beta_k r_{t-k,k} + \epsilon_{t,k}, \quad t = 1, \ldots .T \tag{12}$$

where $r_{t,k}$ is the k-period return from t-k to t. These regressions are run for k's between 1 and up to 8 years. One can also compute the ratio $Var(r_{t,k})/(k\,Var(r_t))$. If log-prices follow a pure random walk, regression slopes should be zero and variance ratios should be one. These constitute typical null hypotheses for the classical approach (see CLM chapter

2). It can be shown that, under the model in (11), β can tend to -0.5 as k increases.

In standard asymptotic analysis T goes to infinity, but the ratio K/T affects the computations. With $K/T \to 0$, the estimator of the ratio converges to 1 if the true ratio is 1. However, Richardson and Stock (1989) assume that $K/T \to c > 0$. This may reflect the fact that as the sample size increases, the investigator may explore larger k's. They show that the classical estimator of the variance ratio converges then to a value smaller than 1, and its variance does not vanish asymptotically. This alternative asymptotic limit under the hypothesis of no predictability is, in fact, consistent with the typical estimates in the literature (see CLM chapter 2). These conflicting asymptotic limits make the interpretation of results for a given fixed sample difficult. Which asymptotic limit is the empiricist who uses one given sample size supposed to adopt?

In contrast, the Bayesian approach delivers the optimal posterior inference for the given sample of size T studied by the econometrician. Lamoureux and Zhou (1996), (LZ) implement Bayesian inference for the model in (11). The likelihood is a function of the parameters $(\phi, \sigma_u, \sigma_\epsilon)$. The posterior densities for functions of interest, such as β_k, do not obtain analytically, but can be simulated.

LZ use data augmentation to generate convenient conditionals densities that are the basis for a Gibbs sampler. They add the vector $v = u/\sigma_u$ to the parameter space, and consider the joint data (v, r). The contemporaneous covariance between v and r is σ_u^2, the covariance matrix of r is a function of ϕ, σ_u. The joint posterior distribution of the parameters is intractable, but it can now be broken into a set of conditional distributions from which one can draw directly. Specifically, LZ show how to cycle between direct draws of $(v|\phi, \sigma_u, \sigma_\epsilon, r)$ and $(\phi, \sigma_u, \sigma_\epsilon|v, r)$, where the second distribution is further broken down into three univariate conditional posteriors. The key here is that not only one can draw from the conditional $(v|r, .)$, but also that the densities of the original parameters are greatly simplified by the conditioning on v. LZ extend the AR(1) in (11) to an AR(4) for quarterly data. An essential identification of the model is that the vector of AR coefficient must imply stationarity. In a Monte Carlo algorithm, this is enforced by rejection of the posterior draws that fail the requirement. For every draw of the AR(4) parameters, one computes the roots of the characteristic equation and rejects the draw if they are inside the unit circle.

These draws of the model parameters yield, by direct computation, draws of the nonlinear functions of interest, e.g., the ratio of the variances of the random walk shock u_t to the total return r_t, σ_u^2/σ_r^2. The persistence of shocks to the stationarity of z_t is of interest, and LZ compute the posterior distribution of its half-life. One can also compute the long-run autocorrelation, the β_ks from (12), directly, as implied by the model. Each draw of the parameters yields a draw of these functions of interest; therefore we obtain their exact posterior distribution for our sample size.

Classical analysis for these long-term models yielded mixed results. The point estimates of β_k were deemed large enough to warrant attention; however, the power of these regressions against the null of no predictability was known to be weak (see CLM ch. 2). The results of the Bayesian analysis are very different. LZ study the inference on β_3, the 3-year beta in (12). For two different proper priors on the parameters $\phi, \sigma_u, \sigma_\epsilon$, they simulate the implied prior on β_3 simply by drawing from these two priors. Both priors allow for sizable probabilities of large negative betas, and allow for a high fraction of the returns variance to come from the predictable component. They reflect the prior views of an investor who believes that a departure from the random walk hypothesis is quite possible. Strikingly, the resulting posteriors on β_3 are tightly centered on 0. Despite the initial priors, the data clearly speak loudly against the presence of a transitory component in stock returns.

Why then does frequentist analysis find large negative estimates of β_k? LZ make the case that frequentist analysis is akin to a flat prior on the parameters, $(\phi, \sigma_u, \sigma_v)$ in (11). They show that with this prior, $p(\phi, \sigma_u, \sigma_\epsilon) \propto \frac{1}{\sigma_u \sigma_\epsilon}$, the posterior density for β has a mean and standard deviation similar to the point estimates in the classical results. They also show that this flat prior implies a very odd prior on β_3, with two large spikes at 0 and -0.5. The key here is that diffuse priors on both variances allow the ratio of stationary to random walk variance, to be very large or very small, implying in turn, about equal probabilities of either a nearly pure random walk, or a very strong transitory component. Note here that, since the base prior is improper, LZ must have truncated or approximated it with a proper prior, to draw from it. One should make sure that the shape of the implied prior is robust to the truncation or approximation chosen.

This result shows that the notion of flat prior on the base parameters of a model, those used to write the likelihood, here $(\phi, \sigma_u, \sigma_\epsilon)$, can imply very informative priors on non-linear functions of these parameters. One must, therefore, be aware of the prior implied on functions of interest, by the prior chosen for the base parameters. This is easily done in a Bayesian framework. If needed, the prior on the base parameters can be modified to imply a reasonable prior on the functions of interest. A small sample analysis in classical econometrics is possible but complicated and rarely seen in the empirical literature. In standard maximum likelihood analysis, functions of an MLE estimator are assumed to be asymptotically normal, their variance typically approximated via the Delta method. A careful Monte Carlo simulation of the sampling properties of the estimator of the function could detect departures from asymptotic approximations. It would then be clear to the careful classical statistician that, as shown in LZ, the classical analysis did input, some undesirable prior views on the function of interest. This point is important, since an advantage often put forth by the proponents of classical analysis is that they do not need to put prior information into the analysis.

3.1.2 Predictability and cyclicality

We continue with another example where it is easy to understand that flat priors on the regression slope coefficients imply tight priors on a function of interest. A simple AR(1) in the stationary component as in (11) can generate the observed long-term negative autocorrelations; however, it can not also generate the shorter term positive autocorrelations discussed in CLM chapter 2. A model allowing for cyclical autocorrelation is required, that is, at least an AR(2). Geweke (1988) shows the posterior probabilities of a cycle and posterior densities of the cycle periods for GDP. Such macroeconomic variables can be state variables for the investment opportunity set, and their cyclicality could permeate to the process of stock returns.

Jacquier (1991) studies the cyclicality of AR(3) models of stock returns. He shows that flat priors on the AR(3) parameters result in an undesirably informative prior for the two main functions of interest: the probability of existence of a cycle and the period of the cycle. As is well known, cyclicality obtains when the roots of the characteristic equations are complex. Consider an AR(2) and flat priors for (ϕ_1, ϕ_2) in the stationary region; known to be a triangle. Cyclicality occurs when $\phi_1^2 + 4\phi_2 < 0$, an area between a parabola and the base of the triangle (see Zellner (1971)). For flat priors, the probability of being in this region is exactly 2/3; therefore, flat priors on (ϕ_1, ϕ_2) in the region of stationarity imply a 2/3 probability of a cycle. For an AR(3), Jacquier (1991) shows that flat priors on ϕ_1, ϕ_2, ϕ_3 in the stationary region imply a probability of 0.934 of existence of a cycle. For the Bayesian econometrician, posterior probabilities of a cycle of up to 2/3 for an AR(2), and 0.93 for an AR(3) represent no evidence whatsoever of cyclicality.

Given that a cycle exists, its period is of interest. Flat priors on ϕ inside the cyclical domain also imply quite an informative prior on the distribution of this period. It is centered at 3 with about 50% of its mass between 2 and 5. The Bayesian econometrician naturally spots the inference problem by inspecting these implied priors, comparing them to the corresponding posteriors. As for possible remedies, one can easily modify the priors on ϕ to produce, if desired, a flatter-looking prior on the period of the cycle, and a prior probability of existence of the cycle closer to 0.5. Setting flat priors on the roots of the characteristic equation, rather than the AR parameters themselves, goes a long way to resolving the issue. In contrast, both detection and remedy for this situation are not practical in the classical framework. The ordinary least squares point estimate of ϕ matches numerically the Bayesian posterior mean resulting from diffuse priors on the ϕ, therefore, the classical analysis can not escape from this problem uncovered in the Bayesian framework.

3.1.3 Model choice and predictability

Studies of the predictability of stock returns can involve a number of competing regressions of stock returns on firm-specific or economy-wide variables. The number of variables and alternative models makes this an ideal ground for model comparison and, better, model averaging, via odds ratios. In contrast, classical analysis is ill-equipped for multiple model comparison. As early as 1991, Connolly (1991) reports odds ratios on the well known weekend effect. Odds ratios also provide a more natural sample-based metric than the potentially severely misleading use of the p-value. See Berger (1985) for extensive discussions. The odds ratio of model 1 to model 2 is the posterior probability that 1 is true relative to 2, given the sample just observed.

Classically-motivated criteria, such as the Akaike information criterion (AIC) allow model ranking. The Schwartz information criterion (SIC) is proposed as a large sample approximation of the odds ratio. Zellner (1978) shows that the AIC can be seen as a truncation of the posterior odds ratio which omits many important terms. Jacquier (1991) shows that the approximation in the SIC can also be unsatisfactory in small samples even for simple AR models. Using these criteria, Bossaerts and Hillion (1999) find evidence of in-sample predictability, but no such evidence remains out-of-sample. It is unclear to what extent this contradiction between the criteria and the out-of-sample evidence is due to a possible over-fitting by these criteria. Even if the approximation was satisfactory, SIC and AIC could only be used for model ranking, not directly for model averaging.

Posterior odds ratios can of course serve to rank competing models, but, more interestingly, they determine the weight of each model for the purpose of model averaging, the optimal combination of all models. Avramov (2002) studies the predictability for monthly stock returns. He considers 14 widely studied candidate predictors, e.g., dividend and earnings yields, momentum, default and term spreads, inflation, and size and value premiums. These 14 candidates define 2^{14} mutually exclusive models for which Avramov computes posterior odds ratios. Model j is a multivariate regression with normal errors:

$$r_t = B_j X^j_{t-1} + \epsilon_t, \quad \epsilon \sim N(0, \Sigma), \tag{13}$$

where r_t is a vector of asset returns. Avramov models a vector 6 portfolio returns, B_j includes the intercept and the slope coefficients for model j, and X^j includes ones and the candidate predictors z_j, as in Zellner's (1962) seemingly unrelated regression framework.

Avramov uses normal- inverse Wishart priors for B_j, Σ_j: the prior mean of B_j is zero for the slopes and $\overline{r}$ for the intercepts. Now consider a hypothetical sample with the same sample statistics as the one studied: $\overline{r}, \overline{z_j}$, and $\hat{V}_r, \hat{V}_{zj}$, the sample covariances matrices of the returns and predictors. Avramov sets the variance covariance matrix of B_j proportional to that which would arise in the multivariate regression in (13). The proportionality coefficient is in effect a notional sample size T_0 that can be used to tighten the prior against predictability. It is a multivariate generalization of the prior used by Kandel and Stambaugh (1996). This is essentially a version of Zellner's (1866) g-prior. Following K&S, Avramov uses a T_0 equal to 50 times the number of predictors. His sample size is $T = 540$. He finds analytical expressions for the posterior odds.

Avramov reports which predictors appear in the highest posterior odds model. He notes that it is easy to add up the posterior probabilities of the mutually exclusive models where each candidate predictor appears. The contrast between the two measures is striking. The best models use at most two predictors, however, far more than two predictors appear in many models. For example, the best model to predict the returns of a portfolio of large firms with medium book-to-market values has only the Term premium as regressor, which appears in 54% of all possible models. Yet, 4 other candidate predictors appear in more than 20% of all models. The best model for another portfolio only includes inflation and earnings, present in 31 and 39% respectively of all models. But Tbill, lagged index returns and January, present in respectively 28%, 48%, and 21% of the models, are not in the best model. Clearly the common practice in standard model identification and classical studies of predictability to choose and work only with the best model, passes by a lot of information. Odds ratios, while they properly rank models, similarly appear to omit essential information when they are solely used for model comparison.

A composite model can be built, using the posterior odds as weights. For each predictor, the posterior mean of its slope coefficient is the odds weighted average of the posterior means for each model. The posterior variance can be shown to incorporate the within-model posterior variance for each model, as well as the measure of cross-model variability. The composite model defines a weighted predictive distribution for future stock returns. This distribution appropriately integrates out both model and within-model parameter uncertainties. Avramov shows that for his 6 portfolios, from 1953 to 1998, the composite model dominates the models chosen as best by any known model selection criterion.

In Avramov, all 2^{14} models have the same prior probability; however, as Cremers (2002) points out, there is a link between the probability of a variable being in the model and the probability of that model. His reasoning is as follows: assume that all variables have equal and independent prior probabilities p of entering the model, then the probability of any one model is $p^k(1-p)^{14-k}$. The only way that models can be equiprobable is if $p = 0.5$. However, this implies a prior probability of no predictability of 0.0001, and a joint probability of having more than 4 variables of 0.91. This is a lot of prior model mass on predictability. The issue of the choice of prior on parameters and on model size is non trivial, and the subject of a rich literature, see Ley and Steel (2009) for recent work on the issue.

Therefore, Cremers's priors are different from Avramov. He also makes the point that diffuse priors imply higher prior R^2's for the models with more regressors. He controls this by tightening to zero the priors of the larger models so that the implied R^2's are the same. On this issue, Avramov does something similar, since he keeps the notional sample size in the prior equal to T_0 times the number of predictors. This does tighten the slope prior towards zero for the larger models, and he shows that his results are robust to values of T_0 between 25 and 100. Another difference is that Cremers predicts a univariate series, a value-weighted index, while Avramov runs a 6-variate model. In contrast with Avramov, Cremers finds that his best models have more variables, but less out-of-sample evidence of predictability.

3.2 Economic relevance of predictability

We now turn to the economic relevance of predictability, measured by the impact of the competing models on optimal allocation. Performance is measured by the realized out-of-sample Sharpe ratios or certainty equivalent. We first discuss Kandel and Stambaugh (1996) (K&S), which set-up is now standard and has been used and generalized. K&S consider the predictive regression of monthly returns on the dividend yield:

$$r_t = x'_{t-1} b + \epsilon_t. \tag{14}$$

They evaluate the relevance of predictability through its effect on a Bayesian investor's optimal allocation between the market and the TBill. Typical R-squares for this regressions are below 5%, therefore, by statistical standard predictability does not appear formidable. Clearly though, as x_t varies through time, an investor may want to vary her optimal asset allocation since the conditional mean forecast of returns $x'_{t-1}b$ changes as well. This is also related to the inherent noise in the regression as well as the conditional mean at the time. Despite the low R^2's, K&S show that the typical monthly variations in the value of the regressors imply notable changes in the optimal asset allocation. They compute certainty equivalent returns to argue that these allocation changes are worth a lot to the investor.

Returns are log-normal and the investor solves for her optimal allocation by using power utility. K&S allow for stochastic regressors in (14), modeling them as a vector autoregression. The system of the VAR and (14) involves a slope coefficient and an error covariance matrices (B, Σ). The R-square is a non-linear function of this covariance matrix and slopes. To each possible value of (B, Σ) corresponds a value for the R-square. Hence, a distribution of (B, Σ) implies a distribution of the R-square. K&S consider two priors for B, Σ, one diffuse and one denoted "no-predictability", centered on zero. This second prior is that described above for Avramov (2002). The notional sample size T_0 is proportional to the number of predictive variables, so that the implied prior R^2 stays about the same for a different number of predictors.

As the variances are unknown, the predictive density is Student t. K&S constrain the optimal allocation w to be below 0.99 so that expected utility remains defined. Further, as they work in a discrete time setup, the exact w^* would need to be obtained numerically. Instead, they use the continuous time analytical optimum as an approximation. Their results are striking. Even for very low sample R^2's, the optimal asset allocation can vary a lot with the current level of the predictor. K& S compute the increase in expected utility due to the ability to change asset allocation: they compute the difference in optimal expected utility between a position where the regressors are at their unconditional mean and a typical low or high values of the regressors. They find that the differences between these allocations amount to notable differences in certainty equivalent, sometimes more than 3% a year, despite the small R-squares of the regressions.

Avramov (2002) also conducts an optimization. His initial set up in (13) is incomplete as he wants to look more than one step ahead. In order to draw several steps ahead, he formulates an AR(1) process for the regressors. Consequently, the predictive density $p(r_{T+K}|R)$ does not have an analytical expression for $K > 1$, due to the need to integrate out the returns at times $T + 1, \ldots T + K - 1$. In turn, the expected utility does not have an analytical integral either; even though he only optimizes a K steps ahead buy-and-hold portfolio, he needs to simulate the expected utility, and optimize these simulated values. Avramov looks at up to 10 periods ahead. He notes that the asset allocation should be less sensitive to the current value of the predictor as the horizon increases, since the predictor is stationary.

The classic framework for predictability in returns includes a regression of the returns r on stochastic regressors x, which themselves follow an AR(1) with strong autocorrelation:

$$\begin{aligned} r_t &= \alpha + \beta x_{t-1} + u_t, \\ x_t &= \theta + \rho x_{t-1} + v_t. \end{aligned} \tag{15}$$

The shocks u_t, v_t have a non-diagonal covariance matrix Σ. Stambaugh (1999) undertakes a detailed Bayesian study of this predictive regression with lagged stochastic regressors, typically the dividend-price ratio or the corporate yield spread. The system is written as a large multivariate regression, and one formulates priors on $(\alpha, \beta, \theta, \rho, \Sigma)$. Stambaugh uses flat priors on the slope coefficients and Σ, and studies posterior inference and asset allocation. A key result is that the posterior mean of β is linearly related to the posterior mean of ρ through the covariance σ_{uv} which is negative. He describes the impact on inference and asset allocation of two key aspects of the modeling; whether ρ is allowed to be in the non-stationary regions, and whether the first observation is considered known or stochastic, which modifies the likelihood function. He then implements these alternative specifications on four subsamples of the data between 1927 and 1996. The ordering of the posterior means for the various specifications varies with the subsamples, consequently, there is no clear evidence that a given specification produces systematically higher, or lower, posterior means, apart from the naive OLS which always produces the lowest β and highest ρ. Stambaugh then shows that these posterior differences lead to sizable differences in Merton allocation.

Barberis (2000) analyzes the effect of predictability and parameter uncertainty on the investor's asset allocation, especially for long horizons. He uses the same classic model as in Stambaugh (1999) or K&S, as in (15) where x_{t-1} is the dividend yield, $(d/p)_{t-1}$. After K&S, we suspect that predictability will have a strong impact on asset allocation. However, section 2.3 has shown that parameter uncertainty compounds enormously as the horizon grows. Without predictability, but with parameter uncertainty on both mean and variance, Barberis finds results similar to Figure 1, panel a). Even though he integrates out variance and rebalances discretely, he finds optimal allocations very close to those in Jacquier (2006).

Because he rebalances discretely, and does not know variance, Barberis does not have an analytical solution to the Merton optimal allocation problem, such as in (10). For each draws of the posterior parameters, he draws from the multiperiod predictive density of returns. Then for a number of candidate values of $w \in [0, 0.99]$, he computes and averages the utility of the asset allocation over the predictive draws. This is feasible since w is univariate and bounded. The optimal allocation is the w that yielded the highest (Monte Carlo estimate of) expected utility.

Barberis then considers predictability. As in Stambaugh (1999), the normal errors are negatively correlated, with the following effect on long-term allocation. Suppose that the dividend yield falls unexpectedly. The negative correlation implies that this is likely to be accompanied by a contemporaneous positive shock to stock returns. However, since the dividend yield is lower, stock returns are forecast to be lower in the future since $\beta > 0$. This contemporaneous rise, followed by a fall in future returns, causes variance to aggregate slower than for i.i.d. returns, leading the investor with a longer time horizon to allocate more to stocks. Assuming known parameters, Barberis confirms this intuition. However, when Barberis allows for parameter uncertainty as well as predictability, a very strong negative demand, as seen in section 2.3, sets in to counter the positive demand due to predictability alone.

Wachter and Warusawitharana (2009) model predictability with a regression of returns on the dividend-price ratio and the corporate yield spread. They model the investor's degree of confidence in predictability via a prior on the coefficients of this regression. A small prior variance for these coefficients implies high skepticism about predictability since the prior mean is 0. As in Stambaugh (1999) and Barberis (2000), the predictors follow an AR(1), and the shocks to the predictors and returns can be correlated. They conclude that the data would convince even a skeptical investor to time the market. Modeling the prior degree of belief in predictability allows them to determine which types of investors would or would not be swayed by the data.

