Stock Return Predictability: A Bayesian Model Selection Perspective

Author(s): K. J. Martijn Cremers

Source: *The Review of Financial Studies*, Autumn, 2002, Vol. 15, No. 4 (Autumn, 2002), pp. 1223-1249

Published by: Oxford University Press. Sponsor: The Society for Financial Studies.

Stable URL: https://www.jstor.org/stable/1262696

Stock Return Predictability: A Bayesian Model Selection Perspective

K. J. Martijn Cremers
Yale University

Attempts to characterize stock return predictability have resulted in little consensus on the important conditioning variables, giving rise to model uncertainty and data snooping fears. We introduce a new methodology that explicitly incorporates model uncertainty by comparing all possible models simultaneously and in which the priors are calibrated to reflect economically meaningful information. Our approach minimizes data snooping given the information set and the priors. We compare the prior views of a skeptic and a confident investor. The data imply posterior probabilities that are in general more supportive of stock return predictability than the priors for both types of investors.

The characterization of stock return predictability is arguably the most hotly debated issue of empirical asset pricing in the last decade and a half. Its importance for understanding the nature of time-varying risk premia or potential market inefficiencies is unquestioned. Not surprisingly, the last two decades of financial research have seen a plethora of articles documenting the ability of various variables to explain movements in conditional expected returns. Though a few naysayers exist, usually employing some type of data snooping argument, the literature is generally in favor of time-varying expected returns.[1]

While there are numerous articles documenting predictability, there is little consensus across these articles on what the important conditioning variables are.[2] This may seem puzzling given the amount of time and effort devoted to this exact task. From a Bayesian perspective, however, the current approach for identifying the factors that explain expected returns is misguided. The search for the variables with the largest t-statistics places all the weight on one specific model, which clearly ignores the most important issue, namely the tremendous uncertainty the researcher has about the correct model.

This article is based on the author's dissertation at the Stern School of Business at New York University. The author thanks his committee members—Matthew Richardson (the chair), Stephen Brown, Frank Diebold, Edwin Elton, Anthony Lynch, Robert Whitelaw—as well as John Heaton (the editor), Aart de Vos, an anonymous referee, and seminar participants at New York University, Tilburg University, and the E.F.M.A. and E.F.A. 2000 annual meeting for many helpful comments and suggestions that greatly improved this article. All errors are mine. Address correspondence to K. J. Martijn Cremers, Yale School of Management, International Center for Finance, Box 20820, 135 Prospect Street, New Haven, CT 06520, or e-mail: kcremers@stern.nyu.edu.

[1] See, for example, Lo and MacKinlay (1990), Richardson (1993), Foster, Smith, and Whaley (1997), and Bossaerts and Hillion (1999) for skeptical views of stock return predictability.

[2] Dividend yields and past returns are probably the most frequently used conditioning variables, but there are an equally large number of studies ignoring these variables and instead picking their own *economically motivated* choices.

The Review of Financial Studies Fall 2002 Vol. 15, No. 4, pp. 1223–1249

This article takes a step back from the current literature by investigating, using a Bayesian model selection perspective, a standard linear model of stock returns in terms of a large collection of candidate predictive variables. Most important, we introduce a new methodology that explicitly accounts for the large degree of model uncertainty and in which the priors are calibrated to reflect economically meaningful and intuitive prior information. We consider the cases of investors who are skeptical and confident about predictability. Our analysis incorporates their respective prior views on the expected R^2 of the predictive regression, of the variance of the residuals, and of the number of included predictors.

The priors for the model parameters are relatively flat, which ensures that the posterior results are dominated by the data. We compute the posterior probability for each model and the posterior probability of inclusion for each explanatory variable for different prior views. Of particular interest, the estimation is performed in the familiar environment of linear, normal models and thus can be related to the existing literature in a straightforward manner.

Specifically, the Bayesian framework allows us to condition on the whole information set while conducting our inferences, as opposed to conditioning on a single *individual* model.[3] It compares all possible models simultaneously by the extent to which they describe the data as given by the posterior probability. This approach severely limits our data-snooping ability relative to using the information contained in the best model only, which often leads to disregarding much of the information and uncertainty.

Here, data snooping is limited to the choice of explanatory variables included in the initial information set. All the variables in this article have been identified as having predictive power by the previous literature. From this literature, we have selected the 14 variables, that, in our view, capture the most important claims. These variables are a priori given equal likelihood to be included.

Several important results emerge from our analysis. First, the data implies posterior probabilities that are generally more supportive of stock return predictability than the priors. The individual models selected by standard statistical model selection criteria (e.g., the adjusted R^2 and Akaike's information criterion) are in general large, and correspond best to the results for investors who are confident about predictability.

Second, the Bayesian methodology identifies six variables that clearly stand out. For these variables the posterior probability of inclusion is larger than the prior probability of inclusion for all choices of ρ (the prior probability of inclusion). Such stalwarts as past returns, dividend yields, and earnings/price ratios perform relatively poorly in this setting. Of particular

[3] The *individual* model refers to a single model as defined by the inclusion of a specific subset of $\kappa \leq K$ explanatory variables, whereas the *overall* model refers to the weighted average of the entire collection of all 2^K single models.

interest, if one is a priori almost certain about predictability, it is very difficult to distinguish the best predictor variables.

Finally, an out-of-sample forecasting analysis using rolling estimation windows suggests that the Bayesian model selection criteria (i.e., the weighted average best model) outperforms the standard classical methods. In particular, the "best" individual models selected by the statistical criteria generally have poorer or no better out-of-sample performance than the constant, unconditional model. This is in sharp contrast to the overwhelming evidence of predictability their in-sample results suggest. We argue that this is due to the stringent nature of classical tests that put 100% weight on inclusion or exclusion. The conditioning on one individual model fails to take the important model uncertainty into account, which severely underestimates the uncertainty associated with the quantity of interest. However, the in-sample and out-of-sample results for the Bayesian analysis are consistent. The overall, posterior probability-weighted average models generally perform slightly better than the constant model for all choices of priors, and show some, albeit small, evidence of predictability.

Some recent articles have started to investigate the influence of model uncertainty on financial models. MacKinlay and Pastor (2000), Pastor (2000), and Pastor and Stambaugh (2000) discuss prior mispricing uncertainty of asset pricing models and the influence on portfolio choice. Independent from our article, Avramov (2002) investigates the role of uncertainty about the return forecasting model in choosing optimal portfolios.

This article is organized as follows. Section 1 presents a Bayesian model selection approach to finding the best model for expected returns in the context of the existing literature on stock return predictability. Section 2 discusses the prior and posterior distributions. Section 3 describes the data used in the study. In Section 4, the main empirical in-sample and out-of-sample results are provided. Section 5 concludes.

1. Bayesian Model Selection

In this article we explore the predictability of S&P 500 index excess returns in a framework that models the time variation in expected returns conditional on specific factors. A broad literature has gathered substantial evidence of stock return predictability and has claimed that many different variables have forecasting power [see Fama (1991) and Hawawini and Keim (1995) for surveys of this literature].

For illustration purposes, in Table 1 we present an overview of which categories of variables are included in the analysis for 11 articles capturing most of the important characteristics of the overall literature, all published in the period 1986–1999. The initial information set in this article consists of variables from all these categories. Several observations regarding this table are in order. First, almost all these articles focus on a given individual model (i.e.,

Table 1
Variables and published articles

Author(s)	Year	I/P	Period	1	2	3	4	5	6	7	8	9	10	11	12	13
Chen, Roll, and Ross	1986	P	1953–1983	1	0	0	0	1	1	1	1	1	0	1	1	1
Campbell	1987	I	1959–1983	1	0	0	0	0	1	1	0	1	0	0	0	0
Harvey	1989	P	1941–1987	1	1	0	0	1	1	0	1	1	1	0	0	0
Ferson	1990	P	1947–1985	1	0	0	0	0	1	1	0	1	0	0	0	0
Ferson and Harvey	1991	P	1959–1986	1	1	0	0	1	1	0	1	1	0	0	0	1
Ferson and Harvey	1993	I	1970–1989	1	0	0	0	0	1	0	0	0	0	1	1	1
Whitelaw	1994	I	1953–1989	0	1	0	0	1	1	0	0	1	0	0	0	0
Pesaran and Timmermann	1995	I	1954–1992	0	1	1	0	0	1	1	0	1	0	1	1	0
Pontiff and Schall	1998	I	1926–1994	0	1	0	0	1	1	0	1	0	0	0	0	0
Ferson and Harvey	1999	P	1963–1994	0	1	0	0	1	1	0	1	1	0	0	0	0
Bossaerts and Hillion	1999	I	1956–1995	1	1	1	0	0	1	1	1	0	1	0	0	0
This paper	2002	I	1954–1998	1	1	1	1	1	1	1	1	1	1	1	1	1

