Journal of Econometrics 186 (2015) 74–93

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

What is the chance that the equity premium varies over time? Evidence from regressions on the dividend-price ratio☆

Jessica A. Wachter[a,b,*], Missaka Warusawitharana[c,1]

[a] Department of Finance, The Wharton School, University of Pennsylvania, 2300 SH-DH, Philadelphia, PA, 19104, United States

[b] NBER, United States

[c] Division of Research and Statistics, Board of Governors of the Federal Reserve System, Mail Stop 97, 20th and Constitution Ave, Washington D.C., 20551, United States

ARTICLE INFO

Article history:
Received 25 January 2012
Received in revised form
22 January 2014
Accepted 2 May 2014
Available online 11 July 2014

JEL classification:
C11
C22
G11
G17

Keywords:
Return predictability
Bayesian statistics
Model uncertainty

ABSTRACT

We examine the evidence on excess stock return predictability in a Bayesian setting in which the investor faces uncertainty about both the existence and strength of predictability. When we apply our methods to the dividend-price ratio, we find that even investors who are quite skeptical about the existence of predictability sharply modify their views in favor of predictability when confronted by the historical time series of returns and predictor variables. Correctly taking into account the stochastic properties of the regressor has a dramatic impact on inference, particularly over the 2000–2005 period.

1. Introduction

In this study, we evaluate the evidence in favor of excess stock return predictability from the perspective of a Bayesian investor. We focus on the case of a single predictor variable to highlight the complex statistical issues that come into play in this deceptively simple problem.

The investor in our model considers the evidence in favor of the following linear model for excess returns:

$$r_{t+1} = \alpha + \beta x_t + u_{t+1}, \tag{1}$$

where r_{t+1} denotes the return on a broad stock index in excess of the riskfree rate, x_t denotes a predictor variable, and u_{t+1} the unpredictable component of the return. The investor also places a finite probability on the following model:

$$r_{t+1} = \alpha + u_{t+1}. \tag{2}$$

Namely, the investor assigns a prior probability q to the state of the world in which returns are predictable (because the prior on β will be smooth, the chance of $\beta = 0$ in (1) is infinitesimal), and a probability $1 - q$ to the state of the world in which returns are completely unpredictable. In both cases, the parameters are unknown. Thus our model allows for both parameter uncertainty and "model uncertainty."[2]

☆ This paper previously circulated under the title "What is the chance that the equity premium varies over time? Evidence from predictive regressions." We are grateful to Sean Campbell, Mark Fisher, Michael Johannes, Matthew Pritsker, Robert Stambaugh, Stijn van Nieuwerburgh, Jonathan Wright, Moto Yogo, Hao Zhou and seminar participants at the 2008 meetings of the American Finance Association, the 2007 CIRANO Financial Econometrics Conference, the 2007 Winter Meeting of the Econometric Society, the 2010 Federal Reserve Conference on Financial Markets, the Federal Reserve Board, the 2013 NBER NSF time series conference, the University of California at Berkeley and the Wharton School for helpful comments. We are grateful for financial support from the Aronson+Johnson+Ortiz fellowship through the Rodney L. White Center for Financial Research. This manuscript does not reflect the views of the Board of Governors of the Federal Reserve System or its staff.

* Corresponding author at: Department of Finance, The Wharton School, University of Pennsylvania, 2300 SH-DH, Philadelphia, PA, 19104, United States. Tel.: +1 215 898 7634.

E-mail addresses: jwachter@wharton.upenn.edu (J.A. Wachter), missaka.n.warusawitharana@frb.gov (M. Warusawitharana).

[1] Tel.: +1 202 452 3461.

http://dx.doi.org/10.1016/j.jeconom.2014.05.018
0304-4076/

[2] However, note that our investor is Bayesian, rather than ambiguity averse (e.g. Chen and Epstein, 2002). Our priors are equivalent to placing a point mass on $\beta = 0$ in (1).

Allowing for a non-zero probability on (2) is one way in which we depart from previous studies. Previous Bayesian studies of return predictability allow for uncertainty in the parameters in (1), but assume uninformative priors (Barberis, 2000; Brandt et al., 2005; Johannes et al., 2002; Skoulakis, 2007; Stambaugh, 1999). As Wachter (2010) shows, flat or nearly-flat priors imply a degree of predictability that is hard to justify economically. Other studies (Kandel and Stambaugh, 1996; Pastor and Stambaugh, 2009; Shanken and Tamayo, 2012; Wachter and Warusawitharana, 2009) investigate the impact of economically informed prior beliefs. These studies nonetheless assume that the investor places a probability of one on the predictability of returns. However, an investor who thinks that (2) represents a compelling null hypothesis will have a prior that places some weight on the possibility that returns are not predictable at all.

Our work also relates to the Bayesian model selection methods of Avramov (2002) and Cremers (2002). In these studies, the investor has a prior probability over the full set of possible linear models that make use of a given set of predictor variables. Thus the setting of these papers is more complex than ours in that many predictor variables are considered. However, these papers also make the assumption that the predictor variables are either non-stochastic, or that their shocks are uncorrelated with shocks to returns. These assumptions are frequently satisfied in a standard ordinary least squares regression, but rarely satisfied in a predictive regression. In contrast, we are able to formulate and solve the Bayesian investor's problem when the regressor is stochastic and correlated with returns.

When we apply our methods to the dividend-price ratio, we find that an investor who believes that there is a 50% probability of predictability prior to seeing the data updates to a 86% posterior probability after viewing quarterly postwar data. We find average certainty equivalent returns of 1% per year for an investor whose prior probability in favor of predictability is just 20%. For an investor who believes that there is a 50/50 chance of return predictability, certainty equivalent returns are 1.72%.

We also empirically evaluate the effect of correctly incorporating the initial observation of the dividend-price ratio into the likelihood (the exact likelihood approach) versus the more common conditional likelihood approach. In the conditional likelihood approach, the initial observation of the predictor variable is treated as a known parameter rather than as a draw from the data generating process. We find that the unconditional risk premium is poorly estimated when we condition on the first observation. However, when this is treated as a draw from the data generating process, the expected return is estimated reliably. Surprisingly, the posterior mean of the unconditional risk premium is notably lower than the sample average.

Finally, when we examine the evolution of posterior beliefs over the postwar period, we find substantial differences between the beliefs implied by our approach, which treats the regressor as stochastic and realistically captures the relation between the regressor and returns, and beliefs implied by assuming non-stochastic regressors. In particular, our approach implies that the belief in the predictability of returns rises dramatically over the 2000–2005 period while approaches assuming fixed regressors imply a decline. We also evaluate out-of-sample performance over the postwar period, and show that our method leads to superior performance both when compared with a strategy based on sample averages, and when compared with a strategy implied by OLS regression.

The remainder of the paper is organized as follows. Section 2 describes our statistical method and contrasts it with alternative approaches. Section 3 describes our empirical results. Section 4 concludes.

2. Statistical method

2.1. Data generating processes

Let r_{t+1} denote continuously compounded excess returns on a stock index from time t to $t+1$ and x_t the value of a (scalar) predictor variable. We assume that this predictor variable follows the process

$$x_{t+1} = \theta + \rho x_t + v_{t+1}. \tag{3}$$

Stock returns can be predictable, in which case they follow the process (1), or unpredictable, in which case they follow the process (2).[3] In either case, errors are serially uncorrelated, homoskedastic, and jointly normal:

$$\begin{bmatrix} u_{t+1} \\ v_{t+1} \end{bmatrix} \mid r_t, \ldots, r_1, x_t, \ldots, x_0 \sim N\left(0, \Sigma\right), \tag{4}$$

and

$$\Sigma = \begin{bmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{bmatrix}. \tag{5}$$

As we show below, the correlation between innovations to returns and innovations to the predictor variable implies that (3) affects inference about returns, even when there is no predictability.

When the process (3) is stationary, i.e. ρ is between -1 and 1, the predictor variable has an unconditional mean of

$$\mu_x = \frac{\theta}{1-\rho} \tag{6}$$

and a variance of

$$\sigma_x^2 = \frac{\sigma_v^2}{1-\rho^2}. \tag{7}$$

These follow from taking unconditional means and variances on either side of (3). Note that these are population values conditional on knowing the parameters. Given these, the population R^2 is defined as

$$\text{Population } R^2 = \frac{\beta^2\sigma_x^2}{\beta^2\sigma_x^2 + \sigma_u^2}.$$

2.2. Prior beliefs

The investor faces uncertainty both about the model (i.e. whether returns are predictable or not), and about the parameters of the model. We represent this uncertainty through a hierarchical prior. There is a probability q that investors face the distribution given by (1), (3) and (4). We denote this state of the world H_1. There is a probability $1-q$ that investors face the distribution given by (2)–(4). We denote this state of the world H_0. As we will show, the stochastic properties of x have relevance in both cases.

The prior information on the parameters is conditional on H_i. Let

$$b_0 = [\alpha,\ \theta,\ \rho]^\top$$

[3] The model we adopt for stock return predictability is assumed by Kandel and Stambaugh (1996), Campbell and Viceira (1999), Stambaugh (1999), Barberis (2000) and many subsequent studies. The idea that the price-dividend ratio can predict returns is motivated by present-value models of prices (see Campbell and Shiller, 1988). We have examined the possibility of adding lagged returns on the right hand side of both the return and the predictor variable regression; however the coefficients are insignificant.

and

$$b_1 = [\alpha,\ \beta,\ \theta, \rho]^\top.$$

Note that $p(b_1, \Sigma|H_1)$ can also be written as $p(\beta, b_0, \Sigma|H_1)$.[4] We set the prior on b_0 and Σ so that

$$p(b_0, \Sigma|H_0) = p(b_0, \Sigma|H_1) = p(b_0, \Sigma).$$

We assume the investor has uninformative beliefs on these parameters. We follow the approach of Stambaugh (1999) and Zellner (1996), and derive a limiting Jeffreys prior as explained in Appendix A. As Appendix A shows, this limiting prior takes the form

$$p(b_0, \Sigma) \propto \begin{cases} \sigma_x \sigma_u |\Sigma|^{-\frac{5}{2}} & \rho \in (-1, 1) \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

Eq. (8) implies that the process for x_t is stationary and that the mean (6) and variance (7) are well defined. Stationarity of x_t is a standard assumption in the return predictability literature. Studies that rely on ordinary least squares make this assumption at least implicitly, since without it standard asymptotic arguments fail. Other recent studies (e.g. Cochrane, 2008; Van Binsbergen and Koijen, 2010) explicitly assume stationarity. In Section 3.6, we discuss how this assumption affects our results.

The parameter that distinguishes H_0 from H_1 is β. One approach would be to write down a prior distribution for β unconditional on the remaining parameters. However, there are advantages to forming priors on β jointly with priors on other parameters. For example, a high variance of x_t might lower one's prior on β, while a large residual variance of r_t might raise it. Rather than placing a prior on β directly, we follow Wachter and Warusawitharana (2009) and place a prior on the population R^2. To implement this prior on the R^2, we place a prior on "normalized" β, that is β adjusted for the variance of x and the variance of u. Let

$$\eta = \sigma_u^{-1} \sigma_x \beta$$

denote normalized β. We assume that prior beliefs on η are given by

$$\eta|H_1 \sim N(0, \sigma_\eta^2). \tag{9}$$

The population R^2 is closely related to η:

$$\text{Population } R^2 = \frac{\beta^2 \sigma_x^2}{\beta^2 \sigma_x^2 + \sigma_u^2} = \frac{\eta^2}{\eta^2 + 1}. \tag{10}$$

Eq. (10) provides a mapping between a prior distribution on η and a prior distribution on the population R^2. Given an η draw, an R^2 draw can be computed using (10).

A prior on η implies a hierarchical prior on β. The prior for η, (9), implies

$$\beta|\alpha, \theta, \rho, \Sigma \sim N(0, \sigma_\beta^2), \tag{11}$$

where

$$\sigma_\beta = \sigma_\eta \sigma_x^{-1} \sigma_u.$$

Because σ_x is a function of ρ and σ_v, the prior on β is also implicitly a function of these parameters. The parameter σ_η indexes the degree to which the prior is informative. As $\sigma_\eta \to \infty$, the prior over β becomes uninformative; all values of β are viewed as equally likely. As $\sigma_\eta \to 0$, the prior converges to $p(b_0, \Sigma)$ multiplied by a point mass at 0, implying a dogmatic view in no predictability. Combining (11) with (8) implies the joint prior under H_1:

$$p(b_1, \Sigma|H_1) = p(\beta|b_0, \Sigma, H_1) p(b_0|H_1)$$
$$\propto \frac{1}{\sqrt{2\pi\sigma_\eta^2}} \sigma_x^2 |\Sigma|^{-\frac{5}{2}} \exp\left\{-\frac{1}{2}\beta^2 \left(\sigma_\eta^2 \sigma_x^{-2} \sigma_u^2\right)^{-1}\right\}. \tag{12}$$

Jeffreys invariance theory provides an independent justification for modeling priors on β as (11). Stambaugh (1999) shows that the limiting Jeffreys prior for b_1 and Σ equals

$$p(b_1, \Sigma|H_1) \propto \sigma_x^2 |\Sigma|^{-\frac{5}{2}}. \tag{13}$$

This prior corresponds to the limit of (12) as σ_η approaches infinity. Modeling the prior for β as depending on σ_x not only has a convenient interpretation in terms of the distribution of the R^2, but also implies that an infinite prior variance represents ignorance as defined by Jeffreys (1961). Note that a prior on β that is independent of σ_x would not have this property.

Fig. 1 shows the resulting distribution for the population R^2 for various values of σ_η. Panel A shows the distribution conditional on H_1 while Panel B shows the unconditional distribution. More precisely, for any value k, Panel A shows the prior probability that the R^2 exceeds k, conditional on the existence of predictability. For large values of σ_η, e.g. 100, the prior probability that the R^2 exceeds k across the relevant range of values for the R^2 is close to one. The lower the value of σ_η, the less variability in β around its mean of zero, and the lower the probability that the R^2 exceeds k for any value of k. Panel B shows the unconditional probability that the R^2 exceeds k for any value of k, assuming that the prior probability of predictability, q, is equal to 0.5. By the definition of conditional probability:

$$p(R^2 > k) = p(R^2 > k|H_1) q.$$

Therefore Panel B takes the values in Panel A and scales them down by 0.5.

2.3. Likelihood

2.3.1. Likelihood under H_1

Under H_1, returns and the predictor variable follow the joint process given in (1) and (3). It is convenient to group contemporaneous observations on returns and on the state variable into a matrix Y and lagged observations on the state variable and a constant into a matrix X. Let

$$Y = \begin{bmatrix} r_1 & x_1 \\ \vdots & \vdots \\ r_T & x_T \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_0 \\ \vdots & \vdots \\ 1 & x_{T-1} \end{bmatrix},$$

and let

$$z = \text{vec}(Y)$$
$$Z_1 = I_2 \otimes X.$$

In the above, the vec operator stacks the elements of the matrix columnwise. Define $D = \{X, Y\}$. It follows that the likelihood conditional on H_1 and on the first observation x_0 takes the form of

$$p(D|b_1, \Sigma, x_0, H_1)$$
$$= |2\pi\Sigma|^{-\frac{T}{2}} \exp\left\{-\frac{1}{2}(z - Z_1 b_1)^\top \left(\Sigma^{-1} \otimes I_T\right)(z - Z_1 b_1)\right\} \tag{14}$$

(see Zellner, 1996).

The likelihood function (14) conditions on the first observation of the predictor variable, x_0. Stambaugh (1999) argues for treating x_0 and $x_1, \ldots, x_T$ symmetrically: as random draws from the data generating process. If the process for x_t is stationary and has run for a substantial period of time, then results in Hamilton (1994, p. 265)

[4] Formally we could write down $p(b_1, \Sigma|H_0)$ by assuming $p(\beta|b_0, \Sigma, H_0)$ is a point mass at zero.