What matters in these studies is how incorporating predictability in returns affects the predictive density, and in turn the optimal asset allocation of a rational investor, not whether autocorrelations or slope coefficients are statistically significant.

4 Asset Pricing

This section surveys the finance literature that directly tests the validity of asset pricing models. Since Roll (1977), it has been understood that tests of the various versions of the CAPM are often equivalent to testing whether some index portfolio was ex-post mean-variance efficient; therefore, we first discuss the Bayesian approach to tests of portfolio efficiency.

Multi-factor models, whether from economic arguments or data-mining, have become a popular way to remedy the shortcomings of the CAPM. Some empirical analysis is based on latent factors, which requires the estimation of the factor as well as the actual test of the model pricing. We discuss some unique contributions of Bayesian methods to this aspect of the literature.

4.1 Asset pricing tests are portfolio efficiency tests

Typical one-pass tests of the CAPM have often used likelihood ratios, Lagrange multiplier or Wald tests, which small sample distributions are not the same (see CLM, chapter 5). The econometrician selects a number of assets to be priced, and a market index portfolio, and tests whether the index prices the assets properly according to the CAPM. Consistent with Roll's (1977) argument, these tests can be written as functions of a measure of the efficiency of the index chosen as the market portfolio with respect to the frontier spanned by the portfolios and the index (see CLM chapter 5).

Shanken (1987) solves and generalizes the problem in a Bayesian framework. First he tests the efficiency, not of a single index, but of the most efficient linear combination of a set of portfolios. This is still with respect to the frontier spanned by N assets and the portfolios. Assume that the correlation between the highest Sharpe ratio portfolio on this frontier and the benchmark portfolio tested is 0.98. With enough data one will can still reject the null of efficiency, even if the difference between 0.98 and 1 is meaningless. This is a standard critique of the tests of point null hypotheses. The critique is even more warranted here, because, as Roll (1977) points out, we do **not** have the exact market portfolio, only a proxy with hopefully high, but not perfect, correlation with the market portfolio. For a given imperfect correlation between the chosen proxy and the portfolio with maximum Sharpe ratio, how much of the distance to 1 comes from the fact that we do not use the true market portfolio? Shanken formalizes this issue of proxy imperfection. It involves an added parameter, the correlation between the proxy and the true portfolio, on which he posits a prior distribution. He then tests the efficiency of the index by computing odds ratios that take into account the fact that we are using a proxy of the market portfolio.

Harvey and Zhou (1990) address the same problem by formulating priors on the mean and covariances of the assets; they do not incorporate beliefs about the imperfection of the proxy. In their crucial paper, Kandel et al. (1995), derive the posterior distribution of the maximum correlation between the portfolio tested and any portfolio on the efficient frontier. This posterior distribution is shown to be very sensitive to the choice of prior. As in Shanken (1987), they incorporate the fact that the portfolio tested is not a perfect proxy for a theoretical portfolio. This in effect makes the sharp null hypothesis of perfect correlation uninteresting. Their approach works for both cases, with and without a risk-free asset. They find that, especially in the presence of a risk-free asset, the choice of priors affects the results. For conventional sample sizes, a diffuse prior on the mean vector of the assets makes it very hard for the posterior of ρ to concentrate close to 1, the value implied by *efficiency*, even if the sample estimate of ρ is close to 1. This is another case where the parameter of interest, the maximized ρ is a non-linear function of the base parameters μ, Σ, which has perverse effects on the prior distribution of ρ.

In related work, Pastor (2000) discusses how to incorporate into the portfolio optimization the investor's degree of belief in an asset pricing model. Assume that expected returns are a linear combination of K factors, a generalization of the CAPM which centers on the efficiency of a single portfolio. If these factors can be replicated by K *benchmark* portfolios, then the frontier is spanned by these portfolios. The degree of belief is modeled by the tightness of the prior of the deviation from the model's prediction, e.g., Jensen's α for the CAPM. Pastor and Stambaugh (2000) use this framework to compare the CAPM, the Fama-French 3-factor model, and a third model, in a one-period, buy-and-hold mean-variance optimization framework. Portfolios are optimized using the predictive density implied by a model, and a degree of margin requirements. To compare models, they compute the loss in certainty equivalent for an investor who believes in one model but is forced to use weights that are optimal under another one. The result is that for realistic margin requirements and prior model uncertainty, the perceived differences between models are far smaller than classical testing lets us believe. Note also, that, from the view point of portfolio optimization, the best strategy would be a composite model according to the posterior odds ratios of each model, an interesting avenue of research.

4.2 Bayesian tests of the APT and factor models

The Arbitrage Pricing Theory (APT) builds on the assumption that returns are generated by a statistical model with latent factors f_t:

$$R_t = E(R_t) + Bf_t + \epsilon_t, \quad \epsilon_t \sim N(0, D), \tag{16}$$

where R_t is the N vector of asset returns, and B is an N (stocks) $\times$ K (factors) matrix of factor loadings. The crucial assumption is that D, the covariance matrix of the idiosyncratic risks ϵ, is diagonal; exposures to the common factors explain all of the stock covariances. The number of free parameters in the $N \times N$ covariance matrix is constrained since it is modeled with the $K \times K$ factor covariance matrix, the $N \times K$ coefficient matrix, and the N error variances. McCulloch and Rossi (1990) show that at most 5 factors suffice to explain the covariances between returns. Therefore, the factor model is by itself a very effective device for inference on large covariance matrices.

In the absence of arbitrage, the APT model follows from (16). Expected returns are linearly related to the factor exposures B:

$$E(R_t) = r_{ft} i + B\gamma_t, . \tag{17}$$

Here, γ is a k-vector of factor premia. Mc Culloch and Rossi (1990, 1991) are the first to implement a Bayesian test of the APT with latent variables. The first step of their procedure, by which they obtain the factor scores f_t is, however, more classical in spirit. They use the method of asymptotic principal components (see Connor and Korajczyk (1986)) to estimate the factor scores from a cross section of more than $N = 1500$ stock returns. The standard principal component methodology extracts the factor loadings from the sample covariance matrix, with precision increasing with the length of the time series. In contrast, Connor and Korajczyk show how to extract the factor scores from the $T \times T$ cross-product matrix of returns, with precision increasing in the number of stocks N. For the typical stock-market asset-pricing application, with very large N and not so large T, Connor and Korajczyk show that the f_t's are incredibly precisely estimated. Mc Culloch and Rossi (1990, 1991) essentially consider these scores f_t as known, when they implement the cross-sectional regression in (16). Therefore, they concentrate on a Bayesian implementation of (16), assured that it is not subject to issues of errors in the variables.

If both the factor model (16) and the APT (17) are correct, the intercept vector α in the multivariate regression:

$$R = \alpha i' + BF + E, \quad E \sim N(0, \Sigma),$$

must be zero. Mc Culloch and Rossi (1990) produce the posterior distributions for α. They use normal-Wishart priors for α, B, Σ. If the APT in (17) is correct, then α is zero. Second, they compare the certainty equivalent returns for a rational investor optimizing her portfolio with and without the constraint of the APT. For tractability, they assume normal returns and exponential utility. Their work represents the first utility based evaluation of an asset pricing model.

As this is a simple case of nested hypotheses, McCulloch and Rossi (1990) can use the Savage density ratio method to compute posterior odds for the null hypothesis of the APT model versus the hypothesis of mis-pricing ($\alpha \neq 0$). The Savage density ratio method allows to write odds ratios without actually performing the integration necessary to obtain the marginal likelihood. Instead, the odds ratio is simply the ratio of a posterior to a prior ordinate, at the value specified under the constrained model, here $\alpha = 0$ (see Dickey (1971).

In small-small (cross-section and time series) sample situations, even the N-asymptotic of Connor and Korajzyk might be inadequate, and suffer from errors in the variables. The standard principal components methods, which sampling precision increases with T, will also be affected by severe problems of errors in the variables. A pure Bayesian framework, optimal for the T and N used by the econometrician, is going to be very useful. Geweke and Zhou (1995) show how to estimate latent factors and their loadings with a pure Gibbs sampler. The intuition is straightforward since both conditional posterior densities $B|F$ and $F|B$ represent a regression, time-series or cross-sectional. Upon convergence of their algorithm, Geweke and Zhou can produce analysis similar to Mc Culloch and Rossi, however, bypassing any reliance to large N or large T.

4.3 Performance evaluation

One can view performance evaluation as a form of asset pricing test where managed funds, rather than individual stocks or portfolios, are investigated for mis-pricing. Performance evaluation is also tied to predictability as the issue of persistence in performance inevitable arises, see for example Avramov and Wermers (2006) for mutual funds .

Baks et al. (2001) approach performance evaluation via its impact on an investor who optimizes a portfolio of the index and one actively managed fund. Most importantly, they propose an innovative prior on a manager's Jensen α. The prior has a point mass on a slightly negative $\underline{\alpha}$, with probability $1 - q$ that the manager is unskilled. The performance $\underline{\alpha} = a - \text{fee} - \text{cost}$, is slightly negative because one subtracts from the raw performance a, the fees and transaction costs due to active management. a is argued to be slightly negative because the unskilled manager trades with skilled managers. Then, the manager is skilled with probability q. Given skill, α is modeled by a normal truncated at the mode $\underline{\alpha}$. Baks et al. effectively remove possibly large negative values from the prior on α, as unskilled managers are not expected to display systematically large negative performance. Baks et al. show how to elicit the prior by specifying q, the fees and costs, and the probability that the manager's α will be above a specific level, for example 25 basis points. An important result is that, unless the investor is a-priori extremely skeptical toward the possibility of skill, she would invest a non-negligible fraction of her wealth in actively managed funds. In contrast, for many of these same funds, the classical approach would fail to reject the null hypothesis of no skill, which would likely be followed by the decision not to invest in the fund. First, Baks et al.'s careful modeling of the prior on α affords them much needed precision in the estimation of α. Second, this application shows that even though two alternatives, here skill and no skill, may appear statistically close, they can still lead to very different investment decisions.

Baks et al. study each of 1400 fund managers separately, and do not allow for interactions or learning across funds. They compute each α over a Fama-French three factor model, and assume that the correlation matrix of idiosyncratic noises is diagonal. Jones and Shanken (2005) maintain this latter assumption but concentrate on learning across funds. The performance α_j of each fund is considered to be a random draw from a general distribution with cross-sectional mean μ_α and variance σ^2_α. Intuitively, consider that the prior on α_j draws from the average and variance of the sample α's of the other funds in the sample. Jones and Shanken point out that this shrinkage approach counters the undesirable unboundedness of the maximum posterior α when the number of funds increases. Also, in contrast with Baks et al. and others, they model different prior beliefs for each fund. The parameter space includes the α_j's, (considered to be random draws of) $\mu_\alpha, \sigma_\alpha$, as well as the individual fund betas and standard deviations. Since μ_α and σ_α are unknown as well, the posterior densities of the parameters are not known analytically. Therefore, Jones and Shanken write a Gibbs algorithm which updates all parameters, especially the α_j's, μ_α, and σ_α. Their empirical results confirm that incorporating learning across funds dramatically reduces the highest posterior means.

5 Volatilities, Covariances and Betas

Estimating and forecasting (co)variances is crucial in about every area of finance, including, risk management, option pricing, and portfolio optimization. At least in the univariate case, the literature has moved very quickly to the modeling of the time variation of volatility for a few reasons. The time variation of volatility has been taken for granted in finance since Officer (1973), and, because of its high autocorrelation, volatility is more successfully predicted than time varying expected returns. The modeling of time-varying volatility goes a long way to help match the fat-tailness of the unconditional density of financial series. In brief, the research has shown that a good parsimonious model of time varying volatility must have three key ingredients: 1) an autoregressive structure, 2) the ability to model asymmetries in returns where negative returns are associated with a greater volatility than same size positive returns, and 3) some additional modeling of fat-tailness for the conditional distribution of returns.

For a long time researchers used ad-hoc time moving windows to allow for time-varying volatility. Engle's ARCH, a quantum jump in variance modeling, triggered a huge literature (see Bollerslev et al. (1994) for a survey). However, while the time-series ARCH literature was mushrooming, theoretical finance was already exploring the more general stochastic volatility (SV) for modeling purpose. One reason for the resilient success of GARCH models may be that they are viewed as good filters of unobserved volatility. For example, Nelson (1994) shows that, as one converges to continuous time records, GARCH models dominate Kalman filters in terms of mean squared errors . Another reason is their ease of implementation, at least in the univariate setup; the multivariate case is far more complicated. While the ML framework works well computationally for the GARCH framework, classical methods can not handle the SV model well. The reason is that the SV model is non-linear and the volatility is a latent variable.

With the advent of Markov Chain Monte Carlo (MCMC) algorithms, Bayesian methods have been able to deliver the optimal estimation for a large class of SV models (see for example Jacquier, Polson and Rossi (1994,2004).[3] Further, Geweke (1994) and Kim et al. (1998) show that a lot can be gained from the added flexibility of the SV over the GARCH model. This section, therefore, first discusses some Bayesian GARCH algorithms. Then we show how to design, implement, and diagnose a simple MCMC Bayesian algorithm for a general univariate SV model with fat tails and asymmetric returns.

Precise inference for large covariance matrices is difficult even if they are assumed to be constant. Realistic finance applications often have many assets relative to the time period; therefore, parsimonious modeling requires some reasonable constraints on the matrix. One such constraint very well adapted to financial modeling is the factor model. We already discussed the Bayesian estimation of a constant-parameter factor model in the asset-pricing

[3] Jacquier, Johannes and Polson (2007) show that a simple adjustment of the Bayesian algorithm delivers the ML estimate and its asymptotic covariance matrix.

section. We complete this discussion here with the implementation of time varying factor models and betas.

5.1 Bayesian GARCH modeling

In a basic GARCH model, returns and their variance are as follows:

$$\begin{aligned} r_t &= \sqrt{h_t}\epsilon_t, \;\; \epsilon_t \sim N(0,1), \\ h_t &= \omega + \alpha r_{t-1}^2 + \beta h_{t-1}. \end{aligned} \tag{18}$$

The model can be extended to allow for fat tails in the shock ϵ_t. A so-called *leverage effect* can be added to allow for variance to be higher when the return is negative. Glosten et al. (1992) add a sign dummy to the variance equation. In contrast, Nelson (1991) introduces the exponential GARCH model (EGARCH), based upon a logarithm formulation. This allows for negative right-hand side variables in the variance equation, and eliminates the need for positivity constraints. For example, a EGARCH(1,1,) can be written as:

$$\log h_t = \omega + \theta\epsilon_{t-1} + \gamma(\epsilon_{t-1}^2 - E(\epsilon_{t-1}^2)) + \beta \log h_{t-1}. \tag{19}$$

Typically, one uses several years of daily returns to estimate a GARCH model. This often leads to precise reported standard errors of estimation for the simplest models. The prevalent technique has been the maximization of the likelihood function, which is conveniently done for the simplest models; however, experience shows that the maximization can be difficult for the more complex models. See Bollerslev, Engle, and Nelson (1994) for a survey.

Bayesian estimation of the GARCH model requires MCMC methods as one can not draw directly from the posterior distribution of the parameters. To see this, consider the posterior density of the parameters $\theta = (\omega, \alpha, \beta)$ in equation (19), a simple GARCH(1,1) model with normal errors:

$$p(\theta|R) \propto p(\theta) \prod \frac{1}{\sqrt{h_t(\theta)}} \exp - \frac{y_t^2}{2h_t(\theta)}.$$

Even breaking θ into its individual components does not permit a simple Gibbs sampler. The model can be extended with a regression function γx_t in the returns equation, which does not pose any further difficulty. One then introduces an added conditional $p(\gamma|\theta)$ to the MCMC sampler, from which direct draws can be made. Geweke (1989) uses importance sampling to draw from the posterior distribution of the parameters, and, in turn, the h_t's for the pure-ARCH model, with no lagged h_t in the variance equation. Kleibergen and Van Dijk (1993) use importance sampling to estimate a GARCH with Student-t errors. Their approximation function for θ is a multivariate Student t with very low degrees of freedom.

The priors used in Bayesian GARCH estimation are usually diffuse, but one wants to impose the positivity $(\omega, \alpha, \beta) > 0$, and, if desired, stationarity $\alpha + \beta < 1$ conditions. Note that Bayesian analysis, unlike maximum likelihood or the method of moments, does not require the existence of unconditional moments such as the variance of r_t. This is done by rejecting posterior draws which do not meet the conditions, which amounts to using a truncated prior since an indicator variable transfers directly to the posterior via Bayes theorem. Consider, for example, a possibly diffuse but proper density $p(\theta)$ defined on the real line. The econometrician wants to use this shape of density as a prior while imposing the condition $\theta \in [a, b]$. The prior is then $\pi(\theta) \propto p(\theta)\mathcal{I}_{[a,b]}$ where $\mathcal{I}$ is the indicator function. By Bayes theorem, the posterior is then $(\theta|D) \propto p(D|\theta)p(\theta)\mathcal{I}_{[a,b]}$.

Depending on the domain restriction, both prior and posterior may require a complicated integration to find the normalization constant due to the truncation. However, if one only need to draw from the posterior, as in a direct MC or MCMC algorithm, this integration is done by drawing from $p(D|\theta)p(\theta)$, and then rejecting draws that do not belong to $[a, b]$. This truncation by rejection is one of the appealing practicalities of Monte Carlo simulation of posterior and predictive densities. Clearly, the effectiveness of this practice, related to the fraction of draws rejected, depends on the amount of information in the likelihood about the domain of θ. An alternative prior strategy is to use a prior which does not require truncation, e.g. a scaled Beta prior; however, such priors may not always lead to simple posterior densities. In Bayesian GARCH estimation, positivity and stationarity conditions are, therefore, enforced by rejecting the inadequate posterior draws. It turns out that only a very small fraction of the draws is rejected.

Bauwens and Lubrano (1998) estimate a GARCH where ϵ_t in (19) is Student-t. with ν degrees of freedom. They note that the posterior of ν does not integrate if the prior is on $(0, \infty)$. Precisely, $p(\nu)$ needs to decrease at a rate faster than $1/\nu$ for large ν's. It also needs to be well behaved at $\nu = 0$. Truncation away from zero solves the problem and may even be desirable for modeling purpose; recall that the unconditional variance is infinite for $\nu \leq 2$. Bauwens and Lubrano choose the prior $p(\nu) \propto 1/(1|\nu^2)$. An alternative specification could be Geweke's (1993) exponential prior $p(\nu) \propto \psi \exp -\psi\nu$. Bauwens and Lubrano use a griddy Gibbs to draw from each element of the conditional posterior. For each element of the parameter vector, the griddy Gibbs computes posterior ordinates on a grid of carefully selected points, and the CDF between these points by numerical integration (see Ritter and Tanner (1992)). The inverse CDF method is then used to draw from the parameter. This is conceptually straightforward but numerically intensive, and requires a fair number of functional evaluations. Also, as Bauwens and Lubrano note, one needs to choose the grid carefully.

Muller and Pole (1998), in contrast, use a Metropolis Hastings (MH) algorithm to estimate GARCH models with regressions or AR parameters in the mean. They first break the parameter vector (θ, γ) into its individual elements. Then they make an MH independence draw for each element. Nakatsuma (2000) extends the model, also using a MH algorithm. He allows for an ARMA in the errors, making use of Chib and Greenberg (1994) who estimate the ARMA model with a MH algorithm.

Nakatsuma's algorithm cycles the ARMA and GARCH parameters, respectively δ_1 and δ_2. The conditional $\delta_1|\delta_2$ uses an ARMA likelihood with heteroskedastic but known variances h_t, a minor modification of the MH algorithm in Chib and Greenberg (1994). The conditional $\delta_2|\delta_1$ uses the well-known ARMA representation of the GARCH model introduced in Bollerslev (1986) (See Bollerslev et al. (1994)). For a GARCH(1,1), we have:

$$\epsilon_t^2 = \alpha_0 + (\alpha_1 + \beta_1)\epsilon_{t-1}^2 + v_t - \beta_1 v_{t-1}, \quad v_t = \epsilon_t^2 - h_t. \tag{20}$$

Note that v_t is non-normal with variance $2h_t^2$. Nakatsuma uses (20) as the basis for the conditional posterior of $\delta_2|\delta_1$. As a direct draw is not available, a feasible proposal density for an independence MH draw of δ_2 obtains by replacing the true distribution of v_t by a normal. The parameter vector draws are, in fact, further broken down, for both the ARMA and the GARCH, into the autoregressive and the moving average parameters. Accounting for the possible regression parameters, these are 5 major blocks of MH independence draws. However, Nakatsuma's method does not extend to the asymmetric GARCH of Glosten et al. (1992) and Nelson's (1991) EGARCH because they do not have an ARMA representation.

