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

John M. Maheu and Thomas H. McCurdy*

March 2006

Abstract

We provide an optimal approach to forecasting the long-run (unconditional) equity premium in the presence of structural breaks. This forecasting procedure determines in real time how useful historical data are in updating our prior belief about the distribution of market excess returns. The value of historical data has varied considerably, implying that ignoring structural breaks or using a rolling window is not optimal. We obtain realistic out-of-sample forecasts for the entire 1885-2003 period; the forecast at the end of the sample is 4.02 for the structural break model and 5.10 for a no-break model. The results are robust to a wide-range of distributional assumptions about excess returns.

*Maheu (jmaheu@chass.utoronto.ca), Department of Economics, University of Toronto; McCurdy (tmccurdy@rotman.utoronto.ca), Joseph L. Rotman School of Management, University of Toronto, and Associated Fellow, CIRANO. We thank David Goldreich, Mark Kamstra, Lisa Kramer, Jan Mahrt-Smith, Kevin Wang, as well as seminar participants of the $(EC)^2$ conference Istanbul 2005 and the Bank of Canada, for helpful comments. Lois Chan provided excellent research assistance. We are also grateful to the SSHRC for financial support.

1 Introduction

An important topic in finance is the forecast of the return premium on a well diversified portfolio of equity relative to a riskfree asset. Accurate forecasts of this market equity premium are required for capital budgeting, investment, and pricing decisions.

There is an extensive literature that seeks to explain the long-run equity premium. Most of this literature takes as given simple point estimates of the premium obtained as the sample average from a long series of excess return data.[1] In addition, many forecasters, including those using dynamic models with many predictors, report the sample average of excess returns as a benchmark.[2]

The use of a sample average as a forecast of the long-run equity premium assumes that excess returns are stationary and that the process governing them does not undergo structural breaks. Once we allow for structural breaks, it is not clear whether or not historical data are useful for forecasting the equity premium. For instance, including data prior to a structural break may result in a biased forecast. The purpose of this paper is to investigate the value of data in updating our beliefs about the long-run equity premium, and to provide forecasts of the premium while allowing for structural breaks.

We focus on the unconditional distribution of excess market returns and define the long-run premium as the mean of that distribution.[3] Investment and capital budgeting decisions often span many years. With this investment horizon, the long-run equity premium is the relevant measure. Jacquier, Kane, and Marcus (2005) discuss the importance of accurate premium estimates for long-orizon portfolio choice. In addition, by focusing on the long-run premium, as opposed to short-run dynamic models of the premium, we may be less susceptible to model misspecification. That is, the existence of a long-run value of the premium is consistent with different underlying models of risk.

Nevertheless, even for the unconditional distribution of excess returns, misspecified models may provide evidence of structural breaks when the underlying data generating process (DGP) is in fact stable. For example, suppose one assumed a Normal distribution for excess returns when in fact the DGP has fat tails. In this case, realizations in the tail of the maintained Normal distribution could be mistakenly interpreted in real time

[1]For example, Table 1 in a recent survey by Mehra and Prescott (2003) lists four estimates of the unconditional premium using sample averages of data from 1802-1998, 1871-1999, 1889-2000, and 1926-2000.

[2]Derrig and Orr (2004) survey a wide range of both academic and practitioner data-based estimates of the equity premium. There are many asset pricing models that have been used to estimate this premium, building on the three-factor model of Fama and French (1992) or the arbitrage pricing theory of Ross (1976). Another approach uses earnings or dividend growth to model the equity premium, for example, Donaldson, Kamstra, and Kramer (2004) and Fama and French (2002). Estimates of the equity premium in the presence of regimes changes include Mayfield (2004) and Turner, Startz, and Nelson (1989). Recent examples of premium forecasts include Campbell and Thompson (2004), and Goyal and Welch (2004).

[3]In this paper we view the full data set as being potentially partitioned into sequences of data generated from different stationary models. Therefore, within each partition there is a well defined unconditional premium.

as evidence of a structural break. To minimize this potential problem, we use a very flexible model to forecast the long-run premium. In particular, our maintained model is a mixture-of-Normals which can capture skewness and excess kurtosis, both of which are well known features of returns. For robustness, we compare our results to the nested Normal distribution case to see if the more general distribution affects our inference about structural change.[4]

The Bayesian approach to prediction integrates out parameter uncertainty. For example, see Barberis (2000), and Kandel and Stambaugh (1996). Important papers by Pastor and Stambaugh (2001) and Kim, Morley, and Nelson (2005) provide smoothed historical estimates of the equity premium in the presence of structural breaks using a dynamic risk-return model.[5] These papers are based on the structural break model of Chib (1998) which provides estimates conditional on a maintained number of breaks in-sample.

A primary objective of our paper is to stress the learning aspect that would occur in real time and its implications for decision making. That is, we investigate how the evidence for structural breaks changes over time and assess the effects on real time forecasts of ignoring this information. Therefore, our forecasts of the premium also incorporate time-varying model uncertainty. Our approach provides period-by-period out-of-sample forecasts of the premium, incorporating the probability of structural breaks in the past data as well as the possibility of breaks in the future. A by-product of our approach is that it generates an estimate of the number of historical observations that are useful at each point in time for forecasting the long-run premium.

In addition, our maintained model of excess returns, which is subject to structural breaks, can capture heteroskedasticity, asymmetry and fat tails. These are features that may be important for forecasts of the equity premium as well as for identifying structural breaks. As noted above, this allows us to assess the impact of outliers on structural break identification.

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 could bias our estimates and forecasts. This might suggest a rolling window estimator that only uses a portion of the available data. However, such an approach will not be optimal. Indeed, some combination of the data that follow a perceived break, and the (biased) data that preceded it may be a better approach.

To formally deal with this issue, we use the methodology of Maheu and Gordon (2005) and assume that structural breaks are exogenous, unpredictable events that result in a change in the parameter vector associated with the maintained model (in this case a mixture-of-Normals model of excess returns). The structural break model is constructed from a series of submodels. Each submodel has an identical parameterization for excess

[4]A second reason to take the maintained specification of excess returns seriously is that our Bayesian approach provides exact finite sample inference only if the model is well specified.

[5]Additional work on structural breaks in finance include Andreou and Ghysels (2002) and Pettenuzzo and Timmermann (2004).

returns but the parameter is estimated with a different history of data. Each of the 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. Submodels are differentiated by when they start and the data they use. New submodels are continually introduced through time to allow for multiple structural breaks, and for a potential break out-of-sample.