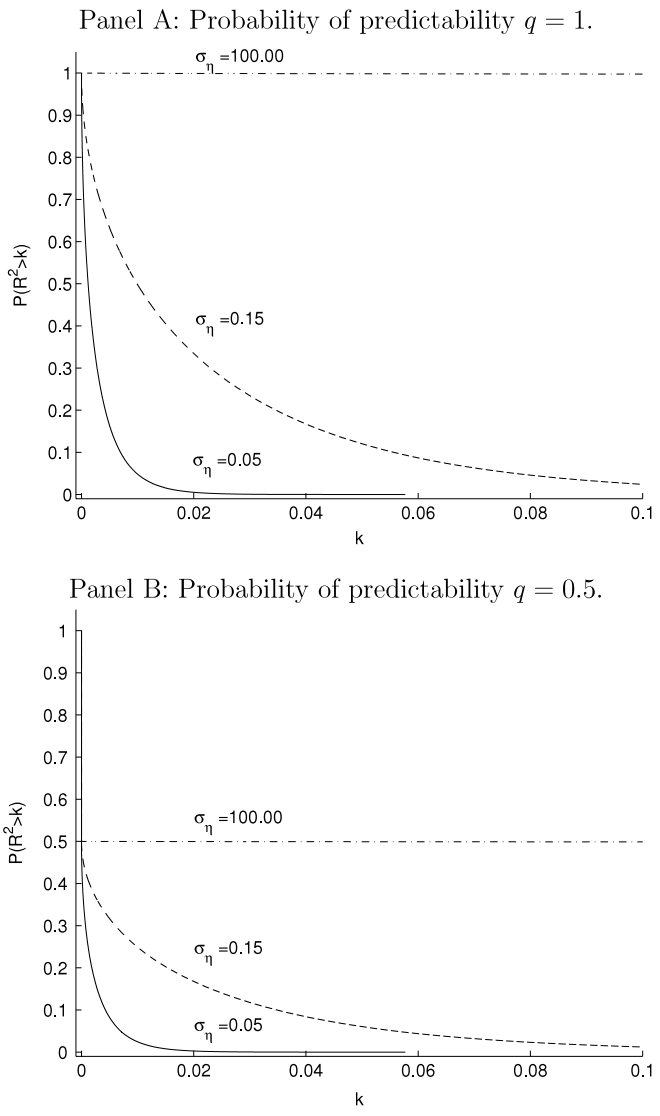

Fig. 1. Prior distribution of the R^2. Notes: The figure shows the prior probability that the R^2 is greater than k for various k. This equals one minus the cumulative prior distribution on the R^2. Panel A shows the distribution conditional on predictability and Panel B shows the full distribution assuming that the prior probability of predictability is $q = 0.5$. The parameter σ_η determines the prior standard deviation of β according to the formula $\sigma_\beta = \sigma_\eta \sigma_x^{-1} \sigma_u$, where σ_x is the standard deviation of the predictor variable and σ_u is the standard deviation of the shock to returns.

imply that x_0 is a draw from a normal distribution with mean μ_x and standard deviation σ_x. Combining the likelihood of the first observation with the likelihood of the remaining T observations produces

$$p(D|b_1, \Sigma, H_1) = |2\pi\sigma_x^2|^{-\frac{1}{2}} |2\pi\,\Sigma|^{-\frac{T}{2}} \exp\left\{-\frac{1}{2}(x_0 - \mu_x)^2 \sigma_x^{-2} - \frac{1}{2}(z - Z_1 b_1)^\top \left(\Sigma^{-1} \otimes I_T\right)(z - Z_1 b_1)\right\}. \quad (15)$$

Following Box and Tiao (1973), we refer to (14) as the *conditional likelihood* and (15) as the *exact likelihood*.

2.3.2. Likelihood under H_0

Under H_0, returns and the predictor variable follow the processes given in (2) and (3). Let

$$Z_0 = \begin{bmatrix} \iota_T & 0_{T\times 2} \\ 0_{T\times 1} & X \end{bmatrix},$$

where ι_T is the $T \times 1$ vector of ones. Then the conditional likelihood can be written as

$$p(D|b_0, \Sigma, x_0, H_0) = |2\pi\,\Sigma|^{-\frac{T}{2}} \times \exp\left\{-\frac{1}{2}(z - Z_0 b_0)^\top \left(\Sigma^{-1} \otimes I_T\right)(z - Z_0 b_0)\right\}. \quad (16)$$

The conditional likelihood takes the same form as in the seemingly unrelated regression model (see Tomohiro and Zellner, 2010). Using similar reasoning as in the H_1 case, the exact likelihood is given by

$$p(D|b_0, \Sigma, H_0) = |2\pi\sigma_x^2|^{-\frac{1}{2}} |2\pi\,\Sigma|^{-\frac{T}{2}} \exp\left\{-\frac{1}{2}(x_0 - \mu_x)^2 \sigma_x^{-2} - \frac{1}{2}(z - Z_0 b_0)^\top \left(\Sigma^{-1} \otimes I_T\right)(z - Z_0 b_0)\right\}. \quad (17)$$

As above, we refer to (16) as the *conditional likelihood* and (17) as the *exact likelihood*.

2.4. Posterior distribution

The investor updates his prior beliefs to form the posterior distribution upon seeing the data. As we discuss below, this posterior requires the computation of two quantities: the posterior of the parameters conditional on the absence or presence of return predictability, and the posterior probability that returns are predictable. Given these two quantities, we can simulate from the posterior distribution.

To compute the posteriors, we apply Bayes' rule conditional on the model:

$$p(b_i, \Sigma|H_i, D) \propto p(D|b_i, \Sigma, H_i) p(b_i, \Sigma|H_i), \quad i = 0, 1. \quad (18)$$

Because σ_x is a nonlinear function of the underlying parameters, the posterior distributions conditional on H_0 and H_1 are nonstandard and must by computed numerically. We can sample from these distributions quickly and accurately using the Metropolis–Hastings algorithm (see Chib and Greenberg, 1995; Johannes and Polson, 2006). See Appendix B for details.

Let $\bar{q}$ denote the posterior probability that excess returns are predictable. By definition,

$$\bar{q} = p(H_1|D).$$

It follows from Bayes' rule, that

$$\bar{q} = \frac{\mathcal{B}_{10} q}{\mathcal{B}_{10} q + (1 - q)}, \quad (19)$$

where

$$\mathcal{B}_{10} = \frac{p(D|H_1)}{p(D|H_0)} \quad (20)$$

is the Bayes factor for the alternative hypothesis of predictability against the null of no predictability. The Bayes factor is a likelihood ratio in that it is the likelihood of return predictability divided by the likelihood of no predictability. However, it differs from the standard likelihood ratio in that the likelihoods $p(D|H_i)$ are not conditional on the values of the parameters. These likelihoods are given by

$$p(D|H_i) = \int p(D|b_i, \Sigma, H_i) p(b_i, \Sigma|H_i)\, db_i\, d\Sigma, \quad i = 0, 1. \quad (21)$$

To form these likelihoods, the likelihoods conditional on parameters (the likelihood functions generally used in classical statistics) are integrated over the prior distribution of the parameters. Under our distributions, these integrals cannot be computed analytically. However, the Bayes factor (20) can be computed directly using the

generalized Savage–Dickey density ratio (Dickey, 1971; Verdinelli and Wasserman, 1995). Details can be found in Appendix C.

Putting these two pieces together, we draw from the posterior parameter distribution by drawing from $p(b_1, \Sigma|D, H_1)$ with probability $\bar{q}$ and from $p(b_0, \Sigma|D, H_0)$ with probability $1-\bar{q}$.

2.5. The exogenous regressor approach

Our likelihood and prior involves not only the process for returns conditional on the lagged predictor, but the process for the predictor variable itself. A common alternative is to form a likelihood function from the return equation only. That is, the likelihood function is taken to be:

$$p(R \mid X, \alpha, \beta, \sigma_u, H_1) = \left(2\pi\sigma_u^2\right)^{-\frac{T}{2}} \exp\left\{-\frac{1}{2}\sum_{t=0}^{T-1}(r_{t+1}-\alpha-\beta x_t)^2\sigma_u^{-2}\right\}, \quad (22)$$

for $R = [r_1, \ldots, r_T]^\top$. This is combined with a prior over α, β and σ_u only.

This approach is appealingly simple, but is it valid? In fact (22) is not a valid likelihood function under reasonable conditions. The reason is that, unless x_t is strictly exogenous, conditioning on the entire time series of x_t, as in (22), implies a different distribution for r_{t+1} than conditioning on x_t alone. Namely, conditional on the future values of x, r_{t+1} is not normally distributed with mean $\alpha+\beta x_t$ and variance σ_u:

$$p(r_{t+1}|x_{t+1}, x_t, \alpha, \beta, \sigma_u, H_1) \neq p(r_{t+1}|x_t, \alpha, \beta, \sigma_u, H_1).$$

The value of x_{t+1} conveys information about the shock v_{t+1}, which in turn conveys information about u_{t+1} (because they are correlated), and u_{t+1} conveys information about r_{t+1}.

Is there perhaps some other way to justify using the right hand side of (22) as a likelihood? The true (conditional) likelihood arises from taking the product of terms

$$p(D|x_0, b_1, \Sigma, H_1) = \prod_{t=0}^{T-1} p(r_{t+1}, x_{t+1}|r_t, x_t, b_1, \Sigma, H_1).^5$$

One could separate out the terms in the product as follows

$$\prod_{t=0}^{T-1} p(r_{t+1}|x_t, \alpha, \beta, \sigma_u)p(x_{t+1}|r_{t+1}, x_t, b_1, \Sigma). \quad (23)$$

However, the second term in (23) depends on α, β and σ_u. It is not, therefore, a constant when one applies Bayes' rule to inference about these parameters. Using the right hand side of (22) thus requires either incorrect conditioning on the time path of x, or an incorrect computation of the posterior.

At the root of the problem is the fact that the similarity between the likelihood in the linear regression model in the time series setting and under OLS is only apparent. In a time series setting, it is not valid to condition on the entire time path of the "independent" variable. The differences ultimately come down to the interpretation of the term u_t. In a standard OLS setting, u_t is an error, and is thus uncorrelated with the independent variable at all leads and lags. In a time series setting, it is not an error, but rather a shock, and this independence does not hold.[6]

Of course, there is a special case in which it is correct to condition on the time path of x_t. This is when the errors u_t and v_t are known to be uncorrelated at all leads and lags. In this case, x_t is strictly exogenous. This is an unrealistic assumption in a time series setting, particularly for the dividend-price ratio (or other scaled measures of market value), because future returns are by definition likely to be correlated with past prices. Indeed, the correlation between u_t and v_t is close to -1. While strict exogeneity could be enforced in the prior, it is clearly counterfactual. Fortunately it is not necessary: our analysis shows how inference can proceed without it. In what follows, we will compare our results to what would happen if x_t were taken to be strictly exogenous, which, for simplicity, we refer to as the *non-stochastic regressor* approach.

3. Results

3.1. Data

We use data from the Center for Research on Security Prices (CRSP). We compute excess stock returns by subtracting the continuously compounded 3-month Treasury bill return from the continuously compounded return on the value-weighted CRSP index at a quarterly frequency. Following a large empirical literature on return predictability, we focus on the dividend-price ratio as the regressor because the present-value relation between prices and returns suggests that it should capture variables that predict stock returns. The dividend-price ratio is computed by dividing the dividend payout over the previous 12 months with the current price of the stock index. The use of 12 months of data accounts for seasonalities in dividend payments. We use the logarithm of the dividend-price ratio as the predictor variable. Data are quarterly from 1952 to 2009.[7]

3.2. Bayes factors and posterior means

Table 1 reports Bayes factors for various priors. Four values of σ_η are considered: 0.051, 0.087, 0.148 and 100. These translate into values of $P(R^2 > .01|H_1)$ (the prior probability that the R^2 exceeds 0.01) equal to 0.05, 0.25, 0.50 and 0.99 respectively. These R^2s should be interpreted in terms of regressions performed at a quarterly frequency. Bayes factors are reported for the exact likelihood, and, to evaluate the importance of including the initial term, the conditional likelihood as well.

Table 1 shows that the Bayes factor is hump-shaped in $P(R^2 > 0.01|H_1)$. For small values, the Bayes factor is close to one. For large values, the Bayes factor is close to zero. Both results can be understood using the formula for the Bayes factor in (20) and for the likelihoods $p(D \mid H_i)$ in (21). For low values of this probability, the investor imposes a very tight prior on the R^2. Therefore the hypotheses that returns are predictable and that returns are unpredictable are nearly the same. It follows from (21) that the likelihoods of the data under these two scenarios are nearly the same and that the Bayes factor is nearly one. This is intuitive: when two hypotheses are close, a great deal of data are required to distinguish one from the other.

The fact that the Bayes factor approaches zero as $P(R^2 > .01|H_1)$ continues to increase is less intuitive. The reduction in Bayes factors implies that, as the investor allows a greater range of values for the R^2, the posterior probability that returns are predictable approaches zero. This effect is known as Bartlett's paradox, and was first noted by Bartlett (1957) in the context of distinguishing between uniform distributions. As Kass and Raftery (1995) discuss, Bartlett's paradox makes it crucial to formulate an informative prior on the parameters that differ between H_0 and H_1. We

[5] Note this likelihood function still conditions on x_0, and so is the conditional rather than the exact likelihood.

[6] This point is also emphasized by Stambaugh (1999).

[7] We obtain very similar results at annual and monthly frequencies.

Table 1
Bayes factors and conditional posterior means.

$P(R^2 > 0.01\|H_1)$	Bayes factor	Posterior means			
		β	ρ	μ_r	μ_x
Panel A: Exact likelihood					
0	Undefined	0	0.997	3.45	−3.25
0.05	4.13	1.07	0.989	3.77	−3.35
0.25	6.48	1.65	0.985	3.85	−3.38
0.50	6.13	1.91	0.983	3.88	−3.39
0.99	0.01	2.06	0.982	3.90	−3.40
Panel B: Conditional likelihood					
0	Undefined	0	0.998	4.48	−6.83
0.05	2.00	0.74	0.993	3.70	−5.28
0.25	2.71	1.36	0.988	3.39	−4.79
0.50	2.56	1.66	0.985	3.11	−4.78
0.99	0.01	1.80	0.984	2.15	−5.03
Panel C: Ordinary least squares					
		2.97	0.973	4.49	−3.54

Notes: The Bayes factor equals the probability of the data D given the predictability model H_1 divided by the probability of the data given the no-predictability model H_0: $p(D|H_1)/p(D|H_0)$. Bayes factors are reported for various priors on the strength of predictability under H_1, indexed by $P(R^2 > 0.01|H_1)$ (namely, the prior probability that the population R^2 exceeds 0.01, assuming H_1). Posterior means are conditional on H_1 and are computed for the predictability coefficient β, the persistence of the dividend-price ratio ρ, the mean of the continuously compounded excess return μ_r, and the mean of the predictor variable μ_x. In Panel C, μ_r and μ_x equal the sample means. Data are quarterly from 7/1/1952 to 3/31/2009.

can understand the paradox based on the form of the likelihoods $p(D \mid H_1)$ and $P(D \mid H_0)$. These likelihoods involve integrating out the parameters using the prior distribution. If the prior distribution on β is highly uninformative, the prior places a large amount of mass in extreme regions of the parameter space. In these regions, the likelihood of the data conditional on the parameters will be quite small. At the same time, the prior places a relatively small amount of mass in the regions of the parameter space where the likelihood of the data is large. Therefore $P(D \mid H_1)$ (the integral of the likelihood under H_1) is small relative to $P(D \mid H_0)$ (the integral of the likelihood under H_0).

Table 1 also shows that there are substantial differences between the Bayes factors resulting from the exact versus the conditional likelihood.[8] The Bayes factors resulting from the exact likelihood are larger than those resulting from the conditional likelihood, thus implying a greater posterior probability of return predictability. This difference reflects the fact that the posterior mean of β, conditional on H_1, is higher for the exact likelihood than for the conditional likelihood, and the posterior mean of ρ is lower.[9]

3.3. The long-run equity premium

For the predictability model, the expected excess return on stocks (the equity premium) varies over time. In the long run, however, the current value of x_t becomes irrelevant. Under our assumptions x_t is stationary with mean μ_x, and therefore r_t is also stationary with mean

$$\mu_r = E[\alpha + \beta x_t + u_{t+1}|b_1, \Sigma] = \alpha + \beta\mu_x.$$

This is a population value that conditions on the value of the parameters. For the no-predictability model, μ_r is simply equal to α. We can think of μ_r as the average equity premium; the fact that it is "too high" constitutes the equity premium puzzle (Mehra and Prescott, 1985), and it is often computed by simply taking the sample average of excess returns.

The posterior expectation of μ_r under various specifications is shown in Table 1. Because differences in the expected return arise from differences in the posterior mean of the predictor variable x, the table also reports the posterior mean of μ_x. The differences in the long-run equity premium are striking. The sample average of the (continuously compounded) excess return on stocks over this period is 4.49%. However, assuming the exact likelihood implies produces a range for this excess return between 3.45% and 3.90% depending on the strength of the prior. Why is the equity premium in these cases as much as a full percentage point lower?