Vrontos et al. (2000) propose a random walk MH algorithm. The algorithm is easy to apply since it does not require the fine tuning of a proposal density. The candidate draw is simply made by adding an increment $N(0, \sigma)$ to the current value; σ is tuned to generate no more than 50% repeats. Small moves generate lower repeat probability, but they do not travel enough in the parameter space. Large moves on the other hand may visit very low probability areas and cause too many repeats. Vrontos et al. initially break the parameter vector into univariate MH draws; however, they show that the numerical efficiency can be increased if one draws jointly the most highly correlated parameters. One should, in general, be cautious with Metropolis draws of high-dimensional vectors, and make sure that they do not lead to a high repeat rate.[4] In their case, however, the dimension is low. The entire parameter space for a GARCH(2,2) is 5 parameters plus the regression or mean parameter.

Possibly the most interesting contribution in Vrontos et al. is their use of the reversible jump algorithm, which allows them to simulate simultaneously a number of competing GARCH or EGARCH models. In addition to the parameter draws, the algorithm jumps to another candidate model with a certain probability. The MCMC algorithm effectively generates the posterior probabilities of the models, as well as the parameter draws of each model (see Greenberg (1995)). So their algorithm produces the posterior odds for each model, as well as the model averaging. A direct by-product of their method can be the posterior distribution of in-sample volatilities and the predictive distribution of future volatilities, for the optimal combination model that is the average of the models considered, with weights the posterior probabilities of each model. For the Greek stock market, Vrontos et al. compare 8 EGARCH models. The best model has a posterior probability of 0.47, but the next three models have posterior probabilities summing up to 0.43. This highlights the potential benefits of model averaging over to the practice of selecting the best model.

These methods do not produce parameter posterior means drastically different from the MLE point estimate for very long series; however, if inference is needed for the parameters, one worries about the use of the Hessian matrix for standard errors, and the asymptotic normality assumption. In contrast, the Bayesian simulation methods produce the expected non-normal posterior distributions. Another, more important, issue is the difference in inference on in-sample and future volatilities (see Geweke (1989)). The MLE estimates the volatilities h_ts at the MLE point estimate of the parameters. The Bayes methods delivers, by simulation, the entire posterior density of each h_t as well as of the parameters. The Bayesian econometrician can then choose the posterior mean as a point estimate optimal under quadratic loss. But having the entire posterior distribution allows proper inference, beyond the use of the posterior mean as a location estimate.

A similar potential problem arises with multi-step ahead forecasting. Again, the MLE simply substitutes parameters estimates, running the variance equation sequentially to compute $\widehat{h}_{t+K}$, replacing r^2_{T+K-1} on its RHS by its forecast $\widehat{h}_{t+K-1}$. In contrast, a simulation-based Bayesian algorithm produces draws from the exact predictive density by running draws of the future shocks and volatilities to time $T+K$ through the volatility equation, for each

[4] One may decrease the repeat rate by reducing the dispersion of the proposal draws, but there may be a risk that the entire domain is not covered properly, especially in a high-dimensional multivariate setting.

draw of the parameters. This naturally produces the correct predictive density of future, time $T+K$ volatilities, that integrate out intermediate volatilities at $T+1,\ldots,T+K-1$, and parameters. See Geweke and Whiteman (2006) for discussions of Bayesian forecasting.

5.2 Stochastic volatility by MCMC

Consider the basic SV model below:

$$\begin{aligned} y_t &= \sqrt{h_t}\,\epsilon_t, \qquad\qquad (21)\\ \log h_t &= \alpha + \delta \log h_{t-1} + \sigma_v v_t, \quad t = 1,\ldots,T \\ (\epsilon_t, v_t) &\sim N(0, I_2). \end{aligned}$$

The key difference with the GARCH model resides in the shock v_t to volatility, which makes it an unobservable latent variable; the knowledge of the parameters, unlike the GARCH model, does not deliver the volatilities. Let $\omega = (\alpha, \delta, \sigma_v)$; the likelihood function $p(y|\omega)$ requires the integration of the T-dimensional vector of volatilities, that is, $p(y|\omega) = \int p(y|h,\omega)p(h|\omega)dh$.

The early literature used the method of moments to estimate the parameters, and the Kalman filter to obtain smoothed or filtered estimates of the volatilities given the parameters. Another approach, the quasi maximum likelihood (QML) was to approximate the SV model by a normal-linear state space model, assuming the normality of $\log \epsilon_t^2$. The likelihood of this approximate model could then be written in terms of the Kalman-filtered volatilities and maximized to obtain the parameters. These methods have been shown to perform poorly, (see Jacquier, Polson and Rossi (1994), hereafter JPR). JPR develop a Bayesian MCMC algorithm to draw from the posterior densities of the parameters and the volatilities, as well as the predictive densities of the future volatilities. The algorithm uses Metropolis-Hastings independence sampling.

5.2.1 A Metropolis-Hastings independence algorithm

Given a prior for the parameters $p(\omega)$, one needs to draw from the posterior density $p(\omega, \mathbf{h}|\mathbf{y})$. Consider first a Gibbs cycle of the two conditional densities, $p(\mathbf{h}|\mathbf{y},\omega)$ and $p(\omega|\mathbf{y},\mathbf{h})$. The second is a draw of the posterior distribution of regression parameters. The prior used in JPR is the conjugate normal-gamma prior, with variances large enough to render it flat over the relevant parameter domain (see JPR). A simple joint draw of the high-dimensional $p(\mathbf{h}|\mathrm{y},\omega)$ is not convenient, so one further breaks it into univariate densities

$p(h_t|h_{-t}, y, \omega)$. JPR show that, by successive applications of Bayes's rule, it can be written as:

$$\begin{aligned} p(h_t|h_{t-1}, h_{t+1}, \omega, \mathbf{y}) &\propto p(y|h_t)\; p(h_{t+1}|h_t, \omega)\; p(h_t|h_{t-1}, \omega), \\ &\propto \frac{1}{h_t^{\frac{3}{2}}} \exp\left(\frac{-y_t^2}{2h_t} - \frac{(\log h_t - \mu_t)^2}{2\sigma_v^2/(1+\delta^2)}\right). \end{aligned} \tag{22}$$

One can not draw directly from (22), but it is well approximated by an inverse gamma density. This is the basis for a MH independence draw. Therefore, the overall MCMC algorithm cycles through the elements of $(\omega, h_1, \ldots, h_T)$. One draws h_t from the inverse gamma density, and accepts the draw with the acceptance probability shown in JPR, otherwise the previous draw is repeated. This is referred to as a single-move algorithm, because it draws the latent variables h_t one at time. Since these volatilities are correlated together, the sequence of draws in this algorithm can exhibit high autocorrelation, especially for σ_v.

Large scale sampling experiments, e.g., demonstrate the behavior of the algorithm. In repeated simulations, as the sample size increases, the Monte Carlo estimate of the posterior mean converges to the true parameters. For example, for the parameters $\delta = 0.95, \sigma_v = 0.26$, and $E(\sqrt{h_t}) = 3.2\%$, JPR simulate 500 samples of $T = 1500$ observations, and compute the posterior means for parameters and volatilities with 50000 draws of the algorithm. They find that the posterior mean of δ is 0.94 on average, with a RMSE of 0.02, while that of σ_v is 0.279 on average with a RMSE of 0.04. Over the 750,000 $\sqrt{h_t}$s of this simulation, the posterior mean exhibits Mean Absolute Error (MAE) of 18.4%. By any standard, the posterior mean is very close to the true value as shown by the absence of bias and the very low RMSE. We also get an idea of the relative precision expected from the smoothed posterior mean of the volatilities, about 18%. For a given sample size, the posterior mean is not necessarily an unbiased estimator of the true parameter. Consequently, some Bayesians may find limited interest in a simulation of the sampling performance of the MCMC estimate of the posterior mean. The question here is whether the MCMC average is a good estimate of the posterior mean. If we find that its sampling behavior exhibits bias or high RMSE, how can we know if because of some failure of the algorithm or if we are actually seeing the sampling behavior of the true posterior mean. Recall that the posterior mean is optimal for a quadratic loss over the posterior distribution of the parameter, for the sample at hand.

Practically, in our case, the MCMC estimate of the posterior means do a great job of coming close to the actual parameters, by any intuitive measure. Second, JPR show that the sampling RMSE of the MCMC estimate of the posterior mean decreases with the sample size. Therefore, it reproduces the behavior expected from the posterior mean as the sample size increases. That is, the posterior mean converges to the MLE estimate as the sample size increases, and the MLE estimator converges to the true parameter. This is consistent with convergence of the algorithm. Another way that sampling experiments can be used is in comparing two algorithms in the same situation. If they both converge, they should produce the same results over and over.

An alternative algorithm has been proposed (see Kim et al. (1998), KSC) for the basic SV model. KSC model $\log \epsilon_t^2$, as a discrete mixture of normals, augmenting the state space accordingly, which allows them to draw directly from the multivariate distribution of $\mathbf{h}$. While the computational burden at each draw is higher, the resulting draws are markedly less autocorrelated, notably for σ_v, than for the single-move sampler. In comparing the two algorithms, KSC and others after, misinterpret the high autocorrelation of the draws in the single-move algorithm as *slow convergence*. One should not confuse a high autocorrelation with a sign that the algorithm does not converge. It is a sign that the algorithm may accumulate information at a slower rate than if the autocorrelations were lower. One then takes the usual precautions to assess the number of draws needed to obtain a desired precision for the MC estimate of, say, the posterior mean. This is done simply by computing standard errors robust to autocorrelation (see Geweke (1992)). In fact, low autocorrelation may not even be a sign that an algorithm has converged; it may be stuck in a region of the parameter space while exhibiting low autocorrelation in that region. With lower autocorrelation, a given desired precision for Monte Carlo estimates requires fewer draws, but this has to be weighted by the required CPU time per draw. In this case, the single-move algorithm is very fast; on a 2.8 Ghz Duo CPU, one generates 100,000 draws in 7 minutes for a sample of $T = 1500$ observations.

Sampling experiments can be used to compare different algorithms, for the same model. Jacquier and Miller (2010) show that the single and multi move algorithms. produce the same output. Table 1 reproduces some results for 500 samples of 1500 observations of the following SV model:

Table 1 here

These performances are nearly identical, especially for the volatilities. Jacquier and Miller run both these algorithms on 809 days of the daily UK pound to US $ exchange rate from January 2, 2006, to February 26, 2009. Table 2 shows the posterior analysis, where the two models produce nearly the same inference.

Table 2 here

Possibly, the multi-move will result in different posterior densities for the volatilities h_t? The sampling analysis in Table 1 only showed the sampling behavior of the posterior mean of h_t, what about the entire posterior distribution? Figure 2 plots the posterior mean and the 5^{th} and 95^{th} quantiles of the posterior distribution of $\sqrt{h_t}$, for both algorithms, they are in fact identical.

Figure 2 here

To conclude, SV models estimated by single-move or multi-move MCMC can deliver, period after period, posterior distributions of smoothed volatilities with a very satisfactory degree of precision, such as below 17% for the posterior mean of $\sqrt{h_t}$. Further results in section 7.2 confirm this for an extended SV model which exploits realized volatility.

5.2.2 SV with correlated errors and fat-tailed conditional returns

This section shows how to extend the basic SV to allow for correlated return and volatility errors, as well as fat-tailed conditional returns. We pay close attention to potential problems arising for the design of the proposal density.

A benefit of the basic single-move algorithm is that it extends readily, without further computing time burden, to the two most desired additional features: fat-tails in the distribution of ϵ_t and correlated errors to generate the so-called leverage effect. We use Student-t errors as with the GARCH, and we introduce a correlation ρ between ϵ_t and v_t. This correlation is in line with the use of SV models in option pricing theory, (see Heston (1993)). The general SV model is:

$$\begin{aligned} y_t &= \sqrt{h_t}\epsilon_t = \sqrt{h_t}\sqrt{\lambda_t}z_t, \\ \log h_t &= \alpha + \delta \log h_{t-1} + \sigma_v v_t, \quad t = 1, \ldots, T, \\ \nu/\lambda_t &\sim \chi^2_\nu, \\ (z_t, v_t) &\sim \mathcal{N}\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right). \end{aligned} \tag{23}$$

The shock to returns ϵ_t is modeled as a Student-t(ν) by setting a prior on λ_t as i.i.d. inverse gamma, that is $\nu/\lambda_t \sim \chi^2_\nu$. Explicitly modeling λ_t allows for a convenient simulation-based diagnostic for each observation. The prior on ν is integer uniform, for example on [5,60], if one wishes to rule out infinite conditional kurtosis. This discreteness of ν is not a problem since it would take a huge sample to deliver precise information for intervals smaller than 1. See Geweke (1993) for a continuous prior for ν. The parameter ω now includes $(\alpha, \delta, \sigma_v, \rho)$. We consider the conditional posterior distributions for the MCMC cycle:

1. $p(\mathbf{h}|\omega, \boldsymbol{\lambda}, \mathbf{y})$, where the conditioning on ν is subsumed by $\boldsymbol{\lambda}$

2. $p(\rho, \sigma_v|\mathbf{h}, \alpha, \delta, \boldsymbol{\lambda}, \mathbf{y})$

3. $p(\alpha, \delta|\sigma_v, \mathbf{h}, \boldsymbol{\lambda}, \mathbf{y}) \equiv \mathrm{p}(\alpha, \delta|\sigma_{\mathrm{v}}, \mathbf{h})$

4. $p(\boldsymbol{\lambda}, \nu|\mathbf{h}, \mathbf{y})$, where the conditioning on ω is subsumed by $\mathbf{h}$

The fourth distribution is the extension for fat tails, it is straightforward and does not increase the computing burden measurably. Given a draw of $(\mathbf{h})$, the model simplifies by considering $y_t^* = y_t/\sqrt{h_t} = \lambda_t z_t$. A direct draw of the posterior $p(\boldsymbol{\lambda}, \nu|.) = p(\boldsymbol{\lambda}|\nu, .)\ p(\nu|.)$ can be made, where $p(\boldsymbol{\lambda}|\mathbf{y}^*, \nu,) = \prod_t p(\lambda_t \mid y_t^*, \nu)$. JPR2 show that each $(\lambda_t|\nu, y_t^*)$ is a direct draw of an inverse gamma draw, and each $(\nu|\mathbf{y}^*)$ is a direct draw from a discrete multinomial distribution.

The third conditional is the posterior distribution of the slope coefficients of the AR(1) regression for $\log h_t$, conditional on the error standard deviation σ_v. We now consider the first two conditionals which contain the correlated errors extension. Their implementation provides a nice example of the care required to design the proposal density for an independence MH algorithm.

We will now see that care must be exercised in choosing the blanket of a Metropolis-Hastings independence algorithm. Given a draw of λ, the model in (23) simplifies by considering $y_t^{**} = y_t/\sqrt{\lambda_t}$. For $\mathbf{y}^{**}$, it is a SV model with correlated normal errors. The correlation ρ is modeled via the regression of $u_t = \sigma_v v_t$ on $z_t = y_t/\sqrt{h_t \lambda_t}$, specifically:

$$u_t = \psi z_t + \Omega \eta_t, \quad (\eta_t, \epsilon_t) \sim N(0, I), \tag{24}$$

where $\psi = \rho \sigma_v$ and $\Omega = \sigma_v^2(1 - \rho^2)$. This reparameterization of (ρ, σ_v) allows direct draws from the regression parameters $(\psi, \Omega|\mathbf{h})$. As the transformation is one-to-one, this yields direct draws for $(\rho, \sigma_v|\mathbf{h})$. Attempts to model ρ directly may require a Metropolis step, an unnecessary complication. JPR2 show how to model the prior on ψ, Ω, so as to have the desired prior on ρ, σ_v. The correlation ρ modifies the conditional posterior of $(h_t|\mathbf{y}, \omega, \mathrm{h_{t-1}}, \mathrm{h_{t+1}})$, JPR2 shows that it becomes:

$$p(h_t|h_{t-1},h_{t+1},\psi,\Omega,\mathbf{y}) \propto \frac{1}{\mathrm{h_t}^{\frac{3}{2}+\frac{\delta\psi \mathrm{y_{t+1}}}{\Omega\sqrt{\mathrm{h_{t+1}}}}}} \exp\left(\frac{-\mathrm{y_t^2}}{2\mathrm{h_t}}(1+\frac{\psi^2}{\Omega}) - \frac{(\log \mathrm{h_t}-\mu_\mathrm{t})^2}{2\Omega/(1+\delta^2)} + \frac{\psi \mathrm{y_t u_t}}{\Omega\sqrt{\mathrm{h_t}}}\right) \tag{25}$$

which modifies (22) for $\psi \neq 0$, mostly by adding a third term in the exponent.[5]

As with the basic model, JPR2 initially approximate and merge the first two terms in the exponent of (25), to design an inverse gamma proposal density, denoted $q_1(h_t)$. It omits the third term in the exponent. Convergence theory, however, suggests that it should not be lost; it will naturally be accounted for when computing the ratio (p/q) needed for the repeat / accept probability. This should, therefore, not affect the theoretical capability of the algorithm to produce draws with invariant distribution p; at worst, one would think that it might affect the rate of convergence.

In fact, JPR2 report that practical convergence did not happen; q_1 produced a very inefficient algorithm that would not close in on ρ no matter how long it would run or where it would start from. If the issue was only about the rate of convergence, it was still too severe for us to wait for it to happen. Autocorrelations in the sequence of draws were not abnormally high, revealing nothing pathological. Recall that a key to performance in accept/reject and Hastings-Metropolis Hastings is how well the blanket approximates the desired posterior p. Specifically, if the ratio p/q is unbounded, the algorithm cannot be uniformly ergodic. It turns out that, in this model, a simple extra work on q dramatically improves performance. JPR2 approximate $\frac{u_t}{\sqrt{h_t}}$ in (25) as a linear function in $\frac{1}{h_t}$, which can, then, be incorporated in the inverse gamma kernel. This yields a new proposal density, q_2, also inverse gamma.

A key diagnostic tool here and for any independence HM algorithm, is a plot of p/q for a wide range of values of h_t as the algorithm evolves. Figure 3 demonstrates this for a given h_t. The right plot shows that the ratio p/q_2 is much more stable than p/q_1 over a wide range of h_t. The left plot, where the kernels are normalized so as to be plotted together, shows that q_2 is right over p, while q_1 misses it. It is worse than q_1 not approximating the shape of p as well as q_2, it is that q_1 is in the wrong place. This is because the third term in the exponent of (25) often does not not have a mode; so it modifies the distribution kernel in the first two terms by shifting them. For independence MH algorithms, one should make sure that the ratio p/q is stable, specifically that it is not unbounded (see Mengersen and Tweedie (1994)).

[5] ρ is the contemporaneous correlation between ϵ_t and v_t. It implies that $E(y_t) \neq 0$; however, the effect is small. Alternatively, ρ can be defined as $\mathrm{Cor}(\epsilon_t, v_{t+1})$, with a minor modification of (25).

Figure 3 here

5.2.3 Volatility predictions for different models

The literature is replete with simulation studies documenting parameter estimation. This is a good first step, but we especially interested in the volatility densities produced by different models, especially future volatilities. Volatility forecasts are vehicles for risk management and option pricing. Do different models produce different forecasts, and in what circumstances? Bayesian MCMC algorithms deliver the marginal posterior density of the vector in-sample volatilities $\mathbf{h}$. One draws the predictive densities of future volatilities, by simply drawing the future shocks v_{T+k}s, and using the AR(1) equation to obtain the future h_{T+k}s, for each draw of $(\mathbf{h}, \omega)$.

While the posterior odds (see below) provide one summary diagnostic, we can use these posterior and predictive densities to compare the outputs of competing models. In this case, we want to know whether differences between the models matter in their economic magnitudes. For the weekly US NYSE index return, JPR2 estimate SV models, basic, with correlated errors, with fat tails, and with both. A first important question is if and when these models produce different posterior densities for volatilities, we concentrate here on the posterior means. Figure 4, left-hand plot, shows the ratio of posterior means of h_t produced by the Fat-tail and basic SV models, $E(h_{t,Fat})/E(h_{t,basic})$ versus the posterior mean of the mixing variable, $\sqrt{E(\lambda_t)}$. It clearly matters what model we choose. The model allowing for fat-tails predicts markedly lower volatilities, especially for a number of observations where it estimates larger lambdas.

Figure 4 here

Consider now adding correlated errors to the SV models with fat-tails. The posterior mean of ρ for these data is -0.4. Figure 4, right-hand side, plots the ratio of posterior means of h_t for the full versus the fat-tailed model versus the posterior mean of ϵ_t (for the full model). Again, the choice of model appears to matter greatly. Observations with negative ϵ_t have larger volatility. The average ratio on the vertical axis is 1.09 for the first decile of ϵ, and only 0.9 for the tenth decile (right tail), a 20% difference.

