University of Toronto
Department of Economics

Working Paper 293

How useful are historical data for forecasting the long-run equity return distribution?

By John M Maheu and Thomas H McCurdy

June 28, 2007

How useful are historical data for forecasting the long-run equity return distribution?

John M. Maheu and Thomas H. McCurdy*

This Draft April 2007

Abstract

We provide an approach to forecasting the long-run (unconditional) distribution of equity returns making optimal use of historical data in the presence of structural breaks. Our focus is on learning about breaks in real time and assessing their impact on out-of-sample density forecasts. Forecasts use a probability-weighted average of submodels, each of which is estimated over a different history of data. The paper illustrates the importance of uncertainty about structural breaks and the value of modeling higher-order moments of excess returns when forecasting the return distribution and its moments. The shape of the long-run distribution and the dynamics of the higher-order moments are quite different from those generated by forecasts which cannot capture structural breaks. The empirical results strongly reject ignoring structural change in favor of our forecasts which weight historical data to accommodate uncertainty about structural breaks. We also strongly reject the common practice of using a fixed-length moving window. These differences in long-run forecasts have implications for many financial decisions, particularly for risk management and long-run investment decisions.

key words: density forecasts, structural change, model risk, parameter uncertainty, Bayesian learning, market returns

*Maheu (jmaheu@chass.utoronto.ca), Department of Economics, University of Toronto and RCEA; McCurdy (tmccurdy@rotman.utoronto.ca), Joseph L. Rotman School of Management, University of Toronto, and Associated Fellow, CIRANO. We thank Bill Schwert for providing equity return data for the 1885-1926 period, and Greg Bauer, Rob Engle, David Goldreich, Stephen Gordon, Eric Jacquier, Mark Kamstra, Lisa Kramer, Jan Mahrt-Smith, Lubos Pastor, Nick Polson, Lukasz Pomorski, Jeroen Rombouts, Kevin Wang, Mike Veall, Benjamin Verschuere, as well as seminar participants at the CIREQ-CIRANO Financial Econometrics conference, the $(EC)^2$ conference Istanbul, the Northern Finance Association annual meetings, the Bank of Canada, HEC Montreal, McMaster University and York University for many helpful comments. Lois Chan provided excellent research assistance. We are also grateful to the SSHRC for financial support.

1 Introduction

Forecasts of the long-run distribution of excess returns are an important input into many financial decisions. For example, Barberis (2000) and Jacquier, Kane, and Marcus (2005) discuss the importance of accurate estimates for long-horizon portfolio choice. Our paper models and forecasts the long-run (unconditional) distribution of excess returns using a flexible parametric density in the presence of potential structural breaks. Our focus is on learning about breaks in real time and assessing their impact on out-of-sample density forecasts. We illustrate the importance of uncertainty about structural breaks and the value of modeling higher-order moments of excess returns when forecasting the return distribution and its moments. The shape of the long-run distribution and the dynamics of the higher-order moments are quite different from those generated by forecasts which cannot capture structural breaks. The empirical results strongly reject ignoring structural change in favor of our forecasts which weight historical data to accommodate uncertainty about structural breaks. We also strongly reject the common practice of using a fixed-length moving window. These differences in long-run forecasts have implications for many financial decisions, particularly for risk management and long-run investment decisions such as those by a pension fund manager.

Existing work on structural breaks with respect to market excess returns has focused on conditional return dynamics and the equity premium. Applications to the equity premium include Pastor and Stambaugh (2001) and Kim, Morley, and Nelson (2005) who provide smoothed estimates of the equity premium in the presence of structural breaks using a dynamic risk-return model. In this environment, model estimates are derived conditional on a maintained number of breaks in-sample. These papers focus on the posterior distribution of model parameters for estimating the equity premium.

Lettau and van Nieuwerburgh (2007) analyze the implications of structural breaks in the mean of the dividend price ratio for conditional return predictability; Viceira (1997) investigates shifts in the slope parameter associated with the log dividend yield. Paye and Timmermann (2006) and Rapach and Wohar (2006) present evidence of instability in models of predictable returns based on structural breaks in regression coefficients associated with several financial variables, including the lagged dividend yield, short interest rate, term spread and default premium.

Additional work on structural breaks in finance includes Pesaran and Timmermann (2002) who investigate window estimation in the presence of breaks, Pettenuzzo and Timmermann (2005) who analyze the effects of model instability on optimal asset allocation, Lettau, Ludvigson, and Wachter (2007) who focus on a regime change in macroeconomic risk, Andreou and Ghysels (2002) who analyze breaks in volatility dynamics, and Pesaran, Pettenuzzo, and Timmermann (2006b) who explore the effects of structural instability on pricing.

To our knowledge, none of the existing applications study the effects of structural

change on forecasts of the unconditional distribution of returns. An advantage to working with the long-run distribution is that it may be less susceptible to model misspecification than short-run conditional models. For example, an unconditional distribution of excess returns can be consistent with different underlying models of risk, allowing us to minimize model misspecification while focusing on the implications of structural change.

We postulate that the long-run or unconditional distribution of returns is generated by a discrete mixture of normals subject to occasional breaks that are governed by an i.i.d. Bernoulli distribution. This implies that the long-run distribution is time-varying and could be non-stationary. We assume that structural breaks partition the data into a sequence of stationary regimes each of which can be captured by a submodel which is indexed by its data history and associated parameter vector. New submodels are introduced periodically through time to allow for multiple structural breaks, and for potential breaks out of sample. The structural break model is constructed from a series of submodels. This approach is based on Maheu and Gordon (2007) extended to deal with multiple breaks out of sample. Short horizon forecasts are dominated by current posterior estimates from the data, since the probability of a break is low. However, long-horizon forecasts converge to predictions from a submodel using the prior density. In other words, in the long run we expect a break to occur and we only have our present prior beliefs on what those new parameters will be.

Our maintained *submodel* of excess returns is a discrete mixture of normals which can capture heteroskedasticity, asymmetry and fat tails. This is the parameterization of excess returns which is subject to structural breaks. For robustness, we compare our results using this flexible submodel specification to a Gaussian submodel specification to see if the more general distribution affects our inference about structural change or our real time forecasts. Flexible modeling of the submodel density is critical in order to avoid falsely identifying an outlier as a break.

Once we allow for structural breaks, it is not clear how useful historical data are for parameter estimation and for out-of-sample density forecasts. Pesaran and Timmermann (2007) and Pastor and Stambaugh (2001) discuss the use of both pre and post-break data. In our case, each submodel identifies a possible break point and is estimated from an associated history of data.

Since structural breaks can never be identified with certainty, submodel averaging provides a predictive distribution, which accounts for past and future structural breaks, by integrating over each of the possible submodels weighted by their probabilities. Individual submodels only receive significant weight if their predictive performance warrants it. We learn in real time about past structural breaks and their effect on the distribution of excess returns. The model average combines the past (potentially biased) data from before the estimated break point, which will tend to have less uncertainty about the distribution due to sample length, with the less precise (but unbiased) estimates based on the more recent post-break data. If a break occurred at 2000 but the submodel in-

troduced in 1990, which uses data from 1990 onward for parameter estimation, provides better predictions, then the latter submodel will receive relatively larger weight. As more data arrive, we would expect the predictions associated with the submodel introduced in 2000 to improve and thus gain a larger weight in prediction. In this sense the model average automatically picks submodels at each point in time based on predictive content. This approach provides a method to combine submodels estimated over different histories of data. Since the predictive density of returns integrates over the submodel distribution, submodel uncertainty (uncertainty about structural breaks) is accounted for in the analysis.

Our empirical results strongly reject ignoring structural change in favor of forecasts which weight historical data to accommodate uncertainty about structural breaks. We also strongly reject the common practice of using a fixed-length moving window. Ignoring structural breaks leads to inferior density forecasts. So does using a fixed-length moving window.

Structural change has implications for the entire shape of the long-run excess return distribution. The preferred structural change model produces kurtosis values well above 3 and negative skewness throughout the sample. Furthermore, the shape of the long-run distribution and the dynamics of the higher-order moments are quite different from those generated by forecasts which cannot capture structural breaks. Ignoring structural change results in misspecification of the long-run distribution of excess returns which can have serious implications, not only for the location of the distribution (the expected long-run premium), but also for risk assessments.

Our evidence clearly supports using a mixture-of-normals submodel with two components over a single-component (Gaussian) submodel. There is another important difference between the alternative parameterizations of the submodel. We show that our discrete mixture-of-normals submodel specification is more robust to *false breaks*. To see this, suppose one assumed a normal distribution for excess returns when in fact the data generating process has fat tails. In this case, realizations in the tail of the maintained normal distribution could be mistakenly interpreted in real time as evidence of a structural break. That is, as we learn about the distribution governing excess returns, sometimes we infer a break that is later revised to be an outlier and not a structural break. The richer specification of the two-component submodel is more robust to these *false breaks*. One reason for this is that the two-component submodel is characterized by a high and low variance state. This allows for heteroskedasticity in excess returns. Therefore, outliers can occur and not be evidence of a break in the distribution of excess returns.

One by-product of our results is inference about probable dates of structural breaks associated with the distribution of market equity excess returns. Using the discrete mixture-of-normals submodel parameterization, we identify breaks in 1929, 1934, 1940, and 1969, as well as possible breaks in the mid-1970s, the early 1990s and sometime

from 1998 through the end of the sample. Note that these breaks are detected in real time and are not the result of a full-sample analysis. For example, using only data up to 1931:04 there is strong evidence (probability over 0.75) that the most recent structural break detectable at that time occurred in 1929. From 1991 forward, however, there is considerable submodel uncertainty with several submodels receiving significant probability weight. Since our model average combines forecasts from the individual submodels, our objective is not to identify specific dates of structural breaks but rather to integrate out break points to produce superior forecasts.

Although our focus is on the distribution of excess returns, we also explore the implications of structural breaks for the predictive distribution of the equity premium. We find that ignoring structural breaks results in substantially different premium forecasts, as well as overconfidence in those forecasts. When a structural break occurs there is a decrease in the precision of the premium forecast which improves as we learn about the new premium level. Uncertainty about the premium comes from two sources: submodel uncertainty and parameter uncertainty. For example, our results show that the uncertainty after the break in 1929 is mainly due to parameter uncertainty, whereas the uncertainty in the late 1990s is from both submodel and parameter uncertainty.

The structural change model produces good density and point forecasts and illustrates the importance of modeling higher-order moments of excess returns. We investigate short (1 month) to long horizon (20 years) forecasts of cumulative excess returns. The structural break model, which accounts for multiple structural breaks, produces superior out-of-sample forecasts of the mean and the variance. These differences will be important for long-run investment and risk management decisions.

The paper is organized as follows. The next section describes the data sources. Section 3 introduces a flexible discrete mixture-of-normals model for excess returns as our submodel parameterization. Section 4 reviews Bayesian estimation techniques for the mixture submodel of excess returns. The proposed method for estimation and forecasting in the presence of structural breaks is outlined in Section 5. Results are reported in Section 6; and conclusions are found in Section 7.

2 Data

The equity data are monthly returns, including dividend distributions, on a well diversified market portfolio. The monthly equity returns for 1885:2 to 1925:12 were obtained from Bill Schwert; details of the data construction can be found in Schwert (1990). Monthly equity returns from 1926:1 to 2003:12 are from the Center for Research in Security Prices (CRSP) value-weighted portfolio, which includes securities on the New York stock exchange, American stock exchange and the NASDAQ. The returns were converted to continuously compounded monthly returns by taking the natural logarithm of the gross monthly return.

Data on the risk-free rate from 1885:2 to 1925:12 were obtained from annual interest rates supplied by Jeremy Siegel. Siegel (1992) describes the construction of the data in detail. Those annual interest rates were converted to monthly continuously compounded rates. Interest rates from 1926:1 to 2003:12 are from the U.S. 3 month T-bill rates supplied by the Fama-Bliss riskfree rate file provided by CRSP.

Finally, the monthly excess return, r_t, is defined as the monthly continuously compounded portfolio return minus the monthly riskfree rate. This monthly excess return is scaled by multiplying by 12. Table 1 reports summary statistics for the scaled monthly excess returns. Both the skewness and kurtosis estimates suggest significant deviations from the normal distribution.

3 Mixture-of-Normals Submodel for Excess Returns

In this section we outline our maintained model of excess returns which is subject to structural breaks. We label this the submodel, and provide more details on this definition in the next section. Financial returns are well known to display skewness and kurtosis and our inferences about forecasts and structural breaks may be sensitive to these characteristics of the shape of the distribution. Our maintained submodel of excess returns is a discrete mixture of normals. Discrete mixtures are a very flexible method to capture various degrees of asymmetry and tail thickness. Indeed a sufficient number of components can approximate arbitrary distributions (Roeder and Wasserman (1997)).

The k-component mixture submodel of excess returns is represented as

$$r_t = \begin{cases} N(\mu_1, \sigma_1^2) & \text{with probability } \pi_1 \\ \vdots & \vdots \\ N(\mu_k, \sigma_k^2) & \text{with probability } \pi_k, \end{cases} \tag{3.1}$$

with $\sum_{j=1}^{k} \pi_j = 1$. It will be convenient to denote each mean and variance as μ_j, and σ_j^2, with $j \in \{1, 2, ..., k\}$. Data from this specification are generated as: first a component j is chosen according to the probabilities $\pi_1, ..., \pi_k$; then a return is generated from $N(\mu_j, \sigma_j^2)$. Note that returns will display heteroskedasticity. Often a two-component specification is sufficient to capture the features of returns. Relative to the normal distribution, distributions with just two components can exhibit fat-tails, skewness and combinations of skewness and fat-tails. We do not use this mixture specification to capture structural breaks, but rather as a flexible method of capturing features of the unconditional distribution of excess returns which is our submodel that is subject to structural breaks.

Since our focus is on the moments of excess returns, it will be useful to consider the implied moments of excess returns as a function of the submodel parameters. The relationships between the uncentered moments and the submodel parameters for a k-

component submodel are:

$$\gamma = Er_t = \sum_{i=1}^{k} \mu_i \pi_i, \tag{3.2}$$

in which γ is defined as the equity premium; and

$$\gamma_2^{'} = Er_t^2 = \sum_{i=1}^{k} (\mu_i^2 + \sigma_i^2)\pi_i \tag{3.3}$$

$$\gamma_3^{'} = Er_t^3 = \sum_{i=1}^{k} (\mu_i^3 + 3\mu_i\sigma_i^2)\pi_i \tag{3.4}$$

$$\gamma_4^{'} = Er_t^4 = \sum_{i=1}^{k} (\mu_i^4 + 6\mu_i^2\sigma_i^2 + 3\sigma_i^4)\pi_i. \tag{3.5}$$

for the higher-order moments of returns. The higher-order centered moments $\gamma_j = E[(r_t - E(r_t))^j]$, $j = 2, 3, 4$, are then

$$\gamma_2 = \gamma_2^{'} - (\gamma)^2 \tag{3.6}$$

$$\gamma_3 = \gamma_3^{'} - 3\gamma\gamma_2^{'} + 2(\gamma)^3 \tag{3.7}$$

$$\gamma_4 = \gamma_4^{'} - 4\gamma\gamma_3^{'} + 6(\gamma)^2\gamma_2^{'} - 3(\gamma)^4. \tag{3.8}$$

As a special case, a one-component submodel allows for normally-distributed returns. Only two components are needed to produce skewness and excess kurtosis. If $\mu_1 = \cdots = \mu_k = 0$ and at least one variance parameter differs from the others the resulting density will have excess kurtosis but not asymmetry. To produce asymmetry and hence skewness we need $\mu_i \neq \mu_j$ for some $i \neq j$. Section 4 discusses a Bayesian approach to estimation of this submodel.