Since structural breaks can never be identified with certainty, Bayesian model 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.[6] Therefore new submodels, which are based on recent shorter histories of data, 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 uncertainty about the premium due to sample length, with the less precise (but unbiased) estimates based on the more recent post-break data. Note that this implies that, in the presence of structural breaks, there does not exist an optimal rolling window estimator.

This approach provides a method to combine submodels estimated over different histories of data. After estimation we can estimate the average number of useful observations at any point in time. In addition, submodel uncertainty is accounted for in the analysis. For example, we show that there is considerable uncertainty as to the number of past observations to use in forecasting the premium toward the end of our sample.

The empirical results provide strong support for structural breaks. In particular, our evidence for structural breaks points towards at least 2 major breaks (1929 and 1940), and possibly a more recent structural break in the late 1990s. 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 1929:11, there is strong evidence (probability .94) that the most recent structural break occurred at 1929:6.

Ignoring structural breaks results at times in substantially different premium forecasts, as well as overconfidence in those estimates. When a structural break occurs there is a decrease in the precision of the premium estimate 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, 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. Differences between premium forecasts which account for structural breaks and those which do not, can be important for many applications. For example, we show that neglecting structural breaks has important implications for a pension fund manager who must finance future liabilities.

Due to the presence of asymmetry and fat tails in excess returns, we favor inference from our structural break model using a mixture-of-Normals submodel with two components. This model produces kurtosis values well above 3 and negative skewness throughout our sample of data. Our statistical measures clearly favor this specification.

[6]Other examples of Bayesian model averaging include Avramov (2002), and Cremers (2002).

Interestingly, the premium forecasts (predictive mean) are quantitatively similar to the structural break model with a single-component submodel. Where they differ is in the shape of the predictive distribution of the premium. In general the two-component model indicates that the predictive distribution of the premium is more disperse. This higher uncertainty associated with the equity premium will be important for investment decisions.

There is another important difference between the alternative parameterizations of the submodel. 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 model 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.

In summary, this paper makes several contributions to the prediction of the equity premium. First, we show that historical data are useful in updating our prior beliefs regarding the equity premium. In the presence of structural breaks, we provide an optimal approach to estimating and forecasting the long-run equity premium using historical data on excess returns. Our structural change model produces realistic forecasts of the premium over the entire 1885-2003 sample. The paper also illustrates the importance of submodel uncertainty and the value of modeling higher-order moments of excess returns when inferring structural breaks and predicting the equity premium. Ignoring structural breaks leads to substantially different premium forecasts as well as overconfidence in the estimates.

The paper is organized as follows. The next section describes the data sources. Section 3 provides an overview of alternative ways to use historical data in order to forecast the equity premium. Included are a case in which all data are used, a fixed-length rolling window of data, and the proposed optimal use of data when structural breaks are taken into account. Section 4 introduces a flexible mixture-of-Normals model for excess returns as our submodel parameterization. Section 5 reviews Bayesian estimation techniques for the mixture model of excess returns. The proposed method for optimal use of data for estimation and forecasting in the presence of structural breaks is outlined in Section 6. Results are reported in Section 7 using data from 1885 to 2003. Conclusions are found in Section 8.

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. It is scaled to an annual excess return by multiplying by 12.

Figure 1 displays a time series plot of the annualized monthly excess returns while Table 1 reports summary statistics for excess returns. Both the skewness and kurtosis estimates suggest significant deviations from the Normal distribution.

3 Forecasting the Equity Premium

We define the long-run equity premium as the expected value of excess returns on a well diversified value-weighted portfolio of securities. In this paper we are concerned with methods of forecasting the long-run equity premium from a series of historical data. If there were no structural breaks, and excess returns were stationary, it would be optimal to use all available data. 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.

To focus on this issue, consider 3 alternative forecasts of the equity premium γ:

$\hat{\gamma}_{ALL,t-1}$ which is based on all available data up to time $t-1$;

$\hat{\gamma}_{W,t-1}$ which is based on a fixed-length rolling window of past data; and

$\hat{\gamma}_{B,t-1}$ uses historical data optimally given the possibility of structural breaks.

The first ignores any structural breaks. Using the average of the entire sample of excess returns is a common example of this approach. The second forecast recognizes that the distribution of excess returns may have undergone a structural break. The method therefore uses a rolling window of historical data for estimation. This has the advantage of dropping past data which may bias the estimate, but with the possible disadvantage of dropping too many data points, resulting in a reduction in the accuracy of the premium estimate. In addition, the second estimator is implicitly assuming that structural breaks are reoccurring by using a fixed window of data at each point in time. The final approach provides optimal use of past data in forecasting the premium. For this estimate, the number of useful data will vary over time and depend on our inference concerning structural breaks.

Section 4 describes our maintained mixture-of-Normals model of excess returns, which is subject to structural breaks. To model the value of historical data for our forecasts of the equity premium, it is natural to use Bayesian methods which stress the learning aspect of statistical inference. That is, how do our beliefs regarding the premium change after observing a set of realizations of excess returns? Section 5 outlines Bayesian estimation of the single-component and the mixture-of-Normals model of excess returns. Once structural breaks are allowed, the usefulness of historical data will be dependent on how recently a break has occurred. Given assumptions about the form of structural breaks, Section 6 provides a methodology to optimally use historical data in this setting. This provides the details of the out-of-sample estimate of $\hat{\gamma}_{B,t-1}$ with comparisons to $\hat{\gamma}_{ALL,t-1}$ and $\hat{\gamma}_{W,t-1}$.

4 Mixture-of-Normals Model for Excess Returns

Financial returns are well known to display skewness and kurtosis and our inference about the market premium may be sensitive to these characteristics of the shape of the distribution. Our maintained model 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)). A k-component mixture model of returns can be represented as

$$r_t = \begin{cases} N(\mu_1, \sigma_1^2) & \text{with probability } \pi_1 \\ \vdots & \vdots \\ N(\mu_k, \sigma_k^2) & \text{with probability } \pi_k, \end{cases} \tag{4.1}$$

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 $N(\mu_j, \sigma_j^2)$. In other words, returns will display heteroskedasticity. Often a two-component specification is sufficient to capture the features of returns. Figure 2 displays examples of excess return distributions that can be obtained from only two components. Relative to the Normal distribution, the distributions exhibit fat-tails, skewness and combinations of skewness and fat-tails