This table reports which variables are included in the analysis for a selection of 11 articles on stock return predictability. For each article we report the authors, the year of publication (Year), which returns are analyzed (I/P/C, where I refer to a well-known index of stocks such as the Dow Jones Industrial Average or an S&P index, and P refers to stock portfolios based on size or industry), what period of returns is studied (Period), and finally which variables are included in the analysis. We distinguish the following categories of variables (the numbers in the table refer to the numbering below):

1. lagged returns
2. dividend yield
3. earnings yield
4. volume of shares traded over price level
5. credit spread between the yields of investment grade and below investment grade bonds
6. yield on a short-term Treasury bill
7. change in the yield on a short-term Treasury bill
8. term spread between the yields on long-term government bonds and the short-term Treasury bill
9. yield spread between the yield on an overnight fixed income security and the short-term Treasury bill
10. January dummy
11. growth rate of industrial production
12. inflation
13. change in inflation or measure of unexpected inflation

Most articles include one or two additional variables in their analysis (such as growth in personal consumption expenditures, oil prices, and a book-to-market ratio). However, in all articles, these additional variables were not reported to be important in the time period under consideration in this article (1954–1998). A "1" means that the relevant variable in included, a "0" indicates exclusion.

a particular number of predictive variables) and then link the predictability of returns to the variables included. Second, it is clear that there is no consensus on what the appropriate model is, as each article focuses on a particular set of explanatory variables. In other words, there is considerable uncertainty in the finance literature not only about the correct model for expected returns, but also about which variables to look at initially. Third, most articles in the table include some variables in their analysis beyond those selected for the initial information set used in this article, such as growth in personal consumption expenditures, oil prices, and book-to-market ratios. However, these additional variables are not reported to be important in the time period under consideration in this article.[4] Fourth, across the articles in the table the selection of

[4] For example, Pontiff and Schall (1998) report that the book-to-market ratio provides important information to predict the Dow Jones Industrial Average from 1926 to 1960, but the importance seems to disappear after 1960.

variables claimed to be important for predictability varies greatly.[5] Finally and essentially, while there is considerable model uncertainty, there is a general consensus in these articles claiming strong evidence for predictability, at least in-sample.

At the same time, other articles question the claims of predictability. The main arguments are data snooping biases and a lack of out-of-sample forecasting power. The data snooping problem is that in-sample predictability could be the result of researchers searching for patterns too studiously [Merton (1987), Ross (1989), Lo and MacKinlay (1990, 1997), Black (1993), and Richardson (1993)]. Therefore any claim of predictability should be backed up by out-of-sample performance statistics. Bossaerts and Hillion (1999) guard against data snooping by using statistical model selection criteria, and find no external validity: the "best" models have no out-of-sample forecasting power.

Foster, Smith, and Whaley (1997) provide a procedure to adjust standard tests for overfitting tendencies. Our approach is even more general: starting with K possible variables, we consider all 2^K different possible linear models. Facing the enormous uncertainty about the correct model, the straightforward solution is to evaluate all these models simultaneously. Our Bayesian framework takes the model uncertainty explicitly into account, by comparing all 2^K models simultaneously by the extent to which they describe the data as given by the posterior probability of the individual model.

In this Bayesian framework, the initial choice of the explanatory variables included in the information set still involves data snooping, but conditional on the selection of these variables, it is clearly minimized. In this study we include a total of 14 different explanatory variables (i.e., one variable in each of the 13 categories of Table 1, and one additional lagged return), leading to $2^{14} = 16{,}384$ different individual models. This preselection of 14 variables is only a fraction of the possible set of explanatory variables, but the number of explanatory variables used is large enough to incorporate the most important claims of predictability and assess their robustness and relative performance.

The classical approach conditions on one "individual" model that is singled out as "best" in some sense. This ignores the major uncertainty about which model is best and possibly leads to the severe underestimation of the uncertainty about any quantity of interest, a criticism that is well known. In his seminal book, Leamer (1978) proposes the standard solution of conditioning on the whole information set as employed in this article. The application of model uncertainty in the context of linear regression models is discussed by, for example, Mitchell and Beauchamp (1988), George and McCulloch (1993), Laud and Ibrahim (1995), Raftery, Madigan, and Hoeting (1997),

[5] For example, while a majority of the articles in the table report that the yield on a short-term Treasury bill adds valuable information at least in-sample, neither Pontiff and Schall (1998) nor Bossaerts and Hillion (1999) confirm this.

and Fernández, Ley, and Steel (2001). As Madigan and Raftery (1994) show, averaging over all possible models also provides optimal predictive ability.[6] Therefore the Bayesian weighted average model should provide superior predictions relative to any single model selected by some information criterion.

2. The Prior and Posterior Distributions

The standard model in the literature for predicting returns is linear in the explanatory variables. Moreover, we will also assume the errors to be normally distributed.[7] In the Bayesian framework employed here, each individual model $M(X_{\kappa+1})$ is a normal linear regression model of the form

$$Y_R = \beta_0 \cdot \iota_T + \sum_{i=1}^{\kappa} \beta_i \cdot X_i + \varepsilon = \beta' \cdot X_{\kappa+1} + \varepsilon, \tag{1}$$

where Y_R is the observed T-vector of equity index excess returns, β_0 is the unknown constant included in all individual models, β is the unknown $(\kappa+1)$-vector consisting of the constant and the coefficients for each of the κ included explanatory variables, $X_{\kappa+1}$ is the $(\kappa+1)\times T$ matrix of a T vector of ones and the κ explanatory variables, and ε is the T vector of identically, independently, and normally distributed disturbances with mean 0 and unknown variance σ^2.

The $2^{14} = 16{,}384$ individual models differ only in the selection of the 14 explanatory variables they contain. The "overall model" is the weighted average of these 2^{14} individual models. The coefficient for or weight on each individual model in the overall model is equal to the posterior probability of that individual model, which sum to one. The derivation of the posterior probability of each individual model is given below.

Two priors need to be specified: the prior of inclusion of each predictive variable in an individual model, and the prior of the distribution of the parameters β and σ^2 given a specific individual model. The choice of the priors should be made carefully, because they will generally affect the posterior weights in the overall model. Also, throughout the article we assume that all variables are transformed to have both a sample mean of zero and a sample variance of one. Furthermore, we use priors for which β and σ^2 can be integrated out analytically, which greatly increases computational speed and clarity of interpretation, but also contains the restriction that the prior of β is dependent of σ^2.

We assume that we have clear prior views on three important characterizations of predictability: the expected sum of explained squared regression

[6] The average model provides better predictive ability than any single model as measured by a logarithmic scoring rule. See Madigan and Raftery (1994).

[7] While this is a standard assumption, other papers employ techniques that do not require explicit distributional assumptions but instead use asymptotic results, e.g., by way of the general method of moments.

residuals, $E[R^2]$, the expected σ^2, and the probability that each variable in the information set is included in Equation (1) and thus contributes to any predictability. The a priori expected R^2 and the prior probability are related, for example, investors with a high prior probability of inclusion for all predictors will have a higher prior expected R^2 than investors with a low prior probability of inclusion.

The prior probability of inclusion of each variable leads directly to the a priori expected number of included predictors, the prior probability of any individual model and thus the prior probability of the i.i.d. model. Since the model selection process here is essentially a variable selection process, it is quite natural to construct the prior probability for each individual model via the prior probability of inclusion for each of the variables. Furthermore, each variable is assigned an equal and independent prior probability of inclusion ρ, because it is assumed that there is a priori no reason to believe that some variables are more likely to be included than others. In this case, the prior probability of an individual $Model(X_\kappa)$ consisting of κ $(0 \leq \kappa \leq 14)$ out of the 14 explanatory variables is

$$P(Model[X_\kappa]) = \rho^\kappa \cdot (1-\rho)^{14-\kappa}. \tag{2}$$

For example, the choice of $\rho = 0.5$ would assign equal prior probability to all models considered, and in case of $\rho = 0.25$, a model including $\kappa - 1$ explanatory variables is a priori $0.75/0.25 = 3$ times more likely than a model including κ variables. The prior probability of no predictability is $(1-\rho)^{14}$, the prior expected number of included variables is $\rho \cdot 14$.

However, the methodology employed easily allows different prior probabilities of inclusion for the various predictors, if an investor has different prior views on some (subset of) variables. For example, an investor might assign a high prior probability of inclusion to variables that she/he can economically motivate to drive changing investment opportunities or risk aversion and a low prior probability of inclusion to variables that are mainly data driven. Specifically, a high prior probability might be given to variables like the dividend yield, which, due to the presence of price in their denominator, will automatically predict returns if there is any time variation in expected returns. Finally, by looking at a wide range of different choices for ρ we are able to examine the views of investors with wide ranging general a priori views on predictability. For example, very optimistic investors will have a high ρ, which results in a prior probability of no predictability close to zero and in a prior expectation of a large model in Equation (1).