4 Estimation of the Submodels

In the next two subsections we discuss Bayesian estimation methods for the discrete mixture-of-normals submodels. This is the parameterization that is subject to structural breaks, as modeled in 5 below. An important special case for the submodel specification is when there is a single component, $k = 1$, which we discuss first.

4.1 Gaussian Case, $k = 1$

When there is only one component our submodel for excess returns reduces to a normal distribution with mean μ, variance σ^2, and likelihood function,

$$p(r|\mu, \sigma^2) = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(r_t - \mu)^2\right) \tag{4.1}$$

where $r = [r_1, ..., r_T]'$. In the last section, this model is included as a special case when $\pi_1 = 1$.

Bayesian methods require specification of a prior distribution over the parameters μ and σ^2. Given the independent priors $\mu \sim N(b, B)I_{\mu>0}$, and $\sigma^2 \sim IG(v/2, s/2)$, where $IG(\cdot, \cdot)$ denotes the inverse gamma distribution, Bayes rule gives the posterior distribution of μ and σ^2 as

$$p(\mu, \sigma^2|r) \propto p(r|\mu, \sigma^2)p(\mu)p(\sigma^2) \tag{4.2}$$

where $p(\mu)$ and $p(\sigma^2)$ denote the probability density functions of the priors. Note that the indicator function $I_{\mu>0}$ is 1 when $\mu > 0$ is true and otherwise 0. This restriction enforces a positive equity premium as indicated by theory.

Although closed form solutions for the posterior distribution are not available, we can use Gibbs sampling to simulate from the posterior and estimate quantities of interest. The Gibbs sampler iterates sampling from the following conditional distributions which forms a Markov chain.

1. sample $\mu \sim p(\mu|\sigma^2, r)$
2. sample $\sigma^2 \sim p(\sigma^2|\mu, r)$

In the above, we reject any draw that does not satisfy $\mu > 0$. These steps are repeated many times and an initial set of the draws are discarded to minimize startup conditions and ensure the remaining sequence of the draws is from the converged chain. See Chib (2001), Geweke (1997), and Robert and Casella (1999) for background information on Markov Chain Monte Carlo methods of which Gibbs sampling is a special case; and see Johannes and Polson (2005) for a survey of financial applications. After obtaining a set of N draws $\{\mu^{(i)}, (\sigma^2)^{(i)}\}_{i=1}^{N}$ from the posterior, we can estimate moments using sample averages. For example, the posterior mean of μ, which is an estimate of the equity premium conditional on this submodel and data, can be estimated as

$$E[\mu|r_T] \approx \frac{1}{N}\sum_{i=1}^{N} \mu^{(i)}. \tag{4.3}$$

To measure the dispersion of the posterior distribution of the equity premium we could compute the posterior standard deviation of μ in an analogous fashion, using sample

averages obtained from the Gibbs sampler in $\sqrt{E[\mu^2|r] - E[\mu|r]^2}$. Alternatively, we could summarize the marginal distribution of the equity premium with a histogram or kernel density estimate.

This simple submodel which assumes excess returns follow a Gaussian distribution cannot account for the asymmetry and fat tails found in return data. Modeling these features of returns may be important to our inference about structural change and consequent forecasts. The next section provides details on estimation for submodels with two or more components which can capture the higher-order moments of excess returns.

4.2 Mixture Case, $k > 1$

In the case of $k > 1$ mixture of normals, the likelihood of excess returns is

$$p(r|\mu, \sigma^2, \pi) = \prod_{t=1}^{T} \sum_{j=1}^{k} \pi_j \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{1}{2\sigma_j^2}(r_t - \mu_j)^2\right) \tag{4.4}$$

where $\mu = [\mu_1, ..., \mu_k]'$, $\sigma^2 = [\sigma_1^2, ..., \sigma_k^2]'$, and $\pi = [\pi_1, ..., \pi_k]$. Bayesian estimation of mixtures has been extensively discussed in the literature and our approach closely follows Diebolt and Robert (1994). We choose conditionally conjugate prior distributions which facilitate our Gibbs sampling approach. The independent priors are $\mu_i \sim N(b_i, B_{ii})$, $\sigma_i^2 \sim IG(v_i/2, s_i/2)$, and $\pi \sim \mathcal{D}(\alpha_1, ..., \alpha_k)$, where the latter is the Dirichlet distribution. We continue to impose a positive equity premium by giving zero support to any parameter configuration that violates $\gamma > 0$.

Discrete mixture models can be viewed as a simpler model if an indicator variable z_t records which observations come from component j. Our approach to Bayesian estimation of this submodel begins with the specification of a prior distribution and the augmentation of the parameter vector by the additional indicator $z_t = [0 \cdots 1 \cdots 0]$ which is a row vector of zeros with a single 1 in the position j if r_t is drawn from component j. Let Z be the matrix that stacks the rows of z_t, $t = 1, ..., T$.

With the full data r_t, z_t the data density becomes

$$p(r|\mu, \sigma^2, \pi, Z) = \prod_{t=1}^{T} \sum_{j=1}^{k} z_{t,j} \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{1}{2\sigma_j^2}(r_t - \mu_j)^2\right). \tag{4.5}$$

Bayes theorem now gives the posterior distributions as

$$\begin{aligned} p(\mu, \sigma^2, \pi, Z|r) &\propto p(r|\mu, \sigma^2, \pi, Z)p(\mu, \sigma^2, \pi, Z) && (4.6) \\ &\propto p(r|\mu, \sigma^2, \pi, Z)p(Z|\mu, \sigma^2, \pi)p(\mu, \sigma^2, \pi). && (4.7) \end{aligned}$$

The posterior distribution has an unknown form, however, we can generate a sequence

of draws from this density using Gibbs sampling. Just as in the $k = 1$ case, we sample from a set of conditional distributions and collect a large number of draws. From this set of draws we can obtain simulation-consistent estimates of posterior moments. The Gibbs sampling routine repeats the following steps for posterior simulation.

1. sample $\mu_i \sim p(\mu|\sigma^2, \pi, Z, r), i = 1, ..., k$
2. sample $\sigma_i^2 \sim p(\sigma_i^2|\mu, \pi, Z, r),\ i = 1, ..., k$
3. sample $\pi \sim p(\pi|\mu, \sigma^2, Z, r)$
4. sample $z_t \sim p(z_t|\mu, \sigma^2, \pi, r),\ t = 1, ..., T.$

Step 1–4 are repeated many times and an initial set of the draws are discarded to minimize startup conditions and ensure the remaining sequence of the draws is from the converged chain. Our appendix provides details concerning computations involved for each of the Gibbs sampling steps.

5 Modeling Structural Breaks

In this section we outline a method to deal with potential structural breaks. Our approach is based on Maheu and Gordon (2007). We extend it to deal with multiple breaks out of sample. Recent work on forecasting in the presence of model instability includes Clark and McCracken (2006) and Pesaran and Timmermann (2007). Recent Bayesian approaches to modeling structural breaks include Koop and Potter (2007), Giordani and Kohn (2007) and Pesaran, Pettenuzzo, and Timmermann (2006a). An advantage of our approach is that we can use existing standard Gibbs sampling techniques and Bayesian model averaging ideas (Avramov (2002), Cremers (2002), Wright (2003), Koop (2003), Eklund and Karlsson (2005)). As such, Gibbs sampling for discrete mixture models can be used directly without any modification. As we discuss in Section 5.3, submodel parameter estimation is separated from estimation of the process governing breaks. Estimation of the break process has submodel parameter uncertainty integrated out, making it a low dimensional tractable problem. Finally, our approach delivers a marginal likelihood estimate that integrates over all structural breaks and allows for direct model comparison with Bayes factors.

5.1 Submodel Structure

Intuitively, if a structural break occurred in the past we would want to adjust our use of the old data in our estimation procedure since those data can bias our estimates and forecasts. We assume that structural breaks are exogenous unpredictable events that result in a change in the parameter vector associated with the maintained submodel, in

this case a discrete mixture-of-normals submodel of excess returns. In this approach we view each structural break as a unique one-time event.

The structural break model is constructed from a series of identical parameterizations (mixture of normals, number of components k fixed) that we label *submodels*. What differentiates the submodels is the history of data that is used to form the posterior density of the parameter vector θ. (Recall that for the $k = 2$ submodel specification, $\theta = \{\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi_1, \pi_2\}$.) As a result, θ will have a different posterior density for each submodel, and a different predictive density for excess returns. Each of the individual submodels assume that once a break occurs, past data are not useful in learning about the new parameter value, only future data can be used to update beliefs. As more data arrives, the posterior density associated with the parameters of each submodel are updated. Our real time approach incorporates the probability of out-of-sample breaks. Therefore, new submodels are continually introduced through time. Structural breaks are identified by the probability distribution on submodels.

Submodels are differentiated by when they start and the number of data points they use. Since structural breaks can never be identified with certainty, submodel averaging provides a predictive distribution, which accounts for past and future structural breaks, by integrating over each of the possible submodels weighted by their probabilities. New submodels only receive significant weights once their predictive performance warrants it. The model average optimally combines the past (potentially biased) data from before the estimated break point, which will tend to have less parameter uncertainty due to sample length, with the less precise (but unbiased) estimates based on the more recent post-break data. This approach provides a method to combine submodels estimated over different histories of data.

To begin, define the information set $I_{a,b} = \{r_a, ..., r_b\}$, $a \leq b$, with $I_{a,b} = \{\emptyset\}$, for $a > b$, and for convenience let $I_t \equiv I_{1,t}$. Let M_i be a submodel that assumes a structural break occurs at time i. The exception to this is the first submodel of the sample M_1 for which there is no prior data. As we have mentioned, under our assumptions the data $r_1, ..., r_{i-1}$ are not informative about parameters for submodel M_i due to the assumption of a structural break at time i, while the subsequent data $r_i, ..., r_{t-1}$ are informative. If θ denotes the parameter vector, then $p(r_t|\theta, I_{i,t-1}, M_i)$ is the conditional data density associated with submodel M_i, given θ, and the information set $I_{i,t-1}$.

Now consider the situation where we have data up to time $t-1$ and we want to consider forecasting out-of-sample r_t. A first step is to construct the posterior density for each of the possible submodels. If $p(\theta|M_i)$ is the prior distribution for the parameter vector θ of submodel M_i, then the posterior density of θ for submodel M_i, based on the information $I_{i,t-1}$, has the form,

$$p(\theta|I_{i,t-1}, M_i) \propto \begin{cases} p(r_i, ..., r_{t-1}|\theta, M_i)p(\theta|M_i) & i < t \\ p(\theta|M_i) & i = t, \end{cases} \tag{5.1}$$

$i = 1, ..., t$. For $i < t$, only data after the assumed break at time i are used, that is, from i to $t-1$. For $i = t$, past data are not useful at all since a break is assumed to occur at time t, and therefore the posterior becomes the prior. Thus, at time $t-1$ we have a set of submodels $\{M_i\}_{i=1}^t$, which use different numbers of data points to produce predictive densities for r_t. For example, given $\{r_1, ..., r_{t-1}\}$, M_1 assumes no breaks in the sample and uses all the data $r_1, ..., r_{t-1}$ for estimation and prediction; M_2 assumes a break at $t = 2$ and uses $r_2, ..., r_{t-1}$; ....; M_{t-1}, assumes a break at $t-1$ and uses r_{t-1}; and finally M_t assumes a break at t and uses no data. That is, M_t assumes a break occurs out-of-sample, in which case, past data are not useful.

In the usual way, the predictive density for r_t associated with submodel M_i is formed by integrating out the parameter uncertainty,

$$p(r_t|I_{i,t-1}, M_i) = \int p(r_t|I_{i,t-1}, \theta, M_i)p(\theta|I_{i,t-1}, M_i)d\theta, \quad i = 1, ..., t. \tag{5.2}$$

For M_t the posterior is the prior under our assumptions. Estimation of the predictive density is discussed in Section 5.6.

5.2 Combining Submodels

As noted in section 1, our structural break model must learn about breaks in real time and combine submodel predictive densities. The usual Bayesian methods of model comparison and combination are based on the marginal likelihood of a common set of data which is not the case in our setting since the submodels $\{M_i\}_{i=1}^t$ are based on different histories of data. Therefore, we require a new mechanism to combine submodels. We consider two possibilities in this paper. First, that the probability of a structural break is determined only from subjective beliefs. For example, financial theory or non-sample information may be useful in forming these beliefs. Our second approach is to propose a stochastic process for the arrival of breaks and estimate the parameter associated with that arrival process. We discuss the first approach in this subsection; in the next subsection we deal with our second approach which requires estimation of the break process.

Before observing r_t the financial analyst places a subjective prior $0 \leq \lambda_t \leq 1$, that a structural break occurs at time t. A value of $\lambda_t = 0$ assumes no break at time t, and therefore submodel M_t is not introduced. This now provides a mechanism to combine the submodels. Let $\Lambda_t = \{\lambda_2, ..., \lambda_t\}$. Note that $\Lambda_1 = \{\emptyset\}$ since we do not allow for a structural break at $t = 1$.

To develop some intuition, we consider the construction of the structural break model for the purpose of forecasting, starting from a position of no data at $t = 0$. If we wish to forecast r_1, all we have is a prior on θ. In this case, we can obtain the predictive density for r_1 as $p(r_1|I_0) = p(r_1|I_0, M_1)$ which can be computed from priors using (5.2). After observing r_1, $p(M_1|I_1, \Lambda_1) = p(M_1|I_1) = 1$ since there is only 1 submodel at this point.

Now allowing for a break at $t = 2$, that is, $\lambda_2 \neq 0$, the predictive density for r_2 is the mixture

$$p(r_2|I_1, \Lambda_2) = p(r_2|I_{1,1}, M_1)p(M_1|I_1, \Lambda_1)(1 - \lambda_2) + p(r_2|I_{2,1}, M_2)\lambda_2.$$

The first term on the RHS is the predictive density using all the available data times the probability of no break. The second term is the predictive density derived from the prior assuming a break, times the probability of a break. Recall that in the second density $I_{2,1} = \{\emptyset\}$. After observing r_2 we can update the submodel probabilities,

$$\begin{aligned} p(M_1|I_2, \Lambda_2) &= \frac{p(r_2|I_{1,1}, M_1)p(M_1|I_1, \Lambda_1)(1 - \lambda_2)}{p(r_2|I_1, \Lambda_2)} \\ p(M_2|I_2, \Lambda_2) &= \frac{p(r_2|I_{2,1}, M_2)\lambda_2}{p(r_2|I_1, \Lambda_2)}. \end{aligned}$$

Now we require a predictive distribution for r_3 given past information. Again, allowing for a break at time $t = 3$, $\lambda_3 \neq 0$, the predictive density is formed as

$$p(r_3|I_2, \Lambda_3) = \left[p(r_3|I_{1,2}, M_1)p(M_1|I_2, \Lambda_2) + p(r_3|I_{2,2}, M_2)p(M_2|I_2, \Lambda_2)\right](1 - \lambda_3) + p(r_3|I_{3,2}, M_3)\lambda_3.$$

In words, this is (predictive density assuming no break at $t = 3$)$\times$(probability of no break at $t = 3$) + (predictive density assuming a break at $t = 3$)$\times$(probability of a break at $t = 3$). Once again $p(r_3|I_{3,2}, M_3)$ is derived from the prior. The updated submodel probabilities are

$$\begin{aligned} p(M_1|I_3, \Lambda_3) &= \frac{p(r_3|I_{1,2}, M_1)p(M_1|I_2, \Lambda_2)(1 - \lambda_3)}{p(r_3|I_2, \Lambda_3)} && (5.3) \\ p(M_2|I_3, \Lambda_3) &= \frac{p(r_3|I_{2,2}, M_2)p(M_2|I_2, \Lambda_2)(1 - \lambda_3)}{p(r_3|I_2, \Lambda_3)} && (5.4) \\ p(M_3|I_3, \Lambda_3) &= \frac{p(r_3|I_{3,2}, M_3)\lambda_3}{p(r_3|I_2, \Lambda_3)}. && (5.5) \end{aligned}$$

In this fashion we sequentially build up the predictive distribution of the break model. As a further example of our model averaging structure, consider Figure 1 which displays a set of submodels available at $t = 10$, where the horizontal lines indicate the data used in forming the posterior for each submodel. The forecasts from each of these submodels, which use different data, are combined (the vertical line) using the submodel probabilities. Since at period $t = 10$, there are no data available for period 11, the point M_{11} on Figure 1 represents the prior density in the event of a structural break at $t = 11$. If there has been a structural break at say $t = 5$, then as new data arrive, M_5 will receive more weight as we learn about the regime change.