Since our focus is on the moments of excess returns, in particular the mean, it will be useful to consider the implied moments of excess returns as a function of the model parameters. The relationships between the uncentered moments and the model parameters for a k-component model are:

$$\gamma = Er_t = \sum_{i=1}^{k} \mu_i \pi_i, \tag{4.2}$$

in which γ is defined as the equity premium; and

$$\gamma_2^{'} = Er_t^2 = \sum_{i=1}^{k}(\mu_i^2 + \sigma_i^2)\pi_i \tag{4.3}$$

$$\gamma_3^{'} = Er_t^3 = \sum_{i=1}^{k}(\mu_i^3 + 3\mu_i\sigma_i^2)\pi_i \tag{4.4}$$

$$\gamma_4^{'} = Er_t^4 = \sum_{i=1}^{k}(\mu_i^4 + 6\mu_i^2\sigma_i^2 + 3\sigma_i^4)\pi_i. \tag{4.5}$$

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

$$\gamma_2 = \gamma_2^{'} - (\gamma)^2 \tag{4.6}$$

$$\gamma_3 = \gamma_3^{'} - 3\gamma\gamma_2^{'} + 2(\gamma)^3 \tag{4.7}$$

$$\gamma_4 = \gamma_4^{'} - 4\gamma\gamma_3^{'} + 6(\gamma)^2\gamma_2^{'} - 3(\gamma)^4. \tag{4.8}$$

As a special case, a one-component model allows for Normally distributed returns. As shown above, 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 5 discusses a Bayesian approach to estimation of this model.

5 Bayesian Estimation

In the next two subsections we review Bayesian estimation methods for the mixture-of-Normals model. An important special case is when there is a single component $k = 1$ which we discuss first.

5.1 Gaussian Case, $k = 1$

When there is only one component our model for excess returns reduces to a Normal distribution with mean μ, variance σ^2, and likelihood function,[7]

$$p(r|\mu, \sigma^2) = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(r_i - \mu)^2\right) \tag{5.1}$$

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

[7]For the one-component case we drop the component subscript on the model parameters.

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 $\sigma^2 \sim IG(v/2, s/2)$,[8] Bayes rule gives the posterior distribution of μ and σ^2 as

$$p(\mu, \sigma^2|r) \propto p(r|\mu, \sigma^2)p(\mu)p(\sigma^2) \tag{5.2}$$

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.

Our object of interest is the long-run equity premium γ defined as the mean of the excess returns distribution. 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 $\sigma^2 \sim p(\sigma^2|\mu, r)$

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.[9] 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 γ, which is an estimate of the equity premium conditional on this model and data, can be estimated as

$$E[\mu|r_T] \approx \frac{1}{N}\sum_{i=1}^{N}\mu^{(i)}. \tag{5.3}$$

To measure the dispersion of the posterior distribution of the equity premium we could compute the posterior standard deviation of γ 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 model 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 the premium. The next section provides details on estimation for models with two or more components which can capture the higher-order moments of excess returns.

[8]Where $IG(,)$ denotes the inverse gamma distribution. See Bernardo and Smith (2000).

[9]See Chib (2001), Geweke (1997), Robert and Casella (1999) for background information on Markov chain Monte Carlo methods of which Gibbs sampling is a special case. See Johannes and Polson (2005) for a survey of financial applications.

5.2 Mixture Case, $k > 1$

In the case of $k > 1$ mixture-of-Normals the likelihood of excess returns is

$$p(r|\mu, \sigma^2, \pi) = \prod_{t=1}^{T} \sum_{j=1}^{k} \pi_j \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{1}{2\sigma_j^2}(r_t - \mu_j)^2\right) \tag{5.4}$$

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 model 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

$$p(r|\mu, \sigma^2, \pi, Z) = \prod_{t=1}^{T} \sum_{j=1}^{k} z_{t,j} \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{1}{2\sigma_j^2}(r_t - \mu_j)^2\right). \tag{5.5}$$

Bayes theorem now gives the posterior distributions as

$$\begin{aligned} p(\mu, \sigma^2, \pi, Z|r) &\propto p(r|\mu, \sigma^2, \pi, Z)p(\mu, \sigma^2, \pi, Z) && (5.6) \\ &\propto p(r|\mu, \sigma^2, \pi, Z)p(Z|\mu, \sigma^2, \pi)p(\mu, \sigma^2, \pi). && (5.7) \end{aligned}$$

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 $\mu \sim p(\mu|\sigma^2, \pi, Z, r)$
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.

Below we detail each of the Gibbs sampling steps. Conditional on z_t we can recast the model as

$$r_t = z_t\mu + u_t, \quad u_t \sim N(0, z_t\sigma^2) \tag{5.8}$$

To jointly sample from the conditional distribution of μ using Gibbs sampling results for the linear regression model, we transform to a homoskedastic model as in

$$y_t = x_t\mu + v_t, \quad v_t \sim N(0, 1) \tag{5.9}$$

with $y_t = r_t/\sqrt{z_t\sigma^2}$, $x_t = z_t/\sqrt{z_t\sigma^2}$. Now the conditional posterior of μ is multivariate normal and a draw is obtained as

$$\begin{aligned} \mu &\sim N(M, V^{-1}) && (5.10)\\ M &= V^{-1}(X^T y + B^{-1}b) && (5.11)\\ V &= X^T X + B^{-1}. && (5.12) \end{aligned}$$

where $b = [b_1 \cdots b_k]^T$, B is a matrix of zeros with diagonal terms B_{ii}, y_t is a row of the vector y, and x_t is a row vector of the matrix X. The conditional posterior of σ_j^2 is,

$$\sigma_j^2 \sim IG\left(\frac{v_j + T_j}{2}, \frac{\sum_{t=1}^{T}(r_t - \mu_j)^2 z_{t,j} + s_j}{2}\right), \quad j = 1, ..., k. \tag{5.13}$$

where $T_j = \sum_{t=1}^{T} z_{t,j}$. Only the observations attributed to component j are used to update the variance σ_j^2.

With the conjugate prior for π, we sample the component probabilities as,

$$\pi \sim \mathcal{D}(\alpha_1 + T_1, ..., \alpha_k + T_k). \tag{5.14}$$

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

$$p(z_{t,i}|r, \mu, \sigma, \pi) \propto \pi_j \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{1}{2\sigma_i^2}(r_t - \mu_i)^2\right), \quad i = 1, ..., k, \tag{5.15}$$

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 (4.3)-(4.8) 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.

5.3 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. This is defined as

$$p(r|M_i) = \int p(r|\mu, \sigma^2, \pi, M_i)p(\mu, \sigma^2, \pi|M_i)d\mu d\sigma^2 d\pi \tag{5.16}$$

where M_i indexes a particular model. For the class of models considered in this paper we can calculate an estimate of this marginal likelihood using output from the posterior simulator. The Bayes factor for model M_0 versus model M_1 is defined as $BF_{01} = p(r|M_0)/p(r|M_1)$. A Bayes factor greater than one is evidence that the data favor M_0. Kass and Raftery (1995) summarize the support for M_0 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.

