On the Intertemporal Risk-Return Relation: A Bayesian Model Comparison Perspective*

Leping Wang†

November 1, 2004

Abstract

The existing empirical studies indicate that inferences on the intertemporal relation between expected return and volatility are highly sensitive to empirical specifications of return dynamics. Glosten, Jagannathan, and Runkle (1993) attempt to resolve this confusing situation by examining several generalizations of the standard GARCH-M model. They conclude a negative risk-return relation solely based on the models that are identified through a variety of diagnostic tests as relatively "better" models. It has *not* been shown, however, whether those selected models outperform the alternatives *decisively or only marginally.* To the extent the strength of sample evidences supporting those selected models is unclear, the inference that is made solely based on those selected models remain questionable because of model uncertainty concern. Our paper propose a Bayesian model comparison approach to explicitly assess the strength of the evidence in support of the models that typically indicate a negative risk-return relation. The empirically computed Bayes factors show that those models indeed outperform, at a decisive degree, the alternative models that suggest a contrary result. Further, with priors that slightly favor return nonpredictability, evidence still indicates a negative relation after model uncertainty is accounted for. Our study, therefore, complements the work of Glosten, Jagannathan, and Runkle (1993).

1 Introduction

Although dozens of papers have studied the time-series relation between the conditional mean and volatility of stock returns, the evidence is inconclusive and leads to inference that is highly sensitive to empirical model specifications (e.g., Harvey (2001)).[1] As Glosten, Jagannathan, and Runkle (1993, p.1780) point out: *Most of the support for a zero or positive relation has come from studies*

*I am grateful to Michael Brandt, Craig MacKinlay, Andrew Metrick, Jessica Wachter, Yihong Xia, and especially my dissertation advisor Robert Stambaugh at the Wharton School of the University of Pennsylvania, for their numerous helpful comments and suggestions. I am fully responsible for any weakness of the paper.

†Correspondence address: School of Business, Singapore Management University, 469 Bukit Timah Road, Singapore 259756. Phone: (65) 6822-0763. Email: lpwang@smu.edu.sg.

[1]Intuition suggests that risk and return should be positively related over time. It has been shown, however, that risk premium on the market portfolio could, in equilibrium, be lower during relatively riskier times if average investors have time varying risk aversion levels.(Abel (1988) and Backus and Gregory (1992)) One intuitive explanation for the negative intertemporal risk-return relation, as suggested by Brandt and Kang (2004), can be seen from habit formation models (Constantinides (1990) and Campbell and Cochrane (1999)). At the peak (trough) of a business cycle, when expected return is typically low (high), the better(worse)-than-habit consumption levels make investors more (less) risk tolerant and thus require a lower (higher) reward-to-risk ratio.

that use the standard GARCH-M model of stochastic volatility. Other studies, using alternative techniques, have documented a negative relation between expected return and conditional variance.[2] This aspect is undesirable as it renders the empirical evidence subject to model misspecification concerns.

This issue appears even more serious as we note that most studies in this area are solely based on one single empirically motivated, but theoretically unjustified, specifications.[3] An exception is Glosten, Jagannathan, and Runkle (1993) who examine various generalizations of the standard GARCH-M approach by incorporating additional information conveyed by certain observable instruments and allowing for asymmetric volatility effects. To resolve the conflict surrounding the sign of the risk-return relation, they use a variety of diagnostic tests to determine whether the estimated residuals of the candidate models are independent and identically distributed with reduced excess skewness and kurtosis, as required in the model assumptions. They then reach a conclusion of a negative risk-return relation by showing that the models which indicate a negative risk-return relation perform better in the diagnostic tests than the alternative models which suggest a contrary result. However, the diagnostic tests they employ can be used only to select the "best" models, but not to assess the strength of the evidence supporting the model selection decision. In other words, it is unclear whether the models, based on which Glosten et al (1993) conclude a negative risk-return relation, outperform the alternative models *significantly or only marginally.* If the latter is the case, further efforts will need to be made to take into account the contrary information conveyed by the alternative models in the inference making because of the increasing concern over model uncertainty in recent finance studies (e.g., Avramov (2002)). Therefore, to the extent the sign of this fundamental relationship is important to finance research, it is important to show that their selected models indeed decisively outperform the alternatives or that their conclusion of a negative risk-return relation is robust to model uncertainty. After all, as it is important to report standard errors or confidence intervals as a measure for accuracy in parameter estimations, it is also important to associate the selected model with a measure for the strength of the supporting evidence.

Our paper addresses this challenge and proposes a full Bayesian specification of model comparisons, which can also be easily applied to account for model uncertainty when necessary. In the investigation of the risk-return relation, the Bayesian methodology is attractive. First, unlike the classical approach, the Bayesian framework has no requirement of nested models, standard probability distributions, or asymptotic regularity, and thus make it possible to compare the various empirical specifications in the risk-return relation literature that typically differ in many aspects. Focusing on the typical model classes used in the literature, we update our prior opinion to the posterior opinion on the uncertainty surrounding the correct model by computing the posterior odds ratio. The posterior odds ratio can be interpreted as the ratio of the posterior model probabilities conditional on the data, and is commonly termed the Bayes factor when two models are equally likely a priori. It summarizes all the sample evidence in favor of one model against its alternative.

[2] Examples falling in the first category include French, Schwert, and Stambaugh (1987), Campbell and Hentschel (1992), Chan, Karolyi, and Stulz (1992), and Bali and Peng (2003); those in the second category include Campbell (1987), Breen, Glosten, and Jagannathan (1989), and Whitelaw (1994), who use some exogenous instruments such as short-term interest rates in the specification of conditional moments.

[3] Note that the existing asset pricing theories are not explicit about how return moments evolve over time.

Motivated by the aforenoted observation made by Glosten, Jagannathan, and Runkle (1993), we particularly focus on comparing the distinct volatility specifications that lead to the conflicting conclusions about the risk-return relation. The first model class includes the models that forecast the return volatility with only the information in the return history; the GARCH-in-mean model of Engle, Lilien, and Robins (1987) is a typical example. It also includes the MIDAS volatility specifications employed in Ghysels, Santa-Clara, and Valkanov (2003), who use past squared daily returns to forecast monthly return volatility to improve estimation accuracy. In general, this class of models yields a positive (if sometimes weakly significant) risk-return relation. The second model class, proposed by Campbell (1987), includes exogenous instrumental variables in the information set for volatility forecasting, and suggests a negative risk-return relation. These two model classes are typically not nested.[4]

Second, the Bayes factor, in terms of posterior model probabilities, is easy to interpret and provides a meaningful scale of the evidence. According to the criteria proposed by Jeffreys (1961), $B_{AB} = 200$, for example, would suggest decisive evidence at odds of two hundred to one that the data favor H_A over H_B. In our context, we find decisive sample evidence in favor of the Campbell's instrumental variables model, which yields a significantly negative risk-return relation.

Third, the Bayesian model averaging approach makes it possible to incorporate the contrary information conveyed by the "rejected" model in our investigation. This turns out especially important if an investigator tends to believe neither return moment is predictable, in which case sample evidences would only marginally but not decisively support the Campbell's instrumental variables model. To be precise, the Bayesian approach assigns posterior probabilities to several competing models, and then obtain an optimally weighted model using the probabilities as weights on the individual models. This weighted model is then used for further analysis. The results still indicate a negative, although weakly significant, risk-return relation. Further, this conclusion of a negative intertemporal risk-return relation is shown to be robust to prior specifications.

We also investigate return dynamics combining the features of both the GARCH-M and instrumental variables models, that is, models predicting future volatility using past volatilities, squared return innovations, and exogenous instruments. Then, we add in terms allowing for an asymmetric volatility effect of positive or negative return shocks. Models framed this way capture the main characteristics of the volatility specifications used in Glosten, Jagannathan, and Runkle (1993). It turns out that these models suggest a negative risk-return relation, and the most importantly, outperform the alternatives at a decisive degree in the sense of data-fitting. As a result, our study fortifies the results of Glosten et al (1993), and therefore, can be viewed as supplementary to their work.

This paper is organized as follows. Section 2 describes the model uncertainty issue and the two typical model classes in the investigation of the risk-return relation. Section 3 presents the Bayesian model comparison framework (i) to empirically identify the "right" model and evaluate the scale of the supporting evidence from data and (ii) to take into account the uncertainty about the true model when necessary. Section 4 provides the main empirical results. It also extends the examined

[4] Although it is possible to implement the classical hypothesis test by extending the parameter space to have a more general specification that can nest other models, the power of the test will be significantly reduced due to the increased number of unknown parameters. Further and more important, in the case of accepting the null in the classical approach, the strength of such evidence is still unknown without the knowledge about the power of the test.

models to capture more well-documented features in the data and identifies the best model among those under consideration. Section 5 concludes.

2 Model Uncertainty

2.1 Intertemporal Risk-Return Relation

An equilibrium relationship between the market risk premium, defined as the expected stock market return in excess of the risk-free interest rate, and risk as measured by the volatility of the stock market, is derived by Merton (1973) in the context of a time varying economy. Its simplest form, under the assumption that a single state variable S_t is sufficient to describe changes in the investment opportunity set, can be written as

$$E_t(R_{t+1}) = \left[\frac{-J_{WW}W}{J_W}\right] V_t(R_{t+1}) + \left[\frac{-J_{WS}}{J_W}\right] COV_t(R_{t+1}, S_{t+1}), \tag{1}$$

where $E_t[\cdot]$, $V_t[\cdot]$ and $COV_t[\cdot]$ are, respectively, the expectation, variance, and covariance operator conditional on the information set at time t, and R_t is the monthly excess stock return over the risk-free interest rate. Subscripts on the derived utility of wealth function, $J(W, S, t)$, denote partial derivatives.[5] If the investment opportunity set is i.i.d. or if investors have log utility, the relation (1) reduces to a simple proportional relation

$$E_t(R_{t+1}) = \left[\frac{-J_{WW}W}{J_W}\right] V_t(R_{t+1}).$$

Hence, assuming $\left[\frac{-J_{WW}W}{J_W}\right]$ to be an intertemporal constant, but unknown, Merton (1980) estimates the proportionate relation

$$E_t(R_{t+1}) = gV_t(R_{t+1}), \tag{2}$$

where g is interpreted as the reward-to-risk ratio.

Furthermore, rather than working in continuous time as Merton (1973), Campbell (1993) takes a different approach by using a loglinear approximation to the intertemporal budget constraint and analytically derives a linear relation

$$E_t(R_{t+1}) = f + gV_t(R_{t+1}). \tag{3}$$

that holds in equilibrium. The parameter g relates the expected return to the conditional volatility, and its sign is what attracts most attention. This linear relation (3) nests the proportionate relation (2) and has been examined by a number of papers, such as French, Schwert, and Stambaugh (1987), Breen, Glosten, and Jagannathan (1989), and Campbell (1987), among others. It thus forms the basis for our empirical work.

Note that the information set available at time t, which the expectation and variance are conditional on, is generally not observable to econometricians. To address any discrepancy between

[5] See Scruggs (1998) for a discussion on the relation (1) for different forms of the function $J(W, S, t)$.

econometricians' and investors' information sets, a variety of assumptions are typically imposed. One conventional assumption, which we adopt as well, is that the econometricians' information set is broad enough to approximate the investors' information set, at least to the degree that the resulting inferences are not sensitive to the difference.

To proceed with estimation of the relation (3), we need to specify how the conditional volatility changes over time. The specifications are typically empirically motivated given the lack of any theoretical guidance, and are formed primarily to replicate documented characteristics of the time-varying return volatility in the stock market (and are hence somewhat ad hoc).