Intuitively, the posterior and predictive density of recent submodels after a break will change quickly as new data arrive. Once their predictions warrant it, they receive larger

weights in the model average. Conversely, posteriors of old submodels will only change slowly when a structural break occurs. Their predictions will still be dominated by the longer and older data before the structural break. Note that our inference automatically uses past data prior to the break if predictions are improved. For example, if a break occurred at $t = 2000$ but the submodel M_{1990}, which uses data from $t = 1990$ onward for parameter estimation, provides better predictions, then the latter submodel will receive relatively larger weight. As more data arrive, we would expect the predictions associated with submodel M_{2000} to improve and thus gain a larger weight in prediction. In this sense the model average automatically picks submodels at each point in time based on predictive content.

Given this discussion, and a prior on breaks, the general predictive density for r_t, for $t > 1$, can be computed as the model average

$$p(r_t|I_{t-1}, \Lambda_t) = \left[\sum_{i=1}^{t-1} p(r_t|I_{i,t-1}, M_i)p(M_i|I_{t-1}, \Lambda_{t-1})\right](1 - \lambda_t) + p(r_t|I_{t,t-1}, M_t)\lambda_t. \quad (5.6)$$

The first term on the RHS of (5.6) is the predictive density from all past submodels that assume a break occurs prior to time t. The second term is the contribution assuming a break occurs at time t. In the latter, past data are not useful and only the prior density is used to form the predictive distribution. The terms $p(M_i|I_{t-1}, \Lambda_{t-1})$, $i = 1, ..., t-1$ are the submodel probabilities, representing the probability of a break at time i given information I_{t-1}, and are updated each period after observing r_t as

$$p(M_i|I_t, \Lambda_t) = \begin{cases} \frac{p(r_t|I_{i,t-1}, M_i)p(M_i|I_{t-1}, \Lambda_{t-1})(1-\lambda_t)}{p(r_t|I_{t-1}, \Lambda_t)} & 1 \leq i < t \\ \frac{p(r_t|I_{t,t-1}, M_t)\lambda_t}{p(r_t|I_{t-1}, \Lambda_t)} & i = t. \end{cases} \quad (5.7)$$

In addition to being inputs into (5.6) and other calculations below, the submodel probabilities also provide a distribution at each point in time of the *most recent structural break* inferred from the current data. Recall that submodels are indexed by their starting point. Therefore, if submodel $M_{t'}$ receives a high posterior weight given I_t with $t > t'$, this is evidence of the most recent structural break at t'.

Posterior estimates and submodel probabilities must be built up sequentially from $t = 1$ and updated as new data become available. At any given time, the posterior mean of some function of the parameters, $g(\theta)$, accounting for past structural breaks can be computed as,

$$E[g(\theta)|I_t, \Lambda_t] = \sum_{i=1}^{t} E[g(\theta)|I_{i,t}, M_i]p(M_i|I_t, \Lambda_t). \quad (5.8)$$

This is an average at time t of the submodel-specific posterior expectations of $g(\theta)$, weighted by the appropriate submodel probabilities. Submodels that receive large posterior probabilities will dominate this calculation.

Similarly, to compute an out-of-sample forecast of $g(r_{t+1})$ we include all the previous t submodels plus an additional submodel which conditions on a break occurring out-of-sample at time $t+1$ assuming $\lambda_{t+1} \neq 0$. The predictive mean of $g(r_{t+1})$ is

$$E[g(r_{t+1})|I_t, \Lambda_{t+1}] = \sum_{i=1}^{t} E[g(r_{t+1})|I_{i,t}, M_i]p(M_i|I_t, \Lambda_t)(1-\lambda_{t+1}) \\ +E[g(r_{t+1})|I_{t+1,t}, M_{t+1}]\lambda_{t+1}. \tag{5.9}$$

Note that the predictive mean from the last term is based only on the prior as past data before $t+1$ are not useful in updating beliefs about θ given a break at time $t+1$.

5.3 Estimation of the Probability of a Break

We now specify the process governing breaks and discuss how to estimate it. We assume that the arrival of breaks is i.i.d. Bernoulli with parameter λ. With this additional structure, and given a prior $p(\lambda)$, we can update beliefs given sample data. From a computational perspective an important feature of this approach is that the break process can be separated from the submodel estimation. The posterior of the submodel parameters (5.1) is independent of λ. Furthermore, the posterior for λ is a function of the submodel predictive likelihoods, which have parameter uncertainty integrated out. Therefore, the likelihood is a function of only 1 parameter, so the posterior for λ is

$$p(\lambda|I_{t-1}) \propto p(\lambda)\prod_{j=1}^{t-1} p(r_j|I_{j-1}, \lambda) \tag{5.10}$$

where $p(r_j|I_{j-1}, \lambda)$ is from (5.6) with $\Lambda_j = \{\lambda_2, ..., \lambda_j\} = \{\lambda, ..., \lambda\}$ which we denote as λ henceforth. To sample from this posterior we use a Metropolis-Hastings routine with a random walk proposal. Given $\lambda = \lambda^{(i)}$, the most recent draw from the Markov chain, a new proposal is formed as $\lambda' = \lambda + e$ where e is a symmetric density. This is accepted, $\lambda^{(i+1)} = \lambda'$, with probability $\min\{\frac{p(\lambda'|I_{t-1})}{p(\lambda|I_{t-1})}, 1\}$ and otherwise rejected, $\lambda^{(i+1)} = \lambda^{(i)}$.

After dropping a suitable burn-in sample, we treat the remaining draws $\{\lambda^{(i)}\}_{i=1}^{N}$ as a sample from the posterior. A simulation-consistent estimate of the predictive likelihood of the break model is

$$p(r_t|I_{t-1}) = \int p(r_t|I_{t-1}, \lambda)p(\lambda|I_{t-1})d\lambda \tag{5.11}$$

$$\approx \frac{1}{N}\sum_{i=1}^{N} p(r_t|I_{t-1}, \lambda^{(i)}). \tag{5.12}$$

Posterior moments, as in (5.8), must have λ integrated out as in

$$E[g(\theta)|I_t] = E_\lambda E[g(\theta)|I_t, \lambda] = \sum_{i=1}^{t} E[g(\theta)|I_{i,t}, M_i] E_\lambda[p(M_i|I_t, \lambda)], \tag{5.13}$$

where $E_\lambda[\cdot]$ denotes expectation with respect to $p(\lambda|I_t)$. Recall that the submodel posterior density is independent of λ. It is now clear that the submodel probabilities after integrating out λ are $E_\lambda[p(M_i|I_t, \lambda)]$ which could be denoted as $p(M_i|I_t)$.

5.4 Forecasts

To compute an out-of-sample forecast of some function of r_{t+1}, $g(r_{t+1})$, we include all the previous t submodels plus an additional submodel which conditions on a break occurring out-of-sample at time $t+1$. The predictive density is derived from substituting (5.6) into the right-hand side of (5.11). Moments of this density are the basis of out-of-sample forecasts. The predictive mean of $g(r_{t+1})$, as in (5.9), after integrating out λ is

$$\begin{aligned} E[g(r_{t+1})|I_t] &= E_\lambda E[g(r_{t+1})|I_t, \lambda] & (5.14)\\ &= \sum_{i=1}^{t} E[g(r_{t+1})|I_{i,t}, M_i] E_\lambda[p(M_i|I_t, \lambda)(1-\lambda)] & (5.15)\\ &\quad + E[g(r_{t+1})|I_{t+1,t}, M_{t+1}] E_\lambda[\lambda]. \end{aligned}$$

$E[g(r_{t+1})|I_{i,t}, M_i]$ is an expectation with respect to a submodel predictive density and is independent of λ. $E_\lambda[\cdot]$ denotes an expectation with respect to $p(\lambda|I_t)$. These additional terms are easily estimated with $E_\lambda[p(M_i|I_t, \lambda)(1-\lambda)] \approx \frac{1}{N}\sum_{i=1}^{N} p(M_i|I_t, \lambda^{(i)})(1-\lambda^{(i)})$, and $E_\lambda[\lambda] \approx \frac{1}{N}\sum_{i=1}^{N} \lambda^{(i)}$.

Multiperiod forecasts are computed in the same way,

$$\begin{aligned} E[g(r_{t+2})|I_t] &= \sum_{i=1}^{t} E[g(r_{t+2})|I_{i,t}, M_i] E_\lambda[p(M_i|I_t, \lambda)(1-\lambda)^2] & (5.16)\\ &\quad + E[g(r_{t+2})|I_{t+1,t}, M_{t+1}] E_\lambda[\lambda(1-\lambda)] + E[g(r_{t+2})|I_{t+2,t}, M_{t+2}] E_\lambda[\lambda] \end{aligned}$$

which allows for a break at time $t+1$ and $t+2$. Note that the last two expectations with respect to returns in (5.16) are identical and derived from the prior. Grouping them together gives the term $E[g(r_{t+2})|I_{t+1,t}, M_{t+1}] E_\lambda[\lambda(1+(1-\lambda))]$. Following this, the $h-$period expectation is

$$\begin{aligned} E[g(r_{t+h})|I_t] = \sum_{i=1}^{t} E[g(r_{t+h})|I_{i,t}, M_i] E_\lambda[p(M_i|I_t, \lambda)(1-\lambda)^h] & (5.17)\\ + E[g(r_{t+h})|I_{t+1,t}, M_{t+1}] E_\lambda[\lambda \sum_{j=0}^{h-1} (1-\lambda)^j]. \end{aligned}$$

As $h \to \infty$ the weight on the prior forecast $E[g(r_{t+1})|I_{t+1,t}, M_{t+1}]$ goes to 1, and the weight from the submodels that use past data goes to 0. In essence, this captures the idea that in the short-run we may be confident in our current knowledge of the return distribution; but in the long-run we expect a break to occur, in which case the only information we have is our prior beliefs.

5.5 Predictive Distribution of the Equity Premium

Although the focus of this paper is on the predictive long-run distribution of excess returns, the 1st moment of this density is the long-run equity premium. There is an extensive literature that uses this unconditional premium. Much of this literature uses a simple point estimate of the premium obtained as the sample average from a long series of excess return data. For example, Table 1 in a recent survey by Mehra and Prescott (2003) lists four estimates of the equity premium using sample averages of data from 1802-1998, 1871-1999, 1889-2000, and 1926-2000. In addition, many forecasters, including those using dynamic models with many predictors, report the sample average of excess returns as a benchmark. For example, models of the premium conditional: on earnings or dividend growth include Donaldson, Kamstra, and Kramer (2006) and Fama and French (2002); on macro variables, Lettau and Ludvigson (2001); and on regime changes Mayfield (2004) and Turner, Startz, and Nelson (1989). Other examples of premium forecasts include Campbell and Thompson (2005), and Goyal and Welch (2007). In this subsection, we explore the implications for the predictive distribution of the unconditional equity premium of our approach to forecasting the long-run distribution of excess returns in the presence of possible structural breaks.

The predictive mean of the equity premium can be computed using the results in the previous section by setting $g(r_{t+1}) = r_{t+1}$. Note, however, that we are interested in the entire predictive distribution for the premium, for example, to assess the uncertainty about the equity premium forecasts. Using the discrete mixture-of-normals specification as our submodel with k fixed, the equity premium is $\gamma = \sum_{i=1}^{k} \mu_i \pi_i$. Given I_{t-1} we can compute the posterior distribution of the premium as well as the predictive distribution. It is important to note that even though our mixture-of-normals submodel is not dynamic, allowing for a structural break at t differentiates the posterior and predictive distribution of the premium. Therefore, since we are concerned with forecasting the premium, we report features of the *predictive distribution of the premium* for period t, given I_{t-1}, defined as,

$$p(\gamma|I_{t-1}) = \sum_{i=1}^{t-1} p(\gamma|I_{i,t-1}, M_i) E_\lambda[p(M_i|I_{t-1}, \lambda)(1-\lambda)] + p(\gamma|I_{t,t-1}, M_t) E_\lambda[\lambda]. \quad (5.18)$$

This equation is analogous to the predictive density of returns (5.11).

From the Gibbs sampling output for each of the submodels, and the posterior of λ,

we can compute the mean of the predictive distribution of the equity premium as,

$$E[\gamma|I_{t-1}] = \sum_{i=1}^{t-1} E[\gamma|I_{i,t-1}, M_i]E_\lambda[p(M_i|I_{t-1},\lambda)(1-\lambda)] + E[\gamma|I_{t,t-1}, M_t]E_\lambda[\lambda]. \quad (5.19)$$

Note that this is the same as (5.15) when $g(r_{t+1})$ is set to r_{t+1} in the latter. In a similar fashion, the standard deviation of the predictive distribution of the premium can be computed from $\sqrt{E[\gamma^2|I_{t-1}] - (E[\gamma|I_{t-1}])^2}$. This provides a measure of uncertainty about the premium.

In Section 6.5 below, we provide results for alternative forecasts of the equity premium. $\hat{\gamma}_{A,t-1}$ uses all available data weighted equally (submodel M_1) and thus assumes no structural breaks occur, $\hat{\gamma}_{W,t-1}$ is analogous to the no-break forecast in that it weights past data equally but uses a fixed-length (10 years of monthly data) moving window of past data rather than all available data, and $\hat{\gamma}_{B,t-1}$ uses all available data optimally after accounting for structural breaks. These forecasts are

$$\hat{\gamma}_{A,t-1} = E[\gamma|I_{t-1}, M_1] \quad (5.20)$$

$$\hat{\gamma}_{W,t-1} = E[\gamma|I_{t-1}, M_{t-120}] \quad (5.21)$$

$$\hat{\gamma}_{B,t-1} = E[\gamma|I_{t-1}]. \quad (5.22)$$

Recall that the $\hat{\gamma}_B$ forecasts integrate out all submodel uncertainty surrounding structural breaks using (5.19).