6 Optimal Use of the Data

6.1 Accounting for Structural Breaks

In this section we outline a method to deal with potential structural breaks. 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. To formally deal with this, we follow the methodology of Maheu and Gordon (2005) and assume that structural breaks are exogenous unpredictable events that result in a change in the parameter vector associated with the maintained model, in this case a mixture-of-Normals model of excess returns.

The structural break model is constructed from a series of identical parameterizations (mixture-of-Normals, 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. 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. Structural breaks are identified by the probability distribution on submodels. Since breaks are permitted out-of-sample, new submodels are continually introduced through time. As more data arrives, the posterior density of the submodel parameter is updated from its prior. This allows for an increasing number of structural breaks through time.

Submodels are differentiated by when they start and the number of data points they use. Since structural breaks can never be identified with certainty, Bayesian model 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 uncertainty about the premium 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, and assess how many historical observations should be used to estimate the premium at any point in time.

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 $I_t = I_{1,t}$. Let M_i be a submodel that assumes a structural break occurs at time i.[10] As we have mentioned, under our assumptions the data $r_1, ..., r_{i-1}$ are not informative about the submodel parameter due to the structural break, 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 for submodel M_i, given θ, and the information set $I_{i,t-1}$. Now consider the situation where we have the data I_{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 $I_{i,t-1}$ has the form,

$$p(\theta|I_{i,t-1}, M_i) \propto \begin{cases} p(r_i, ..., r_{t-1}|\theta, M_i)p(\theta|M_i) & i < t \\ p(\theta|M_i) & i = t, \end{cases} \tag{6.1}$$

$i = 1, ..., t$. In the first case, only data after the assumed break at time $i-1$ are used. 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.[11] For instance, 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. Thus M_t assumes a break occurs out-of-sample, in which case, past data is not useful. In the usual way the predictive density for submodel M_i is

[10] The exception to this is the first submodel of the sample M_1 for which there is no prior data.

[11] In our application, submodels are differentiated only by the assumption of when a break occurred. In addition to this, it is possible to allow for different families of submodels. However, there may not be a common interpretation of θ among different specifications.

formed by integrating out the parameter uncertainty,

$$p(r_t|I_{i,t-1}, M_i) = \int p(r_t|I_{i,t-1}, \theta, M_i)p(\theta|I_{i,t-1}, M_i)d\theta, \quad i = 1, ..., t. \tag{6.2}$$

For M_t the posterior is the prior under our assumptions.

Up to this stage we have said nothing about how to combine these submodels. First note that the usual Bayesian methods of model comparison and combination are based on the marginal likelihood of a common set of data. This cannot be used to compare the submodels $\{M_i\}_{i=1}^t$, since they are based on different histories of data. Therefore we require a new method to combine the submodels. In keeping with our interpretation of a structural break, we assume the occurrence of past structural breaks does not indicate anything about the occurrence of future structural breaks.[12] As such, we only have a subjective prior on the likelihood of a break.[13]

Consistent with this, the financial analyst places a subjective prior $0 \leq \lambda_t \leq 1$, $t = 1, ..., T$ 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.

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. We can obtain the predictive density using (6.2) which gives $p(r_1|I_0) = p(r_1|I_0, M_1)$ and, after observing r_1, we have $P(M_1|I_1) = 1$. Now allow for a break at $t = 2$, with $\lambda_2 \neq 0$, the predictive density is the mixture

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

The first term is the predictive density using all 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.[14] After observing r_2 we can update submodel probabilities,

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

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

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

In words, this is (predictive density assuming no break at $t = 3$)×(probability of no break at $t = 3$) + (predictive density assuming a break at $t = 3$)×(probability of a

[12] If we assumed past breaks told us something about future breaks, then λ_t could be estimated as a function of past data. We do not pursue this extension in this paper.

[13] Non-sample information may be important in forming the prior on breaks.

[14] Recall that in the second density $I_{2,1} = \{\emptyset\}$.

break at $t = 3$). Once again $p(r_3|I_{3,2}, M_3)$ is derived from the prior. The updated submodel probabilities are

$$P(M_1|I_3) = \frac{p(r_3|I_{1,2}, M_1)p(M_1|I_2)(1-\lambda_3)}{p(r_3|I_2)} \tag{6.3}$$

$$P(M_2|I_3) = \frac{p(r_3|I_{2,2}, M_2)p(M_2|I_2)(1-\lambda_3)}{p(r_3|I_2)} \tag{6.4}$$

$$P(M_3|I_3) = \frac{p(r_3|I_{3,2}, M_3)\lambda_3}{p(r_3|I_2)}. \tag{6.5}$$

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 3 which displays a set of submodels available at $t = 10$, where the horizontal lines indicate the data used in forming the posterior. The forecasts from each of these submodels, which use different data, are combined (the vertical line) using the model probabilities. M_{11} represents the prior 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 arrives and once their predictions warrent it they receive larger weights in the model average. Conversely, old submodels will only change slowly when a structural break occurs. Their predictions will still be dominated by the longer and older data prior to the structural break.

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

$$p(r_t|I_{t-1}) = \left[\sum_{i=1}^{t-1} p(r_t|I_{i,t-1}, M_i)p(M_i|I_{t-1})\right](1-\lambda_t) + p(r_t|I_{t,t-1}, M_t)\lambda_t. \tag{6.6}$$

The first term on the RHS of (6.6) 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 this case, past data are not useful and only the prior density is used to form the predictive distribution. The terms $p(M_i|I_{t-1})$, $i = 1, ..., t-1$ are the submodel probabilities, representing the probability of a break at time i give information I_{t-1}, and are updated each period after observing r_t as

$$p(M_i|I_t) = \begin{cases} \frac{p(r_t|I_{i,t-1},M_i)p(M_i|I_{t-1})(1-\lambda_t)}{p(r_t|I_{t-1})} & 1 \le i < t \\ \frac{p(r_t|I_{t,t-1},M_t)\lambda_t}{p(r_t|I_{t-1})} & i = t. \end{cases} \tag{6.7}$$

In addition to being inputs into (6.6) 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 model $M_{t'}$ 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 model probabilities must be built up sequentially from $t = 1$ and updated as a new observation becomes available. At any given time, a posterior moment $g(\theta)$ which accounts for past structural breaks can be computed as,

$$E[g(\theta)|I_t] = \sum_{i=1}^{t} E[g(\theta)|I_{i,t}, M_i]p(M_i|I_i). \tag{6.8}$$

This is an average at time t of the model-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

$$E[g(r_{t+1})|I_t] = \left[\sum_{i=1}^{t} E[g(r_{t+1})|I_{i,t}, M_i]p(M_i|I_t)\right](1-\lambda_{t+1}) + E[g(r_{t+1})|I_{t+1,t}, M_{t+1}]\lambda_{t+1} \tag{6.9}$$

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 give a break at time $t+1$.