2.2 Volatility Specifications

For illustration purpose, in Table I we first present an overview of which volatility conditioning variables are included in the analysis of the risk-return relation for several representative papers and their corresponding conclusions. This list is by no means exhaustive. Several observations regarding this table are in order. First, all these studies conduct estimations and inferences based on one predetermined information set. Although some authors do consider several distinct model specifications in the analysis, no meaningful measure is provided to distinguish among models. Second, it should be clear that there is no agreement on whether the expected stock return is positively or negatively related to the return volatility over time. The parameter g, relating the first two moments of stock returns, is reported as either positive or negative in different studies. Third, as Glosten, Jagannathan, and Runkle (1993) observe, the models that forecast volatility using past volatilities and squared return innovations indicate a positive intertemporal risk-return relation, while those including some exogenous instruments such as short-term risk-free interest rates in the volatility prediction typically give a negative sign. Hence, uncertainty about the correct information set used in volatility forecasting could be a potential source of such inconclusive results.

We thus focus on two often-used model classes categorized by the conditioning information set used in the volatility specification. The first class forecasts return volatility using only the information in past returns (e.g., French et al. (1987)), while the second class uses exogenous predictive variables (e.g., Campbell (1987)). We describe the model specifications as follows.

Hypothesis A:

A widely used model to capture the time-varying volatility in financial data is the ARCH model of Engle (1982) and its various extensions such as the GARCH of Bollerslev (1986) and the EGARCH of Nelson (1991). This approach models the conditional volatility as a nonstochastic function of the unanticipated part of lagged excess stock returns and lagged conditional volatility. One appealing feature of this structure is that it reflects the characteristic observed in financial data that big surprises are often followed by big surprises, a phenomenon commonly termed volatility clustering.

We start with the most generic model of this type — the ARCH(1) given by

Model A1:

$$\sigma_t^2 = \alpha + \beta\varepsilon_t^2,$$

where σ_t^2 stands for the return volatility $V_t(R_{t+1})$, and ε_t is the disturbance given by $R_t - E_{t-1}(R_t)$. At time $t-1$, ε_t is normally distributed with mean zero and variance σ_{t-1}^2.

To allow for possible higher-order volatility persistence, we also examine an ARCH(2) given by:

Model A2:

$$\sigma_t^2 = \alpha + \beta\varepsilon_t^2 + \delta\varepsilon_{t-1}^2,$$

and a more parsimonious GARCH(1, 1) described by:

Model A3:

$$\sigma_t^2 = \alpha + \beta\varepsilon_t^2 + \gamma\sigma_{t-1}^2.$$

In all specifications where relevant, α, β, γ, and δ need to be nonnegative to prevent return volatility from falling below zero. Nelson (1991) points out that the nonnegativity constraints required by these models can sometimes present difficulties in estimation. One example is Engle, Lilien, and Robins (1987), who must impose additional structure on the coefficients to conduct an efficient estimation.

An alternative model that successfully avoids the nonnegativity constraint on the parameters while retaining reasonable return volatilities is the EGARCH(1, 1) of Nelson (1991) denoted by:

Model A4:

$$\ln\sigma_t^2 = \alpha + \beta z_t^2 + \gamma \ln\sigma_{t-1}^2,$$

where z_t denotes $\varepsilon_t/\sigma_{t-1}$, which is used instead of ε_t in the EGARCH specification to ensure a well-behaved volatility process. Nelson (1991) proves by theorem that the conditional volatility process specified in the EGARCH is strictly stationary, ergodic, and covariance-stationary if and only if γ is less than one in absolute value. Note that, as he points out, there is no necessary implication from strict stationarity to covariance stationarity in this case since the conditional moments of a general process may explode even under ergodic strict stationarity.

Each volatility specification together with the linear risk-return relation given in (3) forms a particular case of the GARCH-M model of Engle, Lilien, and Robins (1987), which has been investigated by several researchers (e.g., French, Schwert, and Stambaugh (1987)). We label this category as Model A, denoted by H_A. Glosten, Jagannathan, and Runkle (1993) also examine a generalization of the GARCH-M approach by allowing for seasonal effects, volatility asymmetries, information contained in nominal interest rates, and exponential form of conditional volatility. We will analyze this type of models later in an attempt to identify a better model in the sense of data fitting.

In an anticipation that high-frequency data may improve the accuracy of the volatility estimates, French, Schwert, and Stambaugh (1987) and Ghysels, Santa-Clara, and Valkanov (2003) use past daily returns to forecast monthly return volatilities. The former propose a simple and intuitive one-month rolling-window estimator with equal weights on past squared daily returns, and find a mostly insignificant risk-return relation. The latter use a longer estimation window — roughly one year of trading days— with a relatively more flexible form of parameterization on the weights given to the lagged squared daily returns, and report a significantly positive relation between the expected return and variance of the aggregate stock market. We include their volatility specifications as follows.

Models A5 & A6:

$$\sigma_t^2 = 22 \sum_{d=1}^{D} w_d r_{t-d}^2 \tag{4}$$

where D is the number of days used in the estimation of variance , r_{t-d} denotes the daily return at the date $t-d$ (i.e., d days previous to the first day of month $t+1$), and w_d is the weight given to the squared returns r_{t-d}^2, which sums up to one. Note that we use lower case r to denote daily returns and upper case R to denote monthly returns. Weighted past squared daily returns are normalized by the factor 22 to monthly units since one month typically consists of 22 trading days.

More interestingly, Ghysels, Santa-Clara, and Valkanov (2003) find that when the estimation window size in equation (4) is lengthened from one month to three or four months and equal weights are used, the risk-return coefficient changes from an insignificant estimate in French, Schwert, and Stambaugh (1987) to a significantly positive estimate; the maximum likelihood across window sizes is obtained with a four-month window. For this reason, we investigate this volatility specification using a choice of four-month rolling windows; i.e., $D = 88$.

We examine two forms of weighting functions. In Model A5, equal weights are set on each day in the estimation window. In Model A6, the weight w_d is set to be proportional to $\exp(-0.03d)$, which declines as a function of the number of lags.[6] By putting more weight on recent observations, more attention is paid to capturing the time variation of return volatility than to controlling for estimation error. Following Ghysels, Santa-Clara, and Valkanov (2003), we call these two models the *mixed data sampling* (or MIDAS) approach.

Hypothesis B:

Campbell (1987) suggests a competing class of models that forecasts volatility using certain exogenous instruments, given the empirical evidence that stock market movements can be predicted by variables related to the business cycle (e.g., Chen, Roll, and Ross (1986), Keim and Stambaugh (1986), Campbell and Shiller (1988), Fama and French (1988), and Ferson and Harvey (1991)).

A fairly extensive literature examines the relation between stock market excess returns and interest rates. Breen, Glosten, and Jagannathan (1989), for example, investigate the ability of nominal interest rates to predict stock market excess returns and find a statistically significant negative correlation between the two. This result, under the assumption that interest rates are good proxies for expected inflation, can be explained by a negative relation between stock excess returns and inflation. Stulz (1986) constructs a simple representative agent model and shows that worsening productivity could induce increases in expected inflation associated with declines in excess stock returns. To the extent that changing market risks are correlated with changing market premiums, nominal interest rates could also be a good predictor of future return volatility. Breen, Glosten, and Jagannathan (1989), for example, empirically demonstrate that the one-month interest rate is useful in forecasting the volatility of excess stock returns.

We examine a model specified as

[6] Ghysels, Santa-Clara, and Valkanov (2003) study a more flexible form of model specification postulating weights as a function of unknown parameters, which are estimated jointly with the risk-return coefficient using maximum-likelihood estimation. In order to simplify the model comparisons while retaining the key features of their model, to which they attribute the finding of a significantly positive risk-return relation, we simply take those weight-related parameters as given. The specific factor -0.03 is chosen to obtain the maximum likelihood after several experiments.

Model B:

$$\sigma_t^2 = c + dx_t,$$

where x_t denotes the one-month risk-free interest rate. Following Glosten, Jagannathan, and Runkle (1993), we call this approach Campbell's instrumental variables model, and label it as Model B, denoted by H_B. We investigate only one specific form of such a model using short-term interest rate as the single predictive variable mainly because interest rate is the most often-used instrument in this literature. In addition, because models using other instruments have all been shown to lead to a negative risk-return relation (e.g., Campbell (1987), Whitelaw (1994), and Harvey (2001)), enlarging this model class by including more instruments will only strengthen our conclusion of a negative risk-return relation, a final result reported later.

3 Econometric Approach

The models of interest in comparison are the two most often-used but competing model hypotheses denoted by H_A and H_B. Although differing only in the specifications of the information set used to forecast conditional volatility, these two models lead to paradoxical answers to how the first two return moments move together over time, one of the most fundamental questions in finance. As financial theory has little to say on how stock returns evolve over time, it remains an empirical question as to which model better describes the underlying return dynamics. In other words, we should let the data help us distinguish between the two competing model specifications H_A and H_B.

The most popular data-based model selection techniques between two competing statistical models are based on an interpretation of p-values under the classical hypothesis test framework.[7] The classical approach, however, is not very general in that essentially it requires nested models, standard probability distributions, or asymptotic regularity. Furthermore, it is often arbitrary to take one of the two nonnested models, H_A and H_B, as the null hypothesis, and the two tests taking each model in turn as the null hypothesis may not present consistent conclusions. In particular, both models, H_A and H_B, may be rejected or may fail to be rejected, in which case the tests provide no means of model comparison, not to mention that merely failing to reject the null hypothesis does not indicate the strength of the evidence in favor of the null since the power characteristics of a test set at certain significance level are often unknown and hard to obtain.

To overcome these well-known drawbacks of classical model selection techniques, we use a Bayesian approach, which, rather than testing the validity of one model against another, treats the problem as model comparison (instead of model selection), recognizing that no model is absolutely perfect, and more important, offers a way of evaluating the strength of evidence favoring each hypothesis.

3.1 Bayesian Model Comparison

Let θ_i be the unknown parameter vector of model H_i ($i = A$ or B). The vector θ_i could have common parameters such as g across the models. Conditional on a model H_i and its involved

[7] Examples are the J test (see Davidson and MacKinnon (1981)) and the Cox (1961, 1962) test, to name a few.

parameter vector θ_i, we can express the conditional probability distribution of the data, denoted by D:

$$H_i: \quad p(D|\theta_i, H_i), \quad \text{for } i = A \text{ or } B.$$

To reflect the ex ante opinion of the uncertainty surrounding the models and parameter values, we assign a prior probability $p(H_i)$ to each model, and a prior probability distribution $p(\theta_i|H_i)$ to the parameter vector of each model. This prior formulation induces a complete model specification described by the joint distribution:

$$p(D, \theta_i, H_i) = p(D|\theta_i, H_i)p(\theta_i|H_i)p(H_i),$$

and can be intuitively understood as a three stage hierarchical mixture model for generating the data D; first the model H_i is generated from $p(H_i)$, second the parameter vector θ_i is generated from $p(\theta_i|H_i)$, and third the data D is generated from $p(D|\theta_i, H_i)$.

Conditional on having observed the data D, we can then update our prior opinion to a posterior opinion on model uncertainty by computing the posterior model probability using the Bayes theorem:

$$p(H_i|D) = \frac{p(D|H_i)p(H_i)}{\sum_{i=A,\ B} p(D|H_i)p(H_i)},$$

where

$$p(D|H_i) = \int p(D|\theta_i, H_i)p(\theta_i|H_i)d\theta_i \tag{5}$$

represents the marginal likelihood of H_i. This posterior distribution extracts all the relevant information in the data and provides a complete and coherent summary of post data uncertainty about the correct model that generates the data. The direct probability interpretation of the Bayes factor is readily understandable even by nonstatisticians, while it is very hard to properly interpret the p-values many classical schemes are directly or indirectly based on (See, e.g., Berger and Sellke (1987)).