We will first give the general form of priors and posterior distributions, and then address separately two distinct prior views on predictability for comparison purposes: (i) a skeptic view with $E[R^2] = 1\%$, $E[\sigma^2] = 0.99$ combined with values of $\rho = 0.05, 0.10, 0.20$, and 0.25, and (ii) a confident view with $E[R^2] = 12\%$, $E[\sigma^2] = 0.92$, combined with values of $\rho = 0.25, 0.50, 0.75$,

and 0.90. The skeptic a priori views practically all predictability as spurious and the confident investor is a strong believer in predictability. The latter prior view of R^2 is taken as representative of the results of the literature advocating the existence of predictability, as listed in Table 1. The value of 12% is a reasonable average of the wide range of results in the articles listed there.

2.1 Priors and posterior distributions

We assume that we have no clear prior view or information on the coefficients in β conditional on a specific model. Even when a prior view as described above or general prior information in the form of expert opinion or past experience is available, we assume that this is very difficult to translate into prior beliefs about β. Particularly, the inclusion of most variables in our initial information set is mainly data driven. Moreover, in the absence of a theory that prescribes some prior view, such informative prior could strictly speaking only use information or expert opinion available to the researcher at the beginning of the data sample period.

We will choose the prior distributions of β that reflect this situation of very weak prior information, when there is little reason to give certain (areas of) parameter values higher prior density than other values. The choice of a prior is greatly complicated by the problem that we cannot choose an improper prior for β, because these will render the posterior model probabilities arbitrary.[8] Therefore the proper prior for β conditional on σ^2 is normally distributed with mean zero and very large prior variances relative to the sample variance of the predictors, such that the prior distribution is rather flat over the range of relevant parameter values. In this case, the posterior results conditional on a model are relatively insensitive to changes in the prior distribution. Edwards, Lindman, and Savage (1963) call this the "stable estimation" case.

Particularly, we choose a natural conjugate g-prior specification for β [see Zellner (1986)], a popular choice in Bayesian statistics that adopts a full correlation structure between the included predictors [see also Poirier (1985), Laud and Ibrahim (1995, 1996), and Fernández, Ley and Steel (2001)]. Furthermore, for σ^2 we choose the conjugate inverse-gamma distribution. For a model including κ variables, the prior of β conditional on σ^2 is

$$P(\beta \mid \sigma^2) = N_{\kappa+1}\left(0_{\kappa+1}, \sigma^2 \cdot \varphi \cdot (X_{\kappa+1}{}'X_{\kappa+1})^{-1}\right), \tag{3}$$

$$P(\sigma^2) = IG(s_0^2, T), \tag{4}$$

where $N_{\kappa+1}(\cdot)$ denotes the normal distribution $0_{\kappa+1}$ a $(\kappa+1)$-vector or zeros, φ a scalar that remains to be chosen and IG denotes the inverse gamma

[8] This is because improper priors are only defined up to an arbitrary constant, which depend on the specific model. Therefore each posterior model probability involves a unique unspecified constant [Kass and Raftery (1995)].

distribution, such that the prior expectation and variance of σ^2 are given by

$$E(\sigma^2) = \tfrac{T}{T-2} \cdot s_0^2, \tag{5}$$

$$V(\sigma^2) = \tfrac{T^2}{(T-2)(T-4)} \cdot s_0^2. \tag{6}$$

Furthermore, note that we assume that we do not have particular prior information about the linear coefficients conditional on their inclusion, for example, taken from economics. As argued above, our purpose here is to investigate the evidence in the data, relying on as little prior information as possible. This is motivated by the complete lack of consensus in the literature about what the important predictors are. The minimal information in the prior distribution of β is reflected by its centering around zero, the use of the same φ for all predictors, and by the g-prior, which gives equal scale to all regressors and greatly improves the analytical clarity.

That said, our methodology could incorporate any specific prior view that would consist of choosing different values of β_0 and φ for different predictors. However, it would need careful implementation, as conditional on inclusion, the prior view on β will generally depend on the colinearity with the other variables and thus will be model specific.

The resulting posterior distributions of β and σ^2 conditional on a model are

$$P(\beta \mid Y_R, \sigma^2) = N_{\kappa+1}\left(\tfrac{\varphi}{1+\varphi}\beta_{OLS}, \sigma^2 \cdot \tfrac{\varphi}{1+\varphi} \cdot (X_{\kappa+1}{}'X_{\kappa+1})^{-1}\right), \tag{7}$$

$$P(\sigma^2 \mid Y_R) = IG(s_1^2, 2 \cdot T), \tag{8}$$

where β_{OLS} is the ordinary least squares (OLS) estimator of β, and the scalar s_1^2 is

$$s_1^2 = \frac{1}{2T} \cdot \left[T \cdot s_0^2 + (Y_R - \beta_{OLS}{}'X_{\kappa+1})'(Y_R - \beta_{OLS}{}'X_{\kappa+1}) + \frac{1}{1+\varphi} \cdot B_{OLS}{}'X_{\kappa+1}{}'X_{\kappa+1}\beta_{OLS}\right]. \tag{9}$$

Therefore, s_1^2 can be interpreted as the average of the prior fit s_0^2 and the average of squared OLS residuals plus the "error of the prior prediction guess," or a measure of how far the prior expected β is away from the OLS estimate [see Equation (3)]. However, for the relevant values of φ, this last component is very small in our analysis.

Finally, the posterior probability of a model that includes the κ predictors in X_κ is obtained by combining the full prior distributions of α, β, and σ^2

with the likelihood function, and subsequently integrating out these parameters [see Poirier (1995: chaps. 8 and 9)],

$$P(Model[X_\kappa] \mid R) = c \cdot P(Model[X_\kappa]) \cdot \left(\frac{1}{1+\varphi}\right)^{\kappa/2} \cdot s_1^{-T}. \qquad (10)$$

The posterior model probability in Equation (10) consists of three components. The first component is the prior model probability and is determined by ρ [see Equation (2)]. The second component, $(1+\varphi)^{-\kappa/2}$, is the square root of the ratio of the determinant of the posterior covariance matrix of β over the determinant of the prior covariance matrix of β. The third component is determined by s_1^2.

The choice of φ serves as a "penalty for model size" in Equation (10).[9] Intuitively, choosing a larger φ increases the prior and hence the posterior weight on larger values of β. This increases the expected amount in the posterior that each variable attributes to predictability. Each predictor will have more explanatory power by itself. Therefore if we choose a larger φ, then to arrive at the same amount of predictability we would need fewer predictors. Hence larger models are penalized.

The prior of inclusion ρ of each variable together with the prior densities for the parameters β and σ^2 result in a prior for the overall model as follows. In the overall model, the prior of the coefficient of each predictor has a mass point at zero equal to $1-\rho$ and is relatively flat everywhere else (as determined by φ). This discontinuity in the priors for the overall model is the direct result of the notion of model uncertainty. Any prior view that allows for uncertainty about the inclusion of any regressor gives a mass point at zero in the prior distribution of the relevant parameter. Not one single model can be claimed to be "true," so we assign relative likelihood statements to different models simultaneously, as is common in the Bayesian literature [see Kass and Raftery (1995) for a survey and discussion on Bayesian model comparison].

The prior for the overall model with its mass point at zero also illuminates the respective effects of ρ and φ on the model choice. Both decreasing ρ and increasing φ results in smaller models being favored more. The first decreases the prior probability of inclusion for each predictor. This increases the mass point at zero in the prior of the overall model and increases the prior model probability of smaller versus larger models. The second makes the continuous part of the distribution more flat and gives more prior and thus posterior weight to large values of β. As described above, this increases the explanatory power of each predictor conditional on inclusion and creates the "model penalty" effect.

[9] Technically, increasing φ will decrease the second and increase the third component in Equation (10). For our dataset, the first effect clearly dominates the second. Thus increasing φ lowers the posterior model probability for larger models relative to smaller models.

Specifically, the choice of ρ alone determines the a priori expected average model size and all the prior model probabilities. It reflects the prior view on the number of predictors, but not on how important each will be economically. The choice of φ determines the extent to which each variable is expected to contribute to the predictability. Therefore, for a fixed amount of prior expected predictability, the choice of a larger ρ will lead to a smaller φ. Because of the larger ρ, we expect that more predictors will attribute to the fixed amount of predictability. As a result, each predictor will attribute less, thus less prior weight is given to large values of β or we choose a smaller φ.

Because increasing ρ and increasing φ both lead to a larger prior expected R^2, an investor with a small ρ and high φ could expect the same amount of predictability as an investor with a large ρ and a small φ. The first expects a few predictors to each have a big effect, the second expects many different predictors to each play a minor role.

The posterior for the overall model is constructed as the weighted average of all 2^{14} individual models, with the posterior probabilities of the individual models as weights. Furthermore, the posterior probability of inclusion for each of the 14 explanatory variables is calculated as the total sum of the posterior probabilities of all individual models in which the particular variable is included.

The Bayesian model averaging procedure advocated here could be interpreted as replacing the following alternative procedure, which is arguably intuitively more straightforward. Herein only the largest model would be considered and the discontinuous prior of the overall model would be used. The discontinuity of the prior of β in the overall model is due to the mass points at zero or the prior uncertainty of inclusion for the predictors. Therefore the latter procedure would create a large computational and analytical burden, which is prevented by regarding all individual models and their continuous priors separately and then averaging them.