Return to the fat tails; does it matter if, as Figure 4 shows, the model with fat-tails allocates some of the h_t into λ_t? The top plot in Figure 5 shows the daily change in $UKPound/\$$ in 1985. An agent can implement the basic or the fat-tailed model. The thick line in the middle plot shows $E(\sqrt{h_t})$ for the basic model, the dashed line shows $E(\sqrt{h_t})$ for the fat-tailed model. The ragged line is $E(\sqrt{h_t \lambda_t})$, it shows that λ_t is far more than just a device to implement the Student-t errors. The model allocates a large λ_t to mostly the days that have a high h_t under the basic model. Then, the fat-tailed model will predict lower future volatility than the basic model, because λ_t does not have any persistence. These high volatility days are those when getting the best possible volatility prediction is crucial for a risk manager. The bottom plot shows out-of-sample forecasts originating from September 23^{rd} confirms this intuition. The two models can make very different volatility predictions, especially high volatility days with high volatility.

Figure 5 here

5.2.4 The cost of a model

The loss function approach allows an agent to summarize the cost of making the wrong decision. Our purpose here is illustrative, possibly more than normative, because with actual data, even the best inference will rarely lead to certainty as to the right and wrong models. Therefore, the reader should see the sections of the chapter, where we discuss model averaging.

Consider an investor with quadratic loss, the posterior mean is an optimal location estimate on the posterior distribution because it minimizes expected loss. This decision theoretic aspect can also be used for model choice, whereby the agent computes the expected loss of choosing each model. For J models M_j with posterior probabilities $p(M_j|.)$, the expected loss of model or decision i is:

$$EL(M_i|.) = \sum_j L(M_i|M_j) p(M_j|.), \tag{26}$$

One can take $L(M_i|M_i)$ to be zero, which is a way write $L(M_i|M_j)$ as the incremental loss of using model i if j is true. The losses could for example be the RMSE of variance estimation. If the models are mutually exclusive, then one chooses the model with the lowest expected loss. Posterior model odds ratios, normalized, are the weights that allow us to compute the expected loss. Recall that when models can be combined without added complexity, the odds ratios can be used as weights to determine the optimal combination of models, as seen above with Avramov (2002).

It is however interesting to compare the models along the loss function in (26). Consider an agent contemplating incorporating fat-tails to a basic SV model. She wonders what penalty may result from omitting to incorporate fat-tails when they are present in the data, or from unnecessarily incorporating them if they are not needed. Specifically, JPR (2004) simulate 500 samples of $T = 1000$ observations from the basic SV and 500 from the SV with fat tails. They then estimate both models on each sample. The loss function used is the RMSE and % MAE of variance estimation. Table 3 shows that the cost of unnecessarily extending the model with fat-tails is lower than that of omitting them when they should be used. First, both models fit s_t with equal success, in terms of RMSE or %MAE of their posterior means. If the data originate from the basic SV model, inference from the fat-tailed model is not affected as it correctly allocates very little volatility to the λ_t's. However, if the data originate from the fat-tailed model, inference from the basic SV model is seriously hampered by the fact that it lumps λ_t and h_t into its view of h_t. The third column in Table 3 shows that problem becomes quite severe on days when there was a large λ_t; while using the SV model results in a 24% MAE for $\sqrt{h_t}$, compared to 31% for the basic model. Such errors will transfer into volatility forecasts, as was seen in Figure 5.

Table 3 here

5.2.5 Computing odds ratios directly from the MCMC output

Posterior odds ratios allow a convenient ranking of multiple competing models. Normalized as model probabilities, they can be used to design an optimal model combination. Posterior odds ratios are based upon the marginal likelihood of the data for each model. For a model M_0, we have:

$$p(y|M_0) = \int p(y|\theta_0, M_0)\, p(\theta_0|M_0)\, d\theta_0.$$

This marginal likelihood can be a source of computational difficulties, for any reasonably complex model, it does not have an analytical integral. In addition, with latent variables as $\mathbf{h}$, even the conditional likelihood $p(y|\theta_0, M_0)$ itself does not have an analytical integral. Therefore, as much as possible, one should avoid computing this integral directly. Note also that the integral in the marginal likelihood requires proper priors. For the SV class of models, Jacquier and Polson (2000) follow Newton and Raftery (1994), and show how to compute the odds ratios directly from the MCMC posterior output of $(\mathbf{h}, \omega, \boldsymbol{\lambda}, \nu|\mathbf{y})$, without resorting to a direct evaluation of the marginal likelihood. They use the Savage density ratio method for an odds ratio on correlated errors, and Student's formula for an odds ratio on fat-tailed errors. Quantities are computed at each iteration of the MCMC posterior simulator, their

Monte Carlo average delivers the odds ratios. This is fast because the posterior draws are already available from the posterior parameter and volatility sampler. We now review this methodology.

Odds for correlated errors

In the SV model with correlated errors, the density of ρ follows by direct draw from that of ψ, Ω. Posterior analysis is intuitive, especially since we can formulate a very flat, but proper, prior for ρ. Since the basic SV, where $\rho = 0$, is nested in the correlated errors SV, we can use the efficient Savage density ratio method (see Dickey (1971)). Consider two models $\mathcal{M}_1$: (ϕ, ω) and $\mathcal{M}_0$: $\omega = \omega_0$. If $p_1(\phi|\omega = \omega_0) = p_0(\phi)$, then:

$$\mathrm{BF}_{0/1} = \frac{p_1(\omega_0|y)}{p_1(\omega_0)}.$$

The computation is done under the nesting model, and only requires ordinates of the posterior and prior densities of the parameter being restricted. If the posterior ordinate of $p_1(\omega)$ at ω_0 is larger than the prior ordinate, the Bayes factor favors the restricted model.

The ratio requires the exact ordinates, so that it can't be applied to parameters drawn by Metropolis for which we only know the kernel of the density. This is another reason why algorithms that do not draw ρ or ν directly would be problematic. JPR (2004) use the Savage density ratio method on ψ, the slope coefficient in the regression (24). We have $\psi = 0$ for the basic SV model denoted B, the correlated model is denoted C, and the Bayes factor is the ratio of ordinates:

$$BF_{B|C} \frac{p_C(\psi = 0|\mathbf{y})}{p_C(\psi = 0)}.$$

The marginal posterior ordinate $p_C(\psi = 0|\mathbf{y})$ is obtained by integrating out all the other parameters and state variables. The density of ψ conditional on the other parameters is normally distributed, the slope of a regression; Ω can be integrated analytically, which yields a Student-t for $p_C(\psi|\mathbf{h}, \alpha, \delta, \mathbf{y})$. The integration of the other parameters is done by averaging the Student-t ordinate over the draws of the MCMC sampler. The Bayes factor can be approximated by

$$\widehat{\mathcal{BF}}_{B/C} = \frac{\Gamma(\frac{\nu_0+T}{2})\Gamma(\frac{\nu_0}{2})}{\Gamma(\frac{\nu_0+T-1}{2})\Gamma(\frac{\nu_0+1}{2})} \frac{1}{G} \sum_{g=1}^{G} \sqrt{\frac{1 + a_{11}^{(g)}/p_0}{1 + a_{22.1}^{(g)}/\nu_0 t_0^2}} \left[1 + \frac{\tilde{\psi}^2}{\nu_1 t_1^2/p_1}^{(g)}\right]^{-\frac{\nu_0+T}{2}}, \tag{27}$$

where a_{11} and $a_{22.1}$, vary with the parameter draw (g) (see JPR2 for details). Note the averaging over the G draws of the MCMC sampler. The odds ratio only requires computing and cumulating the quantity on the right of the summation sign at each iteration of the sampler.

Odds for normal versus fat-tailed errors

The posterior for ν is not an convenient vehicle for a formal odds ratio between the SV models with fat-tailed and normal errors. Since ν has a finite upper bound in JPR 2, $\nu \in [5, 60]$, the model with fat-tailed errors does not nest the one with normal errors. The fat-tail model could nest the basic model with Geweke's (1994) parameterization in $1/\nu$; however, another condition necessary to the application of the Savage density ratio, $0 < p(\omega = \omega_0|D)$, would not be met, because the posterior goes to zero as $1/\nu \to 0$.

The following method helps circumvent the direct computation of the marginal likelihood. By Bayes theorem, with simple rearrangement of terms, the marginal likelihood for model $\mathcal{M}_1$ can be written as:

$$p_1(\mathbf{y}) = \frac{\mathrm{p}_1(\mathbf{y}|\omega, \psi)\mathrm{p}_1(\omega, \psi)}{\mathrm{p}_1(\omega, \psi|\mathbf{y})}. \tag{28}$$

This holds for any (ω, ψ) in the parameter space and is known as Student's formula, (see Besag (1989)). Chib (1995) proposes to use (28) directly to compute the marginal likelihood, by averaging the right-hand side over the Monte-Carlo draws. It might be computationally unstable as it involves a high-dimension likelihood. Jacquier and Polson (2000) show instead that (28) can be incorporated into the computation of the Bayes factor, as follows:

$$\begin{aligned} BF_{0|1} &= \int \frac{p_0(\mathbf{y}|\omega)\mathrm{p}_0(\omega)\mathrm{d}\omega}{p_1(\mathbf{y})} = \int \left[\int \frac{p_0(\mathbf{y}|\omega)\mathrm{p}_0(\omega)\mathrm{d}\omega}{p_1(\mathbf{y})} \right] p_1(\psi) d\psi \\ &= \int \int \frac{p_0(\mathbf{y}|\omega)\mathrm{p}_0(\omega)}{p_1(\mathbf{y}|\omega, \psi)\mathrm{p}_1(\omega|\psi)} p_1(\omega, \psi|\mathbf{y})\mathrm{d}\psi\mathrm{d}\omega, \end{aligned}$$

which is:

$$BF_{0|1} = E_{\omega,\psi|\mathbf{y}} \left[\frac{p_0(\mathbf{y}|\omega)}{p_1(\mathbf{y}|\omega, \psi)} \frac{p_0(\omega)}{p_1(\omega|\psi)} \right]. \tag{29}$$

The expectation is taken with respect to the posterior draws of ω, ψ in the larger model. In this general formulation, the domain for ψ in the larger model does not need to contain the values which represent the smaller model. For example, a SV model with Student-t errors with fixed degrees of freedom can be compared with the basic SV model, even though the latter corresponds to infinite degrees of freedom. Jacquier and Polson (2000) and JPR2 apply (29) to compute the Bayes factor for the fat-tailed versus the basic SV model:

$$BF_{B|F} = E_{\theta,\nu}\left[\frac{p_B(\mathbf{y}|\theta)\ \mathrm{p_B}(\theta)}{p_F(\mathbf{y}|\theta,\nu)\ \mathrm{p_F}(\theta|\nu)}\right],$$

where $\theta = (\alpha, \delta, \sigma_v, \mathbf{h})$, and E refers to the expectation over the joint posterior of (θ, ν) in the fat-tailed model. The choice of priors allows to further simplify the result; with ν is independent from the other parameters, we have $p_F(\theta|\nu) = p_F(\theta)$, and $p_F(\theta) = p_B(\theta)$. Therefore, the Bayes factor is only the ratio:

$$BF_{B|F} = E\left[\frac{p_B(\mathbf{y}|\theta)}{p_F(\mathbf{y}|\theta,\nu)}\right]. \tag{30}$$

Given a MCMC sample $\{\theta^g, \nu^g\}_{g=1}^G$ from the joint posterior $p_1(\theta, \nu|y)$, a Monte Carlo estimate of (30) is:

$$\widehat{BF}_{B|F} = \frac{1}{G}\sum_{g=1}^{G}\frac{p_B(\mathbf{y}|\theta^{(\mathrm{g})})}{p_F(\mathbf{y}|\theta^{(\mathrm{g})}, \nu^{(\mathrm{g})})}. \tag{31}$$

Under M_B, $y_t \sim \sqrt{h_t}N(0,1)$, and under M_F, $y_t \sim \sqrt{h_t}$ Student-t(ν). Because we condition on $\mathbf{h}$, the likelihoods in the Bayes factors (30) are simple products of independent univariate densities. Since we only need a ratio of likelihoods conditional on the parameters at every draw, and not the likelihood of each model, their magnitude does not cause computational problems.

Further, (30) easily extends to the computation of $BF_{C|FC}$, the Bayes factor of the model with correlated errors over the full model. A draw of $(\alpha, \delta, \sigma_v, \mathbf{h})$, implies a draw for all v_ts. In the presence of correlation, it also provides information on the ϵ_ts, specifically, $\epsilon_t|v_t \sim N(\rho v_t, 1-\rho^2)$. One, therefore, extends (30) to the computation of $BF_{C|CF}$, by replacing y_t with $y_t^* = (y_t - \rho v_t)/\sqrt{1-\rho^2}$.

Empirical results

JPR (2004) apply this method to compute odds ratios among the models for a number of financial series. They report odds that largely favor the general SV model with fat tails and correlated errors against the basic SV, for all stock indices and most exchange rates. For all exchange rates but the Canadian to US one, the odds, between 3 and 10 to one, moderately favor the model with fat-tailed errors for weekly data. For daily data, the odds overwhelmingly favor the fat-tailed errors. For most indices, the odds very strongly favor the model with correlated errors. The full model is overwhelmingly favored to the basic SV model for all indices and exchange rates.

We report here some additional results on the leverage effect, contrasting indices and individual stocks. Table 4 shows the posterior means of ν and ρ for the full model, and the Bayes factors. For individual stocks, the posterior distributions of ρ are centered very close to 0. The odds ratios moderately favor the models with no correlation. Yet, JPR (2004) report strong odds in favor of correlated errors for all the indices studied. The last row in Table 4 estimates the SV models and odds ratios for a portfolio of the 10 stocks above. There is now a negative correlation, and the odds ratio is in favor of the correlated errors. However, leverage can not be the cause for this effect since it does not appear with the individual stocks that exhibit as much variation in leverage as the portfolio. It is sometimes proposed that the correlation ρ can be driven by a small number of exceptional days in the sample. October 17^{th}, 1987, comes to mind. The last column of table 4 shows the Bayes factors for the 1989-1998 period. The result is opposite from what is expected. Again, the odds are moderately against the correlated model for most stocks; however, they very strongly favor the correlated model for the portfolio, in fact, far more strongly than for the period that contains October 19.

Table 4 here

The hypothesis, of volatility feedback, often advanced as an alternative to the leverage effect, could not either affect portfolios but not the stocks that constitute these portfolios. The negative correlation between return and volatility shocks of indices must be the result of a portfolio effect; arising from a time-variation in the correlation matrix of the stocks. A two-regime model for the correlation matrix of individual stock returns could be a fruitful avenue of research.

5.3 Estimating continuous time models

The theoretical option-pricing literature uses continuous-time processes, mainly because of their ability to produce tractable option-pricing models. Data are however observed at discrete times. To estimate a continuous-time model from discrete data, one uses its Euler discretization, which approximates the continuous trajectory of the process into a discrete one between the successive data intervals. For example, the following continuous-time constant elasticity of variance (CEV) model of the short rate:

$$dY_t = (\theta + \kappa Y_t)dt + \sigma Y_t^{\beta} dw_t,$$

is approximated by the discrete time process:

$$Y_t = \theta + (1+\kappa)Y_{t-1} + \sigma Y_{t-1}^{\beta} w_t, w_t \sim N(0,1).$$

This introduces a bias in the drift and diffusion parameters. We now discuss Bayesian methods that alleviate this discretization bias. Eraker (2001) concentrates on the specification of the diffusion process; Jones (2003a) develops independently a very similar technique, and studies the specification of the drift of the short-term rate.

The discretization bias disappears as the time between observations becomes shorter. This is the motivation for the Bayesian approach in Eraker (2001). He introduces $m-1$ missing data between each observation. Consider the process $Y_i = (X_i, Z_i)$, where X_i is observed every m periods and Z_i can be a latent variable such as stochastic volatility; both X and Y can be multidimensional. If we knew the missing data, the discretization bias would be diminished as we would be converging to continuous time. Given these missing data, denoted $\widehat{X}$, the model is a standard discrete time model. Posterior analysis can be conducted with known Bayesian techniques; that is, we can draw from $p(\theta|\widehat{Y})$, where θ is the vector of parameters.

The final intuition comes from Gibbs sampling: if we can draw from the missing data given θ and the observed data X, we have a complete model to improve upon the discretization bias. Eraker (2001) shows how to do this for the CEV or stochastic volatility models. His method also applies to other processes. Let $\widehat{Y_i}$ be the time i element of the matrix $\hat{Y}$, where X_i is observed or is a missing value $\widehat{X_i}$. Given a draw of θ, Eraker updates $\widehat{Y_i}$ sequentially, drawing from $p(\widehat{Y_i}|\widehat{Y}_{i-1}, \widehat{Y}_{i+1}, \theta)$ for $i \in [1, \ldots, mT]$. Of course one does not update the observed values. By Bayes theorem, this update is shown to be:

$$p(\widehat{Y_i}|\widehat{Y}_{i-1}, \widehat{Y}_{i+1}, \theta) \; \propto \; p(\widehat{Y_i}|\widehat{Y}_{i-1}, \theta) \, p(\widehat{Y}_{i+1}|\widehat{Y_i}, \theta) \tag{32}$$

Both conditional discretized densities on the right hand side are simple Gaussians given θ. Eraker characterizes (32) for a number of underlying processes. For constant drift and diffusion, we obtain $\widehat{Y_i} \sim N(\frac{1}{2}(\widehat{Y}_{i-1} + \widehat{Y}_{i+1}), \frac{1}{2}\frac{\sigma^2}{m})$, from which one can draw directly. For other processes, Eraker uses this density as the proposal for a Metropolis-Hastings draw.

As the number of missing data increases, one converges to the continuous time model, however as one converges to continuous time, the MCMC algorithm slows down due to the larger number of observations. Eraker (2001) shows by simulation that the algorithm still works fairly well even with 20 filled in data. For actual data, posterior densities stabilize quickly as one keeps filling in missing data. Eraker (2001) estimates a CEV model with 2288 weekly T=bill yields from 1954 to 1997. He finds that the three parameter posterior densities are unchanged after filling-in 4 missing data between each observed data. Remarkably, his posterior mean for the CEV parameter is 0.76. He then estimates a SV model with CEV, and again, the parameter posterior densities only require 4 filled-in data to stabilize. An estimation with 8 filled-in data does not show any change in the posteriors. See also Elerian et al. (2001) who study the diffusion case.

Jones (2003a), with a very similar approach, estimates the drift of a continuous time model for the short rate. He also finds that posterior distributions stabilize after the introduction of a few missing data. His posterior characterization of the drift is very different from that obtained with maximum likelihood analysis.

He incorporates this approach in Jones (2003b), where he examines the ability of generalized CEV models of stochastic volatility to generate the needed features of the conditional returns distribution in periods of high variance. An interesting feature of Jones (2003b) is that he incorporates the information on volatility contained in option prices. First, he notes that expected average variance over the remaining life of an option is approximately linear in current variance for the processes used: $E^Q[V_{t,T}] \approx A + BV_t$. The coefficients A, B are known parametric functions of the variance process and the price of volatility risk. Second, he notes that Black-Scholes implied volatility is a proxy for $E^Q[V_{t,T}]$, the better when the option is close to the money. He, therefore, incorporates into the modeling the following link:

$$IV_t = A + BV_t + \epsilon_t, \quad \epsilon_t \sim N(0, \xi^2 V_t^2),$$

where IV_t is obtained from the VIX index. The ensuing MCMC algorithm requires a non-trivial draw for most blocks, because the posterior distributions are complicated by the link between implied and current volatilities, and the fact that A, B are non-linear functions of the process parameters. This is, however, an interesting way to bring in information from the option prices into the estimation of the volatility process, without estimating option prices themselves.

5.4 Jumps

Chib, Nardari, Shephard (2002) model an additive jump process in the returns equation of the discrete time SV model. This can be seen as an alternative to the fat-tailed conditional returns seen in section 5.2.2.