Furthermore, given these posterior probabilities, comparison between H_A and H_B can be summarized by the posterior odds:

$$\frac{p(H_A|D)}{p(H_B|D)} = B_{AB} \cdot \frac{p(H_A)}{p(H_B)}, \tag{6}$$

where the Bayes factor B_{AB} is defined as

$$B_{AB} = \frac{p(D|H_A)}{p(D|H_B)}, \tag{7}$$

and can often be interpreted as the odds provided by the data for model H_A versus H_B. Equation (6) reveals the way the data, through the Bayes factor B_{AB}, update the prior odds $\frac{p(H_A)}{p(H_B)}$ to form the posterior odds. In a form independent of the prior model probabilities, the Bayes factor summarizes the evidence provided by the data in favor of one model as opposed to another. Further, it is a simple and popular choice to use the uniform model prior $p(H_A) = p(H_B) = 0.5$, which is noninformative in the sense of favoring both models equally. In this case, the Bayes factor is identical to the posterior odds, defined as the ratio of the posterior model probabilities. Thus $B_{AB} = 0.1$, for

example, would suggest that the data favor H_B over H_A at odds of ten to one.[8]

Note that the Bayes factor is analogous to the likelihood ratio statistic; the only difference is in the way the parameter vector θ_i is eliminated. The marginal likelihood of H_i involved in the Bayes factor eliminates θ_i by the integration (5), while the likelihood ratio statistic does so by maximization.

Unlike the classical tests that either accept or reject a hypothesized model, the Bayesian approach offers a way to evaluate evidence in favor of any individual model, and thus should be more appropriately called model comparison (rather than model selection). The most important appeal of this feature is that, when the evidence in the data only marginally favor one model over another, we can apply Bayesian model averaging to account for model uncertainty in making inferences on parameters of interest.

Further, Bayesian model comparison and model averaging can be easily applied to cases involving far more than two models, which is quite common in practical data analysis.[9] One example is Avramov (2002), who examines the sample evidence on return predictability in the presence of model uncertainty, particularly uncertainty about the choice of independent predictors. Carrying out classical tests in this case is hard and could produce misleading results (see Freedman (1983)).

3.2 Sensitivity analysis of prior specifications

Computing B_{AB} requires specification of $p(\theta_i|H_i)$ $(i = 1, 2)$. However, it is well known that the Bayes factor can be quite sensitive to prior specifications. Indeed, the dependence of the Bayes factor on the prior distributions in model comparison is much stronger than in, say, parameter estimation conditioning on a single model. As sample sizes grow, the influence of the prior distribution vanishes in parameter estimation, but not in model comparison. Therefore it is important to evaluate the Bayes factor over a range of reasonable priors and examine the robustness of the inference to the prior specifications.

Given our little knowledge about values of the model parameters, either from analysis using other related data or from the implications of finance theory, we begin with priors that are relatively noninformative or reflect little information beyond that already incorporated in the data, which leads to the so-called objective Bayesian model comparison. Improper or diffuse priors, which are intended to be noninformative by construction are, therefore, our first choice of parameter priors.

However, while one can successfully implement diffuse priors in many Bayesian parameter estimations conditioning on a single model, it is problematical to directly insert improper priors into (5) for model comparison because of indeterminacy issue. To see this, suppose that improper parameter priors π_A and π_B are used for model H_A and H_B, and the Bayes factor B_{AB} is then calculated. Because the priors are improper, they are defined only up to an undefined multiplicative constant. Therefore, one could have just as well used $c_A\pi_A$ and $c_B\pi_B$ as noninformative priors, in which case the resulting Bayes factor would be $(c_A/c_B) \cdot B_{AB}$. Since the choice of c_A and c_B is

[8] The framework can be easily applied to include distinct modeling assumptions on other dimensions such as the specific functional forms of the risk-return relation. Also motivated by Merton's (1973) intertemporal capital asset pricing model (ICAPM), Scruggs (1998) investigates a different functional form of risk-return relation — a conditional two-factor model.

[9] For an introduction of the Bayesian model comparison approach applied to the multi-model case, see Kass and Raftery (1995).

arbitrary, the Bayes factor is clearly indeterminate.

One way around this difficulty is to use the intrinsic Bayes factor (IBF) proposed by Berger and Perrichi (1996). The idea is to use part of the data as a training sample to convert the improper noninformative prior to a proper posterior distribution, which is then combined with the remaining data to compute the Bayes factor. The resulting measure, for comparing H_A and H_B, can be expressed as the product of the Bayes factor of model H_A to H_B using the whole sample and the Bayes factor of H_B to H_A using the training sample:[10]

$$B_{AB}(l) = B^N_{AB}(D) \cdot B^N_{BA}(D(l)), \tag{8}$$

where $D(l)$ denotes the training sample of size l, and the superscript N indicates the use of non-informative priors. By construction, $B_{AB}(l)$ no longer depends on the scales of the improper prior $p(\theta_i|H_i)$ $(i = 1, 2)$ as the arbitrary ratio c_A/c_B that multiplies $B^N_{AB}(D)$ is cancelled by the ratio c_B/c_A that multiplies $B^N_{BA}(D(l))$.

We use the first 20 years of return data as the training sample in our analysis. This choice is arbitrary, beyond the requirement that sample size needs to be large enough to guarantee a proper posterior density. Training samples of different sizes were also tried, and the results are qualitatively the same.

Given ongoing debate in the Bayesian literature on the incoherence of Bayesian formulations caused by the use of training samples, we also examine priors in the forms of proper distributions as a robustness check. In particular, the priors are assumed in the form of a multivariate normal distribution as

$$p(\theta_i|H_i) \sim N(\mu_i, \Sigma_i), \tag{9}$$

where the restricted parameters, such as the ARCH coefficients, are appropriately truncated and normalized. μ_i is the mean vector and Σ_i is the covariance matrix, both remaining to be specified.

The specification of the prior dispersion Σ_i is a crucial challenge. On the one hand, it should be large enough to avoid too much prior influence; on the other hand, it should be small enough to avoid producing too low a model probability and arbitrary values for the Bayes factor (see Chipman, George, and McCulloch (2001)). We specify, for our second choice of parameter priors, μ_i and Σ_i as the mean vector and covariance matrix of the posterior distribution derived using diffuse noninformative priors conditioning on the individual model H_i.[11] To be precise, we first obtain the posterior probability distribution with a diffuse prior conditioning on model H_i and compute the posterior mean vector and covariance matrix, denoted by μ^p_i and Σ^p_i, respectively. Then, in calculating the Bayes factor for model comparison, the parameter prior distribution for model H_i is assumed to be distributed as $N(\mu^p_i, \Sigma^p_i)$.[12] This choice is made primarily to incorporate in the prior as little information as possible beyond that already incorporated in the data.

We also entertain prior distributions that slightly favor the view of no predictability of stock

[10] See Berger and Perrichi (1996) for the proof and detailed discussions of IBF.

[11] Since those parameters required to be nonnegative are in fact distributed mostly in the positive range, the truncated normal distributions closely approximate the normal distributions. Thus in practice, we simply ignore the normalizing terms. This choice, given the extremely strong evidence obtained in the model comparisons reported later, is very unlikely to affect the results qualitatively.

[12] The practice of specifying the hyperparameters in the prior distribution with statistics from the actual sample is commonly termed empirical Bayes (see Maritz and Lwin (1989)).

returns by choosing the appropriate values for μ_i and Σ_i in (9). Note that in both models, H_A and H_B, $g = 0$ indicates constant expected excess stock returns over time, and β, γ, δ, $d = 0$ suggest no time variation of return volatility. Thus, in choosing our third form of priors, we let those parameters be centered around zero, i.e., distributed with marginal means of zero. f is assumed to be distributed around $\overline{R} = \frac{1}{T}\sum_{t=1}^{T} R_t$, the sample mean, and α of GARCH(1, 1) (or e^{α} of EGARCH(1, 1)) and c around $\widehat{\sigma}_R^2 = \frac{1}{T}\sum_{t=1}^{T}(R_t - \overline{R})^2$, the sample variance. For simplicity, all the parameters are assumed to be independent of each other in the priors. The priors defined in this way represent, at the point of maximum likelihood, the belief of an *i.i.d.* return series normally distributed with sample mean and sample variance. The prior dispersions are specified in a similar manner as in the second choice of the prior specification. The only difference is that, here, for the common parameters appearing in both models, f and g, the prior marginal variances are chosen to be the average of the corresponding posterior variances given in Σ_1^p and Σ_2^p. The variances of the other parameters that are unique to one model, say, H_i, are assumed to take the values of the corresponding parts in Σ_i^p.

The Bayes factors computed using the above three forms of parameter prior distributions are denoted by BF1, BF2, and BF3, respectively. To provide a more convincing robustness check of our results to the prior specifications, we also report the Schwarz (1978) criterion as an approximation of the Bayes factor, which can be written in the form of

$$B_{AB}^S = \frac{p(D|\widehat{\theta}_1, H_A)}{p(D|\widehat{\theta}_2, H_B)} n^{(d_2 - d_1)/2},$$

where $\widehat{\theta}_i$ is the MLE under H_i, n is the sample size, and d_i is the dimension of θ_i. This quantity, arising from Laplace's asymptotic method, has the advantage of simplicity and freedom from prior assumptions (see Tierney and Kadane (1986)).[13] In this sense, it provides a reasonable reference procedure for model comparison. We denote B_{AB}^S by BF4.

3.3 Bayesian Model Averaging

Selecting a model on the basis of data, and then using the same data for estimation and inference based on the model, is well known to yield (often severely) overoptimistic estimates of accuracy. Such an issue is particularly serious if the model is only marginally favored by the data, but not to a decisive degree, over an alternative model. To address this concern, we investigate the sample evidence on risk-return relation based on a composite weighted model, using the posterior probabilities $p(H_i|D)$ as weights on the individual models H_i.

Much of the sample evidence regarding the parameter g is represented in its posterior probability distribution as follows:

$$p(g|D) = \sum_{i=A,\ B} p(g|D, H_i) p(H_i|D).$$

This is an average of the posterior distributions under each model considered, weighted by their posterior model probabilities. Note that the posterior information is formed on the basis of only the

[13] In the case of large samples, B_{AB}^S should provide a reasonable assessment of the model evidence, but extreme caution needs to be taken in drawing inferences for models with irregular asymptotics or with likelihood concentrating at the boundary of the parameter space.

observed data. As a result, any inference made according to this distribution explicitly incorporates model uncertainty and is thereby robust to model misspecification, at least within the universe of the examined models. The posterior mean and variance of g are given as follows (see Kass and Raftery (1995)):

$$E(g|D) = \sum_{i=A,\ B} E(g|D, H_i)p(H_i|D) \tag{10}$$

and

$$Var(g|D) = \sum_{i=A,\ B} (Var(g|D, H_i) + (E(g|D, H_i))^2)p(H_i|D) - E(g|D)^2. \tag{11}$$

The mean and variance in Equations (10) and (11) follows by iterated expectations, conditioning first on the model space. The posterior mean is merely a weighted average of the estimates from individual models. The posterior variance incorporates the parameter uncertainty attributed to both the estimation error in each competing model and the uncertainty about the correct model, which is reflected by the dispersion in the posterior model probabilities. The latter component distinguishes this measure of estimation error from the well-known classical counterpart in that it explicitly takes into account the inability to identify the true return dynamics.

Recent research using the Bayesian model averaging approach to incorporate model uncertainty includes Avramov (2002) and Cremers (2002). Their work concentrates more on the variable selection problem. Concerned with data-snooping critics as to the use of various predictive variables in the return predictability literature, they attempt to analyze the sample evidence of stock return predictability, and identify the most important predictors by comparing all possible linear predictive regressions simultaneously in a Bayesian framework. Our paper investigates a different and more serious issue, where model misspecification could in essence induce seriously spurious conclusions since two different model classes produce profoundly conflicting indications.