In our case, considering all possible individual models is still feasible for two important reasons. First, the choice of a prior for β that is conditional on σ^2 results in analytically tractable posteriors. In our view, this benefit surpasses its arguable restrictiveness. Second, we limit the number of variables under consideration to 14. For larger numbers of variables, methods such as the "Occam's window" algorithm and the Markov chain Monte Carlo model composition approach are available.[10] Raftery, Madigan, and Hoeting (1997) describe the application of both methods for the case of linear regression models.

[10] The "Occam's window" algorithm involves the averaging over a reduced set of models [see Madigan and Raftery (1994)]. The Markov chain Monte Carlo model composition approach directly approximates the complete solution [see Madigan and York (1995)].

2.2 The calibration of the prior for β

As becomes clear in Equation (10), the choice of φ can potentially have a large influence on the posterior results. An important effect of φ is that it gives rise to a penalty for model size: the larger φ, the smaller the posterior probability of the model.

The choice for the prior of β in the context of weak prior information remains a contentious issue in the literature on model averaging. For recent discussions, see George and McCulloch (1993, 1997), Geweke (1996), Kass and Raftery (1997), Raftery, Madigan, and Hoeting (1997), Clyde (1999), George (1999), George and Foster (2000), and Fernández, Ley, and Steel (2001). The general problem is that if prior information about β is missing, the posterior results can be very sensitive to the specification of the prior of β, as also noted by Pastor and Stambaugh (2000). In this article we propose to solve this issue by assuming that the investor has prior information about the regression's R^2, the variance of the residuals, and the probability of inclusion for all the variables. With these three pieces of prior information combined, we calibrate a value of φ that is consistent with these views using

$$E[R^2] = 1 - E\left[\frac{\sigma^2}{\sigma^2 + \beta' \cdot V_X \cdot \beta}\right], \tag{11}$$

where V_X denotes the variance of the X-matrix, which we approximate by its sample estimate [as proposed by Richard and Steel (1988: appendix D)].

For the calibration we take a pair of ρ and s_0^2 and generate a large sample from the prior distribution of R^2 by subsequently sampling from the prior distributions of X (with the binomial(14, ρ) distribution), σ^2, and β and using Equation (11). We repeat this for many different choices of φ until we find the appropriate prior expected R^2. For the confident investor, we choose $E[R^2] = 12\%$, $E[\sigma^2] = 0.92$ and four different choices $\rho = 0.25, 0.50, 0.75,$ and 0.95, giving a required φ of 16, 7, 3.5, and 2.5, respectively. The prior of β of the skeptic investor with $E[R^2] = 1\%$ and $E[\sigma^2] = 0.99$ requires a value of φ equal to 3.25, 2.25, 1.25, and 1.00 for $\rho = 0.05, 0.10, 0.20,$ and 0.25, respectively.[11]

As discussed above, for a fixed $E[R^2]$ the view that more variables contribute to predictability (implemented by a larger ρ) leads to a lower calibrated φ. Intuitively, if more variables are expected to contribute to some fixed level of $E[R^2]$, each variable will on average contribute less. The smaller a coefficient in β, the smaller that variable's contribution is to the R^2. Therefore, in this case, less prior weight is assigned to higher values of β, thus φ is smaller. Alternatively, for a fixed choice of ρ, the expectation that the model will be able to explain a larger proportion of the return variation (or a higher $E[R^2]$) gives a higher calibrated φ.

[11] Both these simulations and the posterior results are robust to changes in $E[\sigma^2]$ by several percentage points.

3. The Data

The endogenous variable is the S&P 500 index monthly excess return, from January 1954 to December 1998, for a total of 540 monthly observations. The explanatory variables in this study can be divided into the following categories: technical, price level, liquidity, interest rate, and macroeconomic variables.

Given the instant availability of all explanatory variables except the three macroeconomic variables, the three macroeconomic variables are lagged once relative to the other explanatory variables to ensure their actual inclusion in the information set. The return series comes from the CRSP database and the predictors from the Basic Economics database. The information set consists of the following 14 explanatory variables:

- The technical variables included are the S&P 500 index excess return lagged once and twice, and a January dummy.
- The price level variables included are the S&P 500 index dividend yield and the S&P 500 earnings yield.
- The liquidity variable included is the NYSE volume divided by the NYSE price level, as a measure of general liquidity of the market (and not especially of the S&P 500 index, because of a lack of data availability).
- The interest rate variables included are the difference between yields on BAA and AAA Moody's rated corporate bonds (the "credit spread"), the yield on a 3-month maturity Treasury bill and its first difference, the difference between the yield on a 10-year maturity Treasury bond and a 3-month maturity Treasury bill (the "term spread") and the difference between the Federal Funds rate and the yield on a 3-month maturity Treasury bill (the "yield spread").
- The macroeconomic variables included are the monthly, seasonally adjusted year-by-year rate of change of inflation as measured by the producer price index for finished goods and its first difference, and the monthly year-by-year rate of change in industrial production.

Table 2 provides some descriptive statistics for all the variables without any transformation, reporting the mean, standard deviation, first- and second-order autocorrelation, and the correlation with the endogenous variable. Most variables show high persistence, both in first- and second-order autocorrelation. For example, the earnings yield, dividend yield, and credit spread have a first-order autocorrelation of 99.04%, 98.91%, and 97.35%, respectively. The only variables with small first- and second-order autocorrelation are the lagged returns and the variables in first differences. The correlation in absolute terms with the endogenous variable runs from a high of -15.20% for the yield spread to a low of 1.22% for the earnings yield. Other variables with a high correlation with the endogenous variable are inflation, -13.76%; the yield on a 3-month Treasury bill and its first difference, -11.60% and -12.19%, respectively; and industrial production, -11.85%.

Table 2
Descriptive statistics of the data

	Mean	StDev	$\gamma[1]$	$\gamma[2]$	Cor
ExRet	0.0036	0.0417	2.60%	−1.82%	100.00%
ExRet-1	0.0035	0.0417	2.33%	−2.22%	2.60%
ExRet-2	0.0035	0.0417	1.91%	−2.40%	−1.69%
Jan	0.0818	0.2743	−8.92%	−8.94%	4.95%
Div	3.6459	0.9664	98.91%	97.26%	2.00%
Earnings	0.0715	0.0246	99.04%	97.62%	1.22%
Vol/P	11.5694	8.7877	96.43%	96.61%	6.53%
Credit	0.9547	0.4393	97.35%	93.70%	4.94%
TBill	5.5599	2.8305	98.42%	95.97%	−11.60%
ch-TBill	0.0060	0.5022	27.50%	−10.66%	−12.19%
Term	1.3051	1.1729	94.34%	86.91%	9.93%
Yield	0.4961	0.7993	86.75%	76.52%	−15.20%
Inflation	0.0325	0.0369	98.68%	97.01%	−13.76%
ch-Inflation	0.0000	0.0060	13.16%	12.17%	−9.89%
IndusProd	0.0327	0.0541	96.34%	89.70%	−11.85%

Descriptive statistics of the data for the period 1/1954–1/1998. The following explanatory variables are used: the S&P 500 index excess return (*ExRet*), the S&P 500 index excess returns lagged once and twice (*ExRet-1* and *ExRet-2*), a January dummy (*Jan*), the S&P 500 index dividend yield (*Div*), the S&P 500 earnings yield (*Earnings*), the NYSE volume divided by the NYSE price level (*Vol/P*), the difference between yields on BAA and AAA Moody's rated corporate bonds—the "credit spread"—(*Credit*), the yield on a 3-month maturity Treasury bill in the previous month (*TBill*, also used to compute the excess returns) and its first difference (*ch-TBill*), the difference between the yield on a 10-year maturity Treasury bond and a 3-month maturity Treasury bill—the "term spread"—(*Term*), the difference between the Federal Funds rate and the yield on a 3-month maturity Treasury bill—the "yield spread"—(*Yield*), the monthly, seasonally adjusted year-by-year rate of change of inflation as measured by the producer price index for finished goods (*Inflation*) and its first difference (*ch-Inflation*), and the monthly year-by-year rate of change in industrial production (*IndusProd*). The included statistics are the mean, standard deviation (StDev), first-order and second-order autocorrelation ($\gamma[1]$ and $\gamma[2]$), and the correlation with the endogenous variable (Cor).

4. The Results

In this section we first discuss whether the data show evidence for predictability in-sample, which variables get most support, and how these results compare to the "best" models as selected by various statistical model selection criteria. Then we compare the out-of-sample performance of the overall model and the individual models with the highest posterior probability of the "best" models according to the model selection criteria, and finally of the constant, unconditional model.

4.1 In-sample results

In Table 3 we report the prior and posterior probabilities of no predictability for various ρ (the prior probability of inclusion for each of the 14 predictive variables) and the ratios of posterior to prior beliefs. The prior probability of no predictability is equal to the prior probability of the constant, unconditional model, or $(1-\rho)^{14}$, while the posterior probability of no predictability is equal to the posterior probability of the constant model.