5.6 Implementation of the Structural Break Model

Estimation of each submodel at each point in time follows the Gibbs sampler detailed in Section 4. After dropping the first 500 draws of the Gibbs sampler, we collect the next 5000 which are used to estimate various posterior quantities. We also require the predictive likelihood to compute the submodel probabilities (5.7) to form an out-of-sample forecast, for example, using (5.15). To calculate the marginal likelihood of a submodel, following Geweke (1995) we use a predictive likelihood decomposition,

$$p(r_i, ..., r_t|M_i) = \prod_{j=i}^{t} p(r_j|I_{i,j-1}, M_i). \quad (5.23)$$

Given a set of draws from the posterior distribution $\{\theta^{(s)}\}_{s=1}^N$, where $\theta^{(s)} = \{\mu_1, ..., \mu_k, \sigma_1^2, ..., \sigma_k^2, \pi_1, ..., \pi_k\}$, for submodel M_i, conditional on $I_{i,j-1}$, each of the individual terms in (5.23) can be estimated consistently as

$$p(r_t|I_{i,j-1}, M_i) \approx \frac{1}{N}\sum_{s=1}^{N} p(r_t|\theta^{(s)}, I_{i,j-1}, M_i). \quad (5.24)$$

This is calculated at the end of each Gibbs run, along with features of the predictive density. Note that (5.24) enters directly into the calculation of (5.7). For the discrete mixture-of-normals specification, the data density is,

$$p(r_t|\theta^{(s)}, I_{i,t-1}, M_i) = \sum_{j=1}^{k} \pi_j \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{1}{2\sigma_j^2}(r_t - \mu_j)^2\right). \tag{5.25}$$

The predictive likelihood of submodel M_i is used in (5.7) to update the submodel probabilities at each point in time, and to compute the individual components $p(r_j|I_{j-1})$ of the structural break model through (5.11) and hence the marginal likelihood of the structural break model as,

$$p(r_1, ..., r_t) = \prod_{j=1}^{t} p(r_j|I_{j-1}). \tag{5.26}$$

5.7 Model Comparison

Finally, the Bayesian approach allows for the comparison and ranking of models by Bayes factors or posterior odds. Both of these require calculation of the marginal likelihood. The Bayes factor for model B versus model A is defined as $BF_{B,A} = p(r|B)/p(r|A)$, where $p(r|B)$ is the marginal likelihood for model B and similarily for model A. A Bayes factor greater than one is evidence that the data favor B. Kass and Raftery (1995) summarize the support for model B from the Bayes factor as: 1 to 3 not worth more than a bare mention, 3 to 20 positive, 20 to 150 strong, and greater than 150 as very strong.

5.8 Selecting Priors

An advantage of Bayesian methods is that it is possible to introduce prior information into the analysis. This is particularly useful in our context as finance practitioners and academics have strong beliefs regarding the distribution of excess returns and particularly its mean. Theory indicates that this equity premium must be positive and, from the wide range of estimates surveyed by Derrig and Orr (2004), the vast majority of the reported estimates are well below 10%. The average survey response from U.S. Chief Financial Officers for recent years is below 5% (Graham and Harvey (2005)). It is also well known that the distribution of returns displays skewness and kurtosis.

There are several issues involved in selecting priors when forecasting in the presence of structural breaks. Our model of structural breaks requires a proper predictive density for each submodel. This is satisfied if our prior $p(\theta|M_i)$ is proper. Some of the submodels condition on very little data. For instance, at time $t-1$ submodel M_t uses no data and has a posterior equal to the prior. There are also problems with using highly diffuse priors, as it may take many observations for the predictive density of a new submodel

to receive any posterior support. In other words, the rate of learning about structural breaks is affected by the priors. Based on this, we use informative proper priors.

A second issue is the elicitation of priors in the mixture submodel. While it is straightforward for the one-component case, it is not obvious how priors on the component parameters affect features of the excess return distribution when $k > 1$. For two or more components, the likelihood of the mixture submodel is unbounded which make noninformative priors inappropriate (Koop (2003)).

In order to select informative priors based on features of excess returns, we conduct a *prior predictive* check on the submodel (Geweke (2005)). That is, we analyze moments of excess returns simulated from the submodel. We repeat the following steps

1. draw $\theta \sim p(\theta)$ from the prior distribution
2. simulate $\{\tilde{r}_t\}_{t=1}^T$ from $p(r_t|I_{t-1}, \theta)$
3. using $\{\tilde{r}_t\}_{t=1}^T$ calculate the mean, variance, skewness and kurtosis

Table 2 reports these summary statistics after repeating the steps 1–3 many times using the priors listed in the footnote of Table 3. The prior can account for a range of empirically realistic sample statistics of excess returns. The 95% density region of the sample mean is approximately $[0, 0.1]$. The two-component submodel with this prior is also consistent with a wide range of skewness and excess kurtosis. In selecting a prior for the single-component submodel we tried to match, as far as possible, the features of the two-component submodel. All prior specifications enforce a positive equity premium.

Although it is possible to have different priors for each submodel, we use the same calibrated prior for all submodels in our analysis. Our main results estimate λ and use the prior $\lambda \sim Beta(0.05, 20)$. This favors infrequent breaks and allows the structural break model to learn when breaks occur. We could introduce a new submodel for every observation but this would be computationally expensive. Instead, we restrict the number of submodels to one every year of data. Our first submodel starts in February 1885. Thereafter, new submodels are introduced in February of each year until 1914, after which new submodels are introduced in June of each year due to the missing 4 months of data in 1914 (see Schwert (1990) for details). Therefore, our benchmark prior introduces a new submodel every 12 months with $\lambda_t = \lambda$; otherwise $\lambda_t = 0$. We discuss other results for different specifications in Section 6.7.

6 Results

This section discusses the real-time, out-of-sample, forecasts starting from the first observation to the last. First, we report the alternative model specifications, priors, and results as measured by the marginal likelihoods. The preferred specification is the structural break model with λ estimated and using a $k = 2$ submodel, which we focus on for

the remainder of the paper. Then we summarize the results for submodel probabilities from which we can infer probable structural break points and evaluate submodel uncertainty, as well as compute an *ex post* measure of mean useful historical observations. The next subsection summarizes the dynamics of higher-order moments of the excess return distribution implied by our preferred model. This is followed by results for the predictive distribution for the equity premium when structural breaks are allowed versus not. We then present an assessment of multi-period out-of-sample mean and variance forecasts generated by the structural break versus no-break models. Finally, we present results from a robustness analysis.

6.1 Model Specification and Density Forecasts

A summary of the model specifications, including priors, is reported in Table 3. The first panel of this table reports results using the Gaussian submodel specification ($k = 1$); whereas the second panel results refer to the case with the more flexible two-component ($k = 2$) mixture-of-normals specification for submodels. In each panel we report results for the no-break model which uses all historical data weighted equally, a no-break model which uses a 10-year moving window of equally-weighted historical data, and our structural change models that combine submodels in a way that allows for breaks. We report results for several alternative parameterizations of the structural change model depending on how often we introduce new submodels (one versus five years) and whether or not we estimate the probability of structural breaks, or leave it at a fixed value.

Table 3 also records the logarithm of the marginal likelihood values, log(ML), for each of the models based on our full sample of historical observations. Recall that this summarizes the period-by-period forecast densities evaluated at the realized data points. That is, it is equal to the sum of the log predictive likelihoods over the sample. This is the relevant measure of out-of-sample predictive content of a model (Geweke and Whiteman (2006)). According to the criterion summarized in Section 5.7, there is overwhelming evidence in favor of allowing for structural breaks. Based on the log(ML) values reported in Table 3, the Bayes factor for the break model against the no-break alternative is around exp(167) for the one-component submodel specification. Even with the more flexible two-component submodel specification, the Bayes factor comparing the model that allows a structural break every year versus the no-break alternative is a very large number, $\exp(-1191.77 + 1241.09) = \exp(49.32)$. Therefore, we find *very* strong evidence for structural breaks, regardless of the specification of the submodels ($k = 1$ versus $k = 2$).

Note that in each case, the best structural break model is the one that allows a break every year. Figure 2 plots the posterior mean for estimates of λ over the entire sample. The ex ante probability of a break is higher throughout the sample for the less flexible $k = 1$ submodel parameterization. For example, at the end of the sample, the estimated

λ is 0.131 ($k = 1$) versus 0.106 for the $k = 2$ submodel parameterization. This indicates that the less flexible $k = 1$ specification finds more breaks.

Note that using the two-component ($k = 2$ mixture-of-normals) specification for submodels always results in log(ML) values that are significantly higher than using the Gaussian submodel specification ($k = 1$). These results provide very strong support for the two-component submodel specification. Therefore, for the remainder of the paper, we will focus on results for that more flexible submodel specification with λ estimated from the data.

In Figure 3 we illustrate the rejection of the no-break forecasts by plotting, at each point in time, the difference in the cumulative predictive likelihood from the break model versus the no-break alternative. Up to 1930 there was no significant difference. There is a large difference after 1930 but also smaller on-going improvements in the performance of the break model versus the no-break alternative until the end of the sample.

At various points above we mentioned the common practice of using a fixed-length moving window of historical data to reduce the impact of potential structural changes on forecasts. Table 3 reports that our structural change models, which optimally weight historical data, very strongly reject a 10-year moving window of equally-weighted historical data. The Bayes factor is $\exp(-1204.17 + 1281.94) = \exp(77.77)$ using a $k = 1$ submodel specification, and $\exp(29)$ using a $k = 2$ submodel specification.

6.2 Submodel Probabilities: Inferred Structural Breaks

The probability associated with submodel M_i at time t can be interpreted as the probability that there was a break point at date i given data up to time t. The 3-dimensional plots in Figures 4 to 6 illustrate these probabilities over some selected time periods for all available submodels. In these plots, the axis labelled Submodel M_i refers to the submodels identified by their starting observation i. The probability associated with a particular submodel at a point in time can be seen as a perpendicular line from the Time axis. As examples, we plot the submodel probabilities over time for some specific submodels in Figure 7. These time-series plots of selected submodel probabilities, correspond to a perpendicular slice through the submodel axis over time for that particular submodel in the 3-dimensional plots (Figures 4 to 6).

Recall that the number of submodels is increasing with time; a new submodel is introduced every 12 months. The submodel probability distribution is the cross-section of the available submodels at a particular point in time. Figures 8 and 9 illustrate the distribution of submodel probabilities at particular points in time, in this case the start of 1960 and at the end of the sample, respectively.

Submodel probabilites are displayed, for the $k = 2$ case, for three different subperiods in the top panel of Figure 4, and in Figures 5 and 6 respectively. Comparisons with the $k = 1$ case (Gaussian submodel specification) are provided by contrasting the top and

bottom panels of Figure 4 for the subperiod 1885-1910; and in Figure 10 which plots the probability for the 1893 submodel in the $k = 1$ versus $k = 2$ case. This plot illustrates the danger of falsely identifying a break if the submodel specfication is not flexible enough.

As shown in the top panels of Figure 4 and Figure 7, for the first 45 years of the sample the first submodel, M_{1885}, receives most of the probability. There was some preliminary evidence of a possible break in 1893. For example, starting in 1894:1 the 1893 submodel gets a probability weight of 0.45 but it drops the following month to 0.12 with the 1885 submodel returning to a weight of 0.85, although 1893 still gets greater than 0.10 weight until 1902:9. Thus learning as new data arrive can play an important role in revising previous beliefs regarding possible structural breaks. Recall that these probability assessments are based on data available in real time. As such, they represent the inference available to financial analysts at the time.

To illustrate the importance of a flexible parameterization of the submodel for the unconditional distribution of excess returns, consider the time-series of probability for the M_{1893} submodel when we use the Gaussian ($k = 1$) submodel specification. As shown in Figure 10, for the $k = 1$ case the probability of a break in 1893:2 shoots up from 0.003 in 1893:6 to 0.91 by 1893:8. However, by the start of 1903 the probability assigned to submodel M_{1893} has fallen to less than 0.10. whereas the M_{1895} submodel is again assigned the majority of the probability weight. Using a Gaussian submodel specification, that doesn't allow the unconditional distribution of excess returns to have fat tails and/or skewness, can lead to outliers being identified, in real time, as breaks. This inference is later revised as more data becomes available. However, as described above and displayed in Figure 10, our flexible ($k = 2$ mixture-of-normals) parameterization of the submodel is less susceptible to this problem of temporarily identifying false breaks in real time. This example underscores the importance of accurately modeling financial returns prior to an analysis of structural breaks. In other words, misspecified models may provide evidence of structural breaks when the underlying DGP is stable.

The first submodel of the sample, M_{1885}, continues to receive most of the support until 1929. There is very strong evidence of a structural break in 1929. By 1930:10, the M_{1929} submodel has a probability weight of greater than 0.50 and 0.76 by 1931:4, which indicates fast learning about the change in the distribution of excess returns. As discussed further below, the identified break in the excess return distribution in 1929 is primarily due to higher-order moments such as volatility (see Figure 12). However, the break has implications for the predictive distribution for the long-run equity premium, as well as higher-order moments of excess returns.

There is an increase in submodel uncertainty during the 1930s. From 1935 to mid-1943, the 1934 submodel receives some weight, as high as 0.70 by 1937:3. However, this break is short-lived, the next major break occurs in 1940. As shown in the 3rd panel of Figure 7, the M_{1940} submodel receives the most probability weight (in excess of 0.50) until 1970.

In the early 1970s there is evidence of a break in 1969. The M_{1940} submodel lost its position of having the most probability weight for the first time in 1970:04 when the M_{1969} submodel is assigned a weight of 0.62 as opposed to 0.16 for the M_{1940} submodel. However, during the first half of the 1970s there was considerable submodel uncertainty. For example, by 1976:6 the probability weight is almost equally shared by the M_{1969}, M_{1973} and M_{1974} submodels, afterwhich the 1969 and 1975 submodels share the significant probability weight until the early 1990s.

Finally, there is submodel uncertainty again from 1991 to the end of the sample. The probability of a break during this period is about 0.50 with the highest probability assigned to the M_{1991}, M_{1992}, and eventually the M_{1998} submodels. By the end of the sample M_{1999}, M_{2000}, and M_{2003} also receive significant weight. This submodel uncertainty can be seen in the bottom right-hand corner of Figure 6 and, more comprehensively, in Figure 9 which illustrates the entire distribution of submodel probabilites at the end of 2003. Figure 8 shows that usually the structural change model is quite decisive in allocating weight to a particular submodel. This plot shows the submodel probability distribution at time 1960 which assigns most of the weight to the 1940 submodel. However, Figure 9 conveys the submodel uncertainty at the end of the sample. We do not have enough data yet to infer the exact date of recent structural breaks in the distribution of excess returns. However, it does not matter for our real-time forecasts since we use all of the information, appropriately weighted, and integrate out that submodel uncertainty.

In summary, we find evidence for breaks in 1929, 1934, 1940, and 1969, as well as possible breaks in the mid-1970s, the early 1990s and sometime from 1998 through the end of the sample. Our results highlight several important points. First, the identification of structural breaks depends on the data used, and false assessments may occur which are later revised when more data become available. This is an important aspect of learning about structural breaks in real time. Second, our evidence of submodel uncertainty indicates the problem with using only one submodel for any particular forecast. In a setting of submodel uncertainty, the optimal approach is to use the probability-weighted submodel average which integrates out the submodel uncertainty.