4 Empirical Results

4.1 Date, Estimation, and Model-Dependent Inference

The monthly returns on the value-weighted NYSE index available from the Center for Research in Security Prices (CRSP) are used as a proxy for the market return. All the returns are calculated in excess of the 30-day Treasury bill rate obtained from CRSP. The 30-day Treasury bill rate is also used as an instrument in the volatility forecasting in the instrumental variables model. Monthly data are from June 1951 through December 2001. We restrict the data to this post-1951 period to avoid the time before the announcement of the Treasury-Federal Reserve Accord in March 1951, when interest rates were held almost constant by the Federal Reserve Board.

We first evaluate the implication of each specification individually regarding the intertemporal relation of expected return and variance. Although these models have all been examined by earlier studies, at least in analogous forms, we want to replicate those results here not only to demonstrate the confusing situation but also to facilitate the subsequent model comparison analysis. For this purpose, we estimate parameters using the Bayesian approach instead of classical methods such as MLE and linear regression that most earlier studies apply.[14]

[14] Although it is easy to show that MLE provides consistent estimators, it is harder to show the asymptotic

The posterior mean and posterior standard deviation (in parentheses) for each parameter are reported in Table II. We report the posterior standard deviation rather than the t-value because estimation is in a Bayesian framework. The posterior distribution is obtained by updating the standard diffuse and independent prior distribution, aiming to draw "objective" Bayesian inferences that are little affected by information external to the observed data. The use of such convenient noninformative prior distributions can be justified by the asymptotic irrelevance of the prior distributions. Further, since the resulting posterior is far from any typical form, the Markov Chain Monte Carlo (MCMC) approach is used to obtain the desired posterior properties of the unknowns.[15]

As Table II shows, the implications of the models for the risk-return relation, captured by the parameter g, are generally consistent with those reported in studies using classical methods. Models that include only the past returns in the conditioning information set yield mostly positive estimates for g, except for the ARCH(1)-M model (Model A1), which estimates the risk-return coefficient to be -0.57, with posterior standard deviation at 14.54, indicating little significance. It is interesting to note that, when a possibly higher degree of volatility persistence is captured by the higher-order ARCH structures, such as the ARCH(2) (Model A2) or the more parsimonious GARCH(1, 1) (Model A3) and EGARCH(1, 1) (Model A4), the estimate of g turns to positive, ranging from 4.87 to 10.39 (and more than one and a half posterior standard deviations away from zero for the GARCH(1, 1)-M model). Under the MIDAS modeling assumption of the forecasting ability of past squared daily returns on the monthly return volatility, the risk-return coefficient g is estimated to be 2.36 and 2.11 for Model A5 and Model A6, respectively, both about one and a half posterior standard deviations away from zero. This is consistent with the findings of Ghysels, Santa-Clara, and Valkanov (2003). By the instrumental variables model (Model B), however, the posterior mean and standard deviation of g are -11.40 and 4.49, respectively, indicating that the conditional expected return and volatility are negatively related over time. The slope coefficient for the interest rate, d, is positive and more than two posterior standard deviations away from zero.

It is quite evident that the answer to whether risk is positively or negatively related to the risk premium over time depends, to a large extent, on the modeling assumptions. Therefore, the potential issue of model uncertainty should be of serious concern to investigators.

4.2 Model comparisons

However, very few researchers have seriously considered the model specification issue. Glosten, Jagannathan, and Runkle (1993) does apply a specification test to their models, while most others simply use statistical inference based on one rather ad hoc return model. While Glosten, Jagannathan, and Runkle (1993) identify the most satisfactory model through a variety of diagnostic tests, it is, however, difficult to evaluate the strength of the evidence supporting their model selection decision, a formidable obstacle in the classical hypothesis test because of the extreme difficulty

properties needed in the subsequent inferences since the required regularity conditions are quite difficult to verify for general heteroskedastic models. Lee and Hansen (1994) give some results for the GARCH(1, 1) process in this respect, but many other concerns, especially regarding more general models, remain unanswered. In practice, however, the regularity problem is typically ignored, and empirical researchers use standard estimation procedures under the assumption that the usual regularity conditions are satisfied.

[15] See the appendix in Wang (2004) for a brief description of the simulation procedure used. Gilks, Richardson, and Spiegelhalter (1996) provide a textbook treatment of more general MCMC approaches.

of properly interpreting a p-value (see, e.g., Berger and Sellke (1987)).[16] It is, therefore, theoretically possible that their selected model, based on which the conclusion of a negative risk-return relation is made, outperforms the alternative models only marginally. If that is the case, to make their result convincing to finance researchers, it is important and necessary to show that their result is robust to model uncertainty. After all, as it is important to report standard errors or confidence intervals as a measure for accuracy in parameter estimations, it is also important to associate the selected model with a measure for the strength of the supporting evidence.

To address this concern, in the following sections we use the Bayesian framework to explicitly compare and evaluate the data-based evidence for model classes H_A and H_B, and to account for model uncertainty in inference making when evidence favoring one model against the other is not strong enough.

Table III reports the Bayes factors B_{AB} of the model H_A, as opposed to the instrumental variables model, H_B. We calculate the Bayes factor for each specific form of the GARCH-M and MIDAS models (Models A1-A6), against the simplest form of the instrumental variables model (Model B) using the one-month interest rate as the single predictor. The Bayes factors under a variety of prior specifications are BF1-BF4, as discussed in Section 3. The Monte Carlo simulation approach is used in calculation of the Bayes factors. See the appendix for a brief description of this technique. To evaluate the accuracy of the Monte Carlo integrations, we repeated the numerical procedure for a range of replication numbers, and checked the variations of the results across several different simulations. The reported results are obtained with $10,000$ replications for each integration.[17]

The Bayes factor B_{AB} can be interpreted as the posterior probability that the model hypothesis H_A is true, divided by the posterior probability that the alternative H_B is true. Thus a Bayes factor B_{AB} with a value lower than one, for example, is evidence in favor of the model H_B. Table III reports that the Bayes factors, corresponding to distinct specifications within the universe of the GARCH-M and MIDAS models and a variety of prior specifications, are uniformly less than one, indicating consistent evidence in favor of the instrumental variables model. This is favorable evidence that the expected return on the aggregate stock market is negatively related to the volatility over time, which implies that the unconditional return distribution is negatively skewed.

The Bayesian approach is consistent in that the Bayes factor will favor the true model if one of the examined models, H_A and H_B, is the true model and if enough data is observed, while most classical model selection techniques, such as p-values and AIC, does not guarantee consistency (Berger and Pericchi (2001)). Furthermore, Berk (1966) and Dmochowski (1996) show that, even if the true model is not included in the model space, the Bayesian approach will favor the model among the candidates that is closest to the true model under a certain criterion.[18]

More important, the values of those posterior odds give us valuable information to assess the strength of the evidence. This is especially important because drawing inferences from the selected instrumental variables approach alone, with no knowledge of how strong the evidence is in its

[16] There is no reason to expect a p-value to be similar to the posterior probability that the null hypothesis is correct.

[17] The simulations from MCMC are quite stable and the resulting numerical integrations achieve fast convergence with 10,000 simulations.

[18] In contrast, frequentist tests tend to reject a null hypothesis almost systematically in the case of very large samples because no single model could precisely describe the true underlying stochastic process that generates the data.

favor, would suffer from the critique of ignoring model uncertainty, and could lead to overoptimistic estimates of accuracy. Toward this end, we use the criterion proposed by Jeffreys (1961, App. B) as a scale of evidence for the interpretation of the Bayes factor B_{AB}. It suggests substantial evidence against the hypothesis H_A in favor of H_B if the Bayes factor B_{AB} is between 0.1 and 0.3, strong evidence if B_{AB} is between 0.01 and 0.1, and decisive evidence if B_{AB} is less than 0.01.

With the noninformative diffuse parameter priors (BF1), the highest Bayes factor value under the ARCH structures (A1-A4) is only 0.0098, which indicates that data-fitting even the best ARCH-M model specification is significantly poorer than the instrumental variables approach. The evidence in support of the instrumental variables model becomes even stronger with the proper but nearly noninformative prior (BF2), and is further confirmed by the prior-independent Schwarz criterion (BF4), as most of the Bayes factor values are lower than 0.0001. Meanwhile, the MIDAS approach using high-frequency data in volatility forecasting does not perform any better than the GARCH-M model and is thus also decisively rejected by the data.

Given such strong evidence in the model comparisons conveyed by the Bayes factors, investigators with noninformative prior beliefs on the parameter values would reasonably choose the instrumental variables model for subsequent analysis with no need for concern over model uncertainty, and consequently favor the conclusion of the negative risk-return relation seen in Table II. This is some justification of the early practice of drawing inferences on the risk-return relation exclusively from the instrumental variables models (e.g., Breen, Glosten, and Jagannathan (1989)).

When the prior distributions slightly favor the nonpredictability of stock returns as to both expected return and volatility, the evidence (BF3) in support of hypothesis H_B is not as strong. The Bayes factors corresponding to the ARCH-M class in this case range from 0.0023 to 0.6606, and the highest value corresponds to the posterior odds for the ARCH(2)-M model versus the instrumental variables model. To better understand the interpretation of those numbers, we look at its implied posterior model probabilities. Taking the two models to be equally likely a priori, and noting that the posterior model probabilities sum to one, we use a simple transformation to evaluate the updated uncertainty surrounding the modeling assumptions:

$$p(H_A|D) = \frac{B_{AB}}{1 + B_{AB}} \tag{12}$$

and

$$p(H_B|D) = 1 - p(H_A|D).$$

Take, for example, the comparison between the ARCH(2)-M model and the instrumental variables model that yields the BF3 of 0.6606. According to (12), the posterior probability of H_A is 0.3978, certainly not low enough in relation to the posterior probability of H_B, 0.6022, to justify ignoring model H_A in the subsequent analysis.

As a result, researchers with such nonpredictability prior beliefs must acknowledge that neither the ARCH-M model nor the instrumental variables approach is perfect in describing the true underlying data-generating process, and therefore, the information conveyed by both models should be carefully incorporated in the analysis. This case will be further analyzed in section 4.3.

In both model classes, H_A and H_B, conditional returns are assumed to be normally distributed. Although it has been shown that daily returns have more mass in the tail areas than would be

predicted by a normal distribution, in practice the Central Limit Theorem applies and drives longer-horizon returns towards normality. For instance, Campbell, Lo, and MacKinlay (1997) reports an extremely high and statistically significant sample excess kurtosis of 34.9 for daily index returns and a contrastingly low estimate of only 2.42 for monthly index returns. In spite of this evidence, to determine the robustness of our results to the normality assumption we also compare two model classes that are slightly modified to capture the potential fat-tailed return distributions. Specifically, we compute the Schwarz criterion for models H_A and H_B under the assumption that ε_t has a scaled t-distribution with 5 degrees of freedom instead of a normal distribution. All but one of the resulting Bayes factors B_{AB} are lower than 0.0001 (the exception is for the GARCH-M model, which gives a Bayes factor of 0.0013), suggesting that the decisive evidence favoring the model H_B is not attributed to the normality approximation.

4.3 Inference based on model averaging

For simplicity of the computation involved in (10) and (11), the posterior mean and variance of g conditional on either model is computed using the noninformative diffuse prior throughout, regardless of the distinct forms of priors used in obtaining the Bayes factor. This simplicity is achieved at the cost of some coherence of Bayesian analysis. However, the results should not be affected since in the parameter estimation any influence of the prior specification on the posterior is expected to wash out in the large sample.