The posterior:prior odds can be interpreted as the change in belief after inference relative to the a priori beliefs. The table shows that the data decrease the probability of no predictability for all cases. For example, if a priori 48.77% or 22.88% weight (for choices of $\rho = 0.05$ and 0.10, respectively) is given to the unconditional model, then the posterior weight of that model becomes 2.27% or 0.39%, respectively. Any skeptic on predictability

Table 3
Priors and posteriors of no predictability, and posterior:prior odds

	ρ	Prior of no predictability(%)	Posterior of no predictability(%)	Posterior:prior odds of no predictability	Posterior:prior odds of predictability
$E[R^2] = 1\%$	0.05	48.77	2.27	4.65×10^{-2}	1.91
	0.10	22.88	0.385	1.68×10^{-2}	1.29
	0.20	4.40	4.17×10^{-2}	9.49×10^{-3}	1.05
	0.25	1.78	1.62×10^{-2}	9.10×10^{-3}	1.02
$E[R^2] = 12\%$	0.25	1.78	7.05×10^{-4}	3.96×10^{-4}	1.02
	0.50	0.0061	3.94×10^{-8}	6.46×10^{-6}	1.00
	0.75	3.73×10^{-7}	3.03×10^{-13}	8.13×10^{-7}	1.00
	0.90	1.00×10^{-12}	4.93×10^{-19}	4.93×10^{-7}	1.00

Prior versus posterior probabilities of no predictability for different choices of ρ (the prior probability of inclusion of each explanatory variable), and the ratios of posterior:prior probabilities of no predictability and of predictability (called the "posterior:prior odds"). The prior probability of no predictability is equal to $(1-\rho)^{14}$. All probabilities are in percentage terms. The posterior probability of no predictability is equal to the posterior probability of the constant, unconditional model.

would therefore see his skepticism greatly decreased by the data. However, an investor who strongly believes in all predictors, thus with $\rho = 0.75$ or 0.95, would find that the data confirm his views.

Table 4 compares the a priori expected R^2 to the posterior model probability weighted average of the R^2 and the adjusted R^2 and the prior expectations of σ^2 to the posterior expectations. Here we find that the skeptic investor finds his expectations of an R^2 of 1% increased five- to ninefold (depending on ρ), while they are decreased for the confident investor with $E[R^2] = 12\%$ by 0.3–2.5% (again depending on ρ). Furthermore, the expected σ^2 is generally decreased, and the larger ρ, the smaller the posterior expected σ^2 (for fixed $E[R^2]$).

Next, in Table 5 we report the a priori expected number of explanatory variables (equal to $\rho \cdot 14$) and the posterior probability weighted average number of explanatory variables included in the models for different ρ. We compare these to the number of explanatory variables in the "best" models according to various statistical model selection criteria. The numbers of

Table 4
Posterior versus prior expected R^2

	ρ	$E[R^2]$	$E[R^2 \mid \text{data}]$	$E[R^2_{\text{adj}} \mid \text{data}]$	$E[\sigma^2]$	$E[\sigma^2 \mid R]$
$E[R^2] = 1\%$	0.05	1.00%	5.79%	5.34%	0.98	0.979
	0.10	1.00%	6.90%	6.27%	0.98	0.977
	0.20	1.00%	8.16%	7.27%	0.98	0.977
	0.25	1.00%	8.60%	4.79%	0.98	0.978
$E[R^2] = 12\%$	0.25	12.00%	9.53%	8.48%	0.92	0.922
	0.50	12.00%	11.0%	9.41%	0.92	0.915
	0.75	12.00%	11.5%	9.54%	0.92	0.917
	0.90	12.00%	11.7%	9.44%	0.92	0.918

Reported are the a priori expected R^2, $E[R^2]$, the posterior model probability weighted average of the R^2, $E[R^2 \mid \text{data}]$, of the adjusted R^2, $E[R^2_{\text{adj}} \mid \text{data}]$, the a priori expected σ^2, $E[\sigma^2]$, and the posterior expected σ^2, $E[\sigma^2 \mid \text{data}]$, for the various choices of ρ.

Table 5
Number of explanatory variables included

	ρ	Prior exp. # of var.	Posterior average # of var.	Criterion	# of var.
$E[R^2] = 1\%$	0.05	0.70	2.57	R2adj	11
	0.10	1.40	3.59	AIC	10
	0.20	2.80	5.18	BIC	3
	0.25	3.50	6.84	FIC	5
				PIC	5
$E[R^2] = 12\%$	0.25	3.50	6.17		
	0.50	7.00	9.03		
	0.75	10.50	11.84		
	0.90	12.60	13.17		

Reported are the a priori expected number of variables included (Prior exp. # of var.) and the posterior average number of variables (Posterior average # of var.). Reported is the number of variables chosen from the initial 14 preselected variables, thus excluding the constant, which is included in all individual models. Second, for all five statistical model selection criteria considered in this article, the number of variables in the "best" model according to each criterion is reported in the last two columns. The five criteria are the adjusted R^2 (R2adj), Akaike's information criterion (AIC), Schwarz's criterion or the Bayesian information criterion (BIC), Fisher information criterion (FIC), and the posterior information criterion (PIC).

variables included are those from our list of 14 preselected variables, thus excluding the constant that is included in all individual models.

We consider the following five statistical model selection criteria: adjusted R^2, Akaike's information criterion [Akaike (1974)], Schwarz's criterion or the Bayesian information criterion [Schwarz (1978)], the Fisher information criterion [Wei (1992)], and the posterior information criterion [Phillips and Ploberger (1996)]. These criteria have been developed in order to select the "best" model in a set of models. All five criteria try to guard against overfitting tendencies by adjusting a general measure of fit, namely the (log of the) sum of squared errors, to reflect these tendencies. All five criteria include some penalty for larger models and thus a priori favor smaller models. The adjustment made by each criterion can be seen as an alternative to directly incorporating model uncertainty.

From Table 5 we find that for all cases the posterior average number of variables is larger than the a priori expected number of variables. For example, if one a priori expects 1.40 or a tenth of the variables to be included, the posterior belief changes this to 3.59 variables, while prior expected beliefs of a model including 7 or half the variables are transformed into posterior beliefs of 9.03 included variables.

In contrast, the statistical model selection criteria tend to select relatively large models despite the penalties associated with including more variables, with the exception of Schwartz's criterion. Such large models would be chosen within the Bayesian framework only as a result of the confident investor. The difference between the methodologies is the result of the stringent nature of the classical model selection criteria, which put full weight on either the inclusion or the exclusion of a variable and ignore model uncertainty. Therefore the penalties for larger models still adjust for overfitting, but there is no straightforward way to adjust the penalty functions in the criteria for the

amount of model uncertainty faced in the particular application. In other words, the Bayesian results show which priors are consistent with which information criterion for this specific application. They suggest that for skeptic investors, the study of predictability of stock returns is characterized by too much uncertainty to be accounted for in the two most frequently used criteria, the adjusted R^2 and Akaike's information criterion.

In Table 6 we report the posterior probability of inclusion for all 14 variables individually for the various ρ. The posterior probability of inclusion for each variable is computed as the total sum of the posterior probabilities of all individual models including that particular variable (i.e., 2^{13} different individual models). The change between the prior and posterior probability of inclusion can be interpreted as support of the data for each variable.

Table 6 shows that for the skeptic investor with $E[R^2] = 1\%$, seven variables stand out: the liquidity variable, the credit spread, the yield on a 3-month Treasury bill and its first difference, the yield spread, the first difference in the inflation variable, and the January dummy. For these variables, the posterior probability of inclusion is larger than the prior probability of inclusion for all choices of ρ. The other eight variables are less likely to be included than thought a priori for at least half the choices of ρ. Particularly, note the relatively poor performance of such stalwarts as past returns and especially dividend yield. For all choices of ρ, the data reduce the probability of inclusion for this variable.

For the confident investor with $E[R^2] = 12\%$, again seven variables stand out. Relative to the best-performing variables for the skeptic investor, the first different in the inflation variables has been exchanged for the term spread.

Table 6
The posterior probabilities of inclusion for all 14 explanatory variables

	$E[R^2] = 1\%$				$E[R^2] = 12\%$			
Variable	5%	10%	20%	25%	25%	50%	75%	90%
ExRet-1	2.84	7.06	18.14	24.10	18.27	60.65	87.07	95.85
ExRet-2	4.27	10.16	23.25	29.51	27.66	68.38	89.27	96.46
Div	3.99	8.75	19.09	24.32	11.88	35.14	68.26	87.85
Earnings	8.14	15.61	27.86	33.08	25.45	41.02	67.99	87.38
Vol/P	18.94	32.89	47.76	52.36	66.46	82.37	92.01	96.99
Credit	25.60	43.99	61.79	66.44	88.39	98.68	99.36	99.68
TBill	25.48	32.63	44.35	49.01	69.45	94.01	97.48	98.92
ch.TBill	32.51	46.86	62.82	67.57	85.49	98.73	99.68	99.89
Yield	33.17	37.84	42.58	45.29	36.75	60.93	83.40	93.77
Term	4.98	12.91	30.70	38.02	62.00	94.51	98.05	99.17
Jan	60.21	58.41	59.72	61.45	71.19	91.95	97.00	98.82
IndusProd	8.66	14.80	27.14	32.78	22.00	49.56	79.14	92.94
Inflation	3.82	8.04	17.42	22.29	10.65	30.82	61.90	84.02
ch.Inflation	24.12	28.84	35.03	38.21	20.87	31.66	63.67	85.71

Posterior probabilities of inclusion for all 14 explanatory variables, for different choices of ρ (the prior probability of inclusion for each variable). These posterior probabilities are calculated as the total sum of the posterior probabilities of all 2^{13} individual models in which the particular variable is included. For the description of the names of the variables, see Table 2. All posterior probabilities are in percentage terms.