In this paper, our main concern is with the equity premium. Using the mixture-of-Normals specification as our submodel with k fixed, this 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. 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,

$$p(\gamma|I_{t-1}) = \left[\sum_{i=1}^{t-1} p(\gamma|I_{i,t-1}, M_i)p(M_i|I_{t-1})\right](1-\lambda_t) + p(\gamma|I_{t,t-1}, M_t)\lambda_t. \tag{6.10}$$

This equation is analogous to the predictive density of returns (6.6). From the Gibbs sampling output for each of the models we can compute the mean of the predictive distribution of the equity premium as,

$$E[\gamma|I_{t-1}] = \left[\sum_{i=1}^{t-1} E[\gamma|I_{i,t-1}, M_i]p(M_i|I_{t-1})\right](1-\lambda_t) + E[\gamma|I_{t,t-1}, M_t]\lambda_t. \tag{6.11}$$

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.

We can now clarify two of the estimators discussed in Section 3. Recall that $\hat{\gamma}_{ALL}$ uses all available data (submodel M_1) while $\hat{\gamma}_B$ optimally uses data after accounting for structural breaks. These are,

$$\hat{\gamma}_{ALL,t-1} = E[\gamma|I_{t-1}, M_1] \tag{6.12}$$
$$\hat{\gamma}_{B,t-1} = E[\gamma|I_{t-1}] \tag{6.13}$$

where the latter estimator integrates out all model uncertainty surrounding structural breaks through (6.11).

Finally, after estimation we can provide an estimate of the number of historical observations that are used at any given time to estimate the excess return distribution and hence the equity premium. Since submodels M_i define the time of a break, if a break occurs at $i < t$ we would only want to use the $(t - i + 1)$ data points $r_i, r_{i+1}, ..., r_t$ after the break to estimate the premium. In practice, we do not know with certainty when a break occurs. However, we can use the submodel probabilities to infer the *mean useful observations* (MUO_t) defined as

$$\text{MUO}_t = \sum_{i=1}^{t} (t - i + 1) p(M_i|I_t). \tag{6.14}$$

A time series plot of MUO_t against time will indicate the number of useful historical observations at each point in time. If there are no structural breaks, we would expect MUO_t to follow the 45 degree line. In situations when breaks have been inferred, the MUO_t may dip substantially below the 45 degree line.

6.2 Calculations

Estimation of each submodel at each point in time follows the Gibbs sampler detailed in Section 5. 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 submodel probabilities to form an out-of-sample forecast of the equity premium using (6.11). To calculate the marginal likelihood of a submodel, following Geweke (1995) we use a predictive likelihood decomposition,

$$p(r_i, ..., r_t|M_i) = \prod_{j=i}^{t} p(r_j|I_{i,j-1}, M_i). \tag{6.15}$$

Given a set of draws from the posterior distribution $\{\theta^{(i)}\}_{i=1}^{N}$, where $\theta^{(i)} = \{\mu_1, ..., \mu_k, \sigma_1^2, ..., \sigma_k^2, p_1, ..., p_k\}$, for submodel M_i, conditional on $I_{i,t-1}$, each of the individual terms in (6.15) can be estimated consistently as[15]

$$p(r_t|I_{i,t-1}, M_i) \approx \frac{1}{N} \sum_{i=1}^{N} p(r_t|\theta^{(i)}, I_{i,t-1}, M_i). \tag{6.16}$$

[15]This method of estimating the predictive likelihood provides accuracy similar to other methods such as Gelfand and Dey (1994).

This is calculated at the end of each Gibbs run, along with features of the predictive density, such as premium forecasts for each submodel. For the mixture-of-Normals specification, the data density is,

$$p(r_t|\theta^{(i)}, I_{i,t-1}, M_i) = \sum_{j=1}^{k} p_j \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{1}{2\sigma_j^2}(r_t - \mu_j)^2\right). \qquad (6.17)$$

The predictive likelihood of submodel M_i is used in (6.7) 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 (6.6) and hence the marginal likelihood of the structural break model as,

$$p(r_1, ..., r_t) = \prod_{j=1}^{t} p(r_j|I_{j-1}). \qquad (6.18)$$

6.3 Selecting Priors on the Premium

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 equity premium. Theory indicates the premium must be positive and from the wide range of estimates Derrig and Orr (2004) survey 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)).

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.[16] 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 proper informative priors.

A second issue is the elicitation of priors in the mixture model. 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 model 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 (2003)). That is, we analyze moments of excess returns simulated from the submodel. We repeat the following steps

[16] 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.

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 summary statistics for the first four moments of excess returns from repeating the steps 1–3 many times. The prior associated with these results is listed in the second panel 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 model with this prior is also consistent with a wide range of skewness and excess kurtosis. In selecting a prior for the single-component model we tried to match, as far as possible, the features of the two-component model. This prior is listed in the top panel of Table 2. 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. Lastly, we set the probability of a break $\lambda_t = 0.01$. This favors infrequent breaks and allows the 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.[17] That is, our benchmark prior introduces a new submodel only every 12 months with $\lambda_t = 0.01$ and otherwise set $\lambda_t = 0$. This implies an expected duration of 100 years between structural breaks in the equity premium. We discuss other results for different specifications in the next section.

7 Results

This section discusses the out-of-sample model forecasts for the equity premium starting from the first observation to the last. First, we present results for a one component mixture submodel, and then in subsection 7.1 results for a two component mixture submodel. A summary of the model specifications, including priors, is reported in Table 3. The main results for the one-component specification are found in Figures 4 to 6, panel A of Figures 7 to 9, and Figure 10.

The out-of-sample forecasts of the equity premium from the one-component specification are found in Figure 4. For comparison purposes, the mean of the predictive distribution of the premium is displayed for both the structural break model and a no-break alternative. These are the forecasts $\hat{\gamma}_{B,t-1}$, computed from equation (6.13) which optimally uses past data, and $\hat{\gamma}_{ALL,t-1}$, computed from equation (6.12) using all available data at time $t-1$. The premium forecasts are similar until the start of the 1930s where

[17] 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).

they begin to diverge. Thereafter, the premium from the structural break model rises over the 1950s and 1960s with a maximum value of 8.23 in 1962:1. Toward the end of the sample the premium decreases to values lower than the no-break model. The final premium forecast at the end of the sample is 3.53 for the structural break model and 4.65 for the no-break model.