The estimation results regarding g, including the posterior mean and standard deviation, accounting for model uncertainty, are presented in Table IV. Again, the analysis is conducted for each pair of models, formed with one specification from each model class. The numbered rows correspond to the four different Bayes factors BF1-BF4 reported in Table III, from which the posterior model probabilities are computed. We see from rows (1), (2), and (4) that the estimation of g conditional on the data alone produces posterior means around -11.4 and posterior standard deviations around 4.5 consistently across prior distributions and model pairs, indicating a significantly negative risk-return relation. Further, these numbers are all quite close to those obtained conditional on the instrumental variables model reported in Table II. This is to be expected, given the decisive evidence conveyed by the Bayes factors, BF1, BF2, and BF4 in Table III, that the model class H_B outperforms H_A.

The more interesting evidence is in row (3), which corresponds to the nonpredictability prior that leads to evidence only marginally favoring the instrumental variables model over the GARCH-M model. In this case, even with the relatively high values of the Bayes factors corresponding to A1-A4 illustrated in Table III, the posterior means of g are still negative, ranging from -11.4 to -2.7, although the significance levels vary. The comparison between the ARCH(2)-M model and the instrumental variables model yields an estimate of -2.7 for g with a posterior standard deviation of 12.6, which suggests no significance. This large measurement error is caused mainly by the considerable uncertainty surrounding the choice between these two model specifications, as indicated by the Bayes factor, 0.6606. In contrast, the inferences accounting for the uncertainty between the EGARCH(1, 1)-M model and the instrumental variables model, due to their relatively low Bayes factors, still produce strong statistically significant evidence regarding the negative value for g. Here, we borrow terminology from the classical approach here. We say that an estimate

is significantly (weakly significantly) different from zero if the posterior mean of the unknown parameter is about two (one and a half) posterior standard deviations away from zero.

It is apparent that the spurious indication of a positive risk-return relation is a result of the ARCH specification of conditional volatility. Although the linear relation in (3), as a good approximation to Merton's (1973) intertemporal CAPM, is ideally suited to investigating the trade-off between the risk and the expected return, consistent estimation for the parameters in this linear relation requires that the full model be correctly specified since the information matrix of the model is no longer block diagonal between the parameters in the conditional mean and variance (Pagan and Ullah (1988)). Hence, if stock returns evolve according to model H_B, the variance process in H_A is misspecified and leads to the biased and inconsistent estimates for the parameters f and g. Furthermore, Pagan (1986) shows that in the two-stage estimation of models A5 and A6 the conventional standard errors may not be appropriate and could result in misleading conclusions about the true underlying relation between expected return and volatility.

4.4 Model extensions

In any statistical modeling, the ultimate goal is a stochastic process that closely approximates the underlying data-generating mechanism, in that it captures most of the key characteristics observed in the data. Although Berk (1966) and Dmochowski (1996) show that the Bayesian model comparison can guarantee selection of the model among the candidates that is closest to the true model, to the extent that the favored model differs from the true one in a manner that is critical to capture certain important characteristics of data, the resulting inference is still questionable. In our case, although the instrumental variables model that forecasts volatility with exogenous instruments is supported by the evidence of the data, the model is inconsistent with at least one important feature of the data, that is, the well-documented volatility clustering phenomenon.

In an effort to identify the "right" model, we extend the model space by considering several extensions of the instrumental variables model. In particular, when making volatility forecasting, we incorporate not only the information in the instruments but also that reflected in past returns so as to capture persistence in the conditional volatility. We add a GARCH component to the instrumental variables model:

Model C1:

$$\sigma_t^2 = \alpha + \beta\varepsilon_t^2 + \gamma\sigma_{t-1}^2 + dx_t.$$

Researchers beginning with Black (1976) have found evidence supporting a negative correlation between current returns and future return volatilities.[19] This feature, however, is ruled out in the GARCH structure considered above, which is symmetric in that negative and positive shocks ε_t have the same impact on the volatility. If the conditional volatility is related to past returns not only through squared return shocks, a symmetric GARCH structure is misspecified, and any empirical results based on it are not reliable. To handle this, we further extend Model C1 by including a term reflecting this asymmetric volatility effect as:

[19] Two popular explanations for the association of negative return innovations with positive volatility shocks are the leverage hypothesis (Black (1976)) and the volatility-feedback hypothesis (Campbell and Hentschel (1992)). See Campbell, Lo, and MacKinlay (1997) for further discussion.

Model C2:

$$\sigma_t^2 = \alpha + \beta|\varepsilon_t - \eta|^2 + \gamma\sigma_{t-1}^2 + dx_t.$$

Here if the shift parameter $\eta = 0$, we are back to the symmetric Model C1. If η is positive, volatility increases less when there is a positive shock of size η than when there is no shock.[20]

Table V presents the estimation results of these two generalized models. Model C1 reflects a certain degree of persistence in the monthly return volatility, i.e., either β or γ is positive and at least weakly significantly different from zero.[21] Under the more general Model C2, the parameter on the asymmetry of volatility effect, η, is estimated to be 0.0977 and statistically significant with a posterior standard deviation of 0.0328, which, along with the positive β and γ, is consistent with the early finding that negative shocks to stock returns tend to increase volatility more than positive shocks of the same magnitude.

More important, the coefficient on the risk-return relation, g, is negative in both models, and weakly significant in Model C2. The fact that this negative relation remains significant even after adding past returns to the information set suggests the crucial role of the one-month interest rate in the intertemporal relation of stock return moments. This result is in accordance with the findings of Glosten, Jagannathan, and Runkle (1993), who examine models similar to Model C2.

Models C1 and C2 can also be viewed as generalizations based on Model A3, the simple GARCH(1, 1) structure. From this perspective, there are several points worth noting as well. First, the risk-return coefficient g changes from positive in Model A3 to negative as soon as the interest rate is added as an instrument in Model C1, and remains negative after the asymmetric volatility effect is allowed in Model C2. Second, the volatility persistence captured by the ARCH and GARCH components, as roughly measured by the sum of β and γ, quickly declines from around 0.9 in the simple GARCH(1, 1) model to around 0.3 in the generalized versions. Third, both β and γ are estimated to be strongly significant in Model A3, where the posterior mean of β is around four times the posterior standard deviation away from zero and the posterior mean of γ is at least 20 times the posterior standard deviation away from zero, but these significances are much weaker in Model C1 and C2. The latter two observations occur because much of the volatility persistence is now captured by the interest rate series, which is itself highly persistent, with a first-order autocorrelation of 0.95 (note that the slope coefficient of the interest rate is positive and substantially distinguishable from zero).

Table VI compares the generalized models, H_C, and the instrumental variables model, H_B, and Table VII estimates the parameter β by averaging the information conveyed by these two models. The Bayes factors corresponding to C1 versus B seem quite sensitive to the parameter prior distributions, yielding values lower or higher than one, depending on the forms of the priors. This sensitivity is not surprising, given the close similarity of the two model classes, H_B and H_C. Further, because both C1 and B produce a negative estimate of g, our conclusion of a negative risk-return relation is unchanged, regardless of which model outperforms the other, as shown in Table VII. In other words, model uncertainty between models C1 and B is not critical to our ultimate

[20] See Campbell, Lo, and MacKinlay (1997) for a discussion about alternative functional forms to capture volatility asymmetry.

[21] A more serious test on the time variation of the return volatility will be to run a joint test of both β and γ being equal to zero.

inference on the parameter of particular interest.

When the asymmetric volatility effect is incorporated, the data decisively and consistently favor Model C2 over Model B, as the Bayes factors in this case are mostly greater than 10^4 across a variety of prior specifications, and the smallest value is close to 10^3. This demonstrates that the asymmetric volatility effect of return shocks of different signs plays an important role in the time variation of stock return volatility.

Readers may have already noted that the generalized model, H_C, has nested the simpler versions H_A and H_B, and thus the classical hypothesis test can also be applied to test H_A or H_B against the alternative H_C. Because of the aforenoted analogy of the likelihood ratio statistic to the Bayes factor, the classical hypothesis test should be expected to lead to results consistent with those from our Bayesian approach, that is, to reject the null hypothesis and accept H_C. Unlike the Bayes factors, the p-value from classical tests is far from a probability measure and is therefore difficult to interpret by nonstatisticians in an attempt to assess the strength of the supporting evidence. This is one of the most important advantages that the Bayesian model comparison approach has over the classical method.

Since the finding of a negative risk-return relation seems to be primarily a result of the use of the one-month interest rate for the volatility forecasting, this calls for some further empirical investigation of the true forecasting ability of this instrument that is extensively used in the return predictability literature. For this purpose, we first estimate the realized variance of the monthly returns following French, Schwert, and Stambaugh (1987) as:

$$\widehat{\sigma}_t^2 = \sum_{d=1}^{N_t} r_{t-d}^2 + 2\sum_{d=1}^{N_t-1} r_{t-d} r_{t-d-1}, \tag{13}$$

where $\widehat{\sigma}_t^2$ denotes the realized volatility in month t, N_t is the number of trading days in month t, and r denotes daily returns.[22] The second term in (13) is included to adjust for serial correlation in the daily returns induced by nonsynchronous trading.

Figure I shows the time series of realized volatilities (thin line) and one-month interest rates (thick line), along with their correlations. We see that the one-month interest rates do provide somewhat valuable information regarding the realized volatility, as suggested by their similar patterns in the time-varying trend and the correlation of 0.14. This could be partly explained by the fact that return volatility tends to increase with an increase in expected inflation, which is incorporated by the market in the determination of the interest rate.

Figure II plots the time series of realized (thin line) and forecasted (thick line) variance of monthly returns for June 1951 through December 2001 using monthly data and the parameter estimates reported in Table II. The first plot displays the forecasted variance estimated by the GARCH(1, 1)-M model (Model A3), and the second one that estimated by the generalized GARCH-M model (Model C2). For a better view on how the forecasted variance tracks the realized variance over time, we display the same plots in the shorter 15-year period January 1987 through December 2001 in Figure III. Correlations between series for the full sample are given below each plot.

In general, the volatility process forecasted by the simple GARCH(1, 1) structure is too smooth to capture many small oscillations in the realized variance, although it does a good job in reflecting

[22] No adjustment is included with respect to the sample mean since, as French, Schwert, and Stambaugh (1987) note, the impact of such small modifications is minimal.

the long-run trend. The generalized GARCH structure, however, after incorporating the information in interest rates and allowing for an asymmetric volatility effect, produces a conditional volatility series that tracks the realized volatility much more closely. This is partially because of the additional oscillation induced by the interest rate series and the correlation of 0.14 between interest rates and the realized volatilities.

5 Conclusion

This paper proposes a Bayesian model comparison framework for examining how the expected return and volatility of the aggregate stock market move together over time, whether they are positively or negatively related. The mixed results in the literature are due largely to differences in the return moment specifications. In general, models that forecast next-month's return volatility using only past returns data produce a positive, albeit sometimes weakly significant, risk-return relation, while models that make use of exogenous instruments such as one-month interest rates indicate a contrary negative relation.

The Bayesian procedure in our paper complements the work by Glosten, Jagannathan, and Runkle (1993), who find that among several generalizations of the standard GARCH-M model, the models that indicate a negative risk-return relation are better specified than those that suggest a contrary result. The diagnostic tests they employ cannot provide a measure for the strength of the evidence supporting their model selection decision, based on which they make the conclusion of a negative risk-return relation. As it is important to report standard errors or confidence intervals as a measure for accuracy in parameter estimations, it is also important to associate the selected model with a measure for the strength of the supporting evidence.