Eraker et al. (2003), hereafter EJP, compare SV models with additive jump components in both the returns and variance equations. They discretize the continuous time models for daily US index returns, and do not implement the improvement discussed in section 5.3, making the case that the discretization bias is small for these daily returns. EJP start from existing results in the literature, that show how SV models with jumps in returns do not account well for the features of historical returns or option prices. They argue that jumps in volatility create a dynamic for both returns and volatility that is far different from added diffusion factors or jumps in returns. EJP consider several models nested within the general (discretized) specification:

$$\begin{aligned} R_{t+1} &= \mu + \sqrt{V_t}\epsilon^r_{t+1} + \xi^r_{t+1} J^r_{t+1} \\ V_{t+1} &= \kappa\theta + (1-\kappa)V_t + \sigma_v \epsilon^v_{t+1} + \xi^v_{t+1} J^v_{t+1} \end{aligned}$$

The jump density is modeled as $J_t \xi_t$, where a jump occurs when $J_t = 1$. The jump intensities are λ^r, λ^v. Relative to the SV model, the parameter domain is extended to the vectors of jump states J^v, J^r and the jump intensity parameters λ^r, λ^v. EJP distinguish the following nested models: SV is the basic stochastic volatility model; SVJ adds the jumps in returns $\xi^r \sim N(\mu_r, \sigma_r)$; SVJC allows for correlated jumps; $\xi^v \sim \exp(\mu_v)$ for volatility; and $\xi^r|\xi^v \sim N(\mu_r + \rho_J \xi^v, \sigma_r^2)$ for returns. In SVIJ, the jumps are independent.

Bayesian inference allows the use of priors to impose constraints on the parameter space, without which the likelihood could be unbounded. The useful priors are those of jump size, intensity, and volatility. Here the prior is used to model large and rare movements in returns, with a low λ^r and a large σ^r. EJP's prior places low probability of jump standard deviation below 1%, and on more than a 10% chance of a daily jump. The priors of the other parameters are left uninformative. The MCMC algorithm draws iteratively from the following blocks of the posterior density of parameters and state variables:

$$\begin{aligned} \text{1 parameters: } & p(\theta_i|\theta_{-i}, J, \xi, V, R) \;\; i = 1, .., K \\ \text{2 jump times: } & p(J_t = 1|\theta, \xi, V, R), \;\; t = 1, .., T \\ \text{3 jump sizes: } & p(\xi^r|\theta, \xi^v, J, V, R) \\ : \; & p(\xi^r|\theta, \xi^v, J, V, R) \\ \text{4 volatilities: } & p(V|\theta, J, \xi, R) \end{aligned}$$

Blocks $1, 2, 3$, the extensions of the basic SV model, can be drawn directly. The ability to analyze models of this level of complexity testifies to the flexibility of the hierarchical formulation used together with Bayesian MCMC methods. If there are concerns that the data may contain little information about some features of such a complex model, recall that the posterior draws will reflect simply reflect this uncertainty. This is in contrast with the potential computational difficulties in attempting to numerically maximize the likelihood (unavailable analytically) of such a model. See Jacquier et al. (2007) for a maximum likelihood approach that exploits the simplicity of the Bayesian MCMC algorithm.

EJP also compute odds ratios for the various jump extensions. They avoid the computation of the marginal likelihood, which is not readily available for these complex models. Instead, they are able to rewrite the odds ratios in terms of posterior quantities. For example, the odds ratio of SVJ versus SV is a function of the probability that the entire vector of jumps J is equal to zero. The MCMC posterior simulator draws from this known probability, (block 2 above). The odds ratios therefore, involves only a minor additional computation at each step of the MCMC simulator. This extension of the Savage density ratio method is possible because the conditional density of the vectors J^r, J^v is available analytically. For the daily NASDAQ 100 and SP500, the Bayes factors show that the data strongly favor jumps in the volatility rather than in the return equations.

5.5 Estimating covariance matrices and betas

5.5.1 Modeling the covariance matrix

Even when it is not time varying, the estimation of the covariance matrix of a large vector of stock returns poses serious problems. It is full rank mathematically as long as the number of periods T is equal to or larger than N; however, it takes a sample far larger than N to obtain sufficient information on, for instance, the smallest Eigen value. Optimal portfolio weights are often functions of the inverse of the covariance matrix; therefore, in small sample, a lot of uncertainty on the smaller Eigen value will result in possible instability for the inverse of the matrix, in turn, affecting the optimal portfolio weights.

The factor model is a unique way to reduce dimensionality. It is also tightly related to finance modeling, e.g., the APT. When factors are latent, Gibbs sampling makes it possible to draw the factors and their loadings, as shown in Geweke and Zhou (1995). The constraint is effective when the residual covariance matrix is assumed to be diagonal. The factor model is a very effective way to constrain the covariance matrix since it replaces $N(N+1)/2$ parameters by $K(K+1)/2 + (N+1)K$. Aguilar and West (2000) implement an algorithm, suggested in Jacquier et al. (1995), that extends the factor model to allow for stochastic volatility in the factors. See also Chib et al. (2006) for discussions of multivariate SV algorithms.

Another approach is the variance / correlation decomposition, where the covariance matrix is written as $D^{0.5}CD^{0.5}$, (see Barnard et al. (2000) and Engle's (2000) DCC). The individual variances in D can follow univariate GARCH or SV models. The correlation matrix can then be modeled separately from the variances, perhaps with regimes. Regime switching models are conveniently estimated by Bayesian methods, most of the time requiring no more than direct Gibbs draws. See McCulloch and Tsay (1993) and Ghysels et al. (1998) for Bayesian estimation of univariate switching models in means and variances.

5.5.2 Modeling betas

Recall the factor model from the APT section. Geweke and Zhou (1996), Mc Culloch and Rossi (1990, 1991), and many others, assume that B, the matrix of factor loadings, is constant. It may, indeed, be too much to ask from the data, in latent variables models, to deliver precise inference on both time-varying betas and factor scores.

In many cases, however, the factors are considered observable, and it may become practical to allow for time-varing betas. In fact, modern intertemporal asset-pricing models imply that betas time vary and are related to economy-wide or firm-specific state variables. Empirical work has so far mostly used basic rolling window filters to estimate betas. Also, even if one maintains that firm betas are constant, this assumption may be less tenable if one studies managed portfolios. Finally, with very few observable factors, time-varying βs may provide a more flexible specification for a time-varying covariance matrix. Jostova and Philipov (2005) use a MCMC algorithm to draw the betas, considered as unobservable latent state variables.

Cosemans et al. (2009) design a Bayesian estimator of β's that combines the cross-section and the time series. Within this framework they show that the fit is improved by the cross-section, and the resulting out-of-sample forecasts improve portfolio performance. Cosemans et al. is a very interesting example where the simultaneous use of the time-series and the cross-section yields predictive improvements. Their criterion is economic-based, as they assess competing models through their impact on optimal portfolio design.

6 Bayesian Inference in Option Pricing

Option prices depend on a number of factors. Some are known, such as the strike price and the time to maturity, and some are assumed to be known in most models, e.g., interest rates, future dividend yields. Volatility, assumed to be known to the investor in the Black-Scholes and other earlier models, is unknown to the econometrician. Volatility over the remaining life of the option is assumed to be unknown to the investor in most modern models. Modeling the uncertainty in volatility is, therefore, a crucial aspect of the econometrics option pricing.

Econometric methods in option pricing take three main approaches. First, one can obtain information from the historical return process and use it to infer the option price; the likelihood comes from the historical return and variance process. Second, one can use option prices to draw inference directly on the risk-neutral process. In this case, the likelihood is formed from the pricing errors. This approach generally uses panels of options spanning a range of moneyness and maturity. A third segment of the literature attempts to combine the historical and pricing information in the likelihood. In this case, it is generally assumed that the historical and risk-neutral processes, of volatility for example, are of the same family, differing only by a drift shift due to the price of volatility risk.

Another strategic choice to be made is how to compute the option price itself. A relatively easily computed, semi-analytical option-pricing formula may be available, based on the model parameters, in the simpler cases, for example with deterministic volatility. In the more complex cases, the flexibility of the Monte Carlo method pays off. It extends the risk-neutral pricing methodology where the option price of a call of maturity T and exercise price K is:

$$C_T = e^{-r_f(T-t)} E_t^Q \left[Max(S_T - K, 0) \right] \tag{33}$$

where r_f is the appropriate risk free rate, and E^Q is the risk-neutral expectation (see Cox and Ross (1976)). For example, one simulates from the predictive density of the state variables, such as volatility, until maturity, each time computing the option payoff at maturity. The average of these discounted payoffs is the Monte Carlo estimate of the option price. This approach is generally referred to as predictive option pricing. This predictive approach is more effective for the latter class of option pricing models, especially with stochastic volatility or jumps. The Bayesian implementation of (33) naturally incorporate in the predictive draws the posterior uncertainty on the parameter. This is in contrast with a conditional predictive implementation of (33) which would condition on a value of the parameters.

6.1 Likelihood based on underlying return only

Early empirical practice has been to substitute point estimates of σ, either historical or implied, into the Black-Scholes formula. However, even the simplest option pricing model is a non-linear function of σ. Due to the non-linearity and the uncertainty in σ, the substitution of a point estimate into the option formula may lead to biases, in addition to failing to reflect the effect of the uncertainty of σ on the price. Karolyi (1993), the first Bayesian empirical option paper, addresses both issues. It adopts an empirical Bayes approach to reduce the uncertainty in σ_i, which is estimated from the history of the underlying stock return R_i, itself assumed to be log-normally distributed (μ, σ_i) in the model. Precision comes from cross-sectional shrinkage: Karolyi chooses a common conjugate prior for σ_i with location parameter τ and dispersion parameter ν, (see Zellner (1971) appendix B). This yields the posterior density for σ_i given the history of underlying returns R_i:

$$p(\sigma_i|R_i) \propto \left(\frac{1}{\sigma_i^2}\right)^{\frac{\nu_i+\nu+2}{2}+1} \exp\left(-\frac{\nu_i s_i^2 + \nu\tau}{2\sigma_i^2}\right),$$

where $\nu_i = T_i - 1$, T_i is the sample size for the returns of asset i, and $\nu_i s_i^2$ is the sum of squared deviations of the R_i's from their sample mean. He obtains τ by the method of moments, where τ is essentially an average of the individual sample variances. Given τ, he assumes that the individual returns series R_i's are uncorrelated, and obtains ν by maximizing the log-likelihood $\log L(\nu, \tau)$. It can be seen that ν increases as the cross-sectional dispersion of the sample estimates of variance s_i^2 decreases. Here, empirical Bayes is used to obtain tighter posterior densities for the σ_i's by shrinkage.

Karolyi then makes points out that substituting a point estimate of σ_i into the Black-Scholes formula is inappropriate. He computes the Black-Scholes price as the expectation of Black-Scholes prices over the posterior density of σ_i, using the Monte Carlo average of the draws as the Black-Scholes estimate. This approach takes into account the non-linearity of the Black-Scholes formula with respect to σ. Note that Karolyi's approach is consistent with a model with stochastic volatility and no premium for volatility risk. In such a case, the option price can be shown to be the expectation of the Black-Scholes price over the distribution of unknown volatility, (see Hull and White (1987)). Volatility is stochastic for Karolyi because he only observes its posterior distribution.

This Bayesian Monte-Carlo approach extends to draw from the predictive density of volatility over the maturity horizon of the option, for more general option pricing models. In fact, in Hull and White (1987), volatility follows an AR(1). A Bayesian implementation by

MCMC will make posterior draws of the model parameters and in-sample volatilities, and will, for each draw, make a predictive draw of volatility up to the maturity of the option. Each draw will then yield a corresponding draw of the option price. The MCMC estimate of the Hull and White price is the average of these draws. The method generalizes to the Heston (1993) model where volatility and return shocks are correlated. In that case, one needs to make joint draws of future returns and volatilities, which is a minor increase in complexity of the algorithm. The option price computed for each draw makes use of the final underlying asset value for each MC draw.

6.2 Risk-neutral predictive option pricing

When the option-pricing model has a non-zero price of volatility risk, the predictive method must take this into account, by simulating from the risk-neutral predictive density rather than the historical-based density. Bauwens and Lubrano (2002), hereafter BL, price options with risk-neutral GARCH volatility forecasts. Their risk-neutral process is given by:

$$\begin{aligned} r_t &= r_f + v_t; \quad v_t \sim N(0, h_t), \\ h_t &= \omega + \alpha(v_{t-1} - \mu_{t-1} + r_f)^2 + \beta h_{t-1}, \end{aligned} \tag{34}$$

where returns earn the risk-free rate r_f. Consequently, the squared error in the GARCH equation is modified. BL use $\mu_t = \mu + \rho r_{t-1}$ for the historical return conditional expectation. It is important to note that this risk-neutral process, while theoretically motivated, will not be estimated through option prices, but from the historical returns.

The starting point in BL's analysis is a Bayesian estimation of GARCH models from returns data, (see also section 5.1 in this chapter). They report that the posterior means of the GARCH parameters imply considerably less persistence than the ML estimates. Being closer to the boundary of the parameter space, the ML estimates, especially their standard errors, are unreliable.

BL implement a Bayesian predictive pricing as per (33). For a maturity of K, this requires drawing returns $r_{t+k} \sim N(r, \omega + \alpha(r_{t+k-1} - \mu_{t-1}) + \beta h_{t+k-1})$, and computing the GARCH risk-neutral volatility h_{t+k}, for $k \in [1, K]$. These draws of the compounded returns yield prices at maturity that allow the computation of a Monte Carlo estimate of (33). BL make the important point that the convergence of the Monte Carlo estimate to the price occurs if the returns process is stationary; draws in the non-stationary region must, therefore, be rejected. In the GARCH analysis, such draws occur about 2% of the time.

The Bayesian implementation of (33) allows for the integration of parameter uncertainty because each new predictive draw is made given a new draw of the posterior distribution of the parameters, here $(\rho, \alpha, \beta, \mu)$. This is in contrast with alternative approaches that would condition on a point estimate of the parameters. BL make N posterior draws and M predictive draws for each posterior draw. They justify it by the computational cost of a posterior relative to a predictive draw. However, with the computational power now available, one can as well set $M = 1$ and N large (see Geweke (1989)).

There is, however, a distinction between posterior and predictive draws. As per most models, the agent is assumed to know the parameters of the volatility process. Consequently, the N predictive draws reflect the Monte Carlo implementation of (33), but not an uncertainty about the option price. The price *is* the expectation of the payoff along this predictive density. The only use of the spread of these N predictive draws would be to make sure that the Monte Carlo estimate has the desired precision. In contrast, the econometrician does not know the parameters, and draws from the predictive density by mixing the posterior draws of the parameters. Due to the posterior uncertainty, the econometrician faces an option price uncertainty, which she could want to document. Then, setting $M > 1$, finding the option price by Monte-Carlo averaging over N predictive draws for each posterior draw, would allow one to characterize the posterior uncertainty.

Eraker et al. (2003), discussed in section 5.4, decompose the uncertainty on simulated option prices into their predictive and posterior component, but in a different manner. They characterize posterior uncertainty by conditioning the draws on the mean of predictive volatility, and volatility uncertainty by conditioning the draws on the posterior mean of the parameters. Their Figure 8 shows that parameter uncertainty has the largest impact for the longer term options, and for extreme moneyness in shorter term options.

6.3 Likelihood based upon pricing errors

6.3.1 Deterministic volatility functions

In the Black-Scholes model, the return standard deviation σ is assumed to be constant, and is known to the investor, who observes prices in continuous time. The theoretical price is obtained by a no-arbitrage argument, whereby the option can be exactly replicated, in continuous time, by a hedge portfolio of the underlying asset and a riskless bill. A family of extensions of the Black-Scholes model allow σ to vary as a deterministic function of the underlying price. This preserves the possibility of forming a hedge portfolio, since σ is still perfectly predictable conditional on the underlying asset price. These are still pure no-arbitrage option pricing models, where any deviation between model and market price can only be an arbitrage possibility. For example, Rubinstein (1995) shows how to fit flexible deterministic volatility functions of the stock price $\sigma = f(S)$ to panels of options, with trees, in a manner consistent with no-arbitrage. This is, however, viewed as over-fitting by the econometrician; using this method, Dumas et al. (1998) report large out-of-sample pricing errors contrasting with quasi-perfect in-sample fit.

One clearly needs to explicitly allow for model error to obtain a likelihood function from option price data. This is done in Jacquier and Jarrow (2000) for the Black-Scholes model and its deterministic extensions. The pricing error yields the likelihood function and, in turn the posterior distribution of option parameters and prices. Jacquier and Jarrow model option prices as:

$$\log C_i = \beta_1 \log b_i(x_{1i}, \sigma) + \beta_2 x_{2i} + \eta_i, \quad \eta_i \sim N(0, \sigma_\eta), \tag{35}$$

where (x_1, x_2) are known data including the stock price. The prior density for σ comes from the historical data observed prior to the panel of options. It, therefore, incorporates information from the historical returns, but not in the likelihood function. The logarithm formulation eliminates the potential for negative prices while keeping a tractable likelihood.

Jacquier and Jarrow produce the posterior densities of $\beta_1, \beta_2, \sigma, \sigma_\eta$, where σ_η is the standard deviation of the pricing error. They break down the joint posterior density into two Gibbs steps. First, one draws directly from $p(\beta, \sigma_\eta|\sigma, C)$, as (35) is a linear regression given σ. Second, $p(\sigma|\beta, \sigma_\eta)$ is obtained by a univariate Hastings Metropolis step, using a truncated normal as proposal density. Jacquier and Jarrow implement the model on panels of individual US stock options.

The results show that the econometric specification has a great impact, in a way the model itself can not suggest. Allowing for heteroskedasticity in the error η_i greatly improves the pricing performance. Unlike the calibration methods used in Dumas et al., the likelihood based approach allows the econometrician to assess in-sample potential problems with the model and its implementation. The posterior distributions of σ_η, β_2 give clear warning of the potential ineffectiveness of the extensions as the dimension of x_2 increases. As more variables are added in x_2, the posterior mean of σ_η decreases for out-of-the money options, but does not improve for the other options. At the same time, the spread of the posterior distribution of the other parameters (β, σ) increases drastically.

A joint draw from the parameters yields a draw for the model price, model prediction, and hedge ratio. The uncertainty in prices and hedge ratios increases with the model size and becomes quite large. Consequently, even before engaging into prediction, the Bayesian econometrician knows that the larger models may not be effective. Having introduced pricing error, the econometrician faces an uncertain option price, which is at odds with the theoretical foundation of the model. We do not address this tension, the resolution of which goes well beyond the scope of this chapter. It involves theoretical modeling, no doubt resulting in a model more complex than the one being implemented. From the econometric viewpoint, we believe that the benefit of incorporating the imperfection of the model at hand far outweighs the cost of the contradiction with the model being implemented.

Jacquier and Jarrow then show that the out-of-sample performance of the extended models is not superior to the basic Black-Scholes model. However, the decrease in performance from in- to out-of sample is not catastrophic as in Dumas et al. This is because, in contrast to the fitting criteria used in Dumas et al., the Bayesian method does not explicitly find parameters to minimize in-sample pricing errors. Such methods will exacerbate the potential for over-fiting, and are more likely to result in seriously degraded out-of-sample performance. The Bayesian method, in a first step, produces the uncertainty in the posterior density of parameters, and functions such as the option price. One expects therefore its out-of-sample performance to be more robust to over-fitting.

We conclude with two observations. First, in the formulation in (35) one can use the Savage density ratio method to compute model odds since direct draws of β are available. Second, the formulation in (35) would be more effective if the extensions were modeled on σ itself, rather than outside of the Black-Scholes model, as for example in $C_i = BS(x_{1,i}, \sigma(\beta_2 x_{2,i} + \eta_i))$. This parameterization would be more consistent with most generalizations of the Black-Scholes model. Having the error inside the Black-Scholes formula would guarantee the no-arbitrage conditions. It would, however, lead to a more complicated MCMC algorithm.

6.3.2 Option pricing with stochastic volatility and jumps

Eraker et al. (2003), discussed in section 5.4, estimate historical processes allowing for jumps in volatility as well as in returns. Using odds ratios and the fraction of volatility explained by jumps and diffusion, they conclude that jumps in volatility are a crucial extension to the SV model. Then they adjust their historical densities for a plausible price of volatility risk to generate option prices as per (38). They conclude that jumps in volatility have the potential to create a very realistic smile in implied volatilities.

Eraker et al. (2003) do not estimate the price of risk from option data; this is done in Eraker (2004) who derives the likelihood from explicit option pricing errors, as Jacquier and Jarrow (2000), but for the vastly more complex model of Eraker et al.(2003). Recall the process on the stock price S and its volatility V :

$$\begin{aligned} dS_t &= aS_t dt + \sqrt{V_t} S_t dw_t^s + S_t dJ_t^s \\ dV_t &= \kappa(\theta - V_t)dt + \sigma_v \sqrt{V_t} dw_t^v + dJ_t^v, \end{aligned} \tag{36}$$

where the volatility and return shocks can be correlated, and both dS and dV can have jumps. To the SV, SVJ, and SVCJ models seen in section 5.4, Eraker (2004) adds a model in which the jump intensity depends on volatility, i.e., $\lambda = \lambda_0 + \lambda_1 V_t$. He uses the discretized versions of these models. Jump components Z are assumed to be exponential for variance and normally distributed for stocks. Consistent with the option pricing literature, the risk neutral processes corresponding to the above historical processes incorporate drift adjustments for the prices of jump and volatility risk. Specifically, the return and volatility drifts, and the mean jump are adjusted for these prices.