Finally, we can compare dates identified by our real-time approach to those found by Pastor and Stambaugh (2001) and Kim, Morley, and Nelson (2005) who use the whole sample and derive smoothed (ex post) estimates of the equity premium. Note that these papers assume a normal density, which we find strong evidence against, and impose a particular structure between the conditional mean and variance, which we do not. Based on a sample from 1926-1999, Kim, Morley, and Nelson (2005) find a permanent decrease in volatility in the 1940s which induces a structural break in the premium through their risk-return model. In addition to a risk-return link, Pastor and Stambaugh (2001) also impose a prior that the premium and prices (realized returns) move in opposite directions during transition from one level of the premium to the next. Using data from 1834-1999,

they find several breaks including 1940 and one in the early 1990s for which there is also evidence in our case.

6.3 Results for Mean Useful Historical Observations

The evidence in the previous subsection suggests that we should not put equal weights on historical data for optimal forecasts in the presence of possible structural breaks. Although our structural break forecasts consider all of the available historical data, the submodel average assigns probability weights to individual submodels only when their contribution to the marginal likelihood warrants it. Therefore, the distribution of submodel probabilities allows us to derive an *ex post* measure of the average number of useful observations at each point in time. This 'mean useful observations' measure (MUO_t) is defined as

$$\text{MUO}_t = \sum_{i=1885}^{t} (t+1-i)p(M_i|I_t). \tag{6.1}$$

Note that $\sum_{i=1885}^{t} ip(M_i|I_t)$, in equation (6.1), is the mean of the submodel distribution at time t.

For example, Figure 8 illustrates the distribution of submodel probabilities at 1960, at which time a probability of 0.63 was assigned to the 1940 submodel. Therefore, at 1960, the mean of the submodel distribution will be about 1940 and the mean useful observations will be about 21 years. Note, however, that our structural change model considers all of the available historical data but assigns very small weights to submodels prior to 1940 (longer samples) as well as to submodels after 1940 (shorter samples).

Our 'mean useful observations' measure defined by (6.1) is analogous to that in Pesaran and Timmermann (2002) who use a reverse-order CUSUM test to identify the most recent structural break and consequently the number of useful observations. For example, using a sample of monthly observations from 1954:1 to 1997:12, they find breaks in 1969, 1974 and 1990 which is consistent with our results discussed in section 6.2.

Time-series of our MUO_t measure are displayed in Figure 11. The 45-degree line corresponds to the no-break specification which uses all available data at each point in time. Consistent with our discussion in the previous subsection, the structural break model uses most of the data until around 1930 where the average number of useful observations drops dramatically. Around 1940 the useful observations begin to steadily increase till further declining in the 1970s and 1990s. In this figure, a moving window model would be represented as a horizontal line. For example, a moving window estimate using the most recent 10 years of data would be a horizontal line at 120. According to our model, this estimate would not be optimal during any historical time period.

6.4 Higher-Order Moments

As discussed in sections 6.1 and 6.2 above, allowing for asymmetries and fat tails in the submodel specification ($k = 2$) results in some differences in submodel probabilities, and superior density forecasts relative to the special case with $k = 1$. Figure 12 displays the posterior mean of the variance, skewness, and kurtosis of the excess returns distribution at each point in the sample using only information available to that time period. We show the time-variation in these higher-order moments implied by both our structural change model and the no-break alternative, using a $k = 2$ mixture-of-normals submodel specification in both cases. The no-break model cannot accommodate structural changes so the break in 1929 shows up in that case as a large permanent change in both skewness and kurtosis in the long-run distribution of excess returns.

6.5 Predictive Distribution of the Equity Premium

The purpose of our paper is to provide forecasts of the distribution of excess returns that accommodate uncertainty about past and future structural breaks. However, as outlined in section 5.5 above, we can also evaluate the implications for the predictive distribution of the equity premium. If there were no structural breaks, and excess returns were stationary, it would be optimal to use all available data weighted equally. However, in the presence of breaks, our forecast of the premium, and our uncertainty about that forecast, could be very misleading if our modeling/forecasting does not take account of those structural breaks.

Panel A of Figure 13 illustrates out-of-sample forecasts (predictive mean) of the equity premium, period-by-period, for both the structural break model and the no-break alternative. These are the forecasts $\hat{\gamma}_{B,t-1}$, computed from equation (5.19), which optimally use past data in the presence of possible structural breaks, versus $\hat{\gamma}_{A,t-1}$, computed from equation (5.20), which assumes no breaks. Henceforth, we refer to $\hat{\gamma}_{A,t-1}$, which is associated with submodel M_{1885}, as the no-break specification. The premium forecasts are similar until the start of the 1930s where they begin to diverge. The 1940 structural break results in clear differences in the equity premium forecasts for the break and no-break models. The premium forecasts from the structural break model rise through the 1940s to the 1960s. Toward the end of the sample the premium decreases to values substantially lower than the no-break model. The final premium forecast at the end of the sample is 3.79% for the preferred structural break model.

The second panel of Figure 13 displays the standard deviation of the predictive distribution of the premium. This is a measure of the uncertainty of our premium estimate in panel A. For the no-break model, uncertainty about the equity premium forecast originates from parameter uncertainty only, while for the structural break model it comes from both parameter and submodel uncertainty. Here again there are differences in the two specifications. The model that uses all data and ignores structural breaks

shows a steady decline in the standard deviation of the premium's predictive distribution as more data become available. That is, for a structurally stable model, as we use more data we become more confident about our premium forecast. However, the standard deviation of the predictive distribution for the premium from the break model shows that this increased confidence is misleading if structural breaks occur. As the second panel of Figure 13 illustrates, when a break occurs our uncertainty about the premium increases.

In subsection 5.5 above, we referred to an additional method often used to estimate the long-run equity premium. The estimator $\hat{\gamma}_{W,t-1}$, computed as in equation (5.21), recognizes that the distribution of excess returns may have undergone a structural break. However, this method just uses a 10-year moving window with equal weights on historical data for estimation. Relative to the no-break alternative, these forecasts have the advantage of dropping past data which may bias the forecast, but with the possible disadvantage of dropping too many data points, resulting in a reduction in the accuracy of the premium estimate. In addition, this estimator is implicitly assuming that structural breaks are reoccurring at regular intervals by using a fixed-length window of data at each point in time. Figure 14 compares 10-year moving window forecasts, at each point in time, to our forecasts that allow breaks, $\hat{\gamma}_{B,t-1}$ computed from (5.19). Note that the simple moving-window sample average is too volatile to produce realistic results. In some periods the sample average is negative while in other periods it is frequently in excess of 10%.

6.6 Forecasts of Long-Horizon Returns

As illustrated in Figures 12 and 13, the dynamics of the moments of the excess return distribution inferred from the structural break model are substantially different than those for the no-break model. For example, as discussed in section 6.4 above, being unable to accommodate breaks in the variance causes large permanent changes in skewness and kurtosis. These differences are likely to have significant effects on out-of-sample forecasts important for risk management.

To further illustrate this point, we computed out-of-sample mean and variance forecasts for the h-month cumulative return, $\sum_{i=1}^{h} r_{t+i}$. The mean forecast is $\sum_{i=1}^{h} E_t[r_{t+i}]$, and the variance forecast is $\sum_{i=1}^{h} \text{Var}_t[r_{t+i}]$. They are evaluated against the realized cumulative return and the cumulative realized volatility $\sum_{i=1}^{h} RV_{t+i}$. RV_{t+i} is computed using the sum of intra-month squared daily returns. This is done for the no-break and break model. The break model allows for out-of-sample breaks every 12 months and forecasts are calculated as in Subsection 5.4.

Table 5 reports forecast results for the $k = 2$ submodel specification and starting the out-of-sample forecasts at month 701 (half-way through the sample at 1943:9). For an horizon of $h = 120$ months, the root mean squared error (RMSE) for the mean forecast

from the break model is 7.36 versus 7.51 for the no-break model. The variance forecast is 22.5 for the structural change model versus 28 for the no-break alternative. For a forecast horizon of 20 years (240 months), the corresponding RMSE results are 11.47 versus 11.86 for the mean and 56.61 versus 67.71 for the variance. In other words, the out-of-sample mean and variance forecasts using the model that accommodates structural breaks dominate those from a forecasting procedure that ignores breaks. Of course the superior density forecasts for the structural change models reported in Table 3 are not just due to superior mean and variance forecasts but rather due to improved fit of the entire distribution of excess returns. For example, a risk manager may also be interested in the improved fit of the tails of the distribution discussed in section 6.4 above.

6.7 Robustness

Table 2 reports sample statistics for the excess return distribution when parameters are simulated from the assumed distributions for priors described in subsection 4.2. These empirical moments seem reasonable. For robustness, we also tried some alternative priors. For example, as discussed at the end of subsection 5.8, we set the prior probability of breaks, λ_t, to .01 which favors infrequent breaks. As indicated in Table 4, we redid all of our estimation and forecasting favoring more frequent structural breaks by assuming that $\lambda_t = .02$. Recall that we allow for one break per year so that this corresponds to an expected duration of 50 years between breaks. The results were very similar. In particular, the log(ML) for the best model was -1194 when λ_t=.02 instead of -1196 for $\lambda_t = .01$. Table 4 also shows results when we consider more diffuse priors for other parameters. They all provide strong evidence against the no-break model and are consistent with previous results.

Another possibility is to re-set priors each period to the most recent posterior. As an example in this direction, whenever a new submodel is introduced we set the prior parameters for the premium to the previous posterior mean and variance of γ. That is, during any period a new submodel is introduced, the prior on γ begins centered on the most recent posterior for γ based on available data. We did this for the $\lambda = 0.01$ case using the $k = 1$ submodel specification. The main difference in the premium forecasts for this case was that the premium was slightly less variable and also had a reduced standard deviation of the predictive distribution for the premium. However, the marginal likelihood is -1216.18 which is slightly worse than our original prior in Table 3 for $k = 1$, and still inferior to the $k = 2$ specification.

7 Conclusion

In summary, we provide an approach to forecasting the unconditional distribution of excess returns making optimal use of historical data in the presence of possible structural

breaks. We focus on learning about structural breaks in real time and assessing their impact on out-of-sample forecasts. As a byproduct, this procedure identifies, in real time, probable dates of structural change. Since structural breaks can never be identified with certainty, our approach is to use a probability-weighted average of submodels, each of which is estimated over a different history of data. Our forecasts consider all of the available historical data but only assign weight to individual submodels when their contribution to the marginal likelihood warrants it. Since the predictive density of returns integrates over the submodel distribution, uncertainty about structural breaks is accounted for in the analysis. The paper illustrates the importance of uncertainty about structural breaks and the value of modeling higher-order moments of excess returns when inferring structural breaks and forecasting the return distribution and its moments.

We use a two-component discrete mixture-of-normals specification for the submodel. This is the parameterization of excess returns which is subject to structural breaks. For robustness, we compare our results using this flexible submodel specification to the nested Gaussian submodel specification to see if the more general distribution affects our inference about structural change or our real-time forecasts. Our evidence clearly supports a structural break model using the more flexible parameterization of the submodel. This richer two-component submodel is also more robust to false breaks.

The empirical results strongly reject ignoring structural change in favor of our forecasts which weight historical data to accommodate uncertainty about structural breaks. We also strongly reject the common practice of using a fixed-length moving window. Ignoring structural breaks leads to inferior density forecasts. So does using a fixed-length moving window of historical data.

Structural change has implications for the entire shape of the long-run excess return distribution. The preferred structural change model produces kurtosis values well above 3 and negative skewness throughout the sample. Furthermore, the shape of the long-run distribution and the dynamics of the higher-order moments are quite different from those generated by forecasts which cannot capture structural breaks. As we show, ignoring structural change results in misspecification of the long-run distribution of excess returns which can have serious implications for long-run forecasts and risk assessments.

To answer the question in the title of our paper, our paper says that one should use all available data but weight data histories optimally according to their contribution to forecasts at each point in time. For most of our sample, older data tends to get low weights fairly quickly but a critical result is that it is very suboptimal to use a fixed-length moving window to capture this effect. Our results show that the value of historical data varies considerably over time. Our paper provides a way of using all available data but assigning appropriate weights to the component data histories. We show the implications of differences in the no-break versus optimal forecasts. These differences are significant and may be important for risk management and long-horizon investment decisions.

8 Appendix

This appendix provides additional details concerning computations for each of the Gibbs sampling steps for the submodels. Conditional on Z_t and σ^2 the conditional posterior for μ_j $j = 1, ..., k$ is

$$\begin{aligned} \mu_j|Z, \sigma^2, r &\sim N(M, V^{-1}) && (8.1) \\ M &= V^{-1}\left(\sigma_j^{-2}\sum_{t=1}^{T} z_{t,j}r_t + B_{jj}^{-1}b_j\right) && (8.2) \\ V &= \sigma_j^{-2}T_j + B_{jj}^{-1}. && (8.3) \end{aligned}$$

where $T_j = \sum_{t=1}^{T} z_{t,j}$. The conditional posterior of σ_j^2 is,

$$\sigma_j^2|Z, \mu, r \sim IG\left(\frac{v_j + T_j}{2}, \frac{\sum_{t=1}^{T}(r_t - \mu_j)^2 z_{t,j} + s_j}{2}\right), \quad j = 1, ..., k. \tag{8.4}$$

Only the observations attributed to component j are used to update μ_j and σ_j^2. With the conjugate prior for π, we sample the component probabilities as,

$$\pi \sim \mathcal{D}(\alpha_1 + T_1, ..., \alpha_k + T_k). \tag{8.5}$$

Finally, to sample $z_{t,i}$, note that,

$$p(z_{t,i}|r, \mu, \sigma, \pi) \propto \pi_i \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{1}{2\sigma_i^2}(r_t - \mu_i)^2\right), \quad i = 1, ..., k, \tag{8.6}$$

which implies that they can be sampled as a Multinomial distribution for $t = 1, ..., T$.

It is well known that in mixture models the parameters are not identified. For example, switching all states Z and the associated parameters gives the same likelihood value. Identification can be imposed through prior restrictions. However, in our application, interest centers on the moments of the return distribution and not the underlying mixture parameters. The moments of returns are identified. If for example, we switch all the parameters of component 1 and 2 we still have the same premium value $\gamma = \sum_{i=1}^{k} \mu_i \pi_i$. Therefore, we do not impose identification of the component parameters but instead compute the mean, variance, skewness and kurtosis using (3.3)-(3.8) after each iteration of the Gibb sampler. It is these posterior quantities that our analysis focuses on. In the empirical work, we found the Markov chain governing these moments to mix very efficiently. As such, 5000 Gibbs iterations, after a suitable burnin period provide accurate estimates.

Table 1: Summary Statistics for Scaled Monthly Excess Returns

Sample	Obs	Mean	Variance	Stdev	Skewness	Kurtosis
1885:02-2003:12	1423	0.0523	0.4007	0.6330	-0.4513	9.9871

Table 2: Sample Statistics for Excess Returns Implied by the Prior Distribution

	Mean	Median	Stdev	95% HPDI
Sample Mean	0.0369	0.0354	0.0320	(-0.0238, 0.1007)
Sample Variance	0.5808	0.5056	0.3312	(0.1519, 1.1786)
Sample Skewness	-0.3878	-0.3077	0.4718	(-1.4077, 0.3534)
Sample Kurtosis	8.1369	6.4816	5.9317	(2.7169, 18.7218)

This table reports summary measures of the empirical moments from the mixture submodel ($k = 2$) when parameters are simulated from the prior distribution. The priors are $\mu_i \sim N(b_i, B_{ii})$, $\sigma_i^2 \sim IG(v_i/2, s_i/2)$, and $\pi \sim \mathcal{D}(\alpha_1, ..., \alpha_k)$, as described in section 4.2. The hyperparameters are found in the footnote to Table 3. First a draw from the prior distribution gives a parameter vector from which T observations of excess returns are simulated $\{\tilde{r}_t\}_{t=1}^T$. From these data we calculate the sample mean, variance, skewness and kurtosis of excess returns. This process is repeated a large number of times to produce a distribution of each of the excess return moments. Finally, from this empirical distribution we report the mean, median, standard deviation and the 95% highest posterior density interval (HPDI).