This has been achieved in our study. Overall, models that include one-month interest rates in the information set for volatility forecasting generally outperform models that do not, and models that allow for an asymmetric volatility effect outperform models that rule it out. In addition, the evidences that distinguish those models are shown to be decisive. We thus conclude that the sample evidence strongly favors a negative relation between the time-varying conditional means and volatilities of stock returns. This result is robust to model uncertainty, at least within the universe of the models examined.

Several studies have assessed the implications of the time-varying return moments, either expected return or volatility or both, for portfolio decision making.[23] If the negative intertemporal relation between the first two return moments exists, more interesting results are expected regarding portfolio implications, since any impact of either return moment on the optimal portfolio can be magnified by the associated movement of the other return moment. For instance, Kandel and Stambaugh (1996) show significant economic values of predictability on expected return by showing that current values of predictive variables can exhibit a substantial impact on a Bayesian investor's one-month optimal stock allocation. Incorporating the conditional heteroskedasticity and acknowledging the negative relation between the first and second return moments could cause a higher sensitivity of the optimal stock allocation to the current values of predictive variables. Of course,

[23] See, for example, Kandel and Stambaugh (1996), Brennan, Schwartz, and Lagnado (1997), Campbell and Viceira (1999), Barberis (2000), Lynch and Balduzzi (2000), and Wang (2004).

adding these two new features into the return dynamics could also change the manner by which the current values of predictive variables influence future return moments.

Our study illustrates the importance of model comparison in the face of conflicting indications of different empirical specifications, and shows the effectiveness of the Bayesian technique in implementing this idea to resolve a puzzling situation. Besides the volatility specifications, there are certainly many other interesting dimensions for possibly having different modeling approaches. Further, in other areas, such as the term structure models of interest rates, where a variety of empirically motivated specifications exist, special care also needs to be taken in making inference that could potentially depend on the used model. The framework we employ in this study can be easily applied to those situations although it will be computationally more challenging with more complicated models.

Appendix A. A description of the Monte Carlo integration used in computing the Bayes factors

In most cases, evaluating the integral (5) involved in the Bayes factor, which we rewrite here as

$$I = \int p(D|\theta, H)p(\theta|H)d\theta, \tag{A-1}$$

can be challenging in the absence of analytical solutions, which are rarely available. The traditional grid method of numerical integration such as Gaussian or Gauss-Hermite quadrature algorithms can be difficult to implement, especially when the parameter space is high-dimensional. With moderate or large sample sizes, the likelihood function $p(D|\theta, H)$ is likely to be highly peaked around its maximum, so extreme care needs to be taken to ensure the accuracy of the numerical solution by appropriately adjusting the grids around the peak. This difficulty increases rapidly with the dimension of the parameter space.

A Monte Carlo simulation method offers a convenient and efficient way to solve high-dimensional integration problems. This appendix describes how to use the Monte Carlo integration technique in the Bayesian approach where one term in the integrand, $p(\theta|H)$, is the parameter prior distribution, which could be proper or improper. For a textbook treatment of the Bayesian calculation, see Berger (1985).

In the case of a proper parameter prior, it is possible to generate an i.i.d. sequence of random samples $\{\theta_i,\ i = 1, ..., m\}$ from the density $p(\theta|H)$, where m is the replication number. Note that the integral in (A-1) can be written in the expectation form

$$I = E[p(D|\widetilde{\theta}, H)]$$

where the random variable $\widetilde{\theta}$ is distributed according to $p(\theta|H)$. It then follows from the strong law of large numbers in probability theory that, under mild regularity conditions, the simple Monte Carlo integration estimate

$$\widehat{I} = \frac{1}{m}\sum_{i=1}^{m} p(D|\theta_i, H) \tag{A-2}$$

almost surely converges to I.

When the likelihood function is highly peaked, it is nearly zero over all but a small portion of the support of $p(\theta|H)$, so most of the random samples drawn from the density $p(D|\theta, H)$ will contribute little to the integral value and to reducing variability across simulations. Thus, in practice a large number of replications, m, are needed for each simulation, and several different series of the simulated $\{\theta_i\}$ are tried in (A-2) to check its variability so as to ensure accuracy of the estimate.

For improper prior $p(\theta|H)$, we apply a slightly modified version of the above Monte Carlo technique, often referred to as the importance sampling approach. It begins by writing the integral I as

$$I = \int \frac{p(D|\theta, H)p(\theta|H)}{h(\theta)} h(\theta)d\theta,$$

where $h(\theta)$ is some proper density from which an i.i.d. sequence of random samples $\{\theta_i,\ i = 1,\ ...,\ m\}$ can be drawn. The integral I can then be approximated with

$$\widehat{\widehat{I}} = \frac{1}{m} \sum_{i=1}^{m} \frac{p(D|\theta_i, H)p(\theta_i|H)}{h(\theta_i)}.$$

The key issue here is to find a suitable $h(\theta)$. Some guidance may be gained by looking at the variance of the importance sampling estimate:

$$V(\widehat{\widehat{I}}) = \frac{1}{m} V(\frac{p(D|\widetilde{\theta}, H)p(\widetilde{\theta}|H)}{h(\widetilde{\theta})}).$$

where $\widetilde{\theta}$ is distributed according to the density $h(\theta)$.

As one can see, the ideal choice of $h(\theta)$ will be one that is exactly proportional to $p(D|\theta, H)p(\theta|H)$, but this cannot be achieved since it requires us to know the integral value, which is what we are trying to estimate. This still provides a rule of thumb that $h(\theta)$ should try to mimic the posterior distribution as closely as possible. For large sample sizes of D, theory shows that, under commonly satisfied assumptions, the posterior distribution will typically be well approximated by a multivariate normal distribution $N(\mu^p,\ \Sigma^p)$, where μ^p and Σ^p are the posterior mean and covariance matrix (Berger (1985), subsection 4.7.8). Thus this normal density could be a reasonable choice of $h(\theta)$.

References

Abel, Andrew, 1988, Stock prices under time-varying dividend risk, *Journal of Monetary Economics* 22, 375–393.

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

Backus, David, and Allan Gregory, 1992, Theoretical relations between risk premiums and conditional variances, *Journal of Business and Economic Statistics* 11, 177–185.

Bali, Turan, and Lin Peng, 2003, Is there a risk-return tradeoff? Evidence from high-frequency data, Working paper, City University of New York.

Barberis, Nicholas, 2000, Investing for the long run when returns are predictable, *Journal of Finance* 55, 225–264.

Berger, James, 1985, *Statistical Decision Theory and Bayesian Analysis* (Springer-Verlag, New York, NY).

——— , and Luis Pericchi, 1996, The intrinsic Bayes factor for model selection and prediction, *Journal of the American Statistical Association* 91, 109–122.

——— , 2001, Objective Bayesian Methods for Model Selection: Introduction and Comparison, in *Lecture Notes - Monograph Series*, vol. 38 . pp. 135–207 (Institute of Mathematical Statistics).

Berger, J., and T. Sellke, 1987, Testing a point null hypothesis: The irreconcilability of p-values and evidence, *Journal of the American Statistical Association* 82, 112–122.

Berk, R., 1966, Limiting behavior of posterior distributions when the model is incorrect, *Annals of Mathematical Statistics* 37, 51–58.

Black, F., 1976, Studies of stock market volatility changes, *Proceedings of the American Statistical Association, Business and Economic Statistics Section,* pp. 177–181.

Bollerslev, Tim, 1986, Generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics* 31, 307–327.

Breen, William, Lawrence R. Glosten, and Ravi Jagannathan, 1989, Economic significance of predictable variations in stock index returns, *Journal of Finance* 44, 1177–1189.

Brennan, Michael J., Eduardo S. Schwartz, and Ronald Lagnado, 1997, Strategic asset allocation, *Journal of Economic Dynamics and Control* 21, 1377–1403.

Campbell, John, 1993, Intertemporal asset pricing without consumption data, *American Economic Review* 83, 487–512.

Campbell, J., and L. Hentschel, 1992, No news is good news: An asymmetric model of changing volatility in stock returns, *Journal of Financial Economics* 31, 281–318.

Campbell, John, and Luis Viceira, 1999, Consumption and portfolio decisions when expected returns are time varying, *Quarterly Journal of Economics* 114, 433–495.

Campbell, John Y., 1987, Stock returns and the term structure, *Journal of Financial Economics* 18, 373–399.

——— , and John H. Cochrane, 1999, By force of habit: A consumption based explanation of aggregate stock market behavior, *Journal of Political Economy* 107, 205–251.

Campbell, John Y., Andrew Lo, and Craig MacKinlay, 1997, *The Econometrics of Financial Markets* (Princeton University Press: Princeton).

Campbell, John Y., and Robert J. Shiller, 1988, Stock prices, earnings, and expected dividends, *Journal of Finance* 43, 661–676.

Chan, K. C., G. A. Karolyi, and René M. Stulz, 1992, Global financial markets and the risk premium on U.S. equity, *Journal of Financial Economics* 32, 137–167.

Chen, Nai-Fu, Richard Roll, and Stephen A. Ross, 1986, Economic forces and the stock market, *Journal of Business* 59, 383–403.

Chipman, Hugh, Edward I. George, and Robert E. McCulloch, 2001, The practical implementation of Bayesian model selection, in *Model Selection* (Institute of Mathematical Statistics, Beachwood, Ohio).

Constantinides, George M., 1990, Habit formation: A resolution of the equity premium puzzle, *Journal of Political Economy* 98, 519–543.

Cox, D., 1961, Tests of separate families of hypotheses, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* 1 Berkeley: University of California Press.

——— , 1962, Further results on tests of separate families of hypotheses, *Journal of the Royal Statistical Society, Series B* 24, 406–424.

Cremers, Martijn, 2002, Stock return predictability: A Bayesian model selection perspective, *The Review of Financial Studies* 15, 1223–1249.

Davidson, R., and J. MacKinnon, 1981, Several tests for model specification in the presence of alternative hypotheses, *Econometrica* 49, 781–793.

Dmochowski, J., 1996, Intrinsic priors via Kullback-Leibler Geometry. In Bayesian Statistics 5 (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.), Oxford Univ. Press.

Engle, Robert, 1982, Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation, *Econometrica* 50, 987–1008.

——— , David Lilien, and Russell Robins, 1987, Estimating time varying risk premia in the term structure: The ARCH-m model, *Econometrica* 55, 391–407.

Fama, Eugene F., and Kenneth R. French, 1988, Dividend yields and expected stock returns, *Journal of Financial Economics* 22, 3–25.

Ferson, Wayne E., and Campbell R. Harvey, 1991, The variation of economic risk premiums, *Journal of Political Economy* 99, 385–415.

Freedman, D. A., 1983, A note on screening regression equations, *The American Statistician* 37, 152–155.

French, Kenneth, William Schwert, and Robert Stambaugh, 1987, Expected stock returns and volatility, *Journal of Financial Economics* 19, 3–29.

Ghysels, Eric, Pedro Santa-Clara, and Rossen Valkanov, 2004, There is a risk-return tradeoff after all, *Journal of Financial Economics* forthcoming.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter, 1996, *Markov Chain Monte Carlo in Practice* (Chapman and Hall, London).

Glosten, Lawrence, Ravi Jagannathan, and David Runkle, 1993, On the relation between the expected value and volatility of the nominal excess return on stocks, *Journal of Finance* 48, 1779–1801.

Harvey, Campbell, 2001, The specification of conditional expectations, *Journal of Empirical Finance* 8, 573–637.

Jeffreys, H., 1961, *Theory of Probability (3rd Ed.)* (Oxford, U.K.: Oxford University Press.).

Kandel, Shmuel, and Robert F. Stambaugh, 1996, On the predictability of stock returns: An asset-allocation perspective, *Journal of Finance* 53, 385–424.