The second panel of this figure 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 premium's predictive distribution from the break model shows that this increased confidence is misleading if structural breaks occur. As the second panel of Figure 4 illustrates, when a break occurs our uncertainty about the premium increases.

Figure 5 plots the mean and standard deviation of the posterior distribution of submodels for each date. Note that the standard deviation is a measure of submodel uncertainty, one of the two sources of uncertainty about the premium. Recall that submodels are indexed with the time period they start at, and their submodel probabilities identify the most recent structural break. Therefore, for any time period, there is a discrete probability distribution of possible submodels defined through (6.7). The mean and standard deviation of this distribution of submodels are

$$\text{mean}_t = \sum_{i=1885}^{t} iP(M_i|I_t); \quad \text{stdev}_t = \sqrt{\sum_{i=1885}^{t} i^2 P(M_i|I_t) - \text{mean}_t^2}. \tag{7.1}$$

These moments are calculated for each time t given the information set I_t. This calculation is repeated from the start of the sample to the end, and represents the inference that is available in real time.

There is a gradual increase in submodel uncertainty, measured by the standard deviation of the posterior distribution of submodels, starting in 1891 and a subsequent lowering after the 1930s and 1940s. It is interesting to note that in the early 1930s it takes less than one year for the uncertainty to drop by 97% from the highest levels in 1929. This indicates decisive evidence of the most recent structural break identified at 1929:6 and very fast learning about this change.[18] This is supported by the fact that the posterior mean of the submodel distribution jumps to the 1929 submodel at this time. There is a small increase in uncertainty during the 1930s but the posterior mean centers

[18]Therefore, the increase in the total uncertainty about the premium after 1929, shown in Figure 4:B, is mainly due to parameter uncertainty.

the distribution around the 1940 submodel until 1969 after which there is an increase in submodel uncertainty.

Figures 7 to 9 display the submodel probabilities through time for three different subperiods for the one-component specification ($k = 1$) in panels A. Figure 6 shows the probability of some selected submodels over time. These correspond to a slice through the submodel axis in panels A of Figures 7 to 9. The latter are 3-dimensional plots of (6.7) which is the probability of the most recent break point given data up to time t. The axis labelled Submodel M_i refers to the submodels identified by their starting observation i. Recall that the number of submodels is increasing with time, with a new submodel introduced every 12 months. The submodel probabilities at a point in time can be seen as a perpendicular line from the Time axis.

As shown in panel A of Figure 7, in the early part of the sample the first submodel, 1885, has probability close to 1. There was some preliminary evidence of a break early in the sample. For example, by 1902, that is, using data from 1885 to 1902, the first submodel M_{1885} received a probability of only 0.24 while submodel M_{1893} had a probability of 0.51. However, by 1907 the evidence for a break in 1893 diminished to 0.078, while the original submodel M_{1885} strengthened to 0.64. 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.

The first submodel of the sample, M_{1885} continues to receive most of the support in the 1910s and 1920s until 1929. As previously mentioned, there is very strong evidence of a structural break at 1929:6. This submodel has a probability of 0.94 based on data to 1929:11 which indicates fast learning about a change in the distribution of excess returns. The change in regime during this time and the subsequent crash in October of 1929 is likely identified as a sharp increase in volatility. As shown in Figure 4, during the 1930s the premium forecast is very similar to the no-break model, suggesting that the identified break in the excess return distribution in 1929 is due to higher-order moments such as volatility.

As mentioned previously, there is an increase in submodel uncertainty during the 1930s. Using data up to 1937, there is some evidence of a break in 1934[19] and in 1937. However, the next major break occurs in 1940. Until 1974, this submodel receives most of the weight with a probability for most of the time in excess of 0.90.[20] As shown in Figure 4, the 1940 structural break results in clear differences in the equity premium forecasts for the break and no-break models. Accounting for structural breaks indicates a larger equity premium after 1940 and more uncertainty about the premium. Note that by the mid-1950s the premium is almost double that obtained from the no-break model.

In the early 1970s there is weak evidence of a break in 1969, however, this subsequently declines during the mid-1970s, while the evidence for M_{1940} strengthens. By the mid-1970s there is uncertainty about submodels associated with 1969, 1973, and 1974,

[19] $M_{1934:6}$ has probability of 0.77 using data to 1937:6

[20] By 1969:5 the submodel still has a probability of 0.94.

which all receive significant support. By the mid-1980s we have learned that the most likely point of a break was 1969.[21] The strength of evidence for the 1969 submodel as the most recent break point is about 0.5 for the whole decade of the 1980s.

During the latter part of the 1990s there is some evidence of a break at 1988, with weaker evidence for the most recent break at 1991 and 1992. By the end of the sample the results support a recent break occurring sometime from 1996-1998 with the submodels M_{1996}, M_{1997}, and M_{1998}, possessing a combined probability of 0.77. In summary, we identify major breaks in 1929 and 1940, with weaker evidence for structural breaks in 1969 and 1988, and possibly a recent break in 1996-98.

Our results highlight several important points. First, the identification of structural breaks in the premium depends on the data used, and false assessments may occur which are later revised when more data become available.[22] This is an important aspect of learning about structural breaks. Second, our evidence of submodel uncertainty indicates the problem with using only one submodel. In a setting of submodel risk, the optimal approach is to model average as done in (6.11). There is overwhelming evidence for the structural break specification as measured by the marginal likelihood values found in Table 3 for the one-component models. A Bayes factor for the break model against the no-break model is around exp(155).

Finally, our discussion suggests that to forecast the premium we should not use all the data equally. The mean useful observations are displayed in Figure 10. The 45-degree line is the model that uses all data. Consistent with our discussion, the structural break model uses most of the data until around 1930 where the number of useful observations drops dramatically. Around 1940 the useful observations begin to steadily increase till further declining in the 1970s, 80s and 90s. In this figure, a rolling window model would be represented as a horizontal line. For example, a rolling window premium 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.

7.1 Robustness

We now turn to the two-component submodel. Recall that this specification allows for higher-order moments in the distribution of excess returns. The results for this specification are found in panels B of Figures 7 to 9 and in Figures 11 to 13. The predictive mean for the equity premium, the standard deviation of the predictive distribution, and the mean useful observations are all broadly consistent with the one-component results. The two-component specification also identifies breaks in 1929 and 1940, and agrees with the previous analysis concerning a recent break in the late 1990s.