To answer this question, it is helpful to look at the posterior means of the predictor variable, reported in the next column of Table 1. For the exact likelihood specification, the posterior mean of the log dividend yield ranges from −3.25 to −3.40. The sample mean is −3.54. It follows that the shocks v_t over the sample period must be negative on average. Because of the negative correlation between shocks to the dividend price ratio and to expected returns, the shocks u_t must be positive on average. Therefore the posterior mean lies below the sample mean.

Continuing with the exact likelihood case, the posterior mean of μ_x is highest (and hence furthest from the sample mean) in the no-predictability model, and becomes lower as the prior becomes less dogmatic. Excess returns follow this pattern in reverse, namely they are lowest (and furthest from the sample mean) for the no-predictability model and highest for the predictability model with the least dogmatic prior. This effect may arise from the persistence ρ. The more dogmatic the prior, the closer the posterior mean of the persistence is to one. The more persistent the process, the more likely the positive shocks are to accumulate, and the more the sample mean is likely to deviate from the true posterior mean.

The results are very different when the conditional likelihood is used, as shown in Panel B. For the no-predictability model, $\mu_r = \alpha$ is equal to the sample mean. However, as long as there is some predictability, estimation of μ_r depends on μ_x, which is unstable due to the presence of $1 - \rho$ in the denominator. It is striking that, in contrast to our main specification, the conditional likelihood specification has great difficulty in pinning down the mean of expected excess stock returns.

3.4. The posterior distribution

We now examine the posterior probability that excess returns are predictable. For convenience, we present results for our main specification that uses the exact likelihood. As a first step, we examine the posterior distribution for the R^2.

The posterior distribution of the R^2

Fig. 2 displays the prior and posterior distribution of the R^2. For now we assume that prior beliefs are given by $P(R^2 > 1\% \mid H_1) = 0.50$ and $q = 0.5$; below we examine robustness to changes in these values. Panel A shows $P(R^2 > k)$ as a function of k for both the prior and the posterior; this corresponds to 1 minus the cumulative density function of the R^2.[10] Panel A demonstrates a rightward shift

[8] We are not the first to note the importance of the first observation in the time series. See, for example, Poirier (1978).

[9] The source of this negative relation is the negative correlation between shocks to returns and shocks to the predictor variable. Suppose that a draw of β is below its value predicted by ordinary least squares (OLS). This implies that the OLS value for β is "too high", i.e. in the sample shocks to the predictor variable are followed by shocks to returns of the same sign. Therefore shocks to the predictor variable tend to be followed by shocks to the predictor variable that are of different signs. Thus the OLS value for ρ is "too low". This explains why values of the posterior mean of ρ are higher for low values of $P(R^2 > 0.01|H_1)$ (and hence low values of the posterior mean of β) than for high values, and higher than the ordinary least squares estimate.

[10] This figure shows the unconditional posterior probability that the R^2 exceeds k; that is, it does not condition on the existence of predictability.

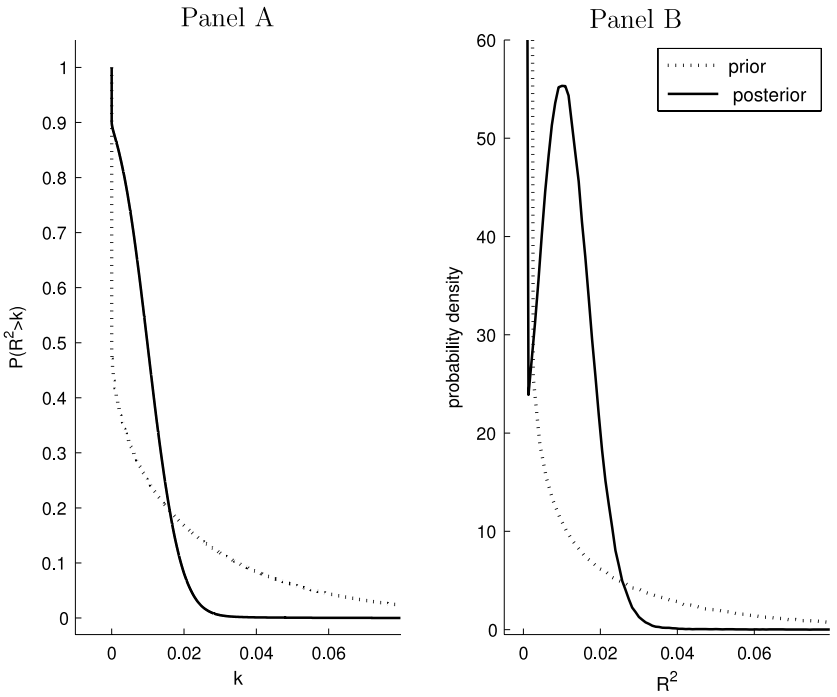

Fig. 2. Posterior Distribution of the R^2. Notes: Panel A shows the prior and posterior probabilities that the R^2 will be greater than k for various k. Panel B shows the prior and posterior density functions of the R^2. Priors are such that $P(R^2 > 0.01|H_1)$ (the probability that the R^2 exceeds 1% conditional on predictability) equals 0.5 and q (the prior probability of predictability) also equals 0.5. Data are quarterly from 7/1/1952 to 3/31/2009.

for the posterior for values of k below (roughly) 2%. While the prior implies $P(R^2 > 1\%) = 0.25$, the posterior implies $P(R^2 > 1\%)$ close to 0.50. Thus, after observing the data, an investor revises his beliefs on the existence and strength of predictability substantially upward.

Panel B shows the probability density function of the R^2. The prior places the highest density on low values of the R^2. The posterior however places high density in the region around 2% and has lower density than the prior for R^2 values close to zero. The evidence in favor of predictability, with a moderate R^2, is sufficient to overcome the investor's initial skepticism.

The posterior probability of return predictability

Table 2 shows how various statistics on the posterior distribution vary as the prior distribution changes. Panel A presents the posterior probabilities of predictability as a function of the investor's prior about the existence of predictability, q, and the prior belief on the strength of predictability. The posterior probability is increasing in q and hump-shaped in the strength of the prior, reflecting the fact that the Bayes factors are hump-shaped in the strength of the prior. An investor with moderate beliefs about the probability that returns are predictable revises these beliefs sharply upward. For example, an investor with $q = 0.5$ and $P(R^2 > .01|H_1) = 0.50$ concludes that the posterior likelihood of predictability equals 0.86. This result is robust to a wide range of choices for $P(R^2 > .01|H_1)$. As the table shows, $P(R^2 > .01|H_1) = 0.25$ implies a posterior probability of 0.87. The posterior probability falls off dramatically for $P(R^2 > .01|H_1) = 0.99$; for these very diffuse priors (which imply what might be considered an economically unreasonable amount of predictability), the Bayes factors are close to zero.[11] Panels B and C show reasonably high means of the β and the R^2, except for the most diffuse prior.

Table 2
Posterior statistics.

$P(R^2 > 0.01\|H_1)$	Prior probability of return predictability q			
	0.20	0.50	0.80	0.99
Panel A: Posterior probability of predictability $\tilde{q}$				
0.05	0.51	0.80	0.94	1.00
0.25	0.62	0.87	0.96	1.00
0.50	0.61	0.86	0.96	1.00
0.99	0.00	0.01	0.05	0.54
Panel B: Posterior mean of predictive coefficient β				
0.05	0.55	0.86	1.01	1.07
0.25	1.02	1.43	1.59	1.65
0.50	1.16	1.64	1.84	1.91
0.99	0.01	0.02	0.09	1.12
Panel C: Posterior mean of R^2 (in percentages)				
0.05	0.30	0.48	0.56	0.59
0.25	0.59	0.83	0.92	0.95
0.50	0.68	0.97	1.08	1.12
0.99	0.00	0.01	0.06	0.68
Panel D: Difference in CER between optimal and no-predictability strategies				
0.05	0.38	0.84	1.10	1.20
0.25	0.85	1.45	1.71	1.81
0.50	1.00	1.72	2.03	2.15
0.99	0.00	0.00	0.02	1.67

Notes: The table reports statistics of the posterior distribution averaged over the models H_1 (predictability) and H_0 (no predictability). The parameter q denotes the prior probability of H_1. Statistics are reported for various value of q and for priors on the strength of predictability under H_1, indexed by $P(R^2 > 0.01|H_1)$ (namely, the prior probability that the population R^2 exceeds 0.01, assuming H_1). CER stands for certainty equivalent return and is annualized by multiplying by four. Data are quarterly from 7/1/1952 to 3/31/2009.

Certainty equivalent returns

We now measure the economic significance of the predictability evidence using certainty equivalent returns. We assume an

[11] See this discussion in Section 3.2 on Bartlett's paradox.

investor who maximizes

$$E\left[\frac{W_{T+1}^{1-\gamma}}{1-\gamma}\,\middle|\,D\right]$$

for $\gamma = 5$, where $W_{T+1} = W_T(we^{r_{T+1}+r_{f,T}} + (1-w)e^{r_{f,T}})$, and w is the weight on the risky asset. The expectation is taken with respect to the predictive distribution

$$p(r_{T+1} \mid D) = \bar{q}p(r_{T+1} \mid D, H_1) + (1-\bar{q})p(r_{T+1} \mid D, H_0),$$

where

$$p(r_{T+1} \mid D, H_i) = \int p(r_{T+1} \mid x_T, b_i, \Sigma, H_i)p(b_i, \Sigma \mid D, H_i)\,db_i\,d\Sigma$$

for $i = 0, 1$. A draw r_{T+1} from the distribution $p(r_{T+1} \mid x_T, b_1, \Sigma)$ is given by (1) with probability $\bar{q}$ and (2) with probability $1-\bar{q}$.

For any portfolio weight w, we can compute the certainty equivalent return (CER) as solving

$$\frac{\exp\{(1-\gamma)\mathrm{CER}\}}{1-\gamma} = E\left[\frac{(we^{r_{T+1}+r_{f,T}} + (1-w)e^{r_{f,T}})^{1-\gamma}}{1-\gamma}\,\middle|\,D\right]. \tag{24}$$

Following Kandel and Stambaugh (1996), we measure utility loss as the difference between certainty equivalent returns from following the optimal strategy and from following a sub-optimal strategy. We define the sub-optimal strategy as the strategy that the investor would follow if he believes that there is no predictability. Note, however, that the expectation in (24) is computed with respect to the same distribution for both the optimal and sub-optimal strategy.

Panel D of Table 2 shows the difference in certainty equivalent returns as described above. These differences are averaged over the posterior distribution for x to create a difference that is not conditional on any specific value. The data indicate economically meaningful economic losses from failing to time the market. Panel D shows that, for example, an investor with a prior on β such that $P(R^2 > .01|H_1) = 0.50$ and a 50% prior belief in the existence of return predictability would suffer a certainty equivalent loss of 1.72% (in annual terms) from failing to time the market. Higher values of q imply greater certainty equivalent losses.

3.5. Evolution of the posterior distribution over time

We next describe the evolution of the posterior distribution over time. This distribution exhibits surprising behavior over the 2000–2005 period. This behavior is a direct result of the stochastic properties of the predictor variable x_t. Unless stated otherwise, the results in this section are for the benchmark specification, namely, the priors given in Section 2.2 combined with the exact likelihood. The prior probability that the R^2 exceeds 1% (conditional on predictability) and the prior probability of predictability are assumed to be 0.5.

Starting in 1972, we compute the posterior distribution conditional on having observed data up to and including that year. We start in 1972 because this allows for twenty years of data for the first observation. The posterior is computed by simulating 500,000 draws and dropping the first 100,000. To save on computation time, we update the posterior every year. For reference, Fig. 3 shows the time series of the log dividend-price ratio. As we will see, much of the behavior of the posterior distribution can be understood based on the time series of this ratio.

Fig. 4 shows the posterior probability of predictability ($\bar{q}$) in Panel A (assuming a prior probability of 0.5). The solid line corresponds to our benchmark specification. This line is above 90% for most of the sample (it is actually at its lowest value at the end of the sample). In the 2000–2005 period, the probability is not distinguishable from one. This is surprising: intuition would suggest that the period in which the dividend-price ratio was falling far below its long-run mean (and during which returns were high regardless) would correspond to an exceptionally low posterior probability of predictability, not a high one. Indeed, it is surprising that data could ever lead the investor to a nearly 100% certainty about the predictability model.

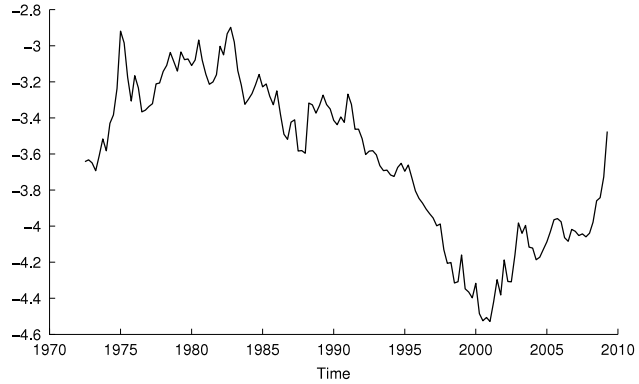

Fig. 3. The log dividend-price ratio. Notes: The figure shows quarterly observations on the log of the dividend-price ratio, computed by dividing the dividend payout over the previous 12 months by the current price. Prices and dividends are for the CRSP value-weighted index.

Panel B, which shows log Bayes factors, gives another perspective on this result. Between 2000 and 2005, the Bayes factor in favor of predictability rises to values that dwarf any others during the sample. The posterior probability takes these Bayes factors and maps them to the [0, 1] interval, so values as high as those shown in the figure are translated to posterior probabilities extremely close to one. Why is it that the Bayes factors rise so high?

An answer is suggested by the time series behavior of β and ρ, shown in Fig. 5. The solid lines show the posterior distributions of β and ρ.[12] The dashed line shows OLS estimates. The posterior for β lies below the OLS estimate for most of the period, while the posterior for ρ lies above the OLS estimator for most of the period. An exception occurs in 2001, when the positions reverse. The posterior for β lies above the OLS estimate and the posterior for ρ lies below it. Note that the OLS estimate of β is biased upwards and the OLS estimate of ρ is biased downwards, so this switch is especially surprising.

The fact that the posterior ρ rises to meet the OLS ρ, and even exceeds it, indicates that the model interprets the rise of the dividend-price ratio as occurring because of an unusual sequence of negative shocks v_t. Namely, negative shocks are more likely to occur after negative shocks during this period. This implies that positive shocks to u_t are also more likely to follow negative shocks v_t than they would in population, so OLS will in fact underestimate the true β (or it will overestimate the true β by less than usual).

This result is similar in spirit to that found in the frequentist analysis of Lewellen (2004) and Campbell and Yogo (2006) (see also the discussion in the survey, Campbell, 2008). It is also an example of how information about shocks that are correlated with errors from a forecasting model can help improve forecasts, as in Faust and Wright (2011). Fig. 4 shows that the consequences of

[12] For the argument below, it makes the most sense, strictly speaking, to examine the posterior distribution of β conditional on the predictability model. However, because the posterior probability of this model is so close to one, this conditional posterior β is nearly indistinguishable from the unconditional posterior β. The same is true for posterior ρ. Therefore, for simplicity, we focus on the unconditional posterior.