Kass, Robert, and Adrian Raftery, 1995, Bayes factors, *Journal of the American Statistical Association* 90, 773–795.

Keim, Donald B., and Robert F. Stambaugh, 1986, Predicting returns in the stock and bond markets, *Journal of Financial Economics* 17, 357–390.

Lee, S., and B. Hansen, 1994, Asymptotic theory for the GARCH(1, 1) quasi-maximum likelihood estimator, *Econometric Theory* 10, 29–52.

Lynch, Anthony, and Pierluigi Balduzzi, 2000, Predictability and transaction costs: The impact on rebalancing rules and behavior, *Journal of Finance* 55, 2285–2309.

Maritz, J. S., and T. Lwin, 1989, *Empirical Bayes Methods* (Chapman and Hall, London).

Merton, Robert, 1980, On estimating the expected return on the market: An exploratory investigation, *Journal of Financial Economics* 8, 323–361.

Merton, Robert C., 1973, An intertemporal capital asset pricing model, *Econometrica* 41, 867–887.

Nelson, Daniel, 1991, Conditional heteroskedasticity in asset returns: A new approach, *Econometrica* 59, 347–370.

Pagan, Adrian, 1986, Two stage and related estimators and their applications, *Review of Economic Studies* 53, 517–538.

——— , and Aman Ullah, 1988, The econometric analysis of models with risk terms, *Journal of Applied Econometrics* 3, 87–105.

Schwarz, G., 1978, Estimating the dimension of a model, *The Annals of Statistics* 6, 461–464.

Scruggs, John, 1998, Resolving the puzzling intertemporal relation between the market risk premium and conditional market variance: A two-factor approach, *Journal of Finance* 53, 575–603.

Stulz, René, 1986, Asset pricing and expected inflation, *Journal of Finance* 41, 209–223.

Tierney, L., and J. B. Kadane, 1986, Accurate approximations for posterior moments and marginal densities, *Journal of the American Statistical Association* 81, 82–86.

Wang, Leping, 2004, Investing when volatility fluctuates, working paper, University of Pennsylvania.

Whitelaw, Robert F., 1994, Time variations and covariations in the expectation and volatility of stock market returns, *The Journal of Finance* 49, 515–541.

Table I: Existing literature on the risk-return relation

The table reports the conditioning variables used in the literature on forming the parametric return variance process to examine the monthly stock risk-return relation:

$$E(R_{t+1}|F_t) = f + gV(R_{t+1}|F_t),$$

where R_t denotes the monthly excess return of the stock index in excess of the risk-free return, F_t is the information set available to investors at time t, and f and g are the parameters, with the sign of g of particular interest. For each paper we report the authors, the year of publication, the information set used in the analysis, and the consequent conclusions on the sign of g. In other notation, ε_t is the disturbance given by $R_t - E(R_t|F_{t-1})$, σ_t^2 denotes the conditional volatility, $R_{i,t}^f$ stands for the ith-month bill rate, OCT_t and JAN_t are the dummy variables for October and January, respectively, ys_t denotes the Baa-Aaa yield spread, dy_t is the excess dividend yield, cs_t is the commercial paper-treasury spread, and I_t is an indicator function that takes the value of one if ε_t is positive.[24] A $+$ and $-$ indicate at least weakly significant positive and negative values for the parameter g, respectively. 0 means that g is statistically indistinguishable from zero.

Authors	Year	Volatility conditioning variables	g
French, Schwert, and Stambaugh[25]	1987	σ_{t-1}^2, ε_t^2, ε_{t-1}^2	$+$
		past daily squared returns	0
Campbell	1987	$R_{1,t}^f$, $R_{k,t}^f - R_{1,t}^f$, $R_{2,t-1}^f - R_{1,t-1}^f$	$-$
Breen, Glosten, and Jagannathan	1989	$R_{1,t}^f$	$-$
Chan, Karolyi, and Stulz	1992	bivariate GARCH-M	0
Campbell and Hentschel	1992	σ_{t-1}^2, $(\varepsilon_t - b)^2$	$+$
Glosten, Jagannathan, and Runkle	1993	σ_{t-1}^2, ε_t^2	$+$
		$R_{1,t}^f$, OCT_t, JAN_t, ε_t^2, $\varepsilon_t^2 I_t$	$-$
Whitelaw	1994	ys_t, dy_t, $R_{12,t}^f$, cs_t	$-$
Harvey	2001	ε_{t-j}^2, $j = 0, ..., 7$	$+$
		ε_t^2, $R_{3,t}^f - R_{1,t}^f$, ys_t, dy_t	$-$
		ε_t^2, $\varepsilon_t/\sigma_{t-1}$	$+$
Ghysels, Santa-Clara, and Valkanov	2003	past daily squared returns	$+$

[24] The rule we use for dating variables differs from that in some other studies. Throughout our paper, we give the variable a time subscript t if its value is observable at the end of month t.

[25] In order to capture the effect of nonsynchronous trading, French, Schwert, and Stambaugh include the moving average term $\theta\varepsilon_{t-1}$in the return innovations, yielding a model slightly different from ours.

Table II: Parameter estimations conditional on a single model of various types

The table reports the Bayesian posterior means and standard deviations of the parameters exclusively conditional on a single model of various types — Models A1-A6 and Model B. Parameters are assumed a priori independent and diffusely distributed on the corresponding parameter spaces. The posterior standard deviation of each parameter is given in the parentheses. The results regarding the parameter g, which relates expected return to volatility, is of particular concern and is emphasized with boldface. The sample period is 1951.6 through 2001.12.

	Model specifications						
	A1	A2	A3	A4	A5	A6	B
f	0.0081	−0.0108	−0.0013	−0.0080	0.0025	0.0029	0.0251
	(0.0244)	(0.0158)	(0.0046)	(0.0152)	(0.0018)	(0.0018)	(0.0071)
g	**−0.5669**	**10.3900**	**4.8715**	**8.7968**	**2.3578**	**2.1058**	**−11.3987**
	(**14.5439**)	(**9.2081**)	(**2.7212**)	(**9.1656**)	(**1.6579**)	(**1.5975**)	(**4.4863**)
α	0.0016	0.0015	0.0001	−1.2398			
	(0.0001)	(0.0001)	(0.0000)	(1.1350)			
β	0.0896	0.0637	0.0872	0.0499			
	(0.0493)	(0.0414)	(0.0227)	(0.0172)			
δ		0.0904					
		(0.0509)					
γ			0.8407	0.8142			
			(0.0409)	(0.1773)			
c							0.0007
							(0.0002)
d							0.2283
							(0.0469)

Table III: Bayes factor of various specifications within model class H_A as opposed to H_B

The table reports the Bayes factor (or equivalently the ratio of posterior model probabilities) of the various specifications within the model class H_A, denoted by Models A1-A6, as opposed to Model B. The Bayes factor, defined as $B_{AB} = \frac{p(D|H_A)}{p(D|H_B)}$, is computed numerically. In a robustness check, we report the Bayes factors computed using three distinct forms of the parameter prior distributions, one noninformative improper prior and two proper priors. The resulting Bayes factors are denoted by BF1, BF2, and BF3. The Schwarz (1978) criterion, denoted by BF4, is also reported as a reference to the Bayes factor.

	Model pairs					
Bayes factor	A1 vs B	A2 vs B	A3 vs B	A4 vs B	A5 vs B	A6 vs B
BF1	0.0020	0.0015	0.0098	0.0024	< 0.0001	0.0024
BF2	< 0.0001	< 0.0001	0.0032	< 0.0001	< 0.0001	< 0.0001
BF3	0.2043	0.6606	0.1972	0.0023	< 0.0001	< 0.0001
BF4	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001

Table IV: Estimations of risk-return relation on the monthly stock index accounting for model uncertainty

The table reports posterior means and standard deviations of the parameter g using the Bayesian model averaging technique. Model uncertainty is incorporated between the various specifications within the model class H_A, denoted by Models A1-A6, and Model B, by reporting an average of the posterior distributions under each model weighted by posterior model probabilities. The numbered rows correspond to the four different Bayes factors BF1-BF4 reported in Table III, from which the model posterior probabilities are computed. Posterior standard deviations are given in parentheses. The sample period is 1951.6 through 2001.12.

	Model pairs					
g	A1 vs B	A2 vs B	A3 vs B	A4 vs B	A5 vs B	A6 vs B
(1)	−11.3775	−11.3657	−11.2409	−11.3508	−11.3983	−11.3667
	(4.5530)	(4.5762)	(4.7482)	(4.6090)	(4.4868)	(4.5294)
(2)	−11.3987	−11.3987	−11.3470	−11.3978	−11.3987	−11.3987
	(4.4863)	(4.4864)	(4.5744)	(4.4887)	(4.4863)	(4.4863)
(3)	−9.5615	−2.7313	−8.7189	−11.3525	−11.3987	−11.3987
	(8.3135)	(12.6323)	(7.3791)	(4.6046)	(4.4863)	(4.4863)
(4)	−11.3987	−11.3987	−11.3966	−11.3982	−11.3987	−11.3987
	(4.4863)	(4.4863)	(4.4899)	(4.4875)	(4.4863)	(4.4863)

Table V: Parameter estimations conditional on a single model of various generalizations of the simple GARCH(1, 1)-M model and instrumental variables model

The table reports the Bayesian posterior means and standard deviations of the parameters exclusively conditional on a single model of two generalizations, Models C1 and C2. Parameters are assumed a priori independent and diffusely distributed on the corresponding parameter spaces. Posterior standard deviations are given in parentheses. The results regarding the parameter g, which relates expected return to volatility, is of particular concern and is emphasized with boldface. The sample period is 1951.6 through 2001.12.

	Model specifications	
	C1	C2
f	0.0168	0.0142
	(0.0058)	(0.0040)
g	$\mathbf{-5.8715}$	-4.8484
	(**3.6614**)	(**2.5819**)
α	0.0005	0.0001
	(0.0002)	(0.0003)
β	0.0664	0.0800
	(0.0376)	(0.0352)
η		0.0977
		(0.0328)
γ	0.2427	0.1934
	(0.1786)	(0.1167)
d	0.1680	0.1204
	(0.0537)	(0.0379)

Table VI: Bayes factor of various specifications within model class H_C as opposed to H_B

The table reports the Bayes factor (or equivalently the ratio of posterior model probabilities) of the various specifications within the model class H_C , denoted by Models C1 and C2, and Model B. The Bayes factor, defined as $B_{CB} = \frac{p(D|H_C)}{p(D|H_B)}$, is computed numerically. In a robustness check, we report the Bayes factors computed using three distinct forms of the parameter prior distributions, one noninformative improper prior and two proper priors. The resulting Bayes factors are denoted by BF1, BF2, and BF3. The Schwarz (1978) criterion, denoted by BF4, is also reported as a reference to the Bayes factor.