Given the option prices Y_t with known characteristics χ_t; the model prices these options with error ϵ_t:

$$Y_t = F(S_t, V_t, \chi_t, \Theta) + \epsilon_t, \tag{37}$$

where Θ is the vector of parameters of the risk-neutral processes of S_t, V_t. Times between transactions can vary. To compute the price F, Eraker uses Fourier inversion methods available due to the affine structure of the model, (see Duffie, Pan and Singleton (2000)).

Equation (37) yields the density of the options data Y conditional on the relevant parameters, observable inputs, and state variables V and S needed for pricing, $p(Y|S, V, \Theta)$. However, these state variables must be integrated out in order to obtain the posterior densities of the parameters and option prices. The formulation of S, V in (36) is conditional on the state vectors of jumps J; those must also, therefore, be integrated out. Consequently, the joint density of options data and all state variables $p(Y, S, V, J, Z, \Theta)$ involves the parameters of the process in (36). This joint density multiplied by the priors on all parameters yields the desired posterior density $p(V, J, Z, \Theta|Y)$. Eraker breaks this posterior into MCMC conditional blocks of volatilities V, jump states J , jump components Z, and parameters Θ.

Eraker estimates the model with 3000 option prices on S&P500 contracts recorded over 1000 days from 1987 to 1990. Due to the computational demands of the model, he uses a small number of randomly selected contracts daily, on average three contracts per day. At the same time the underlying return is recorded daily. The posterior analysis reveals the following. The jump-size parameters are hard to estimate precisely, because option data do not contain information about them. Given the posterior jump intensity which implies very rare jumps, 2 or 3 per 1000 days, one clearly can not have much information in jump size. In contrast, volatility parameters are very precisely estimated. This in turn helps the state-dependent model, in which the jump intensity is linked to the current volatility. Eraker concludes that the jump in volatility dominates the other extensions to the SV model in terms of explaining returns. For these models, return and volatility jumps are negatively correlated. That is, a negative jump in returns is associated with a positive jump in volatility.

The posterior mean of the option pricing function F in (37) is used as the estimate of the option price. Based on this, the in-sample pricing errors of the larger models do not show much improvement over the simpler ones. Eraker notes that this result is in contrast with most of the previous literature. Recall that, unlike least squares, the Bayesian method does not compute estimates to minimize pricing errors. It optimally describes the uncertainty about the parameters; the posterior mean then minimizes a squared error loss along that posterior uncertainty. In contrast, least squares methods are geared at fitting better with larger models. This is why an out-of-sample analysis is a very interesting complement. Eraker's out-of-sample result reveal some performance for the larger jump model but the results are at best mixed.

Overall, the larger models in Eraker (2004) appear to perform better with respect to features of the time series of the stock return than pricing options. Eraker (2004) is an example of the degree of complexity that can be handled by Bayesian MCMC algorithms. It would be interesting to revisit the features of the model with a larger cross-section of options data as computing power allows. Using a small panel of randomly selected data means that only very rough information on the smile is available at any given time, possibly affecting the precision of estimation. Inference on rare and large jumps is definitely a difficult problem, to which option prices do not contribute much information.

7 Particle Filters with Parameter Learning

We conclude the chapter with a methodological approach that appears very promising for Bayesian econometrics in finance: the joint filtering of the latent state variables parameters.

MCMC methods for models with latent variables generally produce the posterior density of the parameters and of the smoothed state variable. For example, the MCMC algorithms for the SV models discussed in section 5, produce $p(h_t|y^T)$ and $p(\theta|y^T)$, where $y^T \equiv (y_1, \ldots, y_T)$, and θ, h_t are the model parameters and the variance at time t. For a given sample of data, however, one may want, for each time t in the sample, the posterior density of both filtered volatilities and parameters $p(h_t|y^t)$ and $p(\theta|y^t)$. Running again the MCMC sampler each time a new observation (y^{t+1}) becomes available, is a feasible but computationally unattractive solution. Recent research has, therefore, been devoted to filtering algorithms for non-linear state space models. Early filtering algorithms solve the problem conditional on a value of θ. This is unattractive for two reasons. First, they do not incorporate the uncertainty on θ into the predictive density of h_t. Second, the most likely value of θ on which to condition, comes from the posterior distribution of a single MCMC algorithm run on the whole sample. However, conditioning on the information from the entire sample is precisely what one wants to avoid when drawing from $p(h_t|y^t)$. Incorporating learning about θ in the filtering algorithm turned out to be quite difficult. Let x_t be the state variable in a general model; earlier attempts to draw from $p(\theta, x_t|y^t)$ suffered from degeneracy problems. Section 8 of the Kohn chapter discusses these earlier methods. Carvalho et al. (2010), hereafter CJLP, resolve the problem. In this section, we briefly outline their method and present an application demonstrating the potential of particle filtering with parameter learning.

7.1 Methodology

Consider a model with observable y_t and latent state variable x_t. The goal is to update the current distribution $p(x_t, \theta|y^t)$ to $p(x_{t+1}, \theta|y^{t+1})$ after observing y_{t+1}. For notational convenience, ignore in a first step the parameter θ. Classic filtering algorithms proceed by first predicting and then updating, as follows:

$$\begin{aligned} p(x_{t+1}|y^t) &= \int p(x_{t+1}|x_t)p(x_t|y^t)dx_t \\ p(x_{t+1}|y^{t+1}) &\propto p(y_{t+1}|x_{t+1})p(x_{t+1}|y^t). \end{aligned}$$

The distribution $p(x_t|y^t)$ is usually not known analytically. Particle filters approximate it by a discrete density $p^N(x_t|y^t)$, consisting of N *particles*, or draws, $x_t^{(i)}$ with weights $w_t^{(i)}$:

$$p^N\left(x_t|y^t\right) = \sum_{i=1}^{N} w_t^{(i)} \delta_{x_t^{(i)}} \rightarrow p(x_t|y^t) \text{ as } N \rightarrow \infty,$$

where δ is the Dirac function. In earlier algorithms, the weights are typically $1/N$. This discretization allows us to replace the integral in the prediction step with a sum. We now have:

$$p^N(x_{t+1}|y^{t+1}) \propto \sum_{i=1}^{N} p(y_{t+1}|x_{t+1})p(x_{t+1}|x_t^{(i)})w_t^{(i)},$$

where $p^N(x_{t+1}|y^{t+1})$ is a finite mixture. There are several classic particle filters to draw from this mixture, such as the exact, the sampling importance resampling (SIR), and the auxiliary particle (APF) filters. The SIR algorithm, for example, relies only on two steps given N samples from $p^N\left(x_t|y^t\right)$:

1. (Propagate) Draw $x_{t+1}^{(i)} \sim p\left(x_{t+1}|x_t^{(i)}\right)$ for $i = 1, ..., N$;
2. (Resample) Draw $x_{t+1}^{(i)} \sim Mult_N\left(\left\{w_{t+1}^{(i)}\right\}_{i=1}^{N}\right)$.

Note how the first step uses the transition density of the state variable, but no information about y_{t+1}. The transition density can be seen as the simplest and most convenient importance density to use in the propagation step. The weights are then based upon the information in y_{t+1}, and a multinomial draw is made from the N x_{t+1}'s obtained from the propagation step. In their auxiliary particle filter, Pitt and Shephard (1999) improve the importance density, and show that the optimal weights are $w_{t+1} \propto p(y_{t+1}|x_t)$. Possible problems with particle filters include degeneracy or sample impoverishment, where the number of particles from which one draws degenerates due to the inability of the transition density to cover the high probability states. This can happen, for example, around extreme values, if the propagation step does not use the information in y_{t+1}. Then the resampling step leads to sample impoverishment as most draws have nearly zero weights. See the Kohn chapter and CJLP for details.

Consider now extending this framework to parameter learning, where one needs to move from $p(\theta, x_t|y^t)$ to $p(\theta, x_{t+1}|y^{t+1})$. CJLP's algorithm combines two contributions. First, they reverse the prediction-updating order, and instead follow a smoothing-prediction sequence:

$$\begin{aligned} p(x_t|y^{t+1}) &\propto p(y_{t+1}|x_t)p(x_t|y^t); \\ p(x_{t+1}|y^{t+1}) &= \int p(x_{t+1}|x_t)p(x_t|y^{t+1})dx_t. \end{aligned}$$

For the discretized distribution used, this sequence leads to a resample-propagate particle algorithm:

1. Resample particles with weights $w_t^{(i)} \propto p\left(y_{t+1}|x_t^{(i)}\right)$:

   Draw an index $z(i) \sim Mult_N\left(\{w_t^{(i)}\}\right)$, and set $x_t^{(i)} = x_t^{z(i)}$ for $i = 1, \ldots, N$;

2. Propagate state particles with $x_{t+1}^{(i)} \sim p\left(x_{t+1}|x_t^{(i)}, y_{t+1}\right)$, for $i = 1, \ldots, N$.

By resampling first, the compounding of approximation errors is reduced because the propagation of the states uses the information in y_{t+1}. The second contribution addresses the issue of parameter learning. Rather than attempting to update $p(\theta|y^t)$ directly, CJLP extend the state with a vector of conditional sufficient statistics s_t, and instead update $p(s_t|y^t)$. The sufficient statistics satisfy the conditions:

$$p(\theta|x^t, y^t) = p(\theta|s_t) \text{ where } s_{t+1} = \mathcal{S}\left(s_t, x_{t+1}, y_{t+1}\right).$$

CJLP show how to use particle methods to find the joint filtering distribution $p(x_t, s_t|y^t)$; the parameter θ is then simulated by $\theta^{(i)} \sim p(\theta|s_t)$. The sufficient statistics s_t are essentially the parameters of the posterior distribution $p(\theta|y_t)$.

Given a particle filtering algorithm with parameter learning, dynamic model comparison or averaging can be performed at each time t via sequential odds ratios. A Bayesian investor can then evaluate the economic benefits of predicting out-of-sample returns by learning about the models, parameters, and state variables sequentially in real time. Bayes rule naturally revises beliefs as new return data is available.

7.2 Incorporating realized volatility into the SV model

We now discuss an application to the SV models. The MCMC algorithms seen in section 5.2 produce draws from $p(h_t|y^T)$, the smoothed posterior density of the volatilities, which uses the entire information in the sample. This smoothed density coincides with the filtered volatility, $p(h_t|\mathbf{y}^t)$, only for the last observation of the sample. An MCMC algorithm, such as in JPR, makes 50,000 draws in 4 minutes for $T = 1500$ observations on a dual-core 2.8 Ghz CPU. Computing draws of the filtered posterior densities $p(h_t, \omega_t|\mathbf{y}^t)$ by running the MCMC algorithm for every subsample y^t for $t \in [500, 1500]$ would require about 45 hours of CPU time. This is not an issue for academics, but practitioners who may want to update many such models everyday would find themselves a bit more pressed for time. Another reason why the filtered density of the latent variables is of interest is that they allow for the computation of the likelihood function. This is convenient if one can not write posterior odds directly through the MCMC algorithm as in section 5.

Jacquier and Miller (2010), JM hereafter, apply both the MCMC and the CJLP algorithm to an extended SV model for the logarithm of variance h_t:

$$\log h_t = \alpha + \delta \log h_{t-1} + \gamma RV_{t-1} + \sigma_v v_t, \tag{38}$$

where RV_{t-1} is a realized volatility measure (see also Brandt and Jones (2005)). The addition of exogenous variables to the volatility equation is a technically simple but potentially very useful extension of the SV model. Other variables of interest can be incorporated in the SV equation, such as implied volatility or the number of non-trading days between observations.

The basic RV measure is computed as the sum of squared intra-day returns: $RV_t = \sum_{j=1}^{m} r_{t,j}^2$. Under ideal assumptions, it is shown to converge to the day's integrated volatility $IV_t = \int_{t-1}^{t} \sigma_\tau^2 d\tau$, as m goes to infinity. However, measurement errors in prices, microstructure effects, and the possibility of jumps have to be taken into account. Therefore, several variations of this basic realized volatility now exist to address these issues (see for example Patton (2008)). The realized volatility literature typically attempts to evaluate these competing measures by their ability to predict integrated volatility, IV_{t+1}. However, since IV_{t+1} is never exactly known, it is typically replaced by a RV_{t+1} measure in predictability regressions.

In contrast, equation (38) takes the view that the daily volatility h_t is the latent variable to be predicted, and that RV_{t-1} is only an observable with information on this latent variable, not the object to be predicted. JM conduct simulations to document what reduction in volatility uncertainty can be expected by incorporating RV_{t-1} in (38). Using the root-mean-squared error of the posterior mean, they show that, for simulated data, RV measures only improve out-of-sample volatility forecasts up to 4 days ahead at the most. They also propose an alternate econometric specification to improve upon (38), which models the fact that RV_t is a noisy estimate of $\log h_t$, and allows for its error η_t to be correlated with v_t. Therefore, instead of having RV_t in the volatility equation as in (38), an additional measurement equation is introduced:

$$\log RV_t = \beta_0 + \beta_1 \log h_t + \eta_t. \tag{39}$$

Competing volatility measures can be introduced, via seemingly unrelated measurement equations, as in (39).

JM apply MCMC to the SV model with and without RV, on the UK Pound, Euro, and Yen daily exchange rate changes, and country index returns over 2006-2009. They find that 90% posterior confidence intervals on $\sqrt{h_t}$ have average widths of 47% on the Pound and 41% on the Euro, relative to $E(\sqrt{h_t})$ when RV is not used. The introduction of RV, as in (38), reduces these to 27% and 32%. However, RV brings no such improvement for the Yen. They report similar improvements due to realized volatility for several country indices, with the exception of the SP500 and the Nikkei.

JM then implement the CJLP algorithm. For these samples of 800 observations, the filtering algorithm with parameter learning requires about 25 minutes of Core-Duo 2.8 Ghz CPU time with 40000 particles. Compared to filtered volatilities, smoothed volatilities benefit from the information contained in future y's; one, therefore, expects the posterior distributions of smoothed volatilities to have a tighter spread than those of filtered volatilities. Figure 6 demonstrates the magnitude of the difference for the British Pound. The top and middle plots show the 90% intervals for the smoothed and filtered volatility densities obtained by MCMC and the CJLP algorithm. The bottom plot demonstrates the evolution of the parameter δ as the filtering algorithm updates its posterior distribution. The filtering algorithm for Figure 6 was run with $N = 40000$ particles.

Figure 6 here

A note of caution is in order with respect to the number of particles used. Jacquier and Miller note that, unless the number of particles is quite large, different runs of the CJLP algorithm can produce very different posterior densities. Consider, for example, the odds ratios for the basic SV model versus the model augmented with realized volatility; which is based on the parameter α_1 in (38). Since both MCMC and the filtering algorithms draw directly from the posterior density of α_1, a simple Savage density ratio can be used.

Figure 7 shows the MCMC odds ratios as a horizontal line, and the CJLP odds ratio updated every period for 5000 to 40000 particles. The MCMC log odds is -35, in favor of the model with realized volatility. The dynamic odds ratio obtained on the last observation from the particle filters should equal the MCMC odds ratios. Figure 7 shows that this only happens when the number of particles used is very large. This is a sign that the posterior distributions in the filtering algorithm may require a very large number of particles to be deemed reliable. Care needs to be exercised when analyzing the output of this type of algorithm. Nevertheless, these algorithms have great potential and are definitely worth exploring.

Figure 7 here

References

Aguilar, O. and M. West (2000). Bayesian Dynamics factor models and portfolio allocation. *Journal of Business and Economics Statistics*, 18, July, 338–357.

Avramov, D., 2002, Stock return predictability and model uncertainty, Journal of Financial Economics 64, 423-458.

Avramov, D., and Wermers, R., 2006, “Investing in Mutual Funds when Returns are Predictable”, Journal of Financial Economics 81, 339–377.

Baks, Klass, Metrick, Andrew, and Jessica Wachter, 2001, Should investors avoid all actively managed mutual funds? A study in Bayesian performance evaluation, Journal of finance 56:45–85.

Barberis, N., 2000. Investing for the Long Run when Returns are Predictable. Journal of finance, 55, 225–264.

Barnard, J., Robert McCulloch and Xiao-Li Meng; 2000, ”Modeling Covariance Matrices in Terms of Standard Deviations and Correlations, with Application to Shrinkage.” Statistica Sinica, 2000, 10(4), 1281–1311.

Bauwens L., and Lubrano M., 1998. Bayesian inference on GARCH models using Gibbs sampler. Econometrics J. v1. 23–46.

Bauwens L., and Lubrano, M., 2002. Bayesian option pricing using asymmetric GARCH models. Journal of Empirical Finance, v9. 321–342.

Bawa, V., Brown S., and Klein, R. 1979 Estimation Risk and Optimal Portfolio Choice. North Holland, Amsterdam.

Berger, J., 1985. Statistical Decision Theory and Bayesian Analysis, SpringerVerlag, New York.

Besag, J. 1989, A Candidate’s Formula: Curious result in Bayesian Prediction, Biometrika, 76, 183.

Black, F. and R. Litterman (1991). Asset Allocation: Combining investor views with market equilibrium. *Journal of Fixed Income* 1, September, 7–18.

Bollerslev, T., R.F. Engle and D.B. Nelson (1994), ARCH Models, in R.F. Engle and D. McFadden (eds.), Handbook of Econometrics, Volume IV, 2959–3038. Amsterdam: North-Holland.

Bossaerts, P., Hillion, P., 1999. Implementing statistical criteria to select return forecasting models: what do we learn? Review of Financial Studies 12, 405-428.

Brandt, Michael W., Portfolio choice problems, in Y. Ait-Sahalia and L.P. Hansen (eds.), Handbook of Financial Econometrics, 2009.

Brandt, Michael W., and Christopher S. Jones, Bayesian Range-Based Estimation of Stochastic Volatility Models, finance Research Letters 2, 2005, 201–209.

Brown, S. Optimal Portfolio Choice Under Uncertainty: A Bayesian Approach. Ph.D, Dissertation, University of Chicago (1976).

Brown, S., 1978, The portfolio choice problem: Comparison of certainty equivalent and optimal Bayes portfolios, Communications in Statistics: Simulation and Computation B7, 321-334.

Campbell, J.Y., A.W. Lo and A.C. MacKinley (1997). The Econometrics of Financial Markets. Princeton University Press.

Carvalho, C., and Johannes, M. and Lopes, H., and Polson, N.G, (2010) Particle learning and smoothing, forthcoming Statistical Science.

Chib, S. (1995), Marginal likelihood from the Gibbs Output. Journal of the American Statistical Association, 90, 432, 1313–1321

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

Chib, S., and Nardari, F., and N. Shephard, 2002, Markov Chain Monte Carlo Methods for Stochastic Volatility Models, Journal of Econometrics, 108, 281–316.

Chib, S., and Nardari, F., and N. Shephard, 2006, Analysis of High Dimensional Multivariate Stochastic Volatility Models, Journal of Econometrics, 134, 341–371.

Connolly, R.A. (1991). A Posterior odds analysis of the weekend effect. Journal of Econometrics, 49, 51–104.

Connor, G., and R. A. Korajczyk, 1986, Performance measurement with the arbitrage pricing theory: A new framework for analysis, Journal of Financial Economics 15, 373–394.

Cosemans, Mathijs, Frehen, Rik G. P., Schotman, Peter C. and Bauer, Rob, (2009) Efficient Estimation of Firm-Specific Betas and its Benefits for Asset Pricing Tests and Portfolio Choice; SSRN: http://ssrn.com/abstract=1342326

Cox, J.C., Ross, S.A. 1976. The valuation of options for alternative stochastic processes, Journal of Financial Economics 3, 145–166.

Cremers, M., 2002. Stock return Predictability: a Bayesian model selection procedure. Review of Financial Studies, 15, 1223–1249.

DeMiguel, V., Garlappi, L., and Uppal, R. 2009 "Optimal versus Naive Diversification: How inefficient is the 1/N Portfolio Strategy." Review of Financial Study, 22(6), 2303–2330.

Dickey, J., 1971. The Weighted Likelihood Ratio, Linear Hypotheses on Normal Location Parameters. Annals of Mathematical Statistics 42, 204–224

Dumas, B., Fleming, J., and Whaley, R., 1998 Implied Volatility Functions: Empirical Tests, Journal of finance, 53, 2059–2106.

Dumas, B., and B. Jacquillat, 1990, Performance of Currency Portfolios Chosen by a Bayesian Technique: 1967-1985, Journal of Banking and finance, 14, 539-58.