Table 3: Model Specifications and Results

model	prior about breaks	log(ML)
$k = 1$	$\lambda_t = 0$, no breaks	-1371.22
$k = 1$	$\lambda_t = 0$, no breaks 10-year moving window	-1281.94
$k = 1$	$\lambda_t = 0.01$, every 5 years otherwise $\lambda_t = 0$	-1235.33
$k = 1$	$\lambda_t = 0.01$, every year otherwise $\lambda_t = 0$	-1216.08
$k = 1$	λ estimated, break every year otherwise $\lambda = 0$	-1204.17
$k = 2$	$\lambda_t = 0$, no breaks	-1241.09
$k = 2$	$\lambda_t = 0$, no breaks 10-year moving window	-1220.78
$k = 2$	$\lambda_t = 0.01$, every 5 years otherwise $\lambda_t = 0$	-1202.01
$k = 2$	$\lambda_t = 0.01$, every year otherwise $\lambda_t = 0$	-1196.30
$k = 2$	λ estimated, break every year otherwise $\lambda = 0$	-1191.77

This tables displays: the number of components, k, in the submodel; the prior on the ocurrence of structural breaks, λ_t; and the logarithm of the marginal likelihood, log(ML), for all specifications based on the full sample of observations used in estimation. The priors are $\mu \sim N(b, B)I_{\mu>0}$, $\sigma^2 \sim IG(v/2, s/2)$ for $k = 1$; and $\mu_i \sim N(b_i, B_{ii})$, $\sigma_i^2 \sim IG(v_i/2, s_i/2)$, $\pi \sim \mathcal{D}(\alpha_1, ..., \alpha_k)$ for $k = 2$. Hyperparameters are $b = 0.03, B = 0.03^2, v = 9.0, s = 4.0$ for the $k = 1$ case; and $b_1 = 0.05, b_2 = -0.30, B_{11} = 0.03^2, B_{22} = 0.05^2, v_1 = 10.0, s_1 = 3, v_2 = 8.0, s_2 = 20.0, \alpha_1 = 7, \alpha_2 = 1$ for the $k = 2$ case. We impose a positive equity premium by giving zero support to any parameter configuration that violates $\gamma = \sum_{i=1}^{2} \mu_i \pi_i > 0$. When λ is estimated, it has a prior of $Beta(0.05, 20)$.

Table 4: Model Robustness, $k = 2$

changes in prior	log(ML)
$\lambda_t = 0.02$	-1194.02
$B_{11} = 0.12^2, B_{22} = 0.2^2$	-1197.21
$v_1 = 5, v_2 = 4$	-1201.43
$B_{11} = 0.12^2, B_{22} = 0.2^2, v_1 = 5, v_2 = 4$	-1203.09

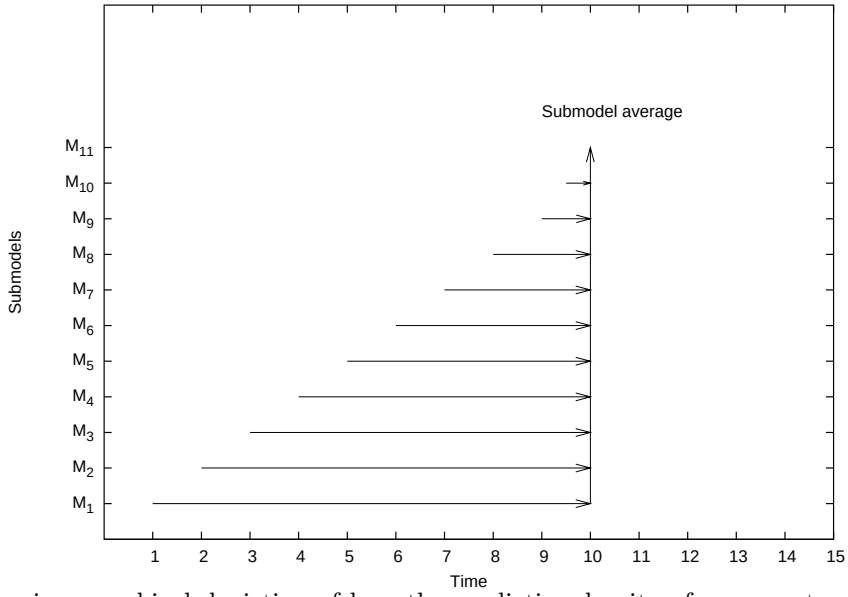
This tables displays, for the $k = 2$ case, the log marginal likelihood for changes to the benchmark prior $b_1 = 0.05, b_2 = -0.30, B_{11} = 0.03^2, B_{22} = 0.05^2, v_1 = 10.0, s_1 = 3, v_2 = 8.0, s_2 = 20.0, \alpha_1 = 7, \alpha_2 = 1$ and $\lambda_t = 0.01$ for one month of every year and otherwise $\lambda_t = 0$.

Table 5: Out-of-sample forecasts, $k = 2$

RMSE		
	$\sum_{i=1}^{h} E_t[r_{t+i}]$ forecast of $\sum_{i=1}^{h} r_{t+i}$	
h months	break	nobreak
1	0.5171	0.5178
12	1.9352	1.9481
120	7.3571	7.5141
240	11.4729	11.8636
	$\sum_{i=1}^{h} \mathrm{Var}_t[r_{t+i}]$ forecast of $\sum_{i=1}^{h} RV_{t+i}$	
1	0.4916	0.5250
12	2.7656	3.3698
120	22.4951	28.0098
240	56.6081	67.7051

Root-Mean-Squared Error (RMSE) associated with period-by-period forecasts of the mean and variance of the h-month cumulative return over the 2nd half of the sample (from month 701 to the end of the sample). Realized volatility, RV_{t+i} is estimated as the sum of intra-month daily squared returns.

Figure 1: Individual Submodels and the Submodel Average



This figure is a graphical depiction of how the predictive density of excess returns is constructed for the structural break model. This corresponds to equation (5.6). The predictive density is computed for each of the submodels $M_1, ..., M_{10}$ given information up to $t = 10$. The final submodel M_{11}, postulates a break at $t = 11$ and uses no data but only a prior distribution. Each submodel is estimated using a smaller history of data (horizontal lines). Weighting these densities via Bayes rule (vertical line) gives the final predictive distribution (model average) of excess returns for $t = 11$.

Figure 2: Estimates of λ through Time

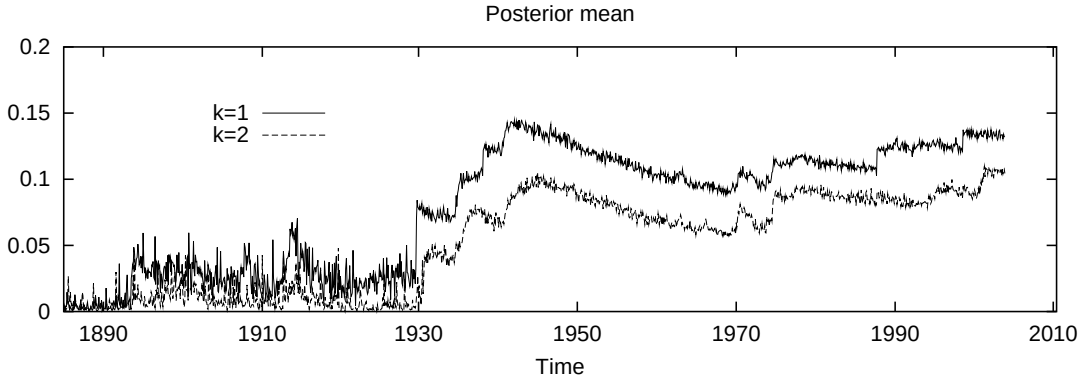
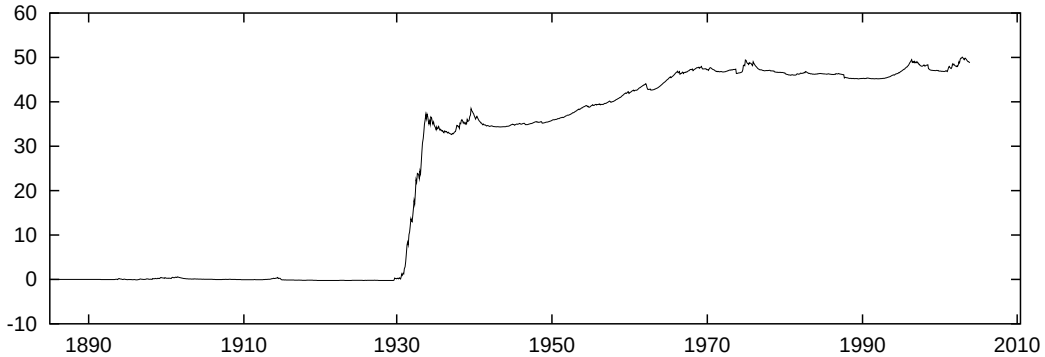


Figure 3: Difference in the Cumulative Sum of the Log Predictive Likelihoods



This figure displays the difference at each point in time in the logarithm of the marginal likelihood (logarithm of equation 5.26) for the break versus the no-break models with $k = 2$ submodel specifications.