Table 3 records the marginal likelihood values of each of the models with and without

[21]For instance, M_{1968}, M_{1969}, M_{1973}, and M_{1974} receive probabilities of 0.13, 0.48, 0.06, and 0.01, respectively, based on data up to 1985:1.

[22]However, this false assessment of a structural break is still the optimal result given the data at hand.

breaks. Both the $k = 1$ and the $k = 2$ specifications provide strong evidence of structural breaks. However, the two-component break model has a log marginal likelihood value about 20 points larger than the one-component break model. According to the criteria in Section 5.3, this is very strong support for the two-component specification.

Figure 13 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. Since the skewness estimates are all less than zero and the kurtosis estimate is always greater than 3, there is clear evidence of higher-order moments that are inconsistent with the one-component specification for excess returns.

Panel B of Figures 7 to 9 display the submodel probabilities through time for the two-component specification. Note that this richer specification is much more decisive in favor of the 1885 submodel than the one-component version in panel A of Figure 7. Figure 9 also suggests that the simpler one-component specification tends to put more weight on more recent submodels. As mentioned earlier, these differences could be due to the fact that the two-component specification is more robust to fat tails (outliers) that, particularly with short samples, can be temporarily identified as probable structural breaks in the more restrictive one-component specification.

The modeling of asymmetries and fat tails results in some differences in submodel probabilities, and hence premium forecasts, mainly near the end of the sample. A comparison of the posterior mean and standard deviation of the distribution of submodels through time for $k = 1$, and $k = 2$ is shown in Figure 14. Both specifications are similar until the 1980s. Here the two-component specification always gives more probability to the 1940 submodel in the range of 0.04-0.15, while the one-component version essentially dismisses this from consideration and weights the submodel associated with 1969 much higher. In the 1990s, the probability of submodel 1940 increases steadily, so that by 1999 M_{1940} has a probability of 0.503.[23] The two-component specification, which can better accommodate outliers by capturing the fat tails and asymmetries in returns, places much more weight on submodel M_{1940}. This example underscores the importance of accurately modeling financial returns prior to an analysis of structural breaks.[24] There is still submodel uncertainty at the end of the sample consistent with a recent structural break. The final significant submodel probabilities, based on the full sample of data, are $M_{1940:6}$ 0.11, $M_{1998:6}$ 0.17, $M_{1999:6}$ 0.16, and $M_{2000:6}$ 0.14. The probability of a break in 1998-2000 is 0.47. The final forecast for the long-run equity premium, which averages over these submodels, is 4.02 percent.

As a further check on our results, Table 3 reports the marginal likelihood values for models which only allow for a structural break every 5 years as opposed to every year. The results favor allowing for structural breaks more frequently.

For the reasons discussed, we favor the structural break model with two-component mixture submodels as our preferred model in forecasting the premium. Our final compar-

[23]Submodel M_{1940} is not displayed in Figure 9.

[24]In other words, misspecified models may provide evidence of structural breaks when the underlying DGP is stable.

ison of the premium estimates from the alternative specifications is shown in Figure 15. Except for the end of the sample, the premium estimates are similar. However, other features of the predictive distribution of the premium do differ. For example, compare the standard deviations in panel B of Figures 4 and 11.

Also included in this figure is a 10-year rolling window based on the sample average. As we discussed above, and as shown in Figure 12, this ad hoc approach to dealing with structural breaks is nowhere optimal for the time period we consider. In addition, the simple rolling-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%.

Although our figures show large differences in the premium forecasts with and without breaks, a natural question is how important these differences are for economic questions. As a simple example, consider a pension fund manager who must make a payment of \$1 twenty years from now. How much does the manager need to invest today in order to expect to meet this future liability? Based on current information, and assuming a zero riskfree rate, the investment required today is $E_t[1/(1+\gamma)^{20}]$, where the expectation is taken with respect to the predictive density of the equity premium at each point in time.[25] This is calculated by taking 1000 draws from the predictive distribution of the premium γ and calculating $1/(1+\gamma)^{20}$ for each. The average of these is the expected required investment. Figure 16 displays the required investment by the pension fund manager for each month through the whole sample for both models. Changes in the nobreak estimate only reflect learning about the model parameter as new data arrives while changes in the break model estimate reflect both learning about model parameters and structural breaks. In general, the shape of the predictive density for the premium affects the calculation of the required investment. This figure shows considerable differences after the first major break in 1929. For example, in 1950:1 the pension fund manager would need to invest 28% less under the structural break model to meet future liabilities.

Finally, it may be that structural breaks only affect the variance of excess returns. To better allow past data to contribute to premium forecasts after a structural break in volatility, we set the prior parameters for the premium in the one component specification to the previous posterior mean and variance of γ when a new submodel is introduced. Therefore, 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. The main difference in the premium forecasts for this case was that the premium was less variable and close to 6% from 1960 on, with a reduced standard deviation of the predictive distribution. 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.

[25] Recall that the forward looking predictive density of the premium allows for breaks out-of-sample.

8 Conclusion

This paper makes several contributions to forecasting the long-run equity premium. First, we show that historical data are useful in updating our prior beliefs regarding the equity premium. In the presence of structural breaks, we provide an optimal approach to estimating and forecasting the equity premium using historical data on excess returns. Our evidence for structural breaks is strong and points toward at least 2 major breaks and possibly a more recent structural break. The paper has also shown the importance of submodel risk and the value of modeling higher-order moments of excess returns when inferring structural breaks and predicting the equity premium. Ignoring structural breaks leads to different premium estimates as well as overconfidence in the estimates.

Due to the presence of asymmetry and fat tails in excess returns, our statistical evidence clearly favors a mixture-of-Normals submodel specification with two components for the unconditional premium. For instance, the structural break model produces kurtosis values well above 3 and negative skewness throughout our sample of data. Interestingly, the premium forecasts (predictive mean) from the two-component model are quantitatively similar to the single-component model. Where they differ is in the shape of the predictive distribution of the premium. In general the two-component specification indicates that the predictive distribution of the premium is more disperse. This higher uncertainty associated with the equity premium will be important for investment decisions.

There is another important difference between the alternative specifications of the maintained submodel for the long-run equity premium. 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 two-component submodel is more robust to these *false breaks*. One reason for this is that the two-component model is characterized by a high and low variance state. This allows for heteroskedasticity in excess returns. Therefore temporary outliers can be consistent with the maintained model and not evidence of a break in the distribution of excess returns.