Elerian, Ola, Sidhartha Chib and Neil Shephard, 2001, Likelihood Inference for Discretely Observed Nonlinear Di?usions, Econometrica 69, 959–994.

Eraker, B. (2001). MCMC Analysis of Diffusion Models with Applications to finance, Journal of Business and Economic Statistics, 19-2, 177–191.

Eraker, B. (2004). Do Stock Prices and Volatility Jump? Reconciling Evidence from Spot and Option Prices, Journal of Finance 59(3), June.

Eraker, B., M. Johannes and N.G. Polson (2003). The Impact of Stochastic Volatility and Jumps in Returns. Journal of finance 58, 1269–1300.

Frost, P.A., and J.E. Savarino, 1986, An empirical Bayes approach to efficient portfolio selection, Journal of Financial and Quantitative Analysis 21, 293-305.

Geweke, John, 1988. The Secular and Cyclical Behavior of Real GDP in 19 OECD Countries, 1957-1983, Journal of Business and Economic Statistics, vol. 6(4), 479–486.

Geweke, John, 1988. Antithetic acceleration of Monte Carlo integration in Bayesian inference, Journal of Econometrics, vol. 38(1-2), 73–89.

Geweke, J. (1989). Exact predictive density for linear models with ARCH disturbances, Journal of Econometrics , 40, 63-86.

Geweke, J. (1993). Bayesian treatment of the student-t linear model, Journal of Applied Econometrics , 8, S19-S40.

Geweke, J. (1994) Bayesian comparison of econometric models. working paper 532, Research Department, Federal Reserve Bank of Minneapolis.

Geweke, J. , 2001, A Note on Some Limitations of CRRA Utility, Economics Letters, 71, 341–346.

Geweke, J., 1992 Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments, in J.O. Berger, J.M. Bernardo, A.P. Dawid, and A.F.M. Smith (eds.), Bayesian Statistics 4, 169–194. Oxford: Oxford University Press.

Geweke, J., and C. Whiteman (2006): Bayesian Forecasting, in Handbook of Economic Forecasting, Volume 1, ed. by G. Elliott, C. W. Granger, and A. Timmermann, pp. 380. North Holland, Amsterdam.

Geweke, J., and Zhou, G. 1995. Measuring the Pricing Error of the Arbitrage Pricing Theory. Review of Financial Studies 9(2), 553-583.

Ghysels, E., McCulloch, R.E., and R.S. Tsay, 1998, "Bayesian Inference for Periodic Regime-Switching Models.", Journal of Applied Econometrics, 13(2), pp. 129–143.

Gibbons, M.R., S.A., Ross, and J. Shanken. 1989. A Test of the Efficiency of a Given Portfolio. Econometrica, vol. 57, no. 5 (September): 1121-1152.

Glosten, L. R., Jagannathan, R. and Runkle, D. E. (1992). On the relation between the expected value and the volatility of the nominal excess return on stocks, Journal of finance

, 48, 1779-1801.

Green, P., 1995, Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika, 82, 711–732.

Harvey, C.R., J.C. Liechty, and M.W. Liechty, 2008, Bayes vs. Resampling: A Rematch, Journal of Investment Management, vol. 6, No. 1.

Harvey, C.R., J.C. Liechty, M.W. Liechty, and P. Mueller, 2010, Portfolio selection with higher moments: A Bayesian Decision Theoretic approach, forthcoming Quantitative Finance.

Harvey, C., and Zhou, G. 1990. Bayesian Inferance in Asset Pricing Tests. Journal of Financial Economics 26, 221-254.

Heston, Steven, 1993, Closed-form solution of options with stochastic volatility with application to bond and currency options, Review of Financial Studies 6, 327-343.

Hull J., and White A. 1987 The Pricing of Options on Assets with Stochastic Volatility, Journal of Finance, 42 (2), 281–300.

Jacquier E., 1991, "Predictability of Long Term Stock Returns and the Business Cycle" PhD. Dissertation, U. of Chicago.

Jacquier, E., 2006, Long-term forecasts of mean returns: Statistical vs. economic rationales, HEC Montreal working paper.

Jacquier, E., and Jarrow, R.A, (2000) "Bayesian Analysis of Contingent Claim Model Error," Journal of Econometrics, 94 (1).

Jacquier, E., and Miller, S., (2010) "The information content of realized volatility", working paper HEC Montreal.

Jacquier, E., Marcus, A., and A., Kane, 2005, Optimal Estimation of the Risk Premium for the Long-Term and Asset Allocation, Journal of Financial Econometrics 3, Winter, 37–56.

Jacquier, E., Polson, N., 2000, Odds Ratios for Non-nested Models: Application to Stochastic Volatility Models. Boston College manuscript.

Jacquier, E., N.G. Polson, and P. Rossi (1994). Bayesian analysis of Stochastic Volatility

Models (with discussion). *Journal of Business and Economic Statistics* ,12, 371–389.

Jacquier, E., N.G. Polson, and P. Rossi (1995), Models and Priors for Multivariate Stochastic Volatility Models, Working paper, University of Chicago.

Jacquier, E., N.G. Polson and P. Rossi (2005). Bayesian Analysis of Stochastic Volatility Models with Fat-Tails and Correlated Errors. Journal of Econometrics, 122, 185–212.

Jacquier, E., M. Johannes and N.G. Polson (2007). MCMC maximum likelihood for latent state models, Journal of Econometrics, Journal of Econometrics 137(2), 615-640

James, W. and Stein, C. (1961), Estimation with Quadratic Loss, in Neyman, J. (ed.), Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, Berkeley: U. of California Press, 361–379.

Jobson, J.D. and Robert Korkie, 1980, Estimation for Markowitz Efficient Portfolios, Journal of the American Statistical Association 75, 544–554.

Jobson, J.D. and Robert Korkie, 1981, Putting Markowtiz theory to work , Journal of Portfolio Management, 70–74, Summer.

Johannes, Polson, Stroud, 2010, Optimal Filtering of Jump-Diffusions: Extracting Latent States from Asset Prices, forthcoming Review of Financial Studies.

Jones, C., 2003, Nonlinear Mean Reversion in the Short-Term Interest Rate, Review of Financial Studies 16: 793–843.

Jones, C., 2003, The Dynamics of Stochastic Volatility: Evidence from Underlying and Options Markets, Journal of Econometrics 116: 181–224.

Jones, C., and J. Shanken, 2005, Mutual Fund Performance with Learning Across Funds. Journal of Financial Economics 78: 507–552.

Jorion, P., 1985, International Portfolio Diversication with Estimation Risk, Journal of Business, 58, 259-278.

Jorion, P., 1986, Bayes-Stein estimation for portfolio analysis, Journal of Financial and Quantitative Analysis 21, 279-292.

Jostova, G., and Philipov A. 2005. Bayesian Analysis of Stochastic Betas, Journal of Financial and Quantitative Analysis 40, No. 4, 747–778.

Kandel, S., McCulloch,R., and R. F. Stambaugh, 1995, Bayesian Inference and Portfolio Efficiency, Review of Financial Studies 8, 1–53.

Kandel, Shmuel, and Robert Stambaugh, 1996, On the Predictability of Stock Returns: An Asset Allocation Perspective, Journal of finance 51, 385–424

Karolyi, A. (1993)A Bayesian Approach to Modeling Stock Return Volatility for Option Valuation. Journal of Financial and Quantitative Analysis 28, December 1993.

Kim S., Shephard N., and Chib, 1998, Stochastic Volatility : Likelihood Inference and Comparison with ARCH Models, Review of Economic Studies 65, 361–393

Kleibergen, F. and H.K. Van Dijk, 1993. Non-stationarity in GARCH models: a Bayesian analysis, Journal of Applied Econometrics, 8, 41–61.

Klein, R.W., and V.S. Bawa, 1976, The effect of estimation risk on optimal portfolio choice, Journal of Financial Economics 3, 215-231.

Lamoureux C., and Guofu Zhou, 1996. "Temporary Components of Stock Returns: What do the Data Tell us? Review of Financial Studies, Winter.

Ley, E., and Steel, M.F.J., 2009, On the Effect of Prior Assumptions in Bayesian Model Averaging With Applications to Growth Regression, Journal of Applied Econometrics, 24(4), 651–674.

Markowitz, H.M., 1952, Portfolio Selection, Journal of finance 7, 77-91.

McCulloch, R. and P.E. Rossi, 1990. Posterior, predictive and utility based approaches to testing arbitrage pricing theory. Journal of Financial Economics 28, 7-38.

McCulloch, R. and P.E. Rossi, 1991, "A Bayesian Approach to Testing the Arbitrage Pricing Theory," Journal of Econometrics 49 (1991), 141-168.

Mc Culloch, R., and R. Tsay, 1993, "Bayesian Inference and Prediction for Mean and Variance Shifts in Autoregressive Time Series", Journal of the American Statistical Association, 88(423), pp. 968-78.

Mengersen, K.L., and Tweedie, R.L., 1996, Rates of Convergence of the Hastings and

Metropolis Algorithms, The Annals of Statistics, Vol. 24, No. 1 (Feb.), 101–121.

Merton, R.C., 1969, Lifetime portfolio selection under uncertainty: The continuous time case, Review of Economics and Statistics 51, 247-257.

Merton, R.C., 1972, An Analytical derivation of the efficient portfolio frontier, Journal of Financial and Quantitative Analysis 7, No. 4, 1851–1872.

Michaud, R.O., 1989, “The Markowitz Optimization Enigma: Is Optimized Optimal”, Financial Analysts Journal, January, 31–42.

Muller and Pole (1998). Monte Carlo posterior integration in GARCH models. Sankhya. v60. 127–144.

Nakatsuma, T. (2000). Bayesian analysis of ARMA-GARCH models: A Markov chain sampling approach, Journal of Econometrics , 95, 57-69.

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach, Econometrica , 59, 347-370.

Nelson, D. (1994), Comment on Bayesian analysis of stochastic volatility models’, Journal of Business and Economic Statistics, 11, 406–410

Newton, M., Raftery, A., 1994, Approximate Bayesian Inference with the Weighted Likelihood Bootstrap. Journal of the Royal Statistical Society Ser. B, 57, 3–48.

Officer, R., 1973, The variability of the market factor of the NYSE, Journal of Busines, 46(3), 434–353.

Pastor, L., 2000, Portfolio selection and asset pricing models, Journal of finance 50, 179-223.

Pastor, L., and R.F. Stambaugh, 2000, Comparing asset pricing models: An investment perspective, Journal of Financial Economics 56, 335-381.

Patton A., (2008). Data-based ranking of realised volatility estimators. Working Paper, University of Oxford.

Pelletier, D., 2006. Regime switching for dynamic correlations, Journal of Econometrics, Elsevier, vol. 131(1-2), 445–473.

Pitt, M., and Shephard, M. (1999), Filtering via simulation: Auxiliary particle filters, Journal of the American Statistical Association 94 (446): 590–599.

Richardson, M. and James H. Stock, 1989. Drawing inferences from statistics based on multiyear asset returns, Journal of Financial Economics, 25(2), 323–348.

Ritter, C. and M. Tanner, 1992, Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy-Gibbs Sampler, Journal of the American Statistical Association, Vol. 87, No. 419 (Sep.), pp. 861-868.

Roll, Richard, 1977. A critique of the asset pricing theory's tests Part I: On past and potential testability of the theory," Journal of Financial Economics, vol. 4(2), pages 129-176, March

Ross, S. A., 1976, The arbitrage theory of capital asset pricing, Journal of Economic Theory 13, 341–360.

Shanken, J. (1986). A Bayesian approach to testing portfolio efficiency. Journal of Financial Economics 19, 195–215.

Sharpe, W., 1963, A simplied model for portfolio analysis, Management Science 9, 277–293

Siegel, J., 1994. Stocks for the long run. Mc Graw-Hill.

Stambaugh, R. F., 1997, Analyzing investments whose histories differ in length, Journal of Financial Economics 45, 285–331.

Stambaugh, R. F., 1999, Predictive Regressions, Journal of Financial Economics 54, 375–421.

Stein, C., 1955, Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution, in 3rd Berkeley Symposium on Probability and Statistics, , vol. 1, pp. 197-206, Berkeley. University of California Press.

Tu , and G. Zhou, 2009, Incorporating Economic Objectives into Bayesian Priors: Portfolio Choice Under Parameter Uncertainty, Journal of Financial and quantitative Analysis.

Vrontos, I. D., Dellaportas, P. and Politis, D. (2000). Full Bayesian inference for GARCH and EGARCH models, Journal of Business and Economic Statistics , 18, 187-198.

Wachter, J.A. and M. Warusawitharana, 2009. Predictable returns and asset allocation: should a skeptical investor time the market? Journal of Econometrics, 148, 162–178.

Zellner, Z.A., and V.K. Chetty, 1965, Prediction and decision problems in regression models from the Bayesian point of view, Journal of the American Statistical Association 60, 608-615.

Zellner, Arnold, 1962. "An efficient method of estimating seemingly unrelated regression equations and tests for aggregation bias. Journal of the American Statistical Association 57: 348-368.

Zellner, A. (1971). An Introduction to Bayesian Inference in Econometrics. New York: Wiley.

Zellner, A. (1978). Jeffreys-Bayes posterior odds ratio and the Akaike information criterion for discriminating between models Economics Letters, Volume 1, Issue 4, 1978, 337-342

Zellner, A. (1986), On assessing prior distributions and Bayesian regression analysis with g -prior distributions. In Bayesian inference and decision techniques, volume 6 of Stud. Bayesian Econometrics Statist., pages 233-243. North-Holland, Amsterdam,

Zellner, A. and W. Vandaele, 1974, Bayes-Stein estimators for k-means regression and simultaneous equations models, in: S. Feinberg and A. Zellner, eds., Studies in Bayesian econometrics and statistics (North-Holland, Amsterdam) 629-653.

Zhou, G., 2008, “An Extension of the Black-Litterman Model: Letting the Data Speak” Working Paper Olin Business School.

Table 1: Comparison of single-move and multi-move MCMC algorithms

	δ	σ_v	$\sqrt{h_t}$
True	0.960	0.210	
	Single-move		
Mean	0.948	0.229	
RMSE	0.021	0.041	0.00219
%MAE			16.27
	Multi-move		
Mean	0.952	0.221	
RMSE	0.017	0.037	0.00219
%MAE			16.19

Table 2: UK Pound to US $, SV posterior analysis

	δ	σ_v
	Single-move	
Mean	0.992	0.108
5% , 95%	0.983 , 0.999	0.075 , 0.146
	Multi-move	
Mean	0.993	0.097
5% , 95%	0.988 , 0.998	0.077 , 0.122

Table 3: Smoothing performance under alternative models

	Data generated by		
	Basic model	Fat-tailed model $\nu = 10$	
	All obs.	All obs	$\lambda_t > \lambda_{.9}$
Estimate fat-tailed SV			
RMSE(s_t)	0.0066	0.0104	0.0198
RMSE($\sqrt{h_t}$)	0.0066	0.0071	0.0076
%MAE($\sqrt{h_t}$)	18.3	21.4	23.7
Estimate basic SV			
RMSE(s_t)	0.0066	0.0107	0.0198
RMSE($\sqrt{h_t}$)	0.0066	0.0082	0.0098
%MAE($\sqrt{h_t}$)	19.2	25.9	30.6

For each observation, we compute the estimation error of the posterior mean of $\sqrt{h_t}$ and $s_t = \sqrt{h_t \lambda_t}$. When the data are generated by the fat-tailed SV, we also report RMSE and %MAE for the observations with a λ_t larger than the 90^{th} percentile of $p(\lambda \mid \nu = 10)$.

Table 4: Odds Ratios for Leverage effect, Daily Stock Returns 1978-1998

		1978-1998		1989-1998
Company	ν	ρ	$BF_{F/FC}$	$BF_{F/FC}$
Merck	10	-0.05	5.3	1.5
Boeing	8	-0.02	7.1	4.1
Dole Food	6	0.02	3.8	1.7
H.P.	8	-0.07	4.4	2.3
Fedex	6	0.08	5.4	0.6
Ford	12	-0.01	8.9	7.1
Sony	8	0.11	1.2	0.4
Fleet Bank	10	0.03	4.3	3.5
Exxon	11	0.01	6.6	3.6
Merrill Lynch	9	0.00	9.3	2.5
Average	9	0.01		
Portfolio	10	-0.23	0.22	1.E-03

F: fat-tail, FC: full model, 25000 posterior draws.

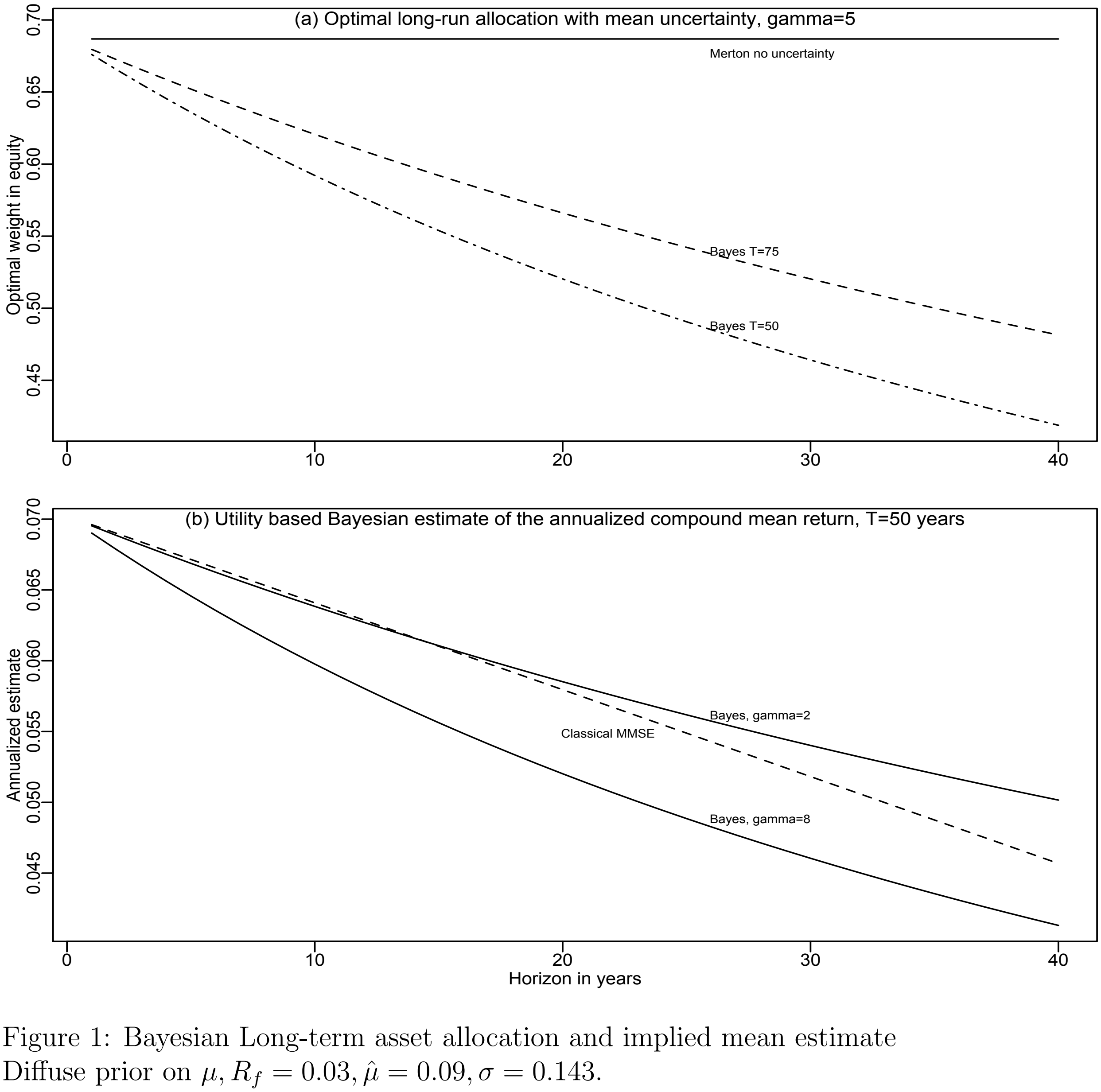

Figure 1: Bayesian Long-term asset allocation and implied mean estimate
Diffuse prior on $\mu, R_f = 0.03, \hat{\mu} = 0.09, \sigma = 0.143$.