	Model pairs	
Bayes factor	C1 vs B	C2 vs B
BF1	3.4307	> 10000
BF2	0.8064	> 10000
BF3	2.3846	1122
BF4	0.0023	1141

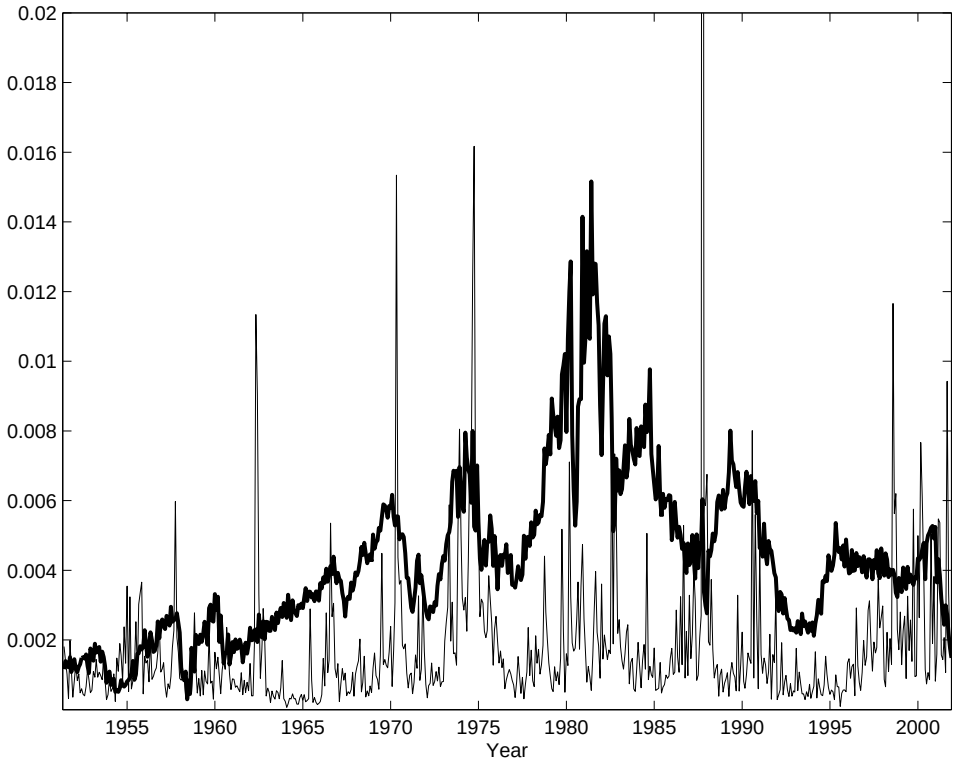
Table VII: Estimations of risk-return relation on the monthly stock index accounting for model uncertainty

The table reports posterior means and standard deviations of the parameter g using the Bayesian model averaging technique. Model uncertainty is incorporated between the various specifications within the model class H_C, denoted by Models C1 and C2, and Model B, by reporting an average of the posterior distributions under each model weighted by posterior model probabilities. The numbered rows correspond to the four different Bayes factors BF1-BF4 reported in Table III, from which the model posterior probabilities are computed. Posterior standard deviations are given in parentheses. The sample period is 1951.6 through 2001.12.

	Model pairs	
g	C1 vs B	C2 vs B
(1)	-7.1189	-4.8484
	(4.5013)	(2.5819)
(2)	-8.9313	-4.8484
	(4.9675)	(2.5819)
(3)	-7.5045	-4.8543
	(4.6638)	(2.5916)
(4)	-11.3862	-4.8542
	(4.4922)	(2.5914)

Figure I: Realized variance of monthly returns and one-month interest rates

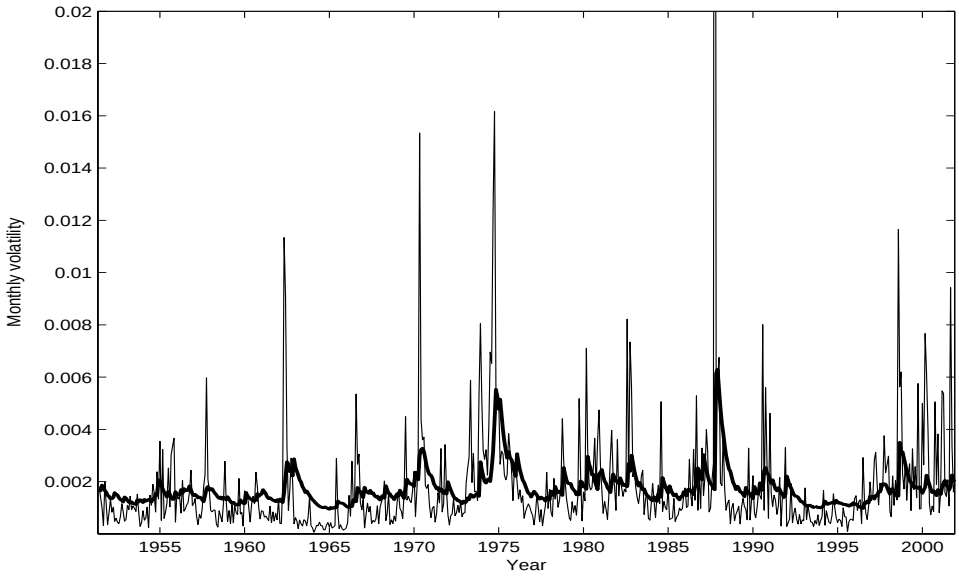
This figure plots the time series of the realized variance of monthly returns (thin line), estimated using equation (13) with within-month daily returns, and the one-month interest rates (thick line) for the period June 1951 through December 2001. To give a clear view of the plot as a whole, we truncate the realized variance of October 1987 at 0.02. The actual value is 0.0689. The correlation between the two series for the full sample is given below the plot.



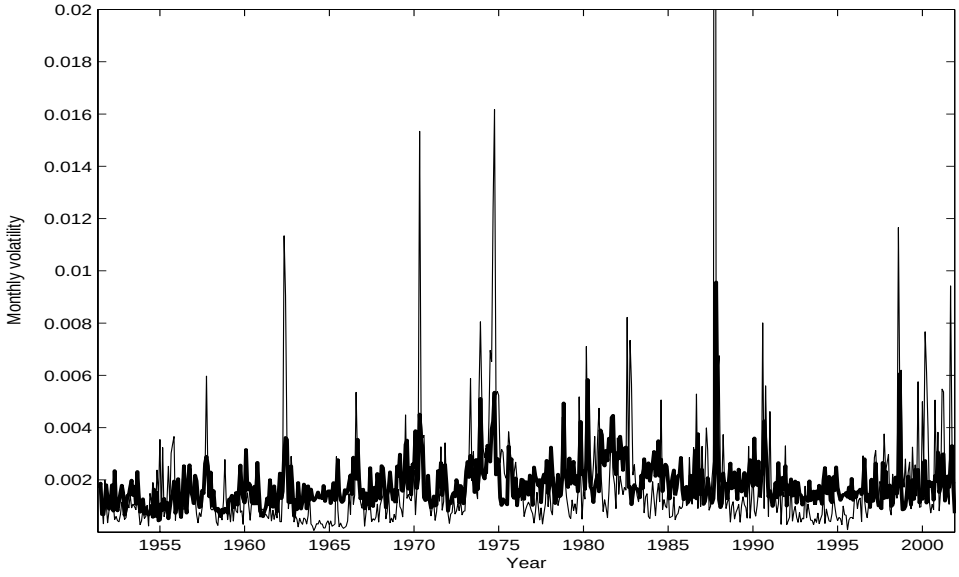
Correlation = 0.14

Figure II: Realized and forecasted variance of monthly returns

This figure plots the time series of the realized (thin line) and forecasted (thick line) variance of monthly returns for the period June 1951 through December 2001. The realized variances are estimated using equation (13) with within-month daily returns. The first plot displays the forecasted variance estimated by the GARCH-M model (Model A3), and the second plot the variance estimated by the generalized GARCH-M model (Model C3), where the estimates reported in Table II are taken as the true values of the parameters. To give a clear view of the plot as a whole, we truncate the realized variance of October 1987 at 0.02. The actual value is 0.0689. The correlations between the two series for the full sample are given below the plots.



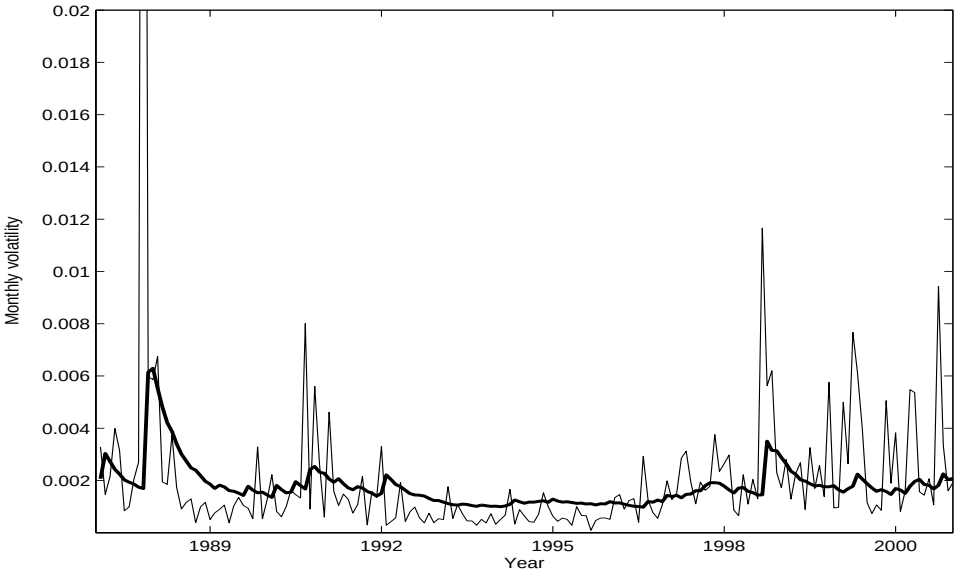
Correlation = 0.24



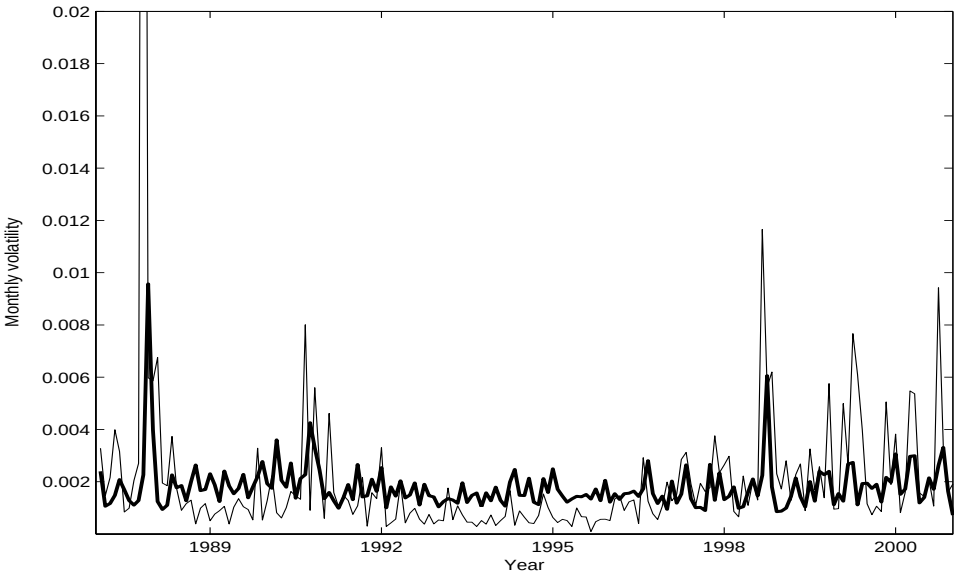
Correlation = 0.28

Figure III: Realized and forecasted variance of monthly returns

This figure plots the time series of the realized (thin line) and forecasted (thick line) variance of monthly returns for the subsample period January 1987 through December 2001. The realized variances are estimated using equation (13) with within-month daily returns. The first plot displays the forecasted variance estimated by the GARCH-M model (Model A3), and the second plot the variance estimated by the generalized GARCH-M model (Model C3), where the estimates reported in Table II are taken as the true values of the parameters. To give a clear view of the plot as a whole, we truncate the realized variance of October 1987 at 0.02. The actual value is 0.0689. The correlations between the two series for the full sample period June 1951 through December 2001 are given below the plots.



Correlation = 0.24



Correlation = 0.28