Our evidence shows at least 2 major breaks (1929 and 1940), and possibly a more recent structural break in the late 1990s. We explicitly characterize the uncertainty with regard to break points which is clearly evident in our 3-dimensional plots (Figures 7 to 9) of the distribution of submodels.

Our model produces realistic forecasts of the premium over the entire 1885-2003 sample. The premium forecasts for the no-break and break alternatives are similar until the start of the 1930s where they begin to diverge. This divergence reflects the fact that the break model uses historical data optimally when breaks occur. In fact, the usefulness of historical data varies considerably over the sample. The premium from the structural break model rises over the 1950s and 1960s with a maximum value of 8.99 in 1961:12. Toward the end of the sample the premium decreases to values lower than the no-break model. The final premium forecast at the end of the sample is 4.02 for the structural break model and 5.10 for the no-break model.

Table 1: Summary Statistics for Annualized 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
γ	0.0369	0.0354	0.0320	(-0.0238, 0.1007)
γ_2	0.5808	0.5056	0.3312	(0.1519, 1.1786)
$\gamma_3/\gamma_2^{3/2}$	-0.3878	-0.3077	0.4718	(-1.4077, 0.3534)
γ_4/γ_2^2	8.1369	6.4816	5.9317	(2.7169, 18.7218)

This table reports summary measures of the empirical moments from the mixture model $k = 2$, when parameters are simulated from the prior distribution. 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: Structural Break Model Specifications and Results

model	breaks	prior	log(ML)
$k=1$	$\lambda_t = 0$ none	$b = 0.03, B = 0.03^2$ $v = 9.0, s = 4.0$	-1371.22
$k=1$	$\lambda_t = 0.01$ every 5 years, otherwise $\lambda_t = 0$	$b = 0.03, B = 0.03^2$ $v = 9.0, s = 4.0$	-1235.33
$k=1$	$\lambda_t = 0.01$ every year, otherwise $\lambda_t = 0$	$b = 0.03, B = 0.03^2$ $v = 9.0, s = 4.0$	-1216.08
$k=2$	$\lambda_t = 0$ none	$b_1 = 0.05, b_2 = -0.30, B_{11} = 0.03^2, B_{22} = 0.05^2$ $v_1 = 10.0, s_1 = 3, v_2 = 8.0, s_2 = 20.0$ $\alpha_1 = 7, \alpha_2 = 1$	-1241.09
$k=2$	$\lambda_t = 0.01$ every 5 years, otherwise $\lambda_t = 0$	$b_1 = 0.05, b_2 = -0.30, B_{11} = 0.03^2, B_{22} = 0.05^2$ $v_1 = 10.0, s_1 = 3, v_2 = 8.0, s_2 = 20.0$ $\alpha_1 = 7, \alpha_2 = 1$	-1202.01
$k=2$	$\lambda_t = 0.01$ every year, otherwise $\lambda_t = 0$	$b_1 = 0.05, b_2 = -0.30, B_{11} = 0.03^2, B_{22} = 0.05^2$ $v_1 = 10.0, s_1 = 3, v_2 = 8.0, s_2 = 20.0$ $\alpha_1 = 7, \alpha_2 = 1$	-1196.30

This tables displays the number of components k, in the mixture model, the prior specification of the submodel parameters as well as the prior on the occurrence of structural breaks λ_t. Finally, the logarithm of the marginal likelihood is reported for all specifications based on the full sample of observations used in estimation.

Figure 1: Annualized Monthly Excess Returns

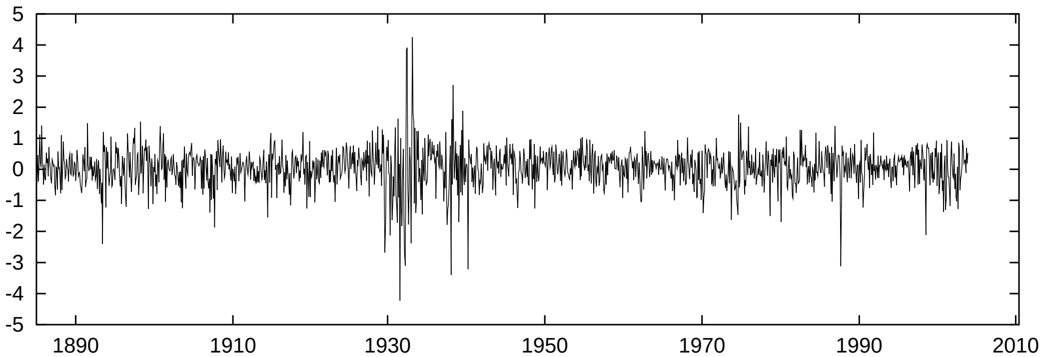
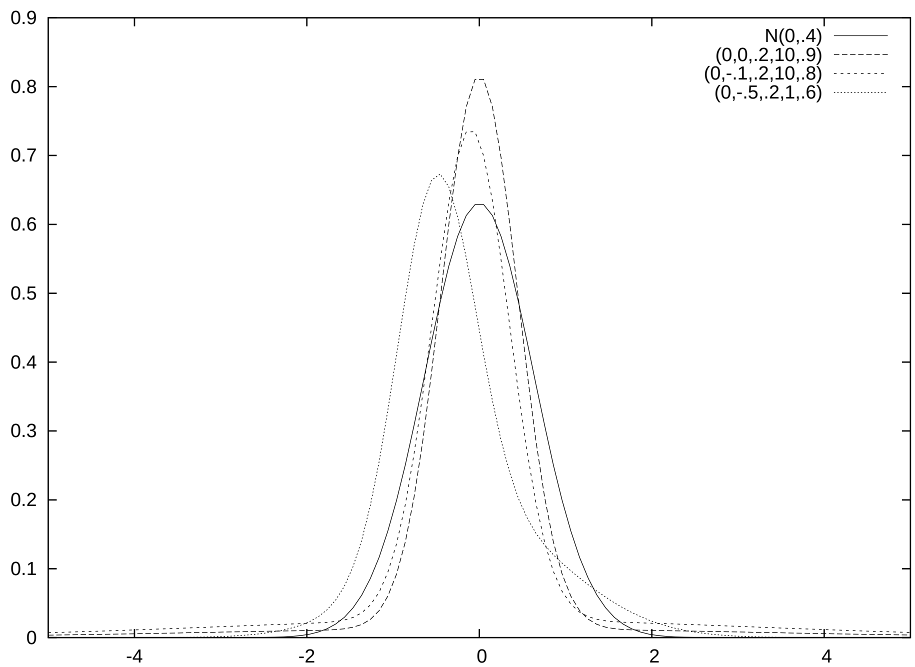