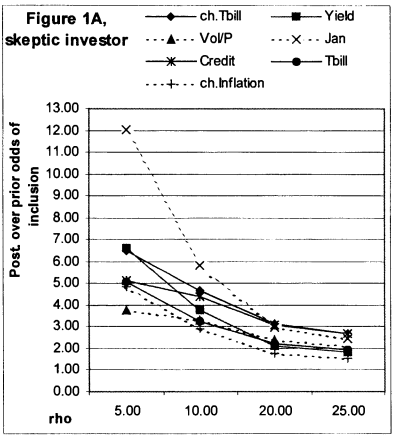
Again, for these variables, the posterior probability of inclusion is larger than the prior probability of inclusion for all choices of ρ. The dividend yield and the inflation variables are again the worst performers.

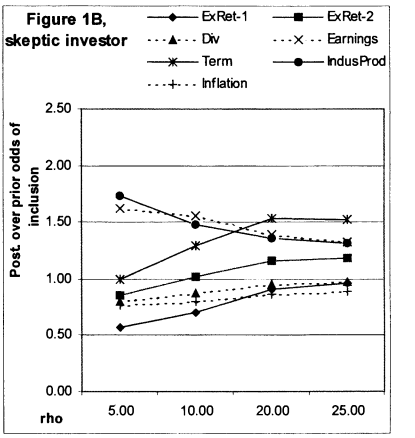
One issue of potential concern is that the performance of, for example, the dividend and earnings yield, may be reduced because of a multicollinearity problem created by the inclusion of many other variables. Especially if an investor assigns higher prior probability to these two variables than to others, this might overly reduce the evidence in favor of these. At the same time, in order to control for data snooping, we want to throw in more variables.

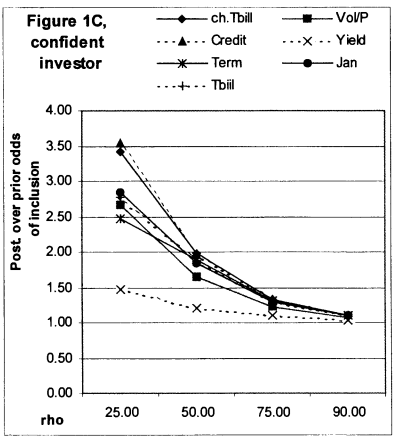
The possible multicollinearity can be investigated by reviewing the correlations between the dividend and earnings yield and the other predictive variables. The highest correlation of the dividend yield with any other variable is 6.1% with the twice-lagged excess return. By contrast, the earnings yield has a correlation greater than half with three variables: a correlation of 92.7% with the liquidity variable, 56.5% with the yield on a 3-month Treasury bill, and 55.4% with industrial production. We conclude that the low performance of the dividend yield does not appear to be caused by any multicollinearity, though this possibly could be the case for the earnings yield, especially because of the good performance of the two variables it is correlated with most. This is also confirmed by the classical model selection criteria, none of which include the dividend yield (see below). For the other variables that perform poorly, the correlations with the other predictors are generally lower than 20%, indicating again that decreasing the initial information set would not appear to increase their posterior probability of inclusion.

Another striking feature of the table is that the higher ρ, the less diversion among the predictors. To further investigate this convergence, we report in Figure 1 the ratio of the posterior probability over the prior probability of inclusion ("posterior over prior odds of inclusion") for all variables. Figure 1A gives the ratios for the seven highest-ranking variables in case of $\rho = 0.05$, and Figure 1B gives the ratios for the seven lowest-ranking variables in case of $\rho = 0.05$, both for the case of the skeptic investor. Figure 1C and 1D gives the ratios for the seven highest and lowest ranking variables, respectively, for $\rho = 0.25$ and the confident investor.

For the skeptic investor and $\rho = 0.05$, the posterior:prior odds of inclusion range from 12.04 for the January dummy to 0.54 for the once-lagged excess return. For $\rho = 0.25$, the posterior:prior odds of inclusion range from 3.54 for the credit spread to 0.42 for the inflation variable. Furthermore, the odds for the seven variables in Figure 1A tend downward toward 1 for higher ρ, while the odds for the seven variables in Figure 1B tend upward, again toward 1. We conclude that not only, and as we would expect, is harder to differentiate between the various variables for higher ρ, but also for $\rho = 0.25$, all variables appear useful. The results for the confident investors are similar,







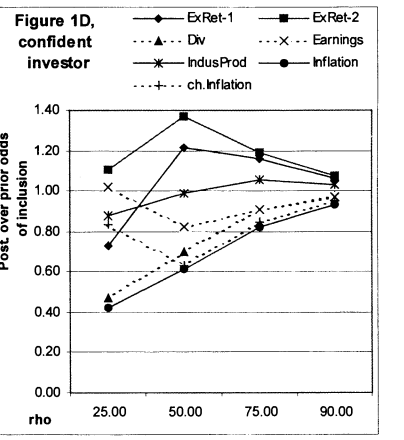


Figure 1
The ratio of the posterior probability over the prior probability of inclusion ("posterior over prior odds of inclusion") of the 14 explanatory variables, for different ρ. (A) The ratios for the seven highest-ranking variables according to Table 6 in case of $\rho = 0.05$ for the skeptic investor. (B) The ratios for the seven lowest-ranking variables according to Table 6 in case of $\rho = 0.05$ for the skeptic investor. (C) The ratios for the seven highest-ranking variables according to Table 6 in case of $\rho = 0.25$ for the confident investor. (D) The ratios for the seven lowest-ranking variables according to Table 6 in case of $\rho = 0.25$ for the confident investor.

and given in Figure 1C and 1D, in which the odds again tend downward and upward, respectively.

It is not clear what causes the relative performance of the various predictors to change in Figure 1. Possible reasons might include nonnormality, nonlinearity, and the fact that there is more multicollinearity if more variables are included for higher ρ. Furthermore, a skeptic could argue that the rankings for the cases with high ρ are the result of overfitting. The drop of

the posterior:prior odds of inclusion for the variables that do best, combined with the increase for the variables that do worst from $\rho = 0.05$ to $\rho = 0.25$ in Figure 1B and from $\rho = 0.25$ to $\rho = 0.90$ in Figure 1D, is the most striking support for this argument, possibly because of the nonstationarity of some of the highly persistent explanatory variables. The decrease in the ranking of the January dummy and the first difference of the inflation variable (both with low persistence) and the increase of the term spread variable (highly persistent) offer further evidence for this explanation (if going from low to high ρ, as discussed before).

In Table 7 we report which variables are included in the individual models with the highest posterior probability and the "best" models according to the five statistical model selection criteria. We can compare the Bayesian results of Table 6 to the "best" individual models as selected by the statistical model selection criteria in Table 7. This shows that the variables picked by the adjusted R^2 and Akaike's information criteria are mostly consistent with those in the models with highest posterior probability for $\rho = 0.75$ and $\rho = 0.90$.

The classical statistical model selection criteria suggest overwhelmingly the existence of in-sample predictability. The "best" individual model according to the adjusted R^2 and Akaike's information criteria differ in only one variable, while the Fisher and posterior information criteria choose the same individual model. However, the models with highest posterior probability for $\rho = 0.05$ and $\rho = 0.10$ indicate that for those priors, most variables are not important for predictability. Only for a choice of $\rho \geq 0.50$ are the majority of variables included in the model with the highest posterior probability.