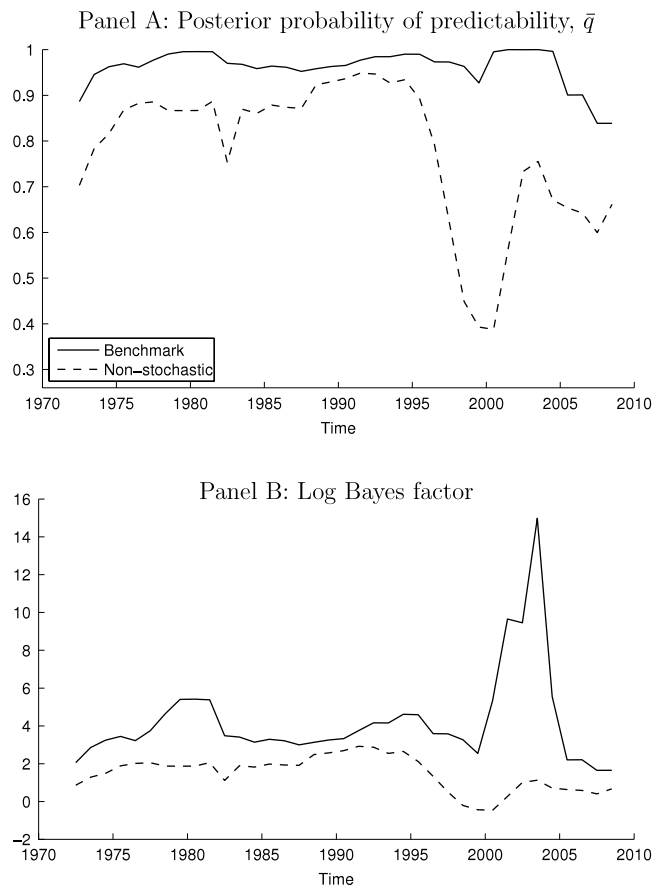

Fig. 4. The Bayes factor and posterior probability of return predictability. Notes: Panel A shows the posterior probability of H_1 (the predictability model), assuming a prior probability of 0.5. Panel B shows the log Bayes factor, equal to the log probability of the data given the predictability model H_1 minus the log probability of the data given the no-predictability model H_0. Both panels assume $P(R^2 > 0.01|H_1)$ (namely, the prior probability that the population R^2 exceeds 0.01, given H_1) equals 0.5. The Bayes factor and the posterior probability are computed using quarterly data beginning in 7/1/1952 and ending at the time shown on the x-axis. The solid line shows results for the benchmark specification. The dashed line shows results for the case of a non-stochastic regressor.

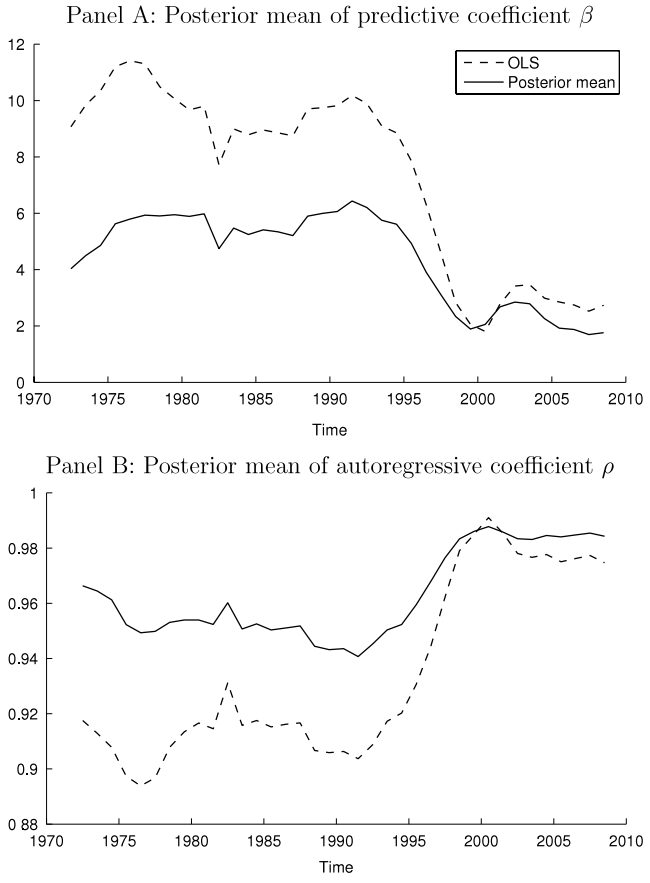

Fig. 5. Posterior means of β and ρ over time. Notes: Panel A shows the posterior mean of β under the benchmark specification (solid line) and the OLS estimate of β (dashed line) using data beginning in 7/1/1952 and ending at the time shown on the x-axis. Panel B shows analogous results for ρ, the autoregressive coefficient on the dividend-price ratio. The posterior distributions are computed assuming q (the prior probability that returns are predictable) equal to 0.50, and assuming $P(R^2 > 0.01|H_1)$ (the prior probability that the population R^2 exceeds 0.01, given H_1) also equal to 0.5.

this result for model selection are quite large. This is because the no-predictability model implies, of course, that β is zero. However, given that OLS finds a positive β, for the no-predictability model to be true, it must be the case that negative shocks to the dividend-price ratio were followed by negative shocks to returns. This is extremely unlikely, given the time series evidence and a stationary predictor variable. Thus the evidence comes to strongly favor the predictability model.

Comparison with the non-stochastic regressor approach

This chain of inference requires knowledge of the behavior of shocks to the predictor variable. The non-stochastic regressor approach described in Section 2.5 eliminates such knowledge and leads to completely different inference over this time period. To fix ideas, we implement this approach using the standard assumption of a conjugate prior distribution. However, our findings do not depend on this assumption, as we discuss in Section 3.6.

We assume the following prior distribution on the return parameters:

$$[\alpha, \beta]^\top \mid \sigma_u^2, H_1 \sim N\left(0, g^{-1}\sigma_u^2 (X^\top X)^{-1}\right) \tag{25}$$

$$\sigma_u^2 \mid H_1 \sim IW(N_0 - 2, s_0), \tag{26}$$

where IW denotes the inverse Wishart distribution, and g^{-1}, N_0 and s_0 are parameters of the prior distribution.[13] Note that prior for β conditional on σ_u is

$$\beta \mid \sigma_u^2, H_1 \sim N(0, g^{-1}\sigma_u^2 T \hat{\sigma}_x^{-2}),$$

where $\hat{\sigma}_x^2$ denotes the sample variance of x:

$$\hat{\sigma}_x^2 = \frac{1}{T}\sum_{t=0}^{T-1}\left(x_t - \frac{1}{T}\sum_{s=0}^{T-1} x_s\right)^2.$$

This allows us to construct these priors so that they are of comparable informativeness to our benchmark priors in Section 2.2 by setting

$$g^{-1}T = \sigma_\eta^2.$$

The prior in (25) and (26) is equivalent to the g-prior of Zellner (1986), and is similar to specifications employed by Fernandez et al. (2001), Chipman et al. (2001), Avramov (2002), Cremers (2002), Wright (2008) and Stock and Watson (2012). As explained in Section 2.5, the likelihood function in the non-stochastic regressor case is given by (22).

[13] We set N_0 equal to 40 and s_0 equal to the sum of squared errors over the sample, multiplied by N_0/T. The results are not sensitive to these choices. See Appendix E for further interpretation of these prior beliefs.

In our time-series setting, (25) relies on incorrect conditioning: the investor must have foreknowledge of the entire time path of the predictor variable. Thus, the approach described here builds in the assumption of a non-stochastic regressor in two ways. First, the terms involving the predictor variable do not appear in the likelihood function. Second, it conditions on the entire time path of x_t in forming the prior distribution.[14]

Appendix E describes the computation of Bayes factors and posterior probabilities in this case. The dashed line in Fig. 4 shows the posterior probabilities and Bayes factors. Notably, the non-stochastic case does not exhibit the large upward spike in Bayes factors, nor do the posterior probabilities approach one in the 2000–2005 period. Rather, the posterior probabilities decline substantially in 1998–2000, and while they increase again after this, they remain at a lower level than earlier in the sample. This behavior stems from the behavior of the OLS predictive coefficients (Fig. 5), which follow a similar pattern. The benchmark case in Fig. 4 combines this information with additional information contained in the shocks v_t, and therefore in u_t.[15] As explained in the paragraphs above, this information makes it very unlikely that the no-predictability model holds over the 2000–2005 period.[16]

3.6. The role of the prior and likelihood in determining Bayes factors

As Section 3.5 shows, whether one models the predictor variable as stochastic or not has a large impact on inference. This section delves more deeply into the reasons for this difference.

Clearly there are many differences between the stochastic (benchmark) and non-stochastic case. Most fundamentally, the benchmark case requires specifying a likelihood function for the data on the predictor variable. This in turn requires a prior over the parameters of this likelihood function. By modeling the predictor as non-stochastic, one appears to avoid this step.

In specifying this prior, we assume that the predictor variable is stationary. Without this assumption, we could not define a prior over the R^2 (because the variance is not well-defined) nor would we have an exact likelihood function (there would be no well-defined distribution for x_0). As we discuss in Section 2.2, this assumption is standard in the return predictability literature, though it is not always stated explicitly. Thus in our setting stationarity is a natural assumption. Here, we seek to understand how it affects our results and why.

We first ask whether it matters if we use the exact or the conditional likelihood. We do this by comparing our benchmark case with one in which we use the conditional likelihood and keep all else the same. Results are shown in Panel A of Fig. 6. Using the conditional likelihood leads to lower Bayes factors, though the Bayes factors still spike up over the 2000–2005 period. The information from the first observation shifts the distribution of ρ toward lower values because the mean of the predictor variable is sufficiently close to the first observation that a high variance of the predictor variable is not necessary to explain the data (a decrease in ρ decreases the unconditional variance of x). Because the distribution of ρ is shifted toward lower values, the distribution of β is shifted toward higher values (see Section 3.2 and Table 1) leading to higher Bayes factors. However, while the exact likelihood does lead to higher Bayes factors, both sets of likelihood functions imply similar time series patterns. Thus the use of the exact likelihood function, by itself, is not the main driver of the difference between the non-stochastic and benchmark case.

We next consider the effect of a prior on the R^2 versus a prior on β. In doing so, we wish to isolate the effect of the prior on β as much as possible. We consider the following:

$$p(\beta|b_0,\, \Sigma,\, H_1) \sim N(0,\, \hat{\sigma}_\beta^2), \tag{27}$$

where

$$\hat{\sigma}_\beta = \sigma_\eta \hat{\sigma}_x^{-1} \hat{\sigma}_u. \tag{28}$$

We compute $\hat{\sigma}_u$ as the standard deviation of the residual from OLS regression of the predictive regression. We assume a standard uninformative prior for the remaining parameters (Zellner, 1996):

$$p(b_0,\, \Sigma|H_1) = p(b_0,\, \Sigma|H_0) \propto |\Sigma|^{-\frac{3}{2}}, \tag{29}$$

for $\rho \in (-1,\, 1)$, and zero otherwise. It follows that

$$p(b_1,\, \Sigma|H_1) \propto \frac{1}{\sqrt{2\pi\hat{\sigma}_\beta^2}} |\Sigma|^{-\frac{3}{2}} \exp\left\{-\frac{1}{2}\beta^2 \hat{\sigma}_\beta^{-2}\right\}. \tag{30}$$

We refer to this as the *empirical Bayes prior* because the data are used to construct $\hat{\sigma}_\beta$. This contrasts with the full Bayes prior that forms our benchmark specification.

The prior over β implied by (27) and (28) is almost identical to the conjugate-g prior used to explore the non-stochastic regressor approach in Section 3.6.[17] When we allow ρ to be above 1 and use the conditional likelihood, the results are nearly identical to the non-stochastic regressor case. This is not surprising: the prior over β is nearly the same in both cases, and the likelihood function is exactly the same. However, we can impose stationarity on the empirical Bayes prior, which we cannot do in the non-stochastic case. Thus we can make the empirical Bayes case more comparable to our benchmark case. Panel B shows the results of this exercise: We use the conditional likelihood, and compare the results of the full Bayes (benchmark) and the empirical Bayes priors. For the empirical Bayes priors, we assume $\rho \in (-1,\, 1)$. The results in Panel B show that, while the use of empirical Bayes raises the Bayes factors somewhat, the effect is relatively small. Replacing full Bayes with empirical Bayes partially cancels out the effect of replacing the exact likelihood with the conditional likelihood, in this sample at least. Thus the use of a prior over the R^2 rather than a prior over β plays at most a minor role in our results.

Finally, in the last panel, we consider the empirical Bayes prior and conditional likelihood with and without stationary. Without stationarity, we are in effect back to the non-stochastic regressor case. We see that whether ρ is restricted to be less than one makes a large difference in the results. As explained in the previous section, the model interprets the rise in the dividend-price ratio as occurring because of an unusual sequence of negative shocks. Because of the negative correlation between the dividend-price

[14] An alternative approach would be to form a conjugate g-prior over a multivariate system that includes the equation for the state variable. Under this approach, terms involving the predictor variable would appear in the prior and likelihood function. However, it would still involve incorrect conditioning in that the entire path of x_t would be used in forming the prior. This approach is described in detail in Appendix D. Comparing the resulting posterior distribution with that from the one-equation conjugate prior case reveals that they differ up to a degrees of freedom adjustment arising from the need to estimate the correlation between the two equations.

[15] For the information in v_t to matter, there must be a non-zero correlation between u and v. As Appendix G shows, in the case of the yield spread, the benchmark and non-stochastic cases look nearly identical in part because the correlation between shocks to the yield spread and shocks to returns is low in magnitude.

[16] The effect is most dramatic over the 2000–2005 period, but holds to some extent in other parts of the sample period as well. This is one of the reasons why Bayes factors for the benchmark case lie above those for the non-stochastic case throughout the sample.

[17] The distinction is whether σ_u is taken from the sample or conditioned on. Because σ_u is estimated very precisely, this distinction makes little practical difference.

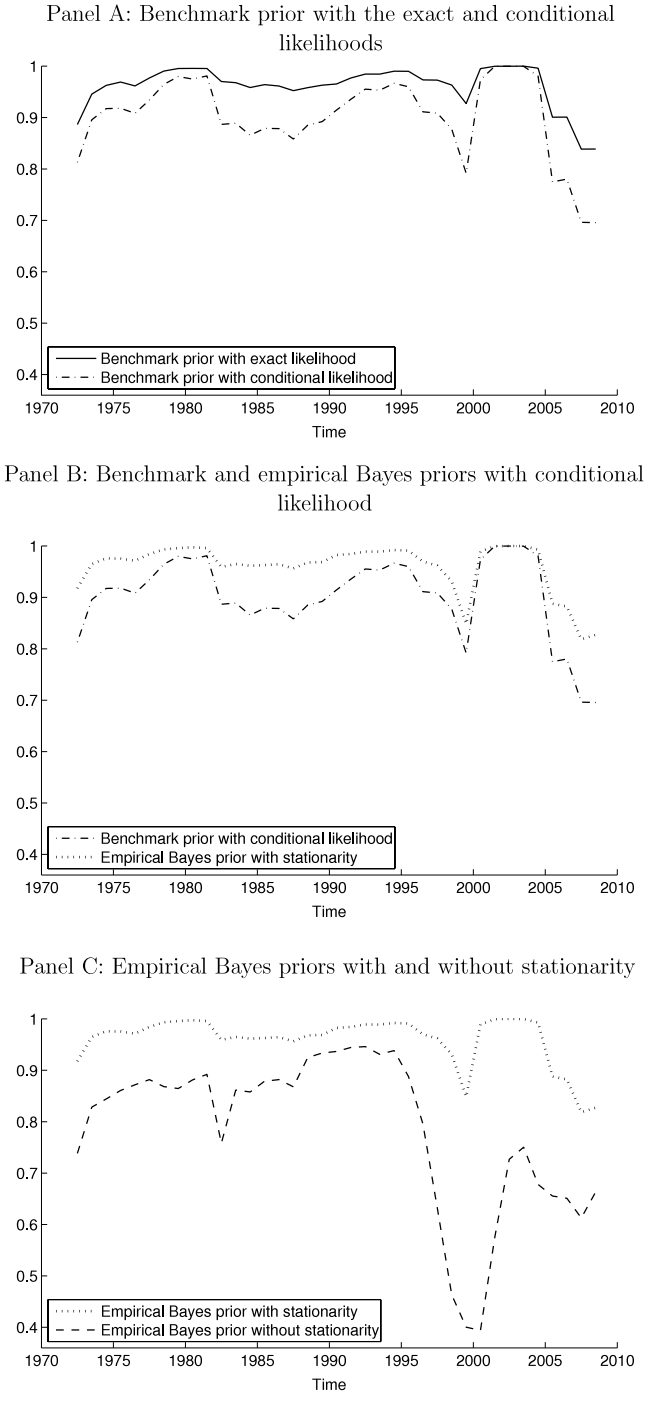

Fig. 6. Posterior probabilities implied by different methods. The figures show the posterior probability of return predictability (see Fig. 4 for more details). Panel A compares our benchmark specification "Exact likelihood", with a specification that uses the conditional likelihood but keeps the prior the same. Panel B compares this latter specification with one that uses the empirical Bayes prior but keeps everything else the same; note both specifications use the conditional likelihood. Panel C compares results from the conditional likelihood and empirical Bayes prior assuming stationarity with results from this same specification without assuming stationarity.

ratio and returns, one would expect positive shocks to returns to follow negative shocks to the dividend-price ratio. In such a sample, OLS would be biased downward, not upward. However, the no-predictability model by definition implies that OLS must be biased upward. Restating somewhat, over this period there is still a negative relation between the lagged dividend-price ratio and returns. The fact that this relation is weakened is not so much evidence against predictability but rather a consequence of an unusual set of shocks. If there truly were no predictability, it would have had to have weakened much further.