Figure 4: Submodel Probabilities through Time, 1885:2-1910:1

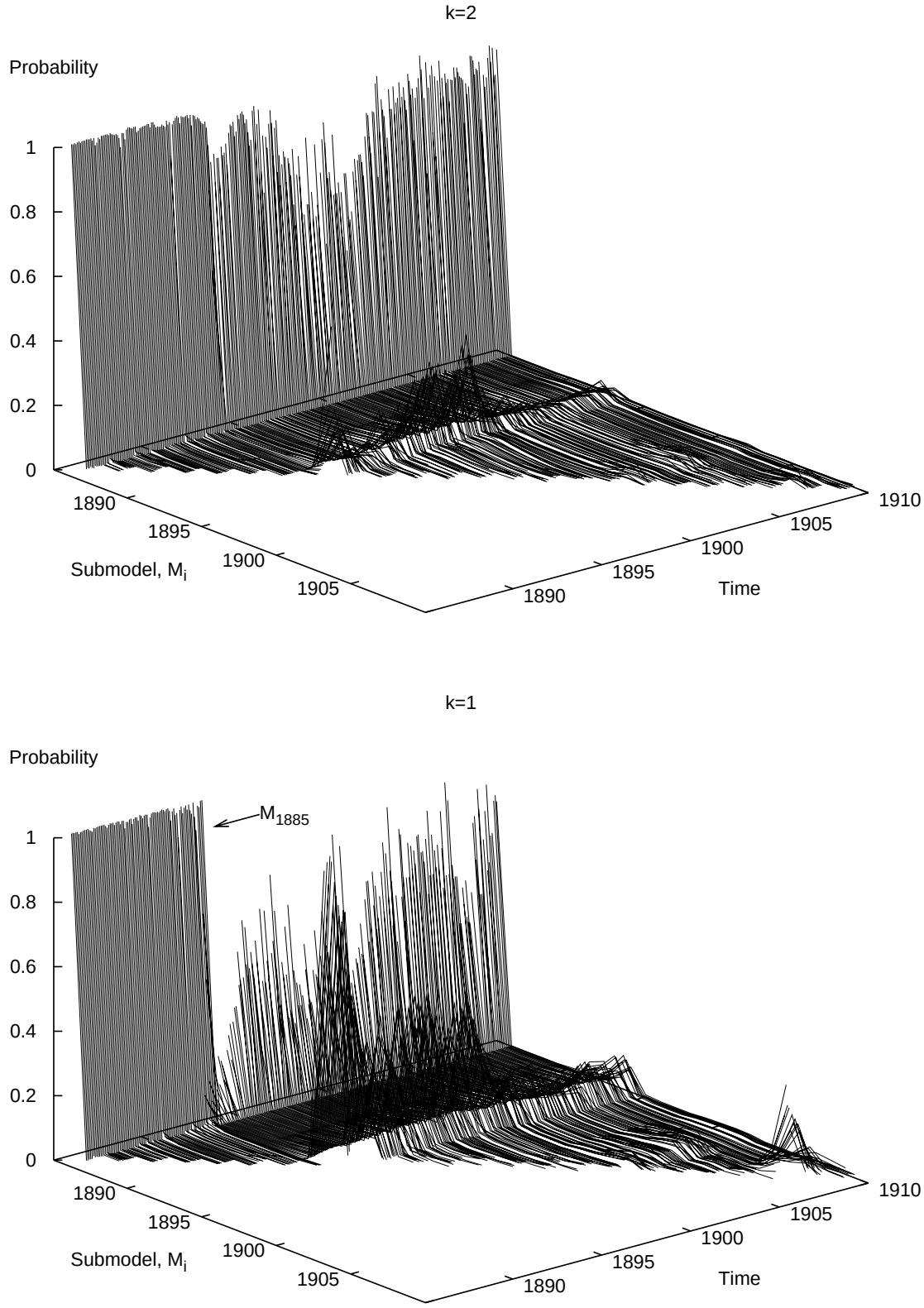


Figure 5: Submodel Probabilities through Time, 1925:1-1945:1

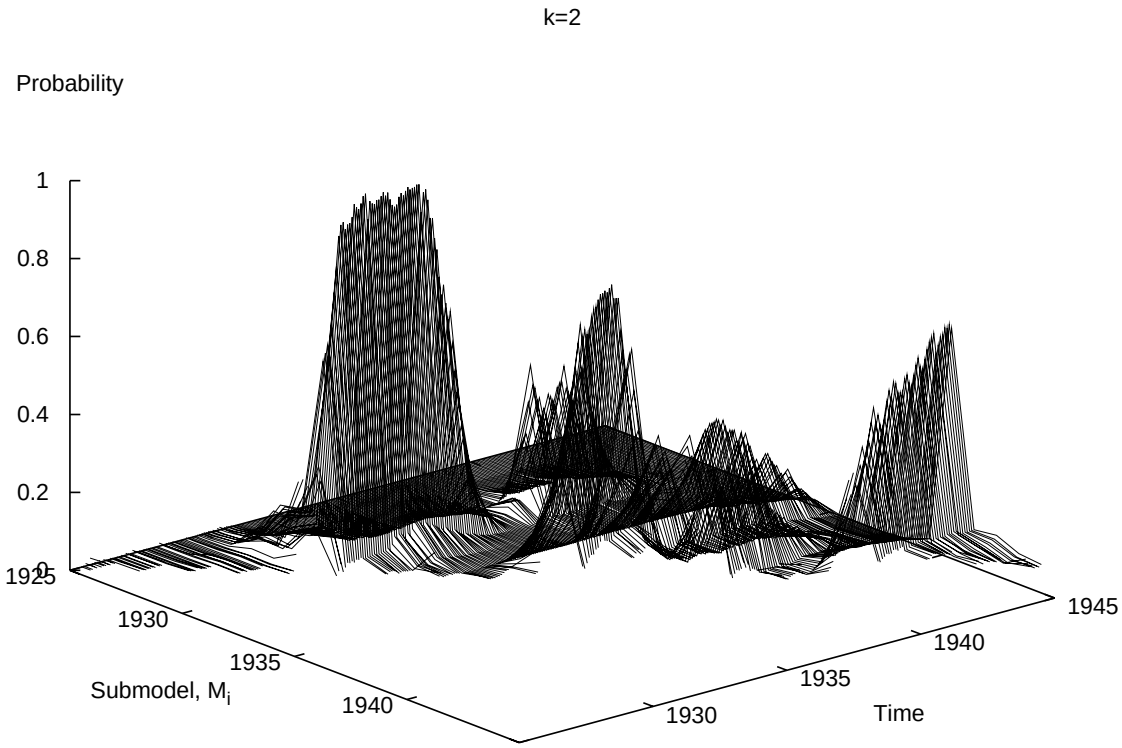


Figure 6: Submodel Probabilities through Time, 1970:1-2003:12

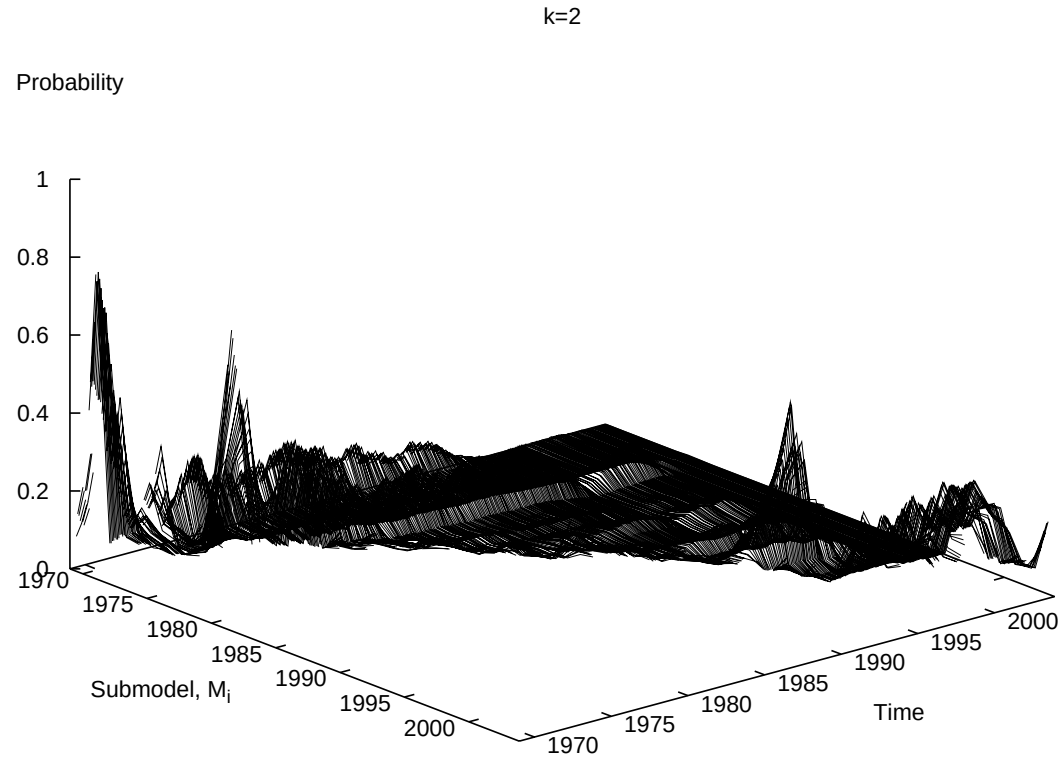


Figure 7: Submodel Probabilities over Time, $k = 2$

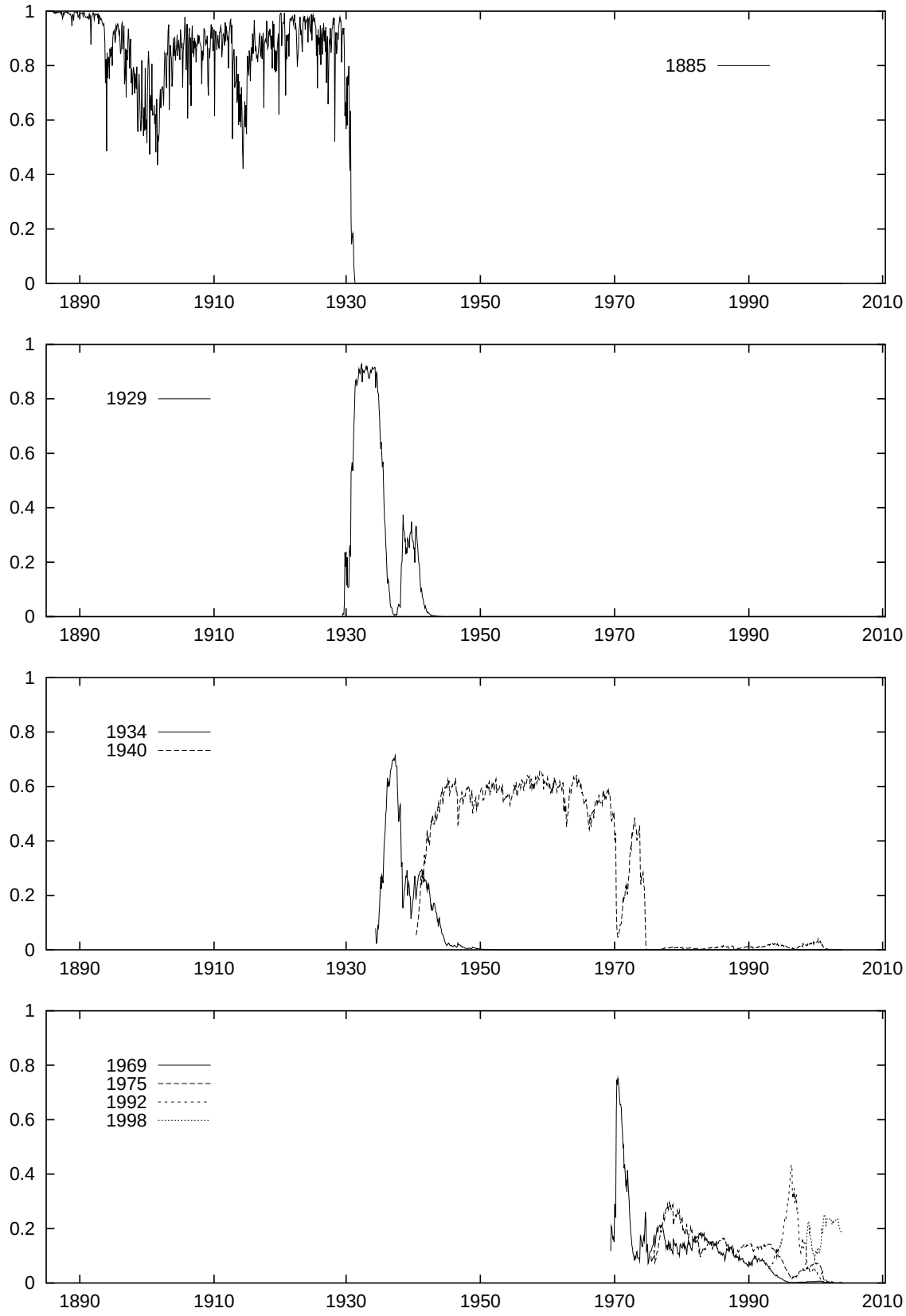


Figure 8: Submodel Probability Distribution at 1960:01, $k = 2$

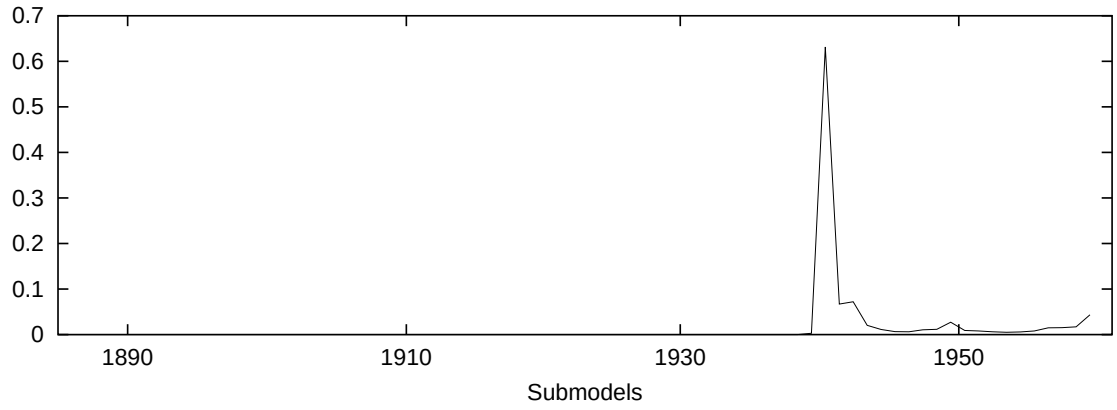


Figure 9: Submodel Probability Distribution at 2003:12, $k = 2$

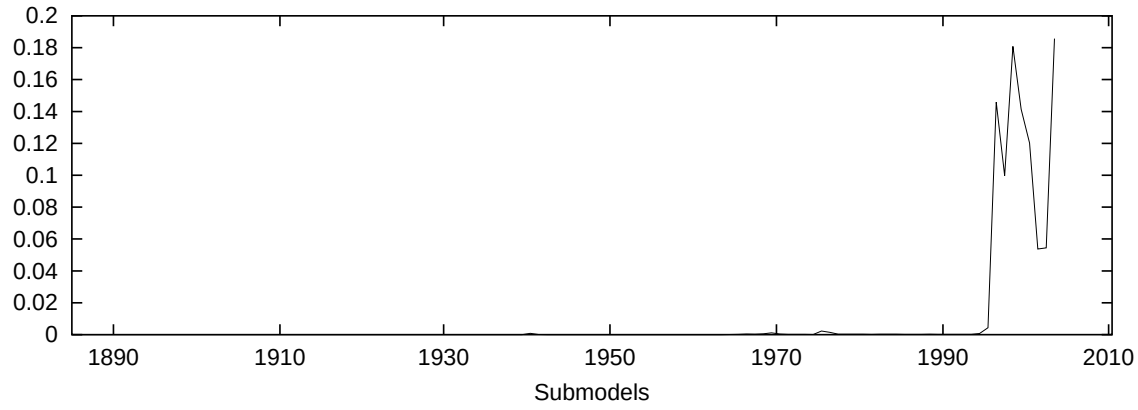


Figure 10: Probabilities for M_{1893}: Example of a False Break

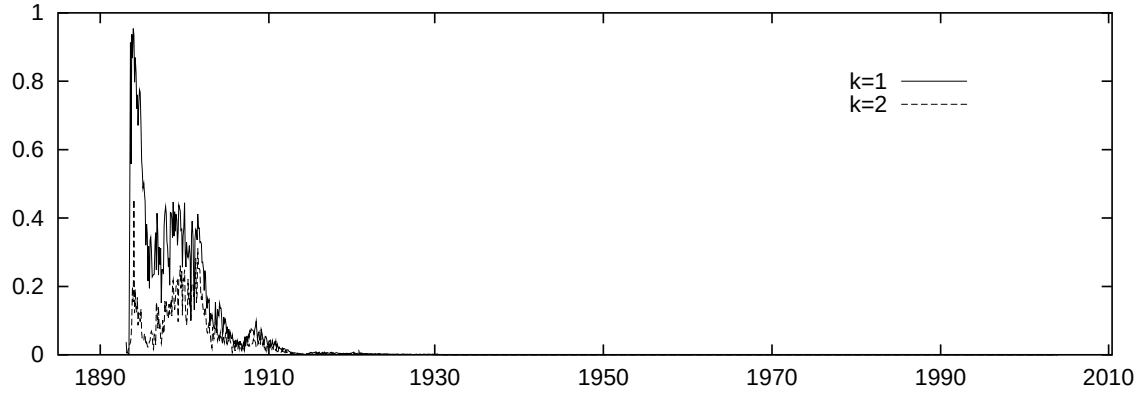
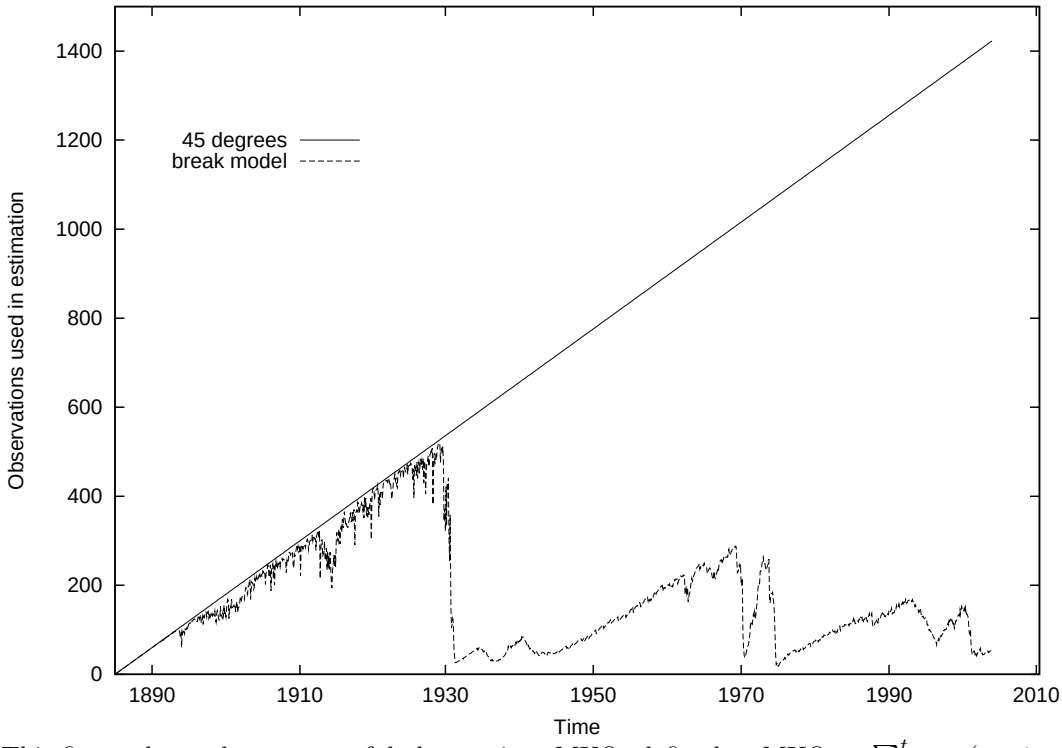
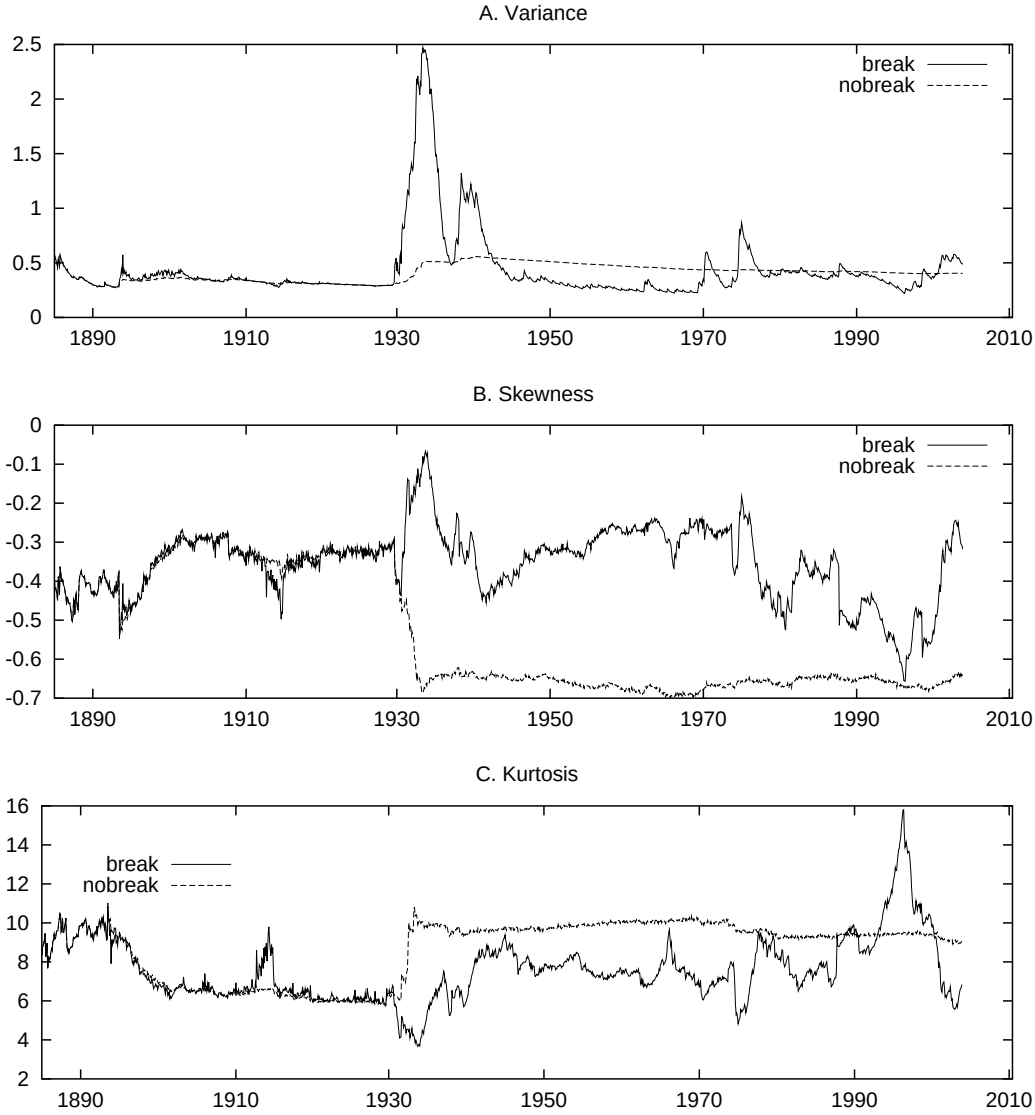