Table 7
Inclusion of the 14 predictive variables in the "best" individual models

	$E[R^2] = 1\%$				$E[R^2] = 12\%$								
Variable	0.05	0.10	0.20	0.25	0.25	0.50	0.75	0.90	R2adj	AIC	BIC	FIC	PIC
ExRet-1	0	0	0	0	0	1	1	1	1	1	0	1	1
ExRet-2	0	0	0	0	0	1	1	1	1	1	0	1	1
Div	0	0	0	0	0	0	1	1	0	0	0	0	0
Earnings	0	0	0	0	0	0	1	1	1	1	1	1	1
Vol/P	0	0	1	1	1	1	1	1	1	1	1	0	0
Credit	0	0	1	1	1	1	1	1	1	1	0	0	0
TBill	0	0	0	1	1	1	1	1	1	1	1	0	0
ch.TBill	0	0	1	1	1	1	1	1	1	1	0	0	0
Yield	1	1	0	0	0	1	1	1	1	1	0	0	0
Term	0	0	0	1	1	1	1	1	1	1	0	0	0
Jan	1	1	0	1	1	1	1	1	1	0	0	0	0
Inflation	0	0	0	0	0	0	1	1	0	0	0	0	0
ch.Inflation	0	0	0	0	0	0	1	1	1	1	0	1	1
IndusProd	0	0	0	0	0	0	1	1	0	0	0	1	1

First, the inclusion is reported for the individual model with highest posterior probability. Second, the individual models as selected by the five statistical model selection criteria are reported. For the description of the names of the variables, see Table 2; for a reference of the abbreviated names of the criteria, see Table 4. A "1" means that the relevant variable is included, a "0" indicates exclusion. For the individual models with highest posterior probability for different ρ, we report the posterior probability. For each individual model, the adjusted R^2 is also reported (in percentage terms, and using the posterior weighted average overall model size to adjust the R^2 for the overall models).

These results will lead to different conclusions for investors with different a priori views on predictability. First, we consider the case for a skeptical investor with $\rho = 0.05$–0.25, who a priori rejects most claims of predictability in the literature using some data snooping argument. Such an investor has found her belief in no predictability considerably strengthened by the data, while the data snooping is minimized given her information set. Such an investor will view about 7 of the 14 variables as adding substantial predictive information.

Second, consider the case for a highly optimistic investor with $\rho = 0.25$–0.90. Such an investor has found his prior belief in predictability confirmed. Such an investor will find it harder to pick variables that are important for predictability, especially for the highest choices of ρ. Still, about the same variables appear useful as in the case of the skeptic investor. For $\rho = 0.90$, only four variables have a posterior probability of inclusion below the prior probability, namely the dividend and earnings yields, and the inflation variable and its first difference.

4.2 Out-of-sample results

In this section we verify whether the in-sample performance is confirmed by consistent out-of-sample performance. To that end we first constructed a series of 300 forecasts (25 years) using five rolling windows, each including 20 years of data for the estimation window and 5 years of forecasts. We are forced to use only five rolling windows instead of a monthly moving window, which would lead to 300 rolling estimation windows, for computational reasons because we have to evaluate all 2^{14} different models in each different window. For consistency, the forecasts using the Bayesian overall models are made with the posterior expected β, and the forecasts using the "best" individual models as selected by the statistical model selection criteria are made with the ordinary least squares β.

Table 8 displays our results, including various out-of-sample statistics for the overall posterior probability weighted average models for various ρ and for both the case of weak and information priors, the "best" individual models according to the statistical model selection criteria, the individual models with highest posterior probability for various ρ and the constant, unconditional model as a reference. We can conclude that the in-sample and out-of-sample results are consistent for the Bayesian model selection procedures, while the in-sample and out-of-sample performance differs dramatically for the "best" individual models according to the statistical criteria.

For the "best" individual models selected by these criteria—although they were developed exactly in order to give the best external validity and their in-sample results overwhelmingly favor predictability—we fail to find any predictive power out of sample. The root mean squared error and the mean

Table 8
Out-of-sample performance

	Overall models, $E[R^2] = 1\%$					Overall models, $E[R^2] = 12\%$			
	0.05	0.25	0.50	0.75		0.25	0.50	0.75	0.90
RMSE	1.080	1.080	1.078	1.078	RMSE	1.084	1.080	1.076	1.076
MAD	0.812	0.812	0.810	0.809	MAD	0.830	0.807	0.797	0.796
Bias	0.104	0.109	0.103	0.098	Bias	0.124	0.067	0.038	0.031
Deviation	1.077	1.076	1.075	1.075	Deviation	1.089	1.080	1.077	1.077

	Indiv. model, $E[R^2] = 1\%$					Indiv. model, $E[R^2] = 12\%$			
	0.05	0.25	0.50	0.75		0.25	0.50	0.75	0.90
RMSE	1.110	1.096	1.125	1.170	RMSE	1.175	1.123	1.102	1.102
MAD	0.828	0.819	0.858	0.907	MAD	0.914	0.833	0.817	0.817
Bias	0.221	0.153	0.146	0.084	Bias	0.133	−0.026	0.008	0.008
Deviation	1.089	1.087	1.117	1.169	Deviation	1.169	1.125	1.104	1.104

"Best" individual models from criteria

	R2adj	AIC	BIC	FIC	PIC	Constant
RMSE	1.123	1.121	1.096	1.112	1.112	1.093
MAD	0.833	0.833	0.819	0.833	0.833	0.818
Bias	−0.052	−0.021	0.153	0.237	0.237	0.072
Deviation	1.124	1.123	1.087	1.088	1.088	1.092

Reported are various out-of-sample statistics for the overall models for various ρ, for the "best" individual models according to the statistical criteria (Ind. Model), for the individual models with highest posterior probability for various ρ, and finally for the constant, unconditional model as a reference. Out-of-sample performance is measured using a series of 300 forecasts (25 years) constructed using five rolling windows, each including 20 years of data for the estimation window and 5 years of forecasts. The reported statistics are the root mean squared error (RMSE, the square root of average of the sum of squared forecast errors), the mean absolute deviation (MAD, the average of the absolute forecast errors), the bias (the average forecast error), and the deviation (the average of the squared difference between the bias and the forecast errors).

absolute deviation statistics are consistently greater (thus worse) than or at best equal to those of the constant, unconditional model.

Therefore, for these individual models there is a sharp contrast between the in-sample evidence of predictability and their out-of-sample performance. However, the differences are minimal. For example, the mean squared error of the adjusted R^2 criterion is about 3% higher than that of the constant model, and its bias is even smaller in absolute value. These results confirm those found in Bossaerts and Hillion (1999), who also fail to find out-of-sample predictive ability by the individual models selected by the statistical criteria, using the years 1956–1990 as their estimation window and forecasting returns for the next 5 years.

For the overall models, the (root) mean squared error and the deviation are generally smaller than those of the constant model. Although the predictive ability found is minimal, this tends to confirm the in-sample conclusions from the overall models found previously. Our results also show the benefit of explicitly taking into account the individual model uncertainty, as the out-of-sample statistics for the overall models perform better than those of the

individual models with highest posterior probability for the various ρ. Therefore, using the posterior model probability weighted average of all models definitely improves forecasts relative to individual models.

However, in many cases the bias and the mean absolute deviation for the overall models are larger than for the constant, unconditional model. Because the in-sample explained proportion of the return variation is not large (see Table 4), we can expect any out-of-sample evidence to be small. Possible explanations of the difference with the in-sample and out-of-sample performance of the overall models are data snooping in the inclusion of the variables in the initial information set, model nonstationarity, and learning in the marketplace.

Thus we argue that the most important reason for the discrepancy found with respect to the individual models selected by the statistical model selection criteria is the severe underestimation of individual model uncertainty. Furthermore, using the average, overall models improve forecasts relative to using the individual models, although out-of-sample predictability remains very small.

5. Conclusion

In this article we employed a Bayesian framework to investigate the claims of predictability of excess stock returns. Our analysis stresses the importance of accounting for the large uncertainty about which variables one should include (what is the "right" model?) and, more specifically, about whether there is, in fact, predictability. Of particular interest, we introduce a new methodology that explicitly accounts for model uncertainty and uses economically meaningful prior information to calibrate the hyperparameters in the prior distributions. This will be especially useful for future research conducted in new markets and using new predictive variables. Specifically, by simultaneously comparing all possible linear models we minimize the data snooping conditional on the initial information set.

Our positive result is that even after controlling for such data snooping, the posterior inference confirms previous findings of in-sample predictability. In particular, the posterior probability weighted average model provides superior forecasts to both the individual model with the highest posterior probability and the models selected by the classical criteria, and gives evidence of some, albeit small, out-of-sample predictability.

The data imply posterior probabilities that are in general more supportive of stock return predictability than the priors. Any skeptic on predictability would, for the choices of priors considered in this article, see his skepticism greatly decreased by the data. However, these posterior results refer to the

probability of predictability, not to its amount. Furthermore, the prior views of a confident investor are mostly confirmed by the data.

For all priors considered, the six or seven variables that do well are strikingly similar. Half of the predictors receive less support in the posterior relative to the prior. In particular, such stalwarts as past returns and dividend yields perform very poorly relative to the other variables, even for optimistic investors, which does not appear to be caused by multicollinearity with the other variables. Furthermore, the poor performance of dividend yield is confirmed by the classical model selection criteria. This is especially worrying for investors who might have stronger views on these popular variables than about others. Finally, if one has much confidence in predictability a priori, it is very hard to pick out specific variables, as almost all variables seem to be important in this case.

Models selected by the classical statistical model selection methods show overwhelming evidence of in-sample predictability, yet their out-of-sample performance is worse or no better than that of the constant model. In contrast, the in-sample and out-of-sample results for the Bayesian analysis are consistent and show some, though minimal, evidence of predictability. Most interesting, the weighting procedure of all individual models by their posterior model probability increases the predictive performance out-of-sample in all cases.