It might seem that the empirical Bayes approach, or indeed the stochastic regressor approach (these are nearly identical), is more robust, as it does not require an assumption of stationarity. However, recall that these approaches rely on incorrect conditioning. They assume not only that the agent can see part of the data but not the rest, but that the agent is not allowed to make full use of this data for inference. Moreover, this apparent robustness is itself concerning. The non-stochastic regressor approach can be shown to be equivalent to the use of ordinary least squares (OLS).[18] Yet, OLS is known to be biased in the time series setting, and invalid when the right-hand-side variable is non-stationary. The fact that OLS (with its known flaws) plays a central role in the non-stochastic case, combined with the fact that this case relies on incorrect conditioning would seem to make the non-stochastic case a less than ideal foundation for Bayesian inference in a time-series setting.

One could generalize the prior distribution that we introduce to allow for a non-stationary distribution for x_t. This would of course admit the possibility that excess returns, too, are non-stationary and the equity premium undefined. We leave this interesting topic to future work.

3.7. The training sample approach

An alternative approach that (like the non-stochastic case) makes use of the principles of conjugacy is to form a prior using a training sample.[19] Unlike the non-stochastic case described in Section 3.5, the training sample approach does not require foreknowledge of the time series of x.[20]

In this section, we evaluate this approach in the setting of model uncertainty. Consider a training sample (an early sub-sample of the data) with $\tilde{T}$ time series observations. Let $\tilde{X}$ and $\tilde{Y}$ denote the analogs to X and Y, defined over this prior sample, $\tilde{b}_1$ the regression coefficients computed over this sample, and $\tilde{S}$ the sum of squared errors. That is:

$$\tilde{B}_1 = (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top \tilde{Y} \tag{31}$$

$$\tilde{b}_1 = \mathrm{vec}(\tilde{B}_1) \tag{32}$$

$$\tilde{S} = (\tilde{Y} - \tilde{X}\tilde{B}_1)^\top (\tilde{Y} - \tilde{X}\tilde{B}_1). \tag{33}$$

Prior beliefs are given as follows:

$$p(b_1|\Sigma, H_1) \propto |\Sigma|^{-1} \exp \times \left\{ -\frac{1}{2}(b_1 - \tilde{b}_1)^\top (\Sigma^{-1} \otimes X^\top X)(b_1 - \tilde{b}_1) \right\} \tag{34}$$

$$p(\Sigma|H_1) \propto |\Sigma|^{-\frac{N+1}{2}} \exp\left\{ -\frac{1}{2} \mathrm{tr} \Sigma^{-1} \tilde{S} \right\}, \tag{35}$$

which implies

$$b_1 \sim N\left(\tilde{b}, \Sigma \otimes (\tilde{X}^\top \tilde{X})^{-1}\right) \tag{36}$$

$$\Sigma \sim IW\left(\tilde{S}, \tilde{T} - 2\right). \tag{37}$$

[18] An apparent alternative would be to allow a flat prior for both β and ρ (thus making the prior over the R^2 unnecessary). As discussed above, this leads to Bayes factors close to zero because of Bartlett's paradox. A second alternative would be to create a training sample. We explore this alternative in detail in the next section.

[19] See Johannes et al. (2014).

[20] Though it does use the conditional rather than the exact likelihood.

This prior distribution can be interpreted as the beliefs the investor would have if starting with a (true) uninformative prior and updated using the conditional likelihood (14) for $\tilde{T}$ observations. The resulting distributions follow from calculations in Zellner (1996, pp. 224–227).[21]

Bayes theorem and the results in Zellner (1996) imply that the posterior distribution takes the same form, but with the training sample quantities replaced by their full-sample counterparts.[22] Let

$$\hat{b}_1 = \text{vec}(\hat{B}_1)$$
$$\hat{B}_1 = (X^\top X)^{-1} X^\top Y$$
$$S = (Y - X\hat{B}_1)^\top (Y - X\hat{B}_1).$$

It follows that

$$p(b_1|\Sigma, H_1, D) \propto |\Sigma|^{-1} \exp\left\{-\frac{1}{2}(b_1 - \hat{b}_1)^\top (\Sigma^{-1} \otimes X^\top X)(b_1 - \hat{b}_1)\right\} \quad (38)$$

$$p(\Sigma|H_1, D) \propto |\Sigma|^{-\frac{T+1}{2}} \exp\left\{-\frac{1}{2}\text{tr}\Sigma^{-1}S\right\}, \quad (39)$$

which implies

$$b_1 \sim N\left(\hat{b}_1, \Sigma \otimes (X^\top X)^{-1}\right) \quad (40)$$

$$\Sigma \sim IW\,(S, T-2)\,. \quad (41)$$

Appendix F describes the computation of Bayes factors.

The disadvantage of this approach is that inference is very sensitive to the choice of the training sample. Fig. 7 shows the implied prior distribution for the coefficient β under different training samples (Panel A) and the corresponding posterior probabilities of predictability (Panel B). We consider priors of length 8, 16 and 40 quarters (Johannes et al., 2014 use monthly data and a training sample length of 24 months). All three prior-likelihood combinations use exactly the same data; the only difference is whether the data is labeled as part of the prior or the likelihood. Nonetheless, the differences in the economic conclusions are striking. A prior formed using 8 quarters of data yields a posterior probability of only 10% at the end of the sample, while assigning 16 quarters to the prior implies a posterior probability of above 50%. Increasing the data in the prior is no guarantee of stability: the posterior probability formed when the prior is 40 quarters is close to 30%.

What is the source of this indeterminacy? As we discuss in Section 3.2, Bartlett's paradox implies that too diffuse a prior will lead to very low Bayes factors, because the mass of the prior is far from what the data suggest. Priors based on a small training sample run into exactly this problem (as can be seen from the prior formed using 8 quarters of data). On the other hand, using a moderate-sized training sample creates its own problems. For example, 40 quarters of data imply a prior distribution that is no longer diffuse. However, because this prior is centered at a different value than that implied by the full sample, the posterior probability is also lower than for the 16-quarter prior. Indeed, Fig. 7 shows that the shortest training sample implies a prior that is diffuse and has little weight on relatively low values of β while the longest training sample implies a prior that is highly informative, but also places little weight on relatively low values of β. In both cases, the Bayes factors are low.

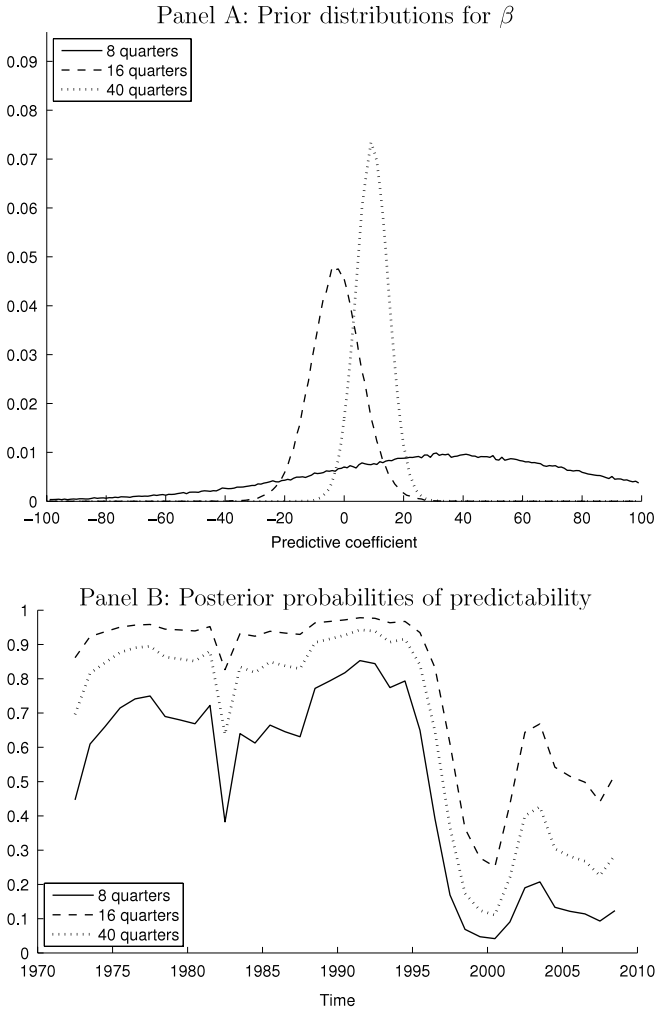

Fig. 7. The conjugate prior and posterior with training samples. Notes: Panel A shows the prior distribution for β (the predictive coefficient) assuming training samples of varying lengths. Panel B shows the posterior probabilities of predictability. The training samples begin on 7/1/1952 and last for the number of quarters given in the legend. The posterior probabilities are computed using the remaining data, ending at the time shown on the *x* axis.

More intuition can be obtained using the formula for the log Bayes factor that applies in this instance:

$$\log \mathcal{B}_{10} = \log p(\beta = 0|H_1) - \log p(\beta = 0|D, H_1) \quad (42)$$

(see Verdinelli and Wasserman, 1995). By definition, altering the end point of the training sample has no effect on the posterior probability of $\beta = 0$, because the posterior is invariant to how the data are divided between the training and the actual sample. However, it will of course affect the prior probability that $\beta = 0$. Eq. (42) shows that the log Bayes factor undergoes a linear shift depending on the training sample. Thus, while the training sample approach avoids some problems with the conjugate prior, it introduces a new one, namely indeterminacy with respect to the choice of the training sample.

3.8. Out-of-sample performance

Goyal and Welch (2008) argue that the out-of-sample performance of predictive regressions, when implemented using standard techniques, is quite poor. This raises the question of whether our approach to predictability leads to superior out-of-sample performance.

[21] We make the standard assumption that true uninformative prior is flat for b_1 and proportional to $|\Sigma|^{-3/2}$ for Σ. Eqs. (31)–(37) then follow from the calculations in Zellner (1996) for the *posterior* given data $\tilde{X}$ and $\tilde{Y}$ of sample length $\tilde{T}$.

[22] For consistency with earlier sections of the paper, we continue to use T as the length of the full sample. The full sample is then comprised of the training sample of length $\tilde{T}$ and an additional sample of length $T - \tilde{T}$.

In this section, we answer this question using the same CRRA utility function used to evaluate in-sample performance in Section 3.4. As in that section, we consider a one-period investor who chooses a weight in the risky asset. We first assume that the investor follows an optimal strategy, that is, he computes expected utility with respect to the predictive distribution of returns (see Section 3.4), and chooses a portfolio strategy to maximize this expected utility. We then compute the out-of-sample certainty equivalent return (CER) associated with this strategy. That is, for each quarter in the sample, we apply the optimal weights computed using information available at that quarter to the actual returns realized over the next quarter. This gives us a time series of quarterly returns; we use this time series to compute the expectation on the right hand side of (24).

We compare the resulting CER to that resulting from a sub-optimal strategy.[23] Motivated by the findings of Goyal and Welch (2008), we first consider the strategy in which the investor computes the distribution of returns assuming no predictability and that the mean and volatility are given by their sample moments.

The results are shown in Panel A of Table 3. We find a positive difference between CERs, indicating superior out-of-sample performance relative to the sample means, for each of the prior beliefs we consider. As elsewhere in the paper, we consider a range of prior beliefs on predictability q and the probability that the R^2 exceeds 0.01. The results are largest for those prior beliefs that lead to a relatively high weight on the predictability model (namely $P(R^2 > .01|H_1) = 0.50$).

Panel A of Table 3 shows that strategies implied by our method outperform a simple strategy based on sample moments. We also assess the statistical significance of this outperformance. That is, we ask: could this outperformance have occurred in a sample with no predictability? Note that outperformance in a no-predictability setting need not be spurious. This is because our strategies not only incorporate evidence on predictability, but allow for Bayesian updating on all the parameters. In performing this exercise, we are assessing the extent to which this outperformance itself constitutes evidence for return predictability.[24]

To accurately capture non-standard features of the portfolio return series, we simulate 400 samples under the null hypothesis of no predictability.[25] For each of these samples, we calculate out-of-sample performance, repeating the procedure we used to calculate performance in actual data. We limit the number of samples to 400 due to the heavy computational requirement of this exercise. Because we have no reason to believe that our method would perform worse under the alternative hypothesis of predictability than under the null, we consider a one-tailed test and report, in brackets, the 95 percent critical value from our simulations. The results show that our out-of-sample values exceed this critical value for 11 out of the 20 priors that we consider. We conclude therefore that the out-of-sample performance our strategies exhibit would have been quite unlikely in a setting with no predictability.

Our results based on sample means raise the question of whether our strategies outperform those constructed using OLS estimates (which were used for evaluation by Goyal and Welch (2008)). We repeat the exercise above, but rather than consider a sub-optimal strategy based on sample means, we consider a sub-optimal strategies constructed using the OLS estimates. Panel B indicates that the OLS strategies do perform worse, reconciling our findings with those of Goyal and Welch. For completeness, we also report the 95% critical value, constructed as described above. However, there is no reason to expect that the difference between our strategies and those based on OLS would be statistically significant, and indeed they are not.[26]

Fig. 8 shows the portfolio weights corresponding to the optimal strategy, the sample mean strategy and the OLS-based strategy. Not surprisingly, the sample mean strategy varies slowly over the period, reflecting changes in the measurement of the mean return. This strategy makes no use of the predictability of stock returns, which, when applied in our Bayesian setting, do turn out to lead to superior out of sample performance. However, the weights implied by the strategy with a 50% prior belief in predictability are notably less volatile than an OLS-based strategy. In fact, the OLS strategy spends much of the time at either 0% or 100% in equities (the discrete-time CRRA investor would never choose to short equities or to invest more than 100% in equities because of the non-zero probability of negative wealth). It is likely that the Bayesian strategy would outperform by an even greater extent if one were to restrict the return distribution to allow for optimal strategies

Table 3
Out of sample certainty equivalent returns (CERs).

$P(R^2 > 0.01\|H_1)$	Prior prob. of return predictability q				
	0.01	0.20	0.50	0.80	0.99
Panel A: Comparison with sample mean					
0.05	1.11*	1.08*	1.06*	1.07*	1.07*
	[0.98]	[1.02]	[1.03]	[1.05]	[1.04]
0.25	1.05	0.91	1.02*	1.08*	1.10*
	[1.07]	[0.98]	[0.99]	[0.99]	[1.00]
0.50	0.85	1.09*	1.20*	1.24	1.25
	[1.08]	[1.08]	[1.20]	[1.32]	[1.35]
0.99	1.17*	1.04*	0.96	0.79	1.12
	[0.99]	[0.99]	[1.01]	[1.05]	[1.13]
Panel B: Comparison with OLS estimates					
0.05	1.06	1.03	1.02	1.02	1.03
	[1.86]	[1.85]	[1.86]	[1.83]	[1.83]
0.25	1.00	0.86	0.98	1.03	1.05
	[1.78]	[1.87]	[1.86]	[1.81]	[1.87]
0.50	0.81	1.04	1.15	1.19	1.20
	[1.91]	[1.83]	[1.79]	[1.80]	[1.77]
0.99	1.12	0.99	0.92	0.74	1.07
	[1.87]	[1.87]	[1.87]	[1.86]	[1.85]

Notes: For each year beginning in 1972, the predictive distribution for returns is computed using all data up to that year. Optimal portfolios are computed quarterly to maximize the utility of an agent with constant relative risk aversion equal to 5; these are combined with the actual returns over the following quarter to create out-of-sample returns on the investment strategy. The CER is the riskfree rate of return that generates the same average utility as this series of returns. Panel A reports the CER for the optimal Bayes strategy using the benchmark approach (the benchmark CER) minus the CER for portfolio weights assuming there is no predictability and that the mean and volatility of returns are equal to their sample counterparts. Panel B reports the benchmark CER minus the CER for portfolio weights computed assuming the process for returns is as estimated using OLS. Statistics are reported for various values of q and for priors on the strength of predictability under H_1 (predictability model), indexed by $P(R^2 > 0.01|H_1)$ (the prior probability that the population R^2 exceeds 0.01, assuming H_1). CERs are annualized by multiplying by four. Numbers in brackets report 95% critical values, generated using Monte Carlo assuming no return predictability. Starred values are significant at the five percent level using a one-tailed test.

[23] As in Section 3.4 and in Kandel and Stambaugh (1996), we measure utility loss by taking the difference between the CER of the optimal strategy and the CER of the suboptimal strategy.

[24] Unlike the rest of the paper, this exercise is purely frequentist in nature. The Bayesian investor would not require such evidence under our framework.

[25] In setting the parameters for this Monte Carlo, we take into account the bias in ρ. We choose ρ to be 0.997, which happens to be its estimate under the no-predictability model. This value of ρ leads to an average OLS estimate of 0.973, similar to that in the data.

[26] In a previous study (Wachter and Warusawitharana, 2009) we found extremely poor performance for an OLS investor. In that study, we assumed mean–variance weights, which allowed for positions of unlimited size. In this study, we assume a CRRA investor, whose weight in the risky asset always falls between 0 and 1. This makes a difference for the OLS strategy, given the extreme nature of the implied beliefs.

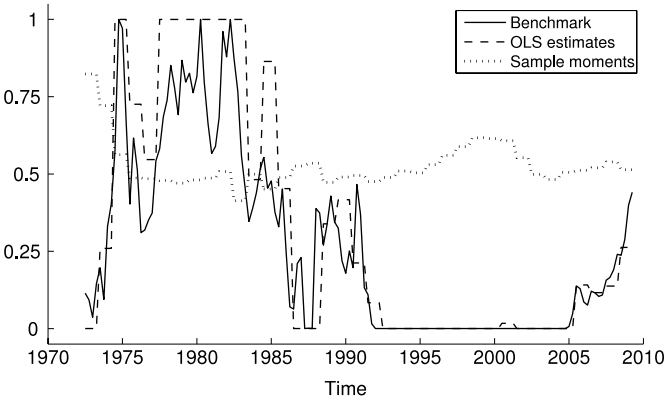

Fig. 8. Portfolio weights for the benchmark approach and implied by sample moments. Notes: The figure shows the time series of weights in the risky asset using the benchmark approach and assuming $q = 0.50$ and $P(R^2 > 0.01|H_1) = 0.50$. The figure also shows the time series of weights assuming that parameters estimated by OLS are known with certainty, as well as the time series of weights computed assuming that returns are not predictable, but that the resulting moments are known with certainty (sample moments). Data are quarterly from 7/1/1952 to the date shown on the x-axis.

outside of these bounds. These results show how economically motivated prior beliefs can improve investment performance out-of-sample, as well as in-sample.

3.9. Allowing for time-varying volatilities

Stochastic volatility is a well-established property of financial returns. Here, we discuss how our approach would generalize to allow for this property.

A critical aspect to our approach is the presence of an informative prior over the predictive coefficient β. This informative prior is what allows us to calculate Bayes factors, and posterior probabilities over models. If this prior were flat, Bartlett's paradox would lead to Bayes factors close to zero. The flat prior is in a sense informative because the predictive coefficient can become very large, leading to implied priors on the R^2 that are unreasonable on economic grounds. The Bayes factors in this case are low, not because predictability is absent, but because the supposedly uninformative prior places too much weight on unreasonable areas of the parameter space. Our approach allows the investor to place an informative prior on the predictive coefficient in a natural and intuitive way.

This insight can be readily generalized to a setting that allows for time-varying volatility. Here, we outline one such approach. Consider a data generating process as in Section 2.1, except allow the volatility of the shocks, and potentially the predictive coefficient, to change over time. That is, we compare the predictive model

$$r_{t+1} = \alpha + \beta_t x_t + u_{t+1}$$

to one without predictability (2), where x_t is given by (3), and the shocks u_{t+1} and v_{t+1} are governed by

$$\begin{bmatrix} u_{t+1} \\ v_{t+1} \end{bmatrix} | \, r_t, \ldots, r_1, x_t, \ldots, x_0 \sim N\left(0, \Sigma_t\right),$$

with

$$\Sigma_t = \begin{bmatrix} \sigma_{u,t}^2 & \sigma_{uv,t} \\ \sigma_{uv,t} & \sigma_{v,t}^2 \end{bmatrix}.$$

We assume that Σ_t follows a multivariate stochastic process such that it is positive definite with probability one.[27] Rather than prior beliefs over Σ itself, the investor would have a prior over the hyperparameters of this process. Because second moments (as opposed to first moments) can generally be accurately measured, the precise form of these priors might not turn out to be important for the conclusions.[28]

As discussed above, the aspect of our approach that one would wish to preserve in this setting is the informative prior on β and its link to the R^2 statistic. The simplest generalization would keep η as a constant parameter with the distribution

$$\eta|H_1 \sim N(0, \sigma_\eta^2). \tag{43}$$

The relation

$$\beta_t = \sigma_{x,t}^{-1}\sigma_{u,t}\eta$$

then gives the prior distribution over β_t. This definition assumes that time-varying parameters are part of the agent's information set at time t, for the purpose of the R^2 calculation.[29] Regardless of time-variation in $\sigma_{u,t}$ and $\sigma_{v,t}$, this would ensure that the amount of predictability remains economically reasonable. Note that the posterior means for β_t could, and most likely would, vary over time.

This system could be generalized still further by allowing η itself to vary over time, replacing (43) with priors on the hyperparameters on the process for η. This prior would allow investor to have the view that predictability could vary over time in a way that is unrelated to the variance of the predictor variable or of returns.

The advantage of either of these approaches is that they would allow the investor to consider both time-varying first and second moments in his investment decision. Given the evidence that second moments vary, this would be useful in improving out of sample performance. However, the qualitative findings of the importance of predictability reported earlier in the manuscript do not rely on homoskedasticity but rather on the negative correlation between shocks to the dividend-price ratio and returns. Thus, while introducing time-varying second moments would be interesting, we expect that our main results would be unaffected.

[27] The difficulties in modeling Σ_t are not unique to our setting, but arise in any multivariate setting with stochastic volatility.

[28] See Johannes et al. (2014) for a recent Bayesian analysis of stochastic volatility in a return predictability setting.

[29] In this, it is analogous to our current calculation for the R^2, which conditions on the true parameters. Note that this only matters for the interpretation of the priors, not for the calculation of the priors themselves.

4. Conclusion

This study takes a Bayesian approach to answering the question of whether the equity premium varies over time. We consider investors who face uncertainty both over whether predictability exists, and over the strength of predictability if it does exist. We find substantial evidence in favor of predictability when the dividend-price ratio is used to predict returns. Moreover, we find large certainty equivalent losses from failing to time the market, even for investors who have strong prior beliefs in a constant equity premium. Our strategies exhibit improved out-of-sample performance when compared with no-predictability strategies and when compared with OLS.

We depart from previous studies in that we model the regressor as stochastic rather than fixed. We show that this raises the probability of predictability in general, and particularly during the 2000–2005 period. Thus the model chosen for the regressor can significantly affect Bayesian inference, often in non-obvious ways. In this study, we model the predictive variable as following a stationary process, and the predictor variable and returns as homoskedastic. Exploring alternative distributional assumptions and their consequences for inference on returns is an interesting topic for further work.

Appendix A. Jeffreys prior under H_0

Given a set of parameters μ, data D, and a log-likelihood $l(\mu; D)$, the limiting Jeffreys prior satisfies

$$p(\mu) \propto \left| -E\left(\frac{\partial^2 l}{\partial \mu \partial \mu^\top} \right) \right|^{1/2}. \tag{A.1}$$

Our derivation for the limiting Jeffreys prior on b_0, Σ generalizes that of Stambaugh (1999). Zellner (1996, pp. 216–220) derives a limiting Jeffreys prior by applying (A.1) to the likelihood (17) and retaining terms of the highest order in T. Stambaugh shows that Zellner's approach is equivalent to applying (A.1) to the conditional likelihood (16), and taking the expectation in (A.1) assuming that x_0 is multivariate normal with mean (6) and variance (7). We adopt this approach.

We derive the prior density for $p(b_0, \Sigma^{-1})$ and then transform this into the density for $p(b_0, \Sigma)$ using the Jacobian. Let

$$l_0(b_0, \Sigma; D) = \log p(D|b_0, \Sigma, H_0, x_0) \tag{A.2}$$

denote the natural log of the conditional likelihood. Let $\zeta = [\sigma^{(11)}\ \sigma^{(12)}\ \sigma^{(22)}]^\top$, where $\sigma^{(ij)}$ denotes element (i, j) of Σ^{-1}. Applying (A.1) implies

$$p(b_0, \Sigma^{-1}|H_0) \propto \left| -E\begin{bmatrix} \frac{\partial^2 l_0}{\partial b_0 \partial b_0^\top} & \frac{\partial^2 l_0}{\partial b_0 \partial \zeta^\top} \\ \frac{\partial^2 l_0}{\partial \zeta \partial b_0^\top} & \frac{\partial^2 l_0}{\partial \zeta \partial \zeta^\top} \end{bmatrix} \right|^{1/2}. \tag{A.3}$$

The form of the conditional likelihood implies that

$$l_0(b_0, \Sigma; D) = -\frac{T}{2}\log|2\pi\,\Sigma| - \frac{1}{2}(z - Z_0 b_0)^\top \left(\Sigma^{-1} \otimes I_T\right)(z - Z_0 b_0). \tag{A.4}$$

It follows from (A.4) that

$$\frac{\partial l_0}{\partial b_0} = \frac{1}{2} Z_0^\top \left(\Sigma^{-1} \otimes I_T\right)(z - Z_0 b_0),$$

and

$$\begin{aligned} \frac{\partial^2 l_0}{\partial b_0 \partial b_0^\top} &= -\frac{1}{2} Z_0^\top \left(\Sigma^{-1} \otimes I_T\right) Z_0 \\ &= -\frac{1}{2}\begin{bmatrix} \iota_T^\top & 0 \\ 0 & X^\top \end{bmatrix}\left(\Sigma^{-1} \otimes I_T\right)\begin{bmatrix} \iota_T & 0 \\ 0 & X \end{bmatrix} \\ &= -\frac{1}{2}\begin{bmatrix} \sigma^{(11)} T & \sigma^{(12)} \iota^\top X \\ \sigma^{(12)} X^\top \iota & \sigma^{(22)} X^\top X \end{bmatrix}. \end{aligned} \tag{A.5}$$

Taking the expectation conditional on b_0 and Σ implies

$$E\left[\frac{\partial^2 l_0}{\partial b_0 \partial b_0^\top}\right] = -\frac{T}{2}\begin{bmatrix} \sigma^{(11)} & \sigma^{(12)}[1\ \mu_x] \\ \sigma^{(12)}\begin{bmatrix} 1 \\ \mu_x \end{bmatrix} & \sigma^{(22)}\begin{bmatrix} 1 & \mu_x \\ \mu_x & \sigma_x^2 + \mu_x^2 \end{bmatrix} \end{bmatrix}. \tag{A.6}$$

Using arguments in Stambaugh (1999), it can be shown that

$$E\left[\frac{\partial^2 l_0}{\partial b_0 \partial \zeta^\top}\right] = 0.$$

Moreover,

$$-\left|E\left(\frac{\partial^2 l_0}{\partial \zeta \partial \zeta^\top}\right)\right| = \left|\frac{\partial^2 \log|\Sigma|}{\partial \zeta \partial \zeta^\top}\right| = |\Sigma|^3$$

(see Box and Tiao (1973, pp. 474–475)). Therefore

$$p(b_0, \Sigma^{-1}|H_0) \propto |\Phi|^{\frac{1}{2}} |\Sigma|^{\frac{3}{2}} \tag{A.7}$$

where

$$\Phi = \begin{bmatrix} \Sigma^{-1} & \mu_x \begin{bmatrix} \sigma^{(12)} \\ \sigma^{(22)} \end{bmatrix} \\ \mu_x \left[\sigma^{(12)}\ \sigma^{(22)}\right] & \left(\sigma_x^2 + \mu_x^2\right)\sigma^{(22)} \end{bmatrix}.$$

This matrix Φ has the same determinant as $-E\left[\frac{\partial^2 l_0}{\partial b_0 \partial b_0^\top}\right]$ because 2 columns and 2 rows have been reversed.

From the formula for the determinant of a partitioned matrix, it follows that

$$|\Phi| = \left|\Sigma^{-1}\right| \left| \left(\sigma_x^2 + \mu_x^2\right)\sigma^{(22)} - \mu_x^2 \left[\sigma^{(12)}\ \sigma^{(22)}\right] \Sigma \begin{bmatrix} \sigma^{(12)} \\ \sigma^{(22)} \end{bmatrix} \right|.$$

Because

$$\Sigma \begin{bmatrix} \sigma^{(12)} \\ \sigma^{(22)} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

it follows that

$$\begin{aligned} |\Phi| &= \left|\Sigma^{-1}\right| \left| \left(\sigma_x^2 + \mu_x^2\right)\sigma^{(22)} - \mu_x^2 \sigma^{(22)} \right| \\ &= |\Sigma|^{-1} \sigma_x^2 \sigma^{(22)}. \end{aligned}$$

The determinant of Σ equals

$$|\Sigma| = \sigma_u^2\left(\sigma_v^2 - \sigma_{uv}^2 \sigma_u^{-2}\right),$$

while $\sigma^{(22)} = \left(\sigma_v^2 - \sigma_{uv}^2 \sigma_u^{-2}\right)^{-1}$. Therefore,

$$|\Phi| = |\Sigma|^{-2} \sigma_u^2 \sigma_x^2.$$

Substituting into (A.7),

$$p(b_0, \Sigma^{-1}|H_0) \propto |\Sigma|^{\frac{1}{2}} \sigma_u \sigma_x.$$

The Jacobian of the transformation from Σ^{-1} to Σ is $|\Sigma|^{-3}$. Therefore,

$$p(b_0, \Sigma|H_0) = |\Sigma|^{-\frac{5}{2}} \sigma_u \sigma_x.$$

Appendix B. Sampling from posterior distributions

This section describes how to sample from the posterior distributions for our benchmark and related models. In all cases, the sampling procedure for the posteriors under H_1 and H_0 involve the Metropolis–Hastings algorithm. Below we describe the case of the exact likelihood and full Bayes prior in detail. The procedures for the conditional likelihood and for the empirical Bayes prior are similar.

B.1. Posterior distribution under H_0

Substituting (8) and (17) into (18) implies that

$$p(b_0, \Sigma|H_0, D) \propto \sigma_u|\Sigma|^{-\frac{T+5}{2}} \exp\left\{-\frac{1}{2}\sigma_x^{-2}(x_0-\mu_x)^2 - \frac{1}{2}(z-Z_0b_0)^\top\left(\Sigma^{-1}\otimes I_T\right)(z-Z_0b_0)\right\}.$$

This posterior does not take the form of a standard density function because of the term in the likelihood involving x_0 (note that σ_x^2 is a nonlinear function of ρ and σ_v). However, we can sample from the posterior using the Metropolis–Hastings algorithm.

The Metropolis–Hastings algorithm is implemented "block-at-a-time", by repeatedly sampling from $p(\Sigma|b_0, H_0, D)$ and from $p(b_0|\Sigma, H_0D)$. To calculate a proposal density for Σ, note that

$$(z-Z_0b_0)^\top\left(\Sigma^{-1}\otimes I_T\right)(z-Z_0b_0) = \operatorname{tr}\left[(Y-XB_0)^\top(Y-XB_0)\Sigma^{-1}\right],$$

where

$$B_0 = \begin{bmatrix}\alpha & \theta\\ 0 & \rho\end{bmatrix}.$$

The proposal density for the conditional probability of Σ is the inverted Wishart with $T+2$ degrees of freedom and scale factor of $(Y-XB_0)^\top(Y-XB_0)$. The target is therefore

$$p(\Sigma|b_0, H_0, D) \propto \sigma_u \exp\left\{-\frac{1}{2}(x_0-\mu_x)^2\sigma_x^{-2}\right\} \times \text{proposal}.$$

Let

$$V_0 = \left(Z_0^\top\left(\Sigma^{-1}\otimes I_T\right)Z_0\right)^{-1}.$$

Let

$$\hat{b}_0 = V_0Z_0^\top\left(\Sigma^{-1}\otimes I_T\right)z.$$

It follows from completing the square that

$$(z-Z_0b_0)^\top\left(\Sigma^{-1}\otimes I_T\right)(z-Z_0b_0) = (b_0-\hat{b}_0)^\top V_0^{-1}(b_0-\hat{b}_0) + \text{ terms independent of } b_0.$$

The proposal density for b_0 is therefore multivariate normal with mean $\hat{b}_0$ and variance–covariance matrix V_0. The accept–reject algorithm of Chib and Greenberg (1995, Section 5) is used to sample from the target density, which is equal to

$$p(b_0|\Sigma, H_0, D) \propto \exp\left\{-\frac{1}{2}(x_0-\mu_x)^2\sigma_x^{-2}\right\} \times \text{proposal}.$$

Note that σ_u and Σ are in the constant of proportionality. Drawing successively from the conditional posteriors for Σ and b_0 produces a density that converges to the full posterior conditional on H_0.

B.2. Posterior distribution under H_1

Substituting (12) and (15) into (18) implies that

$$p(b_1, \Sigma|H_1, D) \propto \sigma_x|\Sigma|^{-\frac{T+5}{2}} \exp\left\{-\frac{1}{2}\beta^2\left(\sigma_\eta^2\sigma_x^{-2}\sigma_u^2\right)^{-2} - \frac{1}{2}\sigma_x^{-2}(x_0-\mu_x)^2\right\} \times \exp\left\{-\frac{1}{2}(z-Z_1b_1)^\top\left(\Sigma^{-1}\otimes I_T\right)(z-Z_1b_1)\right\}.$$

The sampling procedure is similar to that described in Appendix B.1. Details can be found in Wachter and Warusawitharana (2009). To summarize, we first draw from the posterior $p(\Sigma \mid b_1, H_1, D)$. The proposal density is an inverted Wishart with $T+2$ degrees of freedom and scale factor $(Y-XB_1)^\top(Y-XB_1)$, where

$$B_1 = \begin{bmatrix}\alpha & \theta\\ \beta & \rho\end{bmatrix}. \tag{B.1}$$

We then draw from $p(\theta, \rho \mid \alpha, \beta, \Sigma, H_1, D)$. The proposal density is multivariate normal with mean and variance determined by the conditional normal distribution. Finally, we draw from $p(\alpha, \beta \mid \theta, \rho, \Sigma, H_1, D)$. In this case, the target and the proposal are the same, and are also multivariate normal.

Appendix C. Computing the Bayes factor

This section describes computation of Bayes factors for the benchmark and related models. Verdinelli and Wasserman (1995) show

$$\mathcal{B}_{10}^{-1} = p(\beta=0|H_1, D) \times E\left[\frac{p(b_0, \Sigma|H_0)}{p(\beta=0, b_0, \Sigma|H_1)}\,\middle|\,\beta=0, H_1, D\right]. \tag{C.1}$$

To compute $p(\beta=0 \mid H_1, D)$, note that

$$p(\beta=0 \mid H_1, D) = \int p(\beta=0 \mid b_0, \Sigma, H_1, D) \times p(b_0, \Sigma \mid H_1, D)\, db_0\, d\Sigma. \tag{C.2}$$

As discussed in Appendix B.2, the posterior distribution of α and β conditional on the remaining parameters is normal. We can therefore compute $p(\beta=0 \mid b_0, \Sigma, H_1, D)$ in closed form by using the properties of the conditional normal distribution. Consider N draws from the full posterior: $((b_1^{(1)}, \Sigma^{(1)}), \ldots, (b_1^{(N)}, \Sigma^{(N)}))$, where we can write $(b_1^{(i)}, \Sigma^{(i)})$ as $(\beta^{(i)}, b_0^{(i)}, \Sigma^{(i)})$. We use these draws to integrate out over b_0 and Σ. It follows from (C.2) that

$$p(\beta=0|H_1, D) \approx \frac{1}{N}\sum_{i=1}^{N} p(\beta=0|b_0^{(i)}, \Sigma^{(i)}, H_1, D),$$

where the approximation is accurate for large N.

To compute the second term in (C.1), we observe that

$$\frac{p(b_0, \Sigma \mid H_0)}{p(\beta=0, b_0, \Sigma \mid H_1)} = \frac{p(b_0, \Sigma \mid H_0)}{p(\beta=0|b_0, \Sigma, H_1)p(b_0, \Sigma \mid H_1)} = \sqrt{2\pi}\sigma_\beta,$$

because $p(b_0, \Sigma \mid H_0) = p(b_0, \Sigma \mid H_1)$. Note that $\sigma_\beta = \sigma_\eta\sigma_x^{-1}\sigma_u$. We require the expectation taken with respect to the posterior distribution conditional on the existence of predictability and the realization $\beta = 0$. To calculate this expectation, we draw $((b_0^{(1)}, \Sigma^{(1)}), \ldots, (b_0^{(N)}, \Sigma^{(N)}))$ from $p(b_0, \Sigma \mid \beta = 0, H_1, D)$. This involves modifying the procedure for drawing from the posterior for b_1, Σ given H_1 (see Appendix B.2). We sample from $p(\Sigma \mid \alpha, \beta = 0, \theta, \rho, H_1, D)$, then from $p(\rho, \theta \mid \alpha, \beta = 0, \Sigma, H_1, D)$ and finally

from $p(\alpha \mid \beta = 0, \Sigma, \theta, \rho, H_1, D)$, and repeat until the desired number of draws are obtained. All steps except the last are identical to those described in Appendix B.2 (the value of β is identically zero rather than the value from the previous draw). For the last step we derive $p(\alpha \mid \beta = 0, \Sigma, \theta, \rho, H_1, D)$ from the joint distribution $p(\alpha, \beta \mid \Sigma, \theta, \rho, H_1, D)$, making use of the properties of the conditional normal distribution.

Given these draws from the posterior distribution, the second term equals

$$E\left[\frac{p(b_0, \Sigma|H_0)}{p(\beta = 0, b_0, \Sigma|H_1)} \,\middle|\, \beta = 0, H_1, D\right] \approx \frac{1}{N}\sum_{i=1}^{N} \sqrt{2\pi}\sigma_\eta (\sigma_x^{(i)})^{-1}\sigma_u^{(i)}. \tag{C.3}$$

Appendix D. The posterior distribution and Bayes factor for the conjugate g-prior and conditional likelihood

This section generalizes results in Zellner (1996) to the case of a multivariate regression system with an informative conjugate prior. We assume a multivariate version of the conjugate g-prior as follows:

$$b_1|\Sigma, H_1 \sim N(0, g^{-1}\left(\Sigma \otimes (X^\top X)^{-1}\right)), \tag{D.4}$$

$$\Sigma|H_1 \sim IW(S_0, N_0 - 2), \tag{D.5}$$

where g^{-1} is a scale parameter that determines the degree of precision of the prior, IW denotes the inverse-Wishart distribution, and N_0 and S_0 can be interpreted as the length of a hypothetical no-predictability prior sample and the sum of squared errors of this sample, respectively.[30]

Given B_1 defined as in (B.1), it follows from Zellner (1996, Eq. 8.14) that

$$p(B_1|\Sigma, H_1) \propto |\Sigma|^{-1} \exp\left\{-\frac{1}{2}\mathrm{tr}\left(gB_1^\top (X^\top X) B_1 \Sigma^{-1}\right)\right\}. \tag{D.6}$$

Note that the variance of b_1 equals $\Sigma \otimes (X^\top X)^{-1}$, and that

$$|\Sigma \otimes (X^\top X)^{-1}|^{-\frac{1}{2}} \propto |\Sigma|^{-1}$$

because $X^\top X$ can be regarded as a constant when calculating the distribution of B_1. Further, the density for the inverse Wishart distribution (D.5) equals

$$p(\Sigma|H_1) \propto |\Sigma|^{-(N_0+1)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(\Sigma^{-1}S_0)\right\}. \tag{D.7}$$

Therefore the joint prior is given by

$$p(B_1, \Sigma|H_1) = |2\pi\,\Sigma|^{-\frac{N_0+3}{2}} \times \exp\left\{-\frac{1}{2}\mathrm{tr}\left(gB_1^\top (X^\top X) B_1 \Sigma^{-1} + S_0\Sigma^{-1}\right)\right\}. \tag{D.8}$$

Note that this prior imposes a particular structure on the covariance matrix of the parameters that mimics the likelihood specification. It is this structure that is responsible for this specification's tractability. Note also that the data enter into the prior through the term $(X^\top X)^{-1}$, so that this prior requires incorrect conditioning. The entire time path of the state variable must be known when prior beliefs are formulated.

We combine this prior with the conditional likelihood function.[31] Let

$$\hat{B}_1 = (X^\top X)^{-1} X^\top Y \tag{D.9}$$

$$S = (Y - X\hat{B}_1)^\top (Y - X\hat{B}). \tag{D.10}$$

Given this notation, we can rewrite (14) as follows:

$$p(D|B_1, \Sigma, H_1) \propto |\Sigma|^{-\frac{T}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left((B_1 - \hat{B}_1)^\top X^\top \times X(B_1 - \hat{B}_1)\Sigma^{-1} + S\Sigma^{-1}\right)\right\}, \tag{D.11}$$

where $\propto$ in (D.11) should be taken to mean that we have eliminated multiplicative terms that do not depend on B_1 and Σ. For more detail, see Zellner (1996, Chapter 8.1).

Define sufficient statistics for the posterior as follows

$$\bar{B}_1 = \left(X^\top X(1+g)\right)^{-1}(X^\top Y)$$

$$\bar{S} = S_0 + Y^\top Y - (Y^\top X)(X^\top X(1+g))^{-1}(X^\top Y).$$

Note that $\bar{S}$ can be rewritten as

$$\bar{S} = S_0 + S + \hat{B}_1^\top X^\top X \hat{B}_1 - \bar{B}_1^\top (X^\top X)(1+g)\bar{B}_1. \tag{D.12}$$

Bayes rule implies that the posterior is given by

$$p(B_1, \Sigma|D, H_1) \propto p(D|B_1, \Sigma, H_1) p(B_1, \Sigma|H_1)$$

where the first and second terms on the right hand side are given by (D.11) and (D.8) respectively. Completing the square and using (D.12) implies that the posterior density equals

$$p(B_1, \Sigma|D, H_1) = |\Sigma|^{-\frac{T+N_0+3}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left((B_1 - \bar{B}_1)^\top \times (X^\top X(1+g))(B_1 - \bar{B}_1)\Sigma^{-1} + \bar{S}\Sigma^{-1}\right)\right\}. \tag{D.13}$$

We now factor the joint posterior (D.13) into a posterior for B_1 conditional on Σ and the marginal posterior for Σ. This is an important step in computing the Bayes factor, as will be apparent in what follows. By definition,

$$p(B_1, \Sigma|D, H_1) = p(B_1|\Sigma, D, H_1) p(\Sigma|D, H_1). \tag{D.14}$$

Define

$$\bar{b}_1 = \mathrm{vec}(\bar{B}_1).$$

The factorization in (D.14) is accomplished as follows:

$$\begin{aligned} p(B_1|\Sigma, D, H_1) &\propto |\Sigma|^{-1} \exp\left\{-\frac{1}{2}\mathrm{tr}\left((B_1 - \bar{B}_1)^\top \times (X^\top X(1+g))(B_1 - \bar{B}_1)\Sigma^{-1}\right)\right\}, \\ &= |\Sigma|^{-1} \exp\left\{-\frac{1}{2}(b_1 - \bar{b}_1)^\top \times \left(\Sigma^{-1} \otimes X^\top X(1+g)\right)(b_1 - \bar{b}_1)\right\} \end{aligned} \tag{D.15}$$

and

$$p(\Sigma|D, H_1) \propto |\Sigma|^{-\frac{T+N_0+1}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\bar{S}\Sigma^{-1}\right)\right\}. \tag{D.16}$$

The distribution (D.15) represents a multivariate normal distribution.[32]

[30] This interpretation is consistent with having a standard uninformative "prior" before viewing this no-predictability "prior sample" of $p(b_1) \propto$ constant and $p(\Sigma) \propto |\Sigma|^{-3/2}$. See Zellner (1996, Chapter 8.1). A prior sample of length greater than 2 is necessary for a well-defined posterior distribution, since the data also need to be sufficient to identify b_1.

[31] We cannot use the exact likelihood function because the prior does not lead to a well-defined distribution for the predictor variable.

[32] As in (D.6), for (D.15) to be multivariate normal, $|\Sigma|$ must be raised to the power -1.

Our ultimate goal is to calculate the marginal posterior for β, which is the second element of b_1. Let $\bar{\beta}$ be the second element of $\bar{b}_1$ and define

$$\bar{\nu}_x = \left[(X^\top X)^{-1}\right]_{22},$$

namely the second diagonal element of $(X^\top X)^{-1}$.[33] It follows from (D.15) and properties of the multivariate normal distribution that

$$p(\beta|\Sigma, D, H_1) \propto \frac{1}{\sigma_u} \exp\left\{-\frac{1}{2}\sigma_u^{-2}(1+g)\bar{\nu}_x(\beta-\bar{\beta})^2\right\}, \quad \text{(D.17)}$$

where we have used the fact that $(\Sigma^{-1} \otimes (X^\top X))^{-1} = \Sigma \otimes (X^\top X)^{-1}$.

Further, note that (D.17) depends only on σ_u. Therefore, to calculate the marginal prior for β, we only need to integrate out σ_u. It follows from (D.16) and properties of the inverse Wishart distribution that

$$p(\sigma_u^2|D, H_1) \propto \frac{1}{\sigma_u^{T+N_0-1}} \exp\left\{-\frac{\bar{S}_{11}}{2\sigma_u^2}\right\}, \quad \text{(D.18)}$$

where $\bar{S}_{11}$ is the first diagonal element of $\bar{S}$ (see Zellner (1996, p. 227–228)). It follows that

$$\begin{aligned} &p(\beta|D, H_1) \\ &= \int_0^\infty p(\beta|\sigma_u^2, D, H_1)p(\sigma_u^2|D, H_1)\, d\sigma_u^2 \\ &\propto \int_0^\infty \frac{1}{\sigma_u^{T+N_0}} \exp\left\{-\frac{1}{2\sigma_u^2}\left((1+g)\bar{\nu}_x(\beta-\bar{\beta})^2 + \bar{S}_{11}\right)\right\} d\sigma_u^2 \\ &\propto \left((1+g)\bar{\nu}_x(\beta-\bar{\beta})^2 + \bar{S}_{11}\right)^{-\frac{T+N_0-2}{2}} \quad \text{(D.19)} \\ &\propto \left(1 + \frac{1}{T+N_0-3}\right. \\ &\quad \left. \times \left(\frac{(1+g)\bar{\nu}_x(T+N_0-3)}{\bar{S}_{11}}\right)(\beta-\bar{\beta})^2\right)^{-\frac{T+N_0-2}{2}}. \quad \text{(D.20)} \end{aligned}$$

Therefore, β has a t-distribution with location parameter $\bar{\beta}$, scale parameter

$$((1+g)T\bar{\nu}_x(T+N_0-3))^{-1/2}\bar{S}_{11}^{1/2},$$

and $T+N_0-3$ degrees of freedom.

Under the condition

$$p(b_0, \Sigma|H_0) = p(b_0, \Sigma|\beta = 0, H_1), \quad \text{(D.21)}$$

the Bayes factor can be computed using the marginal prior and posterior distributions for β:

$$\mathcal{B}_{10} = \frac{p(\beta = 0|H_1)}{p(\beta = 0|D, H_1)} \quad \text{(D.22)}$$

(see Verdinelli and Wasserman, 1995). The value of $p(\beta = 0|D, H_1)$ can be computed based on (D.20) using the formula for the density of a t-distribution. We can perform the analogous calculation for the prior distribution to find

$$p(\beta|H_1) \propto \left(1 + \frac{1}{N_0-3}\left(\frac{g\bar{\nu}_x(N_0-3)}{S_{0,11}}\right)\beta^2\right)^{-\frac{N_0-2}{2}},$$

where $S_{0,11}$ is the first diagonal element of S_0. This is a central t distribution with scale parameter

$$(g\bar{\nu}_x(N_0-3))^{-1/2}S_{0,11}^{1/2},$$

and N_0-3 degrees of freedom.

[33] Note that this element is also equal to $T\hat{\sigma}_x^2$, namely T (the number of time series observations on the return variable) multiplied by the sample variance of the predictor taken from time 0 to time $T-1$.

Appendix E. The posterior distribution and Bayes factor for the conjugate g-prior when the regressor is strictly exogenous

When the regressor is strictly exogenous, it is correct to use only the return equation. With some abuse of notation, let $b = [\alpha, \beta]^\top$. The prior distribution takes the form

$$p(b|\sigma_u, H_1) \propto \frac{1}{\sigma_u^2} \exp\left\{-\frac{1}{2\sigma_u^2}b^\top(gX^\top X)b\right\} \quad \text{(E.23)}$$

$$p(\sigma_u^2|H_1) \propto \frac{1}{\sigma_u^{N_0}} \exp\left\{-\frac{1}{2}\sigma_u^{-2}s_0\right\} \quad \text{(E.24)}$$

where s_0 and N_0 are constants. We can rewrite this system in terms of familiar distributions:

$$b|\sigma_u, H_1 \sim N\left(0, g^{-1}\sigma_u^2(X^\top X)^{-1}\right), \quad \text{(E.25)}$$

$$\sigma_u^2|H_1 \sim IW(s_0, N_0-2). \quad \text{(E.26)}$$

As in the previous section, it is as if we have a "true" uninformative prior of $p(\sigma_u^2) \propto \sigma_u^{-2}$ and $p(b) \propto$ constant before seeing a "prior sample" with N_0 observations. Because σ_u^2 is scalar in this case, its distribution can also be characterized as an inverse-Gamma.

Define

$$\hat{b} = (X^\top X)^{-1}X^\top R \quad \text{(E.27)}$$

$$s = (R - X\hat{b})^\top(R - X\hat{b}). \quad \text{(E.28)}$$

Note that $s = S_{11}$ in the previous section. The likelihood function is

$$p(R|X, b, \sigma_u^2) \propto \sigma_u^{-T} \exp\left\{-\frac{1}{2\sigma_u^2}\left((b-\hat{b})^\top X^\top X(b-\hat{b}) + s\sigma_u^{-2}\right)\right\}$$

where, as in the previous section $\propto$ should be taken to mean that we have eliminated multiplicative terms that do not depend on b and σ_u.

Analogously to the previous section, define

$$\bar{b} = \left(X^\top X(1+g)\right)^{-1}(X^\top R)$$

and

$$\begin{aligned} \bar{s} &= s_0 + R^\top R - (Y^\top X)(X^\top X(1+g))^{-1}(X^\top R) \\ &= s_0 + s + \hat{b}^\top X^\top X\hat{b} - \bar{b}^\top(X^\top X)(1+g)\bar{b}. \quad \text{(E.29)} \end{aligned}$$

Note that if g is the same, $\bar{b}$ will equal the first column of $\bar{B}_1$, and $\bar{s}$ will equal $\bar{S}_{11}$ (assuming that $s_0 = S_{0,11}$). Completing the square and using (E.29) imply

$$\begin{aligned} &p(b, \sigma_u^2|R, X, H_1) \propto \sigma_u^{-(T+N_0+2)} \\ &\quad \times \exp\left\{-\frac{1}{2\sigma_u^2}\left((b-\bar{b})^\top(X^\top X(1+g))(b-\bar{b}) + \bar{s}\right)\right\}. \quad \text{(E.30)} \end{aligned}$$

The posterior for b conditional on σ_u is multivariate normal:

$$\begin{aligned} &p(b|\sigma_u, R, X, H_1) \\ &\quad \propto \frac{1}{\sigma_u^2} \exp\left\{-\frac{1}{2\sigma_u^2}(b-\bar{b})^\top X^\top X(1+g)(b-\bar{b})\right\} \quad \text{(E.31)} \end{aligned}$$

while the marginal distribution for σ_u^2 is an inverse-Wishart (or, in this case, inverse-Gamma):

$$p(\sigma_u^2|R, X, H_1) \propto \sigma_u^{-(T+N_0)} \exp\left\{-\frac{1}{2\sigma_u^2}\bar{s}\right\}.$$

It follows from (E.31) and properties of the multivariate normal distribution that the distribution for β (the second element of b) is given by

$$p(\beta|\sigma_u^2, R, X, H_1) \propto \sigma_u^{-1} \exp\left\{-\frac{1}{2\sigma_u^2}(1+g)\bar{\nu}_x(\beta-\bar{\beta})^2\right\}, \quad \text{(E.32)}$$

where $\bar{\beta}$ is the second element of $\bar{b}$. Finally, we compute

$$p(\beta|R,X,H_1) \propto \int_0^\infty p(\beta|\sigma_u^2,R,X,H_1)p(\sigma_u^2|R,X,H_1)\,d\sigma_u^2$$
$$\propto \int_0^\infty \frac{1}{\sigma_u^{T+N_0+1}} \exp\left\{-\frac{1}{2\sigma_u^2}\left((1+g)\bar{\nu}_x(\beta-\bar{\beta})^2+\bar{s}\right)\right\} d\sigma_u^2$$
$$\propto \left((1+g)\bar{\nu}_x(\beta-\bar{\beta})^2+\bar{s}\right)^{-\frac{T+N_0-1}{2}}.$$

Arguing by analogy with (D.20), we see that β has a t-distribution with location parameter $\bar{\beta}$, scale parameter

$$((1+g)\bar{\nu}_x(T+N_0-2))^{-1/2}\,\bar{s}^{1/2},$$

and $T+N_0-2$ degrees of freedom. The prior distribution for β will be a central t with scale parameter

$$(g\bar{\nu}_x(N_0-2))^{-1/2}\,s_0^{1/2},$$

and N_0-2 degrees of freedom. Bayes factors can then be computed using (D.21) and (D.22).

It is instructive to compare these results with those of Appendix D. The marginal prior and posterior for β is nearly the same in the one-equation setting as in the two-equation setting, except for the degrees of freedom in the t-distribution. There is an additional degree of freedom in the one-equation setting, corresponding to a t-distribution that is somewhat less fat-tailed. As Zellner (1996, Chapter 8.1) discusses, this change in the degrees of freedom arises because of the need to estimate an additional parameter in the two-equation case, namely the correlation between shocks to u and shocks to v. Because, in effect, the same data needs to work harder in the two-equation case, the distributions are more diffuse. Mathematically, the difference arises from the fact that the marginal distribution of σ_u^2 in (D.18) is not the same as the marginal distribution of σ_u^2 in the single-equation case. However, if the regressor is strictly exogenous, namely if u and v are assumed to be independent, the one-equation case and the two-equation case will yield identical Bayes factors, a manifestation of the general principle discussed in Section 2.5.

Appendix F. Bayes factors for the training sample approach

Bayes factors for the training sample approach (described in Section 3.7) can be computed as a special case of those in Appendix D. Define

$$\tilde{\nu}_x = \left[(\tilde{X}^\top\tilde{X})^{-1}\right]_{22}$$

where we use the notation of Section 3.7, namely variables with a tilde on top correspond to quantities computed over the training sample. Then the prior distribution for β can be computed using results for the *posterior* distribution calculated in Appendix D, for an uninformative prior ($g=0$, $N_0=0$), and with full-sample quantities replaced by their training sample counterparts. That is, (D.19) becomes

$$p(\beta|H_1) \propto \left(\tilde{\nu}_x(\beta-\tilde{\beta})^2+\tilde{S}_{11}\right)^{-\frac{\tilde{T}-2}{2}}, \tag{F.33}$$

where $\tilde{\beta}$ is the second element of $\tilde{b}_1$ and $\tilde{S}_{11}$ is the first diagonal element of $\tilde{S}$. Similarly, the posterior can be calculated in the same way (again, $g=0$ and $N_0=0$), keeping in mind that the full-sample quantities in this case are as in OLS regression. That is (D.19) becomes

$$p(\beta|D,H_1) \propto \left(\tilde{\nu}_x(\beta-\hat{\beta})^2+S_{11}\right)^{-\frac{T-2}{2}}. \tag{F.34}$$

The calculation of the Bayes factor of course requires the true prior and posterior densities of β at zero, not just these values up to a constant that does not depend on β. These densities can be calculated by observing, as in Appendix D, that (F.33) and (F.34) imply t-distributions, with known density functions.

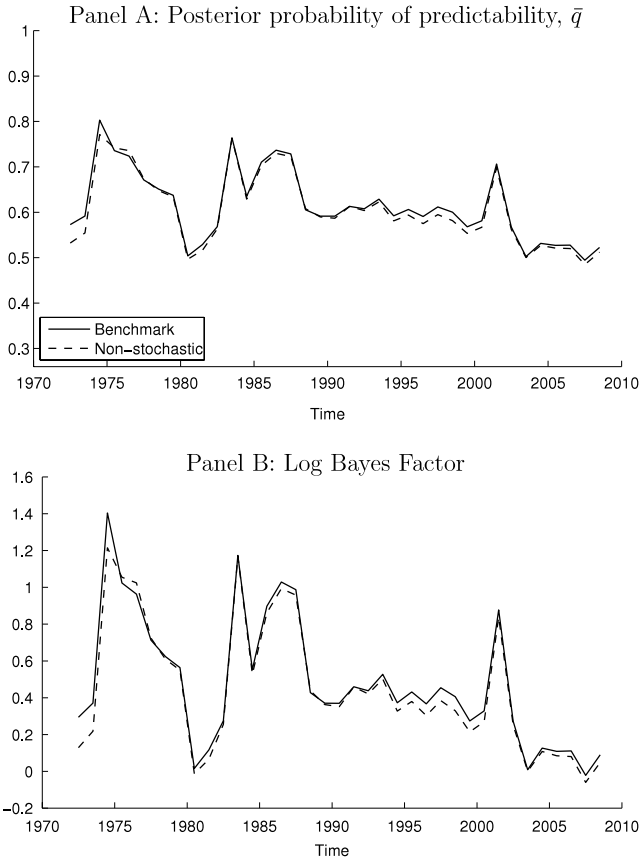

Fig. 9. The Bayes factor and posterior probability of return predictability for the yield spread. Notes: Panel A assumes the posterior probability of predictability and Panel B shows the log Bayes factor, assuming the predictive variable is the yield spread, namely the continuously-compounded yield on the five-year zero coupon bond less the continuously-compounded yield on the 3-month Treasury bill. The solid line shows results for the benchmark specification. The dashed line shows results assuming a non-stochastic regressor.

Appendix G. Results for the yield spread

In Fig. 9, we report results in which the predictor variable is the difference between the continuously-compounded 5-year zero-coupon bond yield and the yield on the 3-month Treasury Bill. Panel A shows that, while the yield spread had significant predictive power for returns in the early part of the sample, its power has been steadily declining. At the end of the sample, the posterior probability of predictability with the yield spread is about 50%, close to the prior. The yield spread has a lower autocorrelation than the dividend yield, and innovations to the yield spread have low correlation with innovations to returns. Both of these facts suggest that the non-stochastic and benchmark analyses would imply very similar results, which indeed they do.

References

Ando, Tomohiro, Zellner, Arnold, 2010. Hierarchical Bayesian analysis of the seemingly unrelated regression and simultaneous equations models using a combination of direct Monte Carlo and importance sampling techniques. Bayesian Anal. 5, 65–96.

Avramov, Doron, 2002. Stock return predictability and model uncertainty. J. Fianc. Econom. 64, 423–458.

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

Bartlett, M.S., 1957. Comment on 'a statistical paradox' by D.V. Lindley. Biometrika 44, 533–534.

Box, George E.P., Tiao, George C., 1973. Bayesian Inference in Statistical Analysis. Addison-Wesley Pub. Co., Reading, MA.
Brandt, Michael W., Goyal, Amit, Santa-Clara, Pedro, Stroud, Jonathan R., 2005. A simulation approach to dynamics portfolio choice with an application to learning about return predictability. Rev. Financial Studies 18, 831–873.
Campbell, John Y., 2008. Viewpoint: estimating the equity premium. Can. J. Econom. 41, 1–21.
Campbell, John Y., Shiller, Robert J., 1988. The dividend-price ratio and expectations of future dividends and discount factors. Rev. Financial Studies 1, 195–228.
Campbell, John Y., Viceira, Luis M., 1999. Consumption and portfolio decisions when expected returns are time-varying. Quart. J. Econom. 114, 433–495.
Campbell, John Y., Yogo, Motohiro, 2006. Efficient tests of stock return predictability. J. Financ. Econom. 81, 27–60.
Chen, Zengjing, Epstein, Larry, 2002. Ambiguity, risk and asset returns in continuous time. Econometrica 70, 1403–1443.
Chib, Siddhartha, Greenberg, Edward, 1995. Understanding the Metropolis–Hastings algorithm. Amer. Statist. 49, 327–335.
Chipman, Hugh, George, Edward I., McCulloch, Robert E., 2001. The practical implementation of Bayesian model selection, in: Lahiri P., (Ed.) Model Selection, IMS Lecture Notes, Bethesda, MA.
Cochrane, John H., 2008. The dog that did not bark: a defense of return predictability. The Rev. Financial Studies 21, 1533–1575.
Cremers, K.J. Martjin, 2002. Stock return predictability: a Bayesian model selection perspective. Rev. Financial Studies 15, 1223–1249.
Dickey, James M., 1971. The weighted likelihood ratio, linear hypotheses on normal location paramaters. Ann. Math. Statist. 42, 204–223.
Faust, Jon, Wright, Jonathan H., 2011. Efficient prediction of excess returns. Rev. Econ. Statist. 93, 647–659.
Fernandez, Carmen, Ley, Eduardo, Steel, Mark F.J., 2001. Benchmark priors for Bayesian model averaging. Journal of Econometrics 100, 381–427.
Goyal, Amit, Welch, Ivo, 2008. A comprehensive look at the empirical performance of equity premium prediction. Rev. Financial Studies 21, 1455–1508.
Hamilton, J.D., 1994. Time Series Analysis. Oxford University Press, Princeton, NJ.
Jeffreys, Harold, 1961. Theory of Probability. Oxford University Press, Clarenden.
Johannes, Michael S., Korteweg, Arthur G., Polson, Nicholas G., 2014. Sequential learning, predictability, and optimal portfolio returns. J. Finance 69, 611–644.
Johannes, Michael, Polson, Nicholas, 2006. MCMC methods for financial econometrics. In: Ait-Sahalia, Yacine, Hansen, Lars (Eds.), Handbook of Financial Econometrics. Elsevier, North-Holland.
Johannes, Michael, Polson, Nicholas, Stroud, Jonathan R., Sequential Optimal Portfolio Performance: Market and Volatility Timing, Working Paper, Columbia University, University of Chicago, and University of Pennsylvania, 2002.
Kandel, Shmuel, Stambaugh, Robert F., 1996. On the predictability of stock returns: an asset allocation perspective. J. Finance 51, 385–424.
Kass, R., Raftery, A.E., 1995. Bayes factors. J. Amer. Statist. Assoc. 90, 773–795.
Lewellen, Jonathan, 2004. Predicting returns with financial ratios. J. Financ. Econom. 74, 209–235.
Mehra, Rajnish, Prescott, Edward, 1985. The equity premium puzzle. J. Monetary Econ. 15, 145–161.
Pastor, Lubos, Stambaugh, Robert F., 2009. Predictive systems: living with imperfect predictors. J. Finance 64, 1583–1628.
Poirier, Dale J., 1978. The effect of the first observation in regression models with first-order autoregressive disturbances. J. R. Stat. Soc. Ser. C. Appl. Stat. 27, 67–68.
Shanken, Jay A., Tamayo, Ane, 2012. Payout yield, risk and mispricing, a Bayesian analysis. J. Financ. Econom. 105, 131–152.
Skoulakis, Georgios, Dynamic Portfolio Choice with Bayesian Learning, Working Paper, University of Maryland, 2007.
Stambaugh, Robert F., 1999. Predictive regressions. J. Financ. Econom. 54, 375–421.
Stock, James H., Watson, Mark W., 2012. Generalized shrinkage methods for forecasting using many predictors. J. Bus. Econom. Statist. 30, 481–493.
Van Binsbergen, Jules H., Koijen, Ralph S.J., 2010. Predictive regressions: a present-value approach. The Journal of Finance 65, 1439–1471.
Verdinelli, Isabella, Wasserman, Larry, 1995. Computing Bayes factors using a generalization of the Savage–Dickey density ratio. J. Amer. Statist. Assoc. 90, 614–618.
Wachter, Jessica A., 2010. Asset allocation. Ann. Rev. Fin. Econ. 2, 175–206.
Wachter, Jessica A., Warusawitharana, Missaka, 2009. Predictable returns and asset allocation: should a skeptical investor time the market? Journal of Econometrics 148, 162–178.
Wright, Jonathan H., 2008. Bayesian model averaging and exchange rate forecasts. Journal of Econometrics 146, 329–341.
Zellner, Arnold, 1996. An Introduction to Bayesian Inference in Econometrics. John Wiley and Sons, Inc., New York, NY.
Zellner, Arnold, 1986. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In: Goel, P.K., Zellner, A. (Eds.), Bayesian Inference and Decision Techniques: Essays in Honour of Bruno de Finetti. North-Holland, Amsterdam, The Netherlands.