Figure 11: Mean useful Observations, $k = 2$.



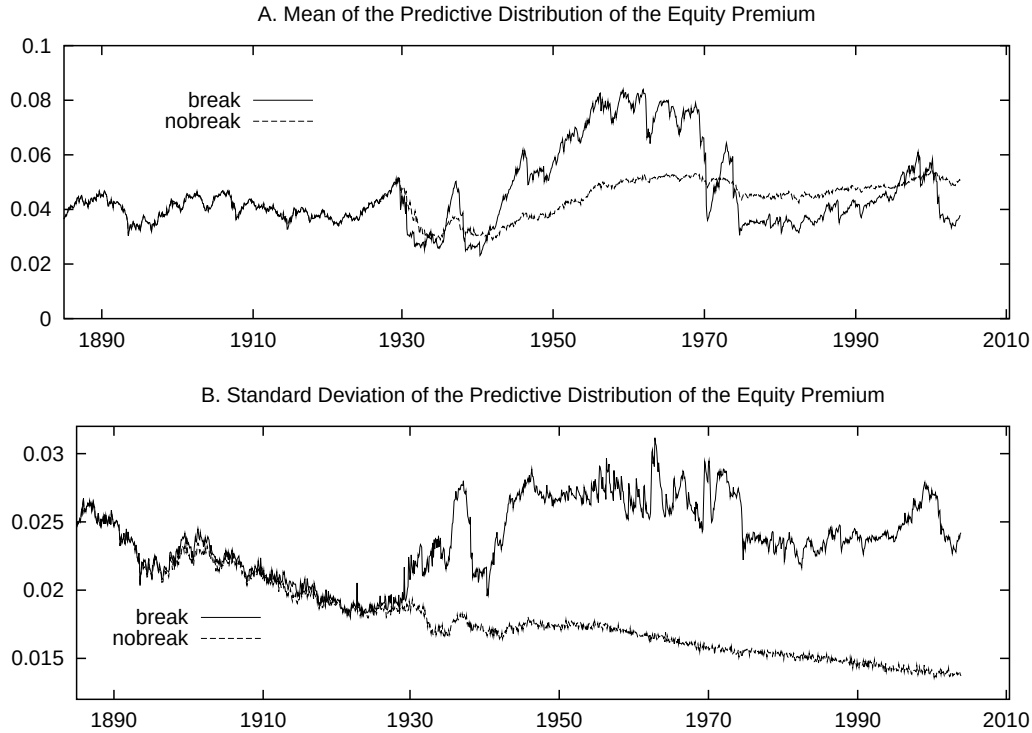
This figure shows the mean useful observations MUO_t defined as $\text{MUO}_t = \sum_{i=1885}^{t}(t+1-i)p(M_i|I_t)$, which is the expected number of useful observation for model estimation at each point in time. $p(M_i|I_t)$ is the posterior submodel probability for M_i given the information set I_t and $\sum_{i=1885}^{t} ip(M_i|I_t)$ is the mean of the submodel distribution at time t. If there are no structural breaks then MUO_t would follow the 45-degree line. A fixed-length moving window would correspond to a horizontal line at the window length number of observations.

Figure 12: Higher-Order Moments of Excess Returns through Time



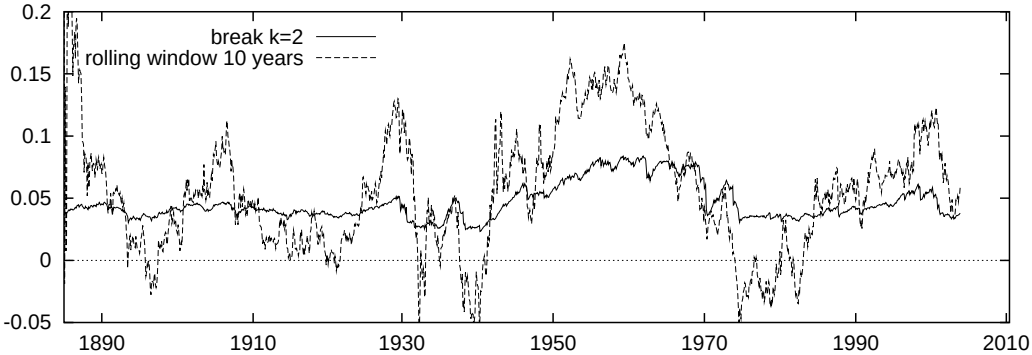
Displayed are the posterior means of the moments of the excess return distribution as inferred from the structural break model with $k = 2$ submodel specification. Each moment is estimated using only information in I_t at each point in time. The moments in equations (3.6)-(3.8) are computed for each Gibbs draw from the posterior distribution for each of the submodels M_i. The submodel specific moments are averaged using equation (5.13). This is repeated at each observation in the sample starting from $t = 1$. The evolution of the excess return moments reflect both learning (as more data arrive) and the effect of structural breaks.

Figure 13: Predictive Distribution of the Premium through Time, $k = 2$.



Panel A displays the out-of-sample forecasts (predictive mean) of the equity premium period-by-period for both the structural break model, as in equation (5.19), and the no-break alternative. Panel B displays the corresponding standard deviation of the predictive distribution of the equity premium.

Figure 14: Comparison of Premium Forecasts



This figure compares the forecasts (predictive mean) of the long-run equity premium from the structural break model, along with the sample average that uses a moving window of 10 years of data. The sample average at time t is defined as $\frac{1}{120}\sum_{i=1}^{120} r_{t-i+1}$.

References

Andreou, E., and E. Ghysels (2002): "Detecting Multiple Breaks in Financial Market Volatility Dynamics," *Journal of Applied Econometrics*, 17(5), 579–600.

Avramov, D. (2002): "Stock return predictability and model uncertainty," *Journal of Financial Economics*, 64(423-458).

Barberis, N. C. (2000): "Investing for the Long Run when Returns are Predictable," *Journal of Finance*, 55, 225–264.

Campbell, J. Y., and S. B. Thompson (2005): "Predicting the Equity Premium Out of Sample: Can Anything Beat the Historical Average?," NBER Working Paper 11468.

Chib, S. (2001): "Markov Chain Monte Carlo Methods: Computation and Inference," in *Handbook of Econometrics*, ed. by Heckman, and Leamer. Elsevier Science.

Clark, T. E., and M. W. McCracken (2006): "Averaging Forecasts from VARs with Uncertain Instabilities," RWP 06-12, The Federal Reserve Bank of Kansas City.

Cremers, K. J. M. (2002): "Stock return predictability: A bayesian model selection perspective," *Review of Financial Studies*, 15(1223-1249).

Derrig, R. A., and E. D. Orr (2004): "Equity Risk Premium: Expectations Great and Small," *North American Actuarial Journal*, 8(1), 45–69.

Diebolt, J., and C. P. Robert (1994): "Estimation of Finite Mixture Distributions through Bayesian Sampling," *Journal of the Royal Stistical Society, Series B*, 56, 363–375.

Donaldson, R. G., M. Kamstra, and L. Kramer (2006): "Estimating the *Ex Ante* Equity Premium," manuscript, University of Toronto.

Eklund, J., and S. Karlsson (2005): "Forecast Combination and Model Averaging using Predictive Measures," Sveriges Riksbank Working Paper No. 191.

Fama, E. F., and K. R. French (2002): "The Equity Premium," *Journal of Finance*, 57(2), 637–659.

Geweke, J. (1995): "Bayesian Comparison of Econometric Models," University of Iowa.

——— (1997): "Using Simulation Methods for Bayesian Econometric Models: Inference, Development, and Communication," *Econometric Reviews*, 18(1), 1–73.

——— (2005): *Contemporary Bayesian Econometrics and Statistics*. John Wiley & Sons.

Geweke, J., and C. Whiteman (2006): "Bayesian Forecasting," in *Handbook of Economic Forecasting*, ed. by G. Elliot, C. Granger, and A. Timmermann. Elsevier, Amsterdam.

Giordani, P., and R. Kohn (2007): "Efficient Bayesian Inference for Multiple Change-Point and Mixture Innovation Models," forthcoming, *Journal of Business and Economic Statistics*.

Goyal, A., and I. Welch (2007): "A Comprehensive Look at the Empirical Performance of Equity Premium Prediction," forthcoming, Review of Financial Studies.

Graham, J. R., and C. R. Harvey (2005): "The long-run equity risk premium," *Finance Research Letters*, 2, 185–194.

Jacquier, E., A. Kane, and A. Marcus (2005): "Optimal Estimation of the Risk Premium for the Long Run and Asset Allocation: A Case of Compounded Estimation Risk," *Journal of Financial Econometrics*, 3, 37–55.

Johannes, M., and N. Polson (2005): "MCMC Methods for Financial Econometrics," Handbook of Financial Econometrics, Elsevier.

Kass, R. E., and A. E. Raftery (1995): "Bayes Factors," *Journal of the American Statistical Association*, 90(420), 773–795.

Kim, C.-J., J. C. Morley, and C. R. Nelson (2005): "The Structural Break in the Equity Premium," *Journal of Business & Economic Statistics*, 23(2), 181–191.

Koop, G. (2003): *Bayesian Econometrics*. Wiley, Chichester, England.

Koop, G., and S. Potter (2007): "Estimation and Forecasting in Models with Multiple Breaks," forthcoming, *The Review of Economic Studies*.

Lettau, M., and S. Ludvigson (2001): "Consumption, Aggregate Wealth, and Expected Stock Returns," *Journal of Finance*, 56(3), 815–849.

Lettau, M., S. C. Ludvigson, and J. A. Wachter (2007): "The Declining Equity Premium: What Role Does Macroeconomic Risk Play?," forthcoming, *Review of Financial Studies*.

Lettau, M., and S. van Nieuwerburgh (2007): "Reconciling the Return Predictability Evidence," forthcoming, *Review of Financial Studies*.

Maheu, J. M., and S. Gordon (2007): "Learning, Forecasting and Structural Breaks," forthcoming, *Journal of Applied Econometrics*.

Mayfield, E. S. (2004): "Estimating the Market Risk Premium," *Journal of Financial Economics*, 73(3), 465–496.

Mehra, R., and E. C. Prescott (2003): "The Equity Premium in Retrospect," in *Handbook of the Economics of Finance*, ed. by G. M. Constantinides, M. Harris, , and R. Stulz. North Holland, Amsterdam.

Pastor, L., and R. F. Stambaugh (2001): "The Equity Premium and Structural Breaks," *Journal of Finance*, 4, 1207–1231.

Paye, B. S., and A. Timmermann (2006): "Instability of Return Prediction Models," *Journal of Empirical Finance*, 13, 274–315.

Pesaran, M. H., D. Pettenuzzo, and A. Timmermann (2006a): "Forecasting Time Series Subject to Multiple Structural Breaks," *Review of Economic Studies*, 73, 1057–1084.

——— (2006b): "Learning, Structural Instability and Present Value Calculations," forthcoming, *Econometric Reviews*.

Pesaran, M. H., and A. Timmermann (2002): "Market Timing and Return Prediction under Model Instability," *Journal of Empirical Finance*, 9, 495–510.

Pesaran, M. H., and A. Timmermann (2007): "Selection of Estimation Window in the Presence of Breaks," *Journal of Econometrics*, 137, 134–161.

Pettenuzzo, D., and A. Timmermann (2005): "Predictability of Stock Returns and Asset Allocation under Structural Breaks," Department of Economics, University of California, San Diego.

Rapach, D. E., and M. E. Wohar (2006): "Structural Breaks and Predictive Regression Models of Aggregate U.S. Stock Returns," *Journal of Financial Econometrics*, 4(2), 238–274.

Robert, C. P., and G. Casella (1999): *Monte Carlo Statistical Methods*. Springer, New York.

Roeder, K., and L. Wasserman (1997): "Practical Bayesian Density Estimation Using Mixtures of Normals," *Journal of the American Statistical Association*, 92(439), 894–902.

Schwert, G. W. (1990): "Indexes of U.S. Stock Prices from 1802 to 1987," *Journal of Business*, 63(3), 399–426.

Siegel, J. (1992): "The Real Rate of Interest from 1800-1990: A Study of the U.S. and the U.K.," *Journal of Monetary Economics*, 29, 227–252.

Turner, C., R. Startz, and C. Nelson (1989): "A Markov Model of Heteroskedasticity, Risk, and Learning in the Stock Market," *Journal of Financial Economics*, 25, 3–22.

Viceira, L. M. (1997): "Testing for structural change in the predictability of asset returns," Harvard University, Manuscript.

Wright, J. H. (2003): "Forecasting US Inflation by Bayesian Model Averaging," International Finance DP 780, Board of Governors of the Federal Reserve System.