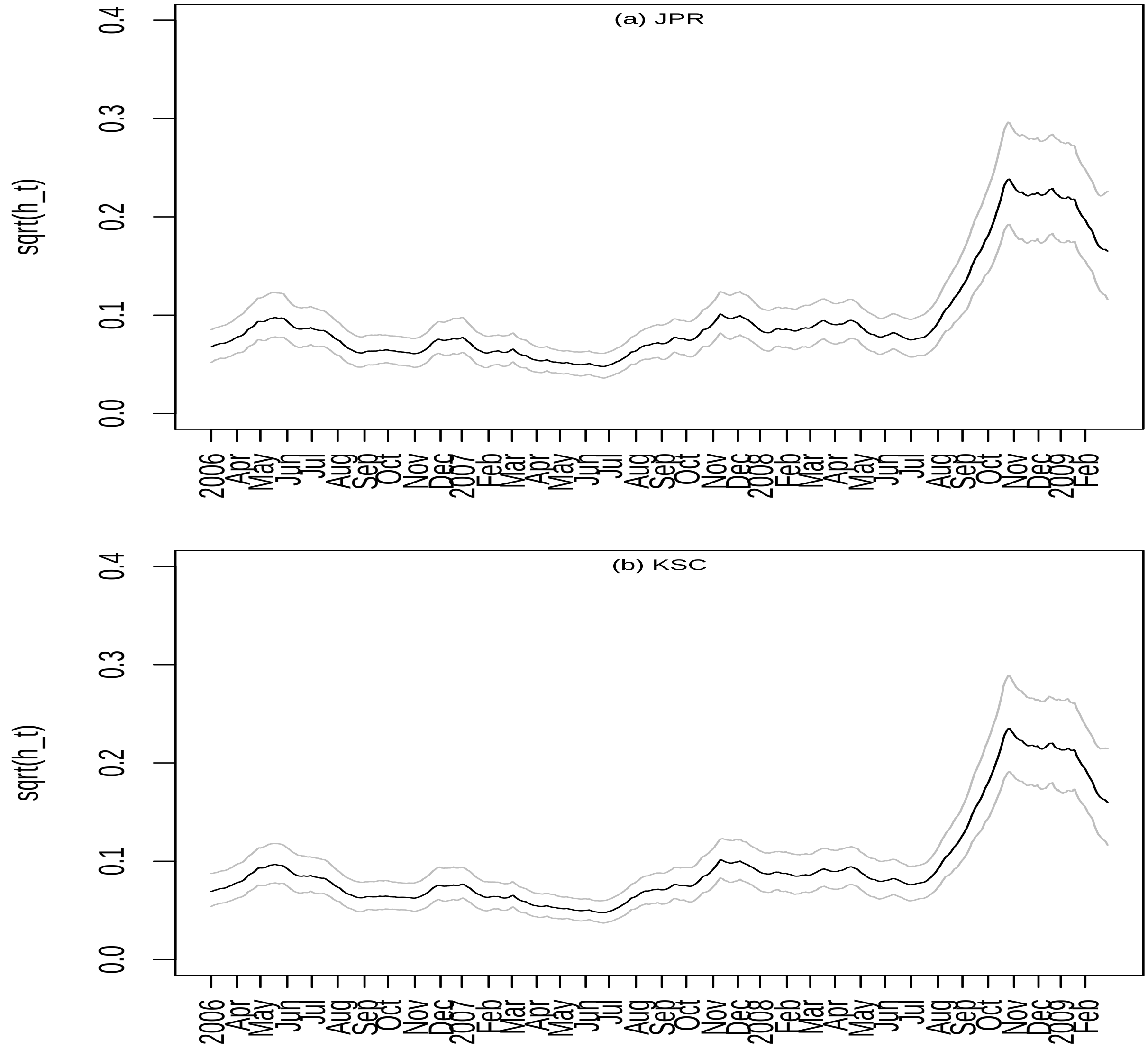

Figure 2: Posterior distribution of $\sqrt{h_t}$, mean and 5%, 95% quantiles

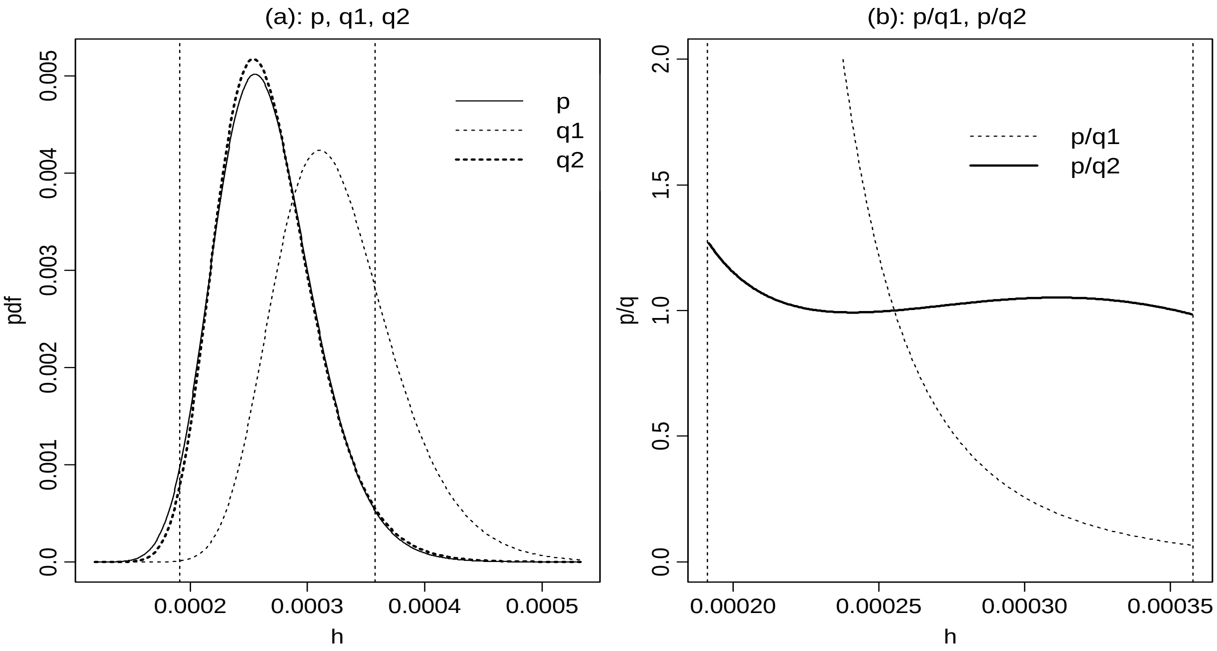

Figure 3: Improving the blanket density for the correlated model

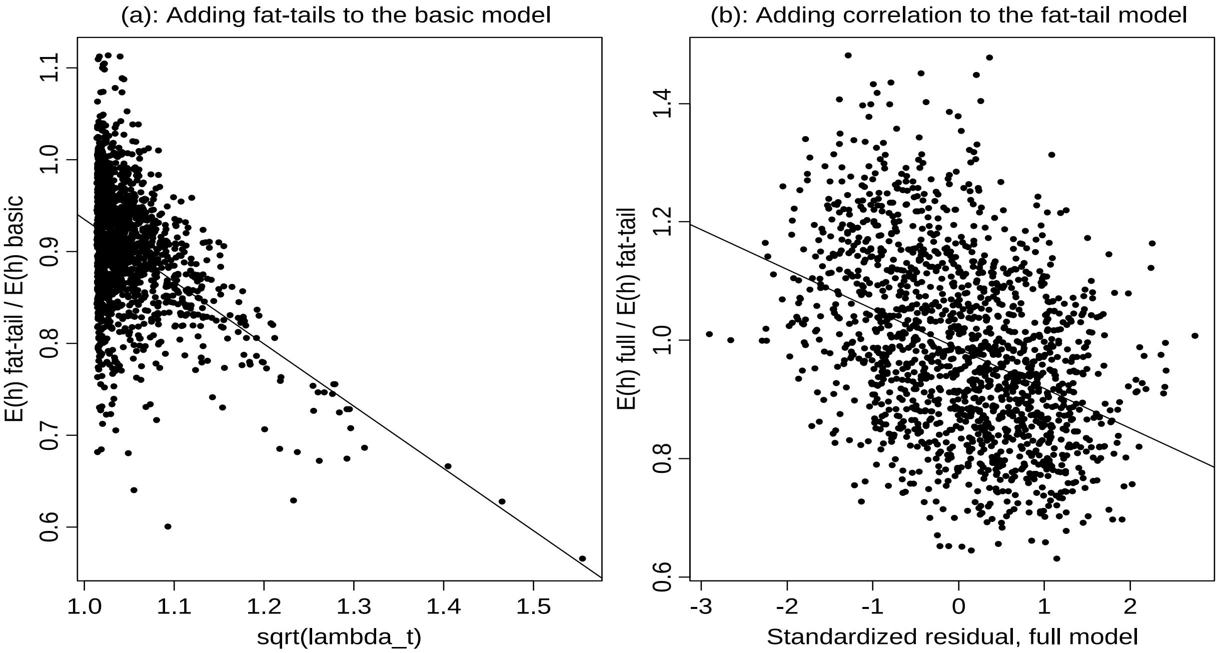

Figure 4: Effect of fat-tails and correlation on E(h_t); Weekly EW index

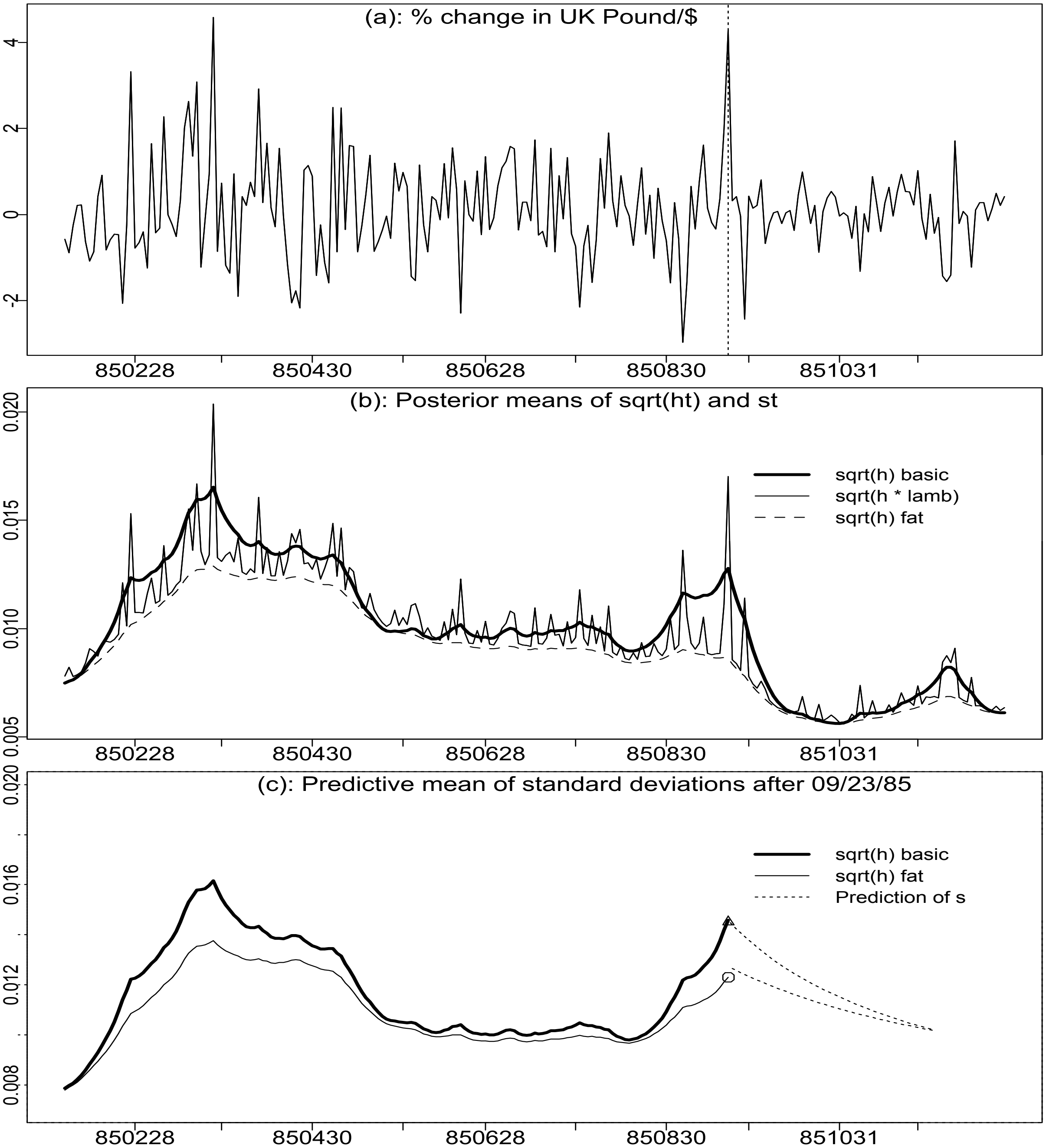

Figure 5: Differences between fat-tail and basic SV volatility forecasts

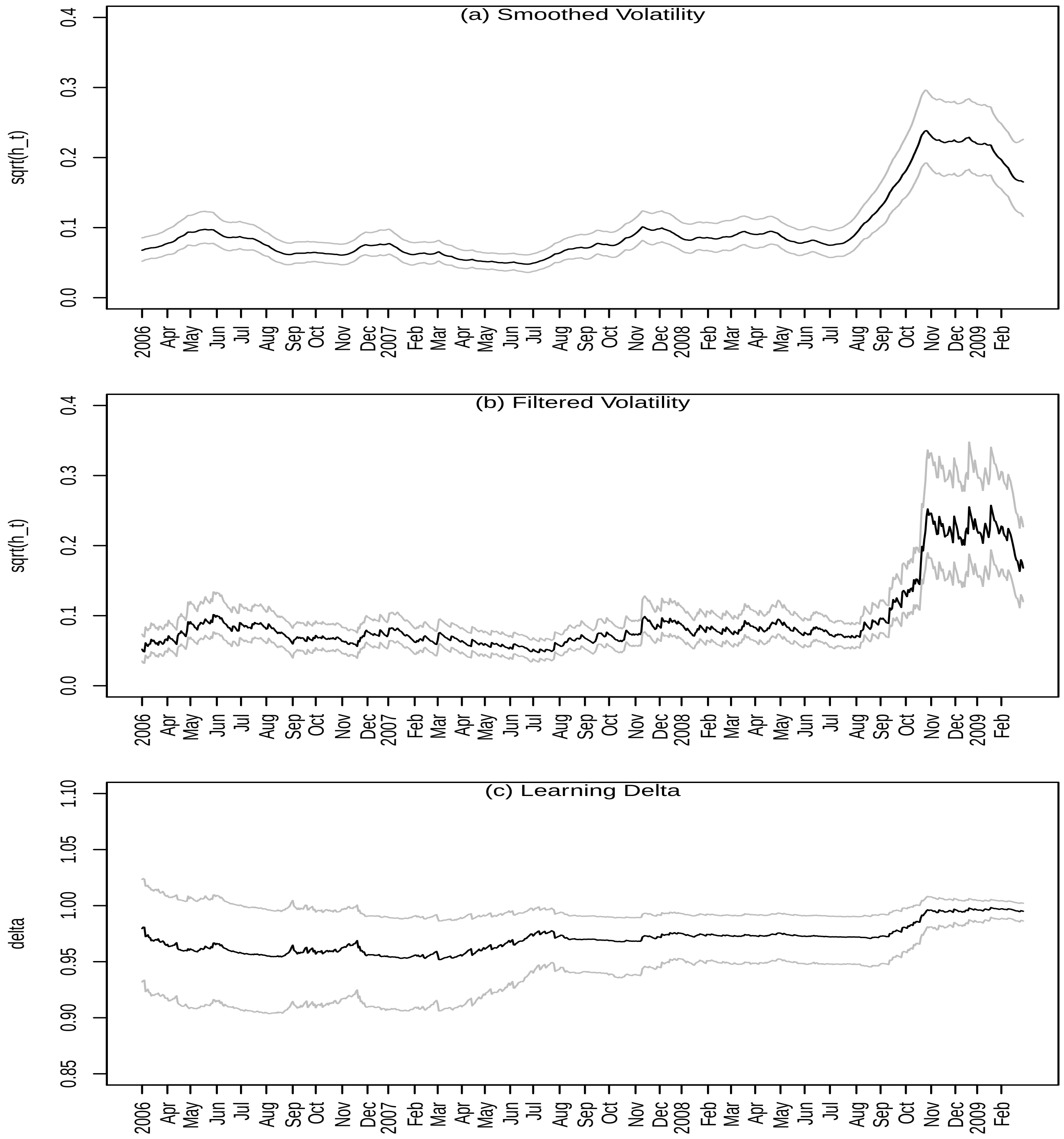

Figure 6: Smoothed $\sqrt{h_t}|R^T$, filtered $\sqrt{h_t}|R^t$, learning $\delta|R^t$
SV model, £/\$, Jan. 2, 2006 to Feb. 26, 2009, Mean and 5%, 95% quantiles.

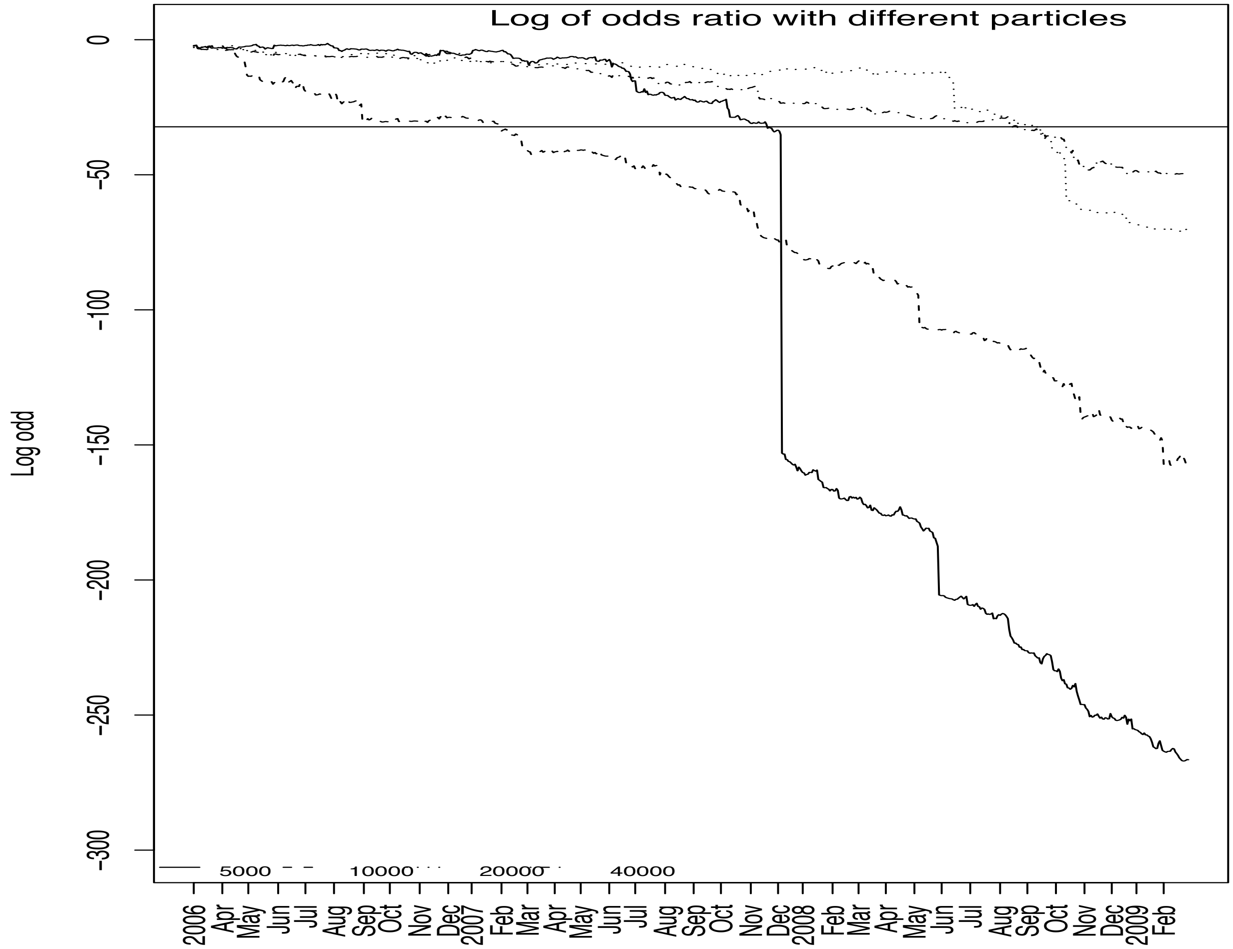

Figure 7: Log of odds ratios for basic SV versus SV with RV
Horizontal line: MCMC odds ratio. Particle filtering based odds ratios for 5K, 10K, 20K, and 40K particles. SV model, £/$, Jan. 2, 2006 to Feb. 26, 2009.