However, our approach has some limitations. First, we incorporated several restrictions such as normality, linearity, parameter stability, and in the choice of our priors. Second, while we argue that we have severely limited our data snooping ability given our initial information set, the data snooping leading to this preselection of our 14 variables from the previous literature could possibly be large.[12] This is important as we have few data points; even though we use a period of 45 years, most explanatory variables are very persistent. Third, similar to the previous literature, we treat the explanatory variables as fixed, while in reality most are lagged stochastic variables.[13]

We deem it unlikely that the results from our arguably simple framework are dominated by this bias or by the other limitations. Also, any added complexity comes at a cost, which may increase predictive power, although there may also be a corresponding increase in overfitting. In any case, our approach allows another interpretation of the previous literature by using the same (however limited) framework, and could be extended in future research.

[12] The most obvious example might be the January dummy, as it is the only calendar variable included.

[13] As Stambaugh (1999) shows under the assumption of a VAR model describing the stochastic process of the explanatory variables, this may cause a bias in the parameter estimates, especially in small samples. The effect of the bias on the posterior model probabilities, the main focus of this article, remains unclear.

References

Akaike, H., 1974, "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, AC-19, 716–723.

Avramov, D., 2002, "Stock-Return Predictability and Model Uncertainty," forthcoming in *Journal of Financial Economics*.

Black, F., 1993, "Estimating Expected Returns," *Financial Analysts Journal*, 49, 36–38.

Bossaerts, P., and P. Hillion, 1999, "Implementing Statistical Criteria to Select Return Forecasting Models: What Do We Learn?" *Review of Financial Studies*, 12, 405–428.

Campbell, J. Y., 1987, "Stock Returns and the Term Structure," *Journal of Financial Economics*, 18, 373–399.

Chen, N., R. Roll, and S. Ross, 1986, "Economic Forces and the Stock Market," *Journal of Business*, 59, 383–403.

Clyde, M., 1999, "Model Averaging and Model Search Strategies," in J. O. Berger, J. M. Bernardo, A. P. Dawid, and A. F. M. Smith (eds.), *Bayesian Statistics 6*, Oxford University Press, Oxford.

Edwards, W., H. Lindman, and L. J. Savage, 1963, "Bayesian Statistical Inference for Psychological Research," *Psychological Review*, 70, 193–242.

Fama, E. F., 1991, "Efficient Capital Markets: II," *Journal of Finance*, 46, 1575–1618.

Fernández, C., E. Ley, and M. F. J. Steel, 2001, "Benchmark Priors for Bayesian Model Averaging," *Journal of Econometrics*, 100, 381–427.

Ferson, W., 1990, "Are the Latent Variables in Time-Varying Expected Returns Compensation for Consumption Risk?" *Journal of Finance*, 45, 397–430.

Ferson, W., and C. Harvey, 1991, "The Variation of Economic Risk Premiums," *Journal of Political Economy*, 99, 385–415.

Ferson, W., and C. Harvey, 1993, "The Risk and Predictability of International Equity Returns," *Review of Financial Studies*, 6, 527–566.

Ferson, W., and C. Harvey, 1999, "Conditioning Variables and the Cross-Section of Stock Returns," *Journal of Finance*, 54, 1325–1360.

Foster, F. D., T. Smith, and R. E. Whaley, 1997, "Assessing Goodness-of-Fit of Asset Pricing Models: The Distribution of the Maximal R^2," *Journal of Finance*, 52, 591–607.

George, E. I., 1999, "Bayesian Model Selection," in *Encyclopedia of Statistical Sciences Update*, vol. 3, Wiley, New York.

George, E. I., and D. P. Foster, 2000, "Calibration and Empirical Bayes Variable Selection," *Biometrika*, 87, 731–747.

George, E. I., and R. E. McCulloch, 1993, "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–890.

George, E. I., and R. E. McCulloch, 1997, "Approaches for Bayesian Variable Selection," *Statistica Sinica*, 7, 339–373.

Geweke, J. F., 1996, "Variable Selection and Model Comparison in Regression," in J. O. Berger, J. M. Bernardo, A. P. Dawid, and A. F. M. Smith (eds.), *Bayesian Statistics 5*, Oxford University Press, Oxford.

Harvey, C. R., 1989, "Time-Varying Conditional Covariances in Tests of Asset Pricing Models," *Journal of Financial Economics*, 24, 289–317.

Hawawini, G., and D. B. Keim, 1995, "On the Predictability of Common Stock Returns: Worldwide Evidence," in R. A. Jarrow, V. Maksimovic, and W. T. Ziemba (eds.), *Handbooks in OR and MS*, vol. 9, Elsevier Science, North-Holland, 497–545.

Kass, R. E., and A. E. Raftery, 1995, "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.

Laud, P. W., and J. G. Ibrahim, 1995, "Predictive Model Selection," *Journal of the Royal Statistical Society*, Series B, 57, 247–262.

Laud, P. W., and J. G. Ibrahim, 1996, "Predictive Specification of Prior Model Probabilities in Variable Selection," *Biometrika*, 83, 267–274.

Leamer, E. E., 1978, *Specification Searches*, Wiley, New York.

Lo, A., and A. C. MacKinlay, 1990, "Data-Snooping Biases in Tests of Financial Asset Pricing Models," *Review of Financial Studies*, 3, 431–467.

Lo, A., and A. C. MacKinlay, 1997, "Maximizing Predictability in the Stock and Bond Markets," *Macroeconomic Dynamics*, 1, 102–134.

MacKinlay, A. C., and L. Pastor, 2000, "Asset Pricing Models: Implications for Expected Returns and Portfolio Selection," *Review of Financial Studies*, 13, 883–916.

Madigan, D., and A. E. Raftery, 1994, "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window," *Journal of the American Statistical Association*, 89, 1535–1546.

Madigan, D., and J. York, 1995, "Bayesian Graphical Models for Discrete Data," *International Statistical Review*, 63, 215–232.

Merton, R., 1987, "On the Current State of the Stock Market Rationality Hypothesis," in R. Dornbusch, S. Fisher, and J. Bossons (eds.), *Macroeconomics and Finance: Essays in Honor of Franco Modigliani*, MIT Press, Cambridge, MA.

Mitchell, T. J., and J. J. Beauchamp, 1988, "Bayesian Variable Selection in Linear Regression," *Journal of the American Statistical Association*, 83, 1023–1036.

Pastor, L., 2000, "Portfolio Selection and Asset Pricing Models," *Journal of Finance*, 55, 179–224.

Pastor, L., and R. F. Stambaugh, 2000, "Comparing Asset Pricing Models: An Investment Perspective," *Journal of Financial Economics*, 56, 353–381.

Pesaran, M. H., and A. Timmermann, 1995, "Predictability of Stock Returns: Robustness and Economic Significance," *Journal of Finance*, 50, 1201–1228.

Phillips, P. C. B., and W. Ploberger, 1996, "Posterior Odds Testing for a Unit Root With Data Based Model Selection," *Econometrica*, 64, 381–412.

Poirier, D. J., 1985, "Bayesian Hypothesis Testing in Linear Models With Continuously Induced Conjugate Priors Across Hypotheses," in J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith (eds.), *Bayesian Statistics 2*, University Press, Valencia, 711–722.

Poirier, D. J., 1995, *Intermediate Statistics and Econometrics: A Comparative Approach*, MIT Press, Cambridge, MA.

Pontiff, J., and L. D. Schall, 1998, "Book-to-Market Ratios as Predictors of Market Returns," *Journal of Financial Economics*, 49, 141–160.

Raftery, A. E., D. Madigan, and J. A. Hoeting, 1997, "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association*, 92, 179–191.

Richard, J. F., and M. F. J. Steel, 1988, "Bayesian Analysis of Systems of Seemingly Unrelated Regression Equations Under a Recursive Extended Natural Conjugate Prior Density," *Journal of Econometrics*, 38, 7–37.

Richardson, M., 1993, "Temporary Components of Stock Returns: A Skeptic's View," *Journal of Business and Economics Statistics*, April, 199–207.

Ross, S. A., 1989, "Regression to the Max," working paper, Yale University.

Rothenberg, T. J., 1963, "A Bayesian Analysis of Simultaneous Equation Systems," Report 6315, Econometric Institute, Netherlands School of Economics, Rotterdam.

Schwarz, G., 1978, "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 416–464.

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

Wei, C., 1992, "On Predictive Least Squares Principles," *Annals of Statistics*, 20, 1–42.

Whitelaw, R., 1994, "Time Variations and Covariations in the Expectation and Volatility of Stock Market Returns," *Journal of Finance*, 49, 515–541.

Zellner, A., 1986, "On Assessing Prior Distributions and Bayesian Regression Analysis With g-Prior Distributions," in P. K. Goel and A. Zellner (eds.), *Bayesian Inference and Decision Techniques: Essays in Honour of Bruno de Finetti*, North-Holland, Amsterdam, 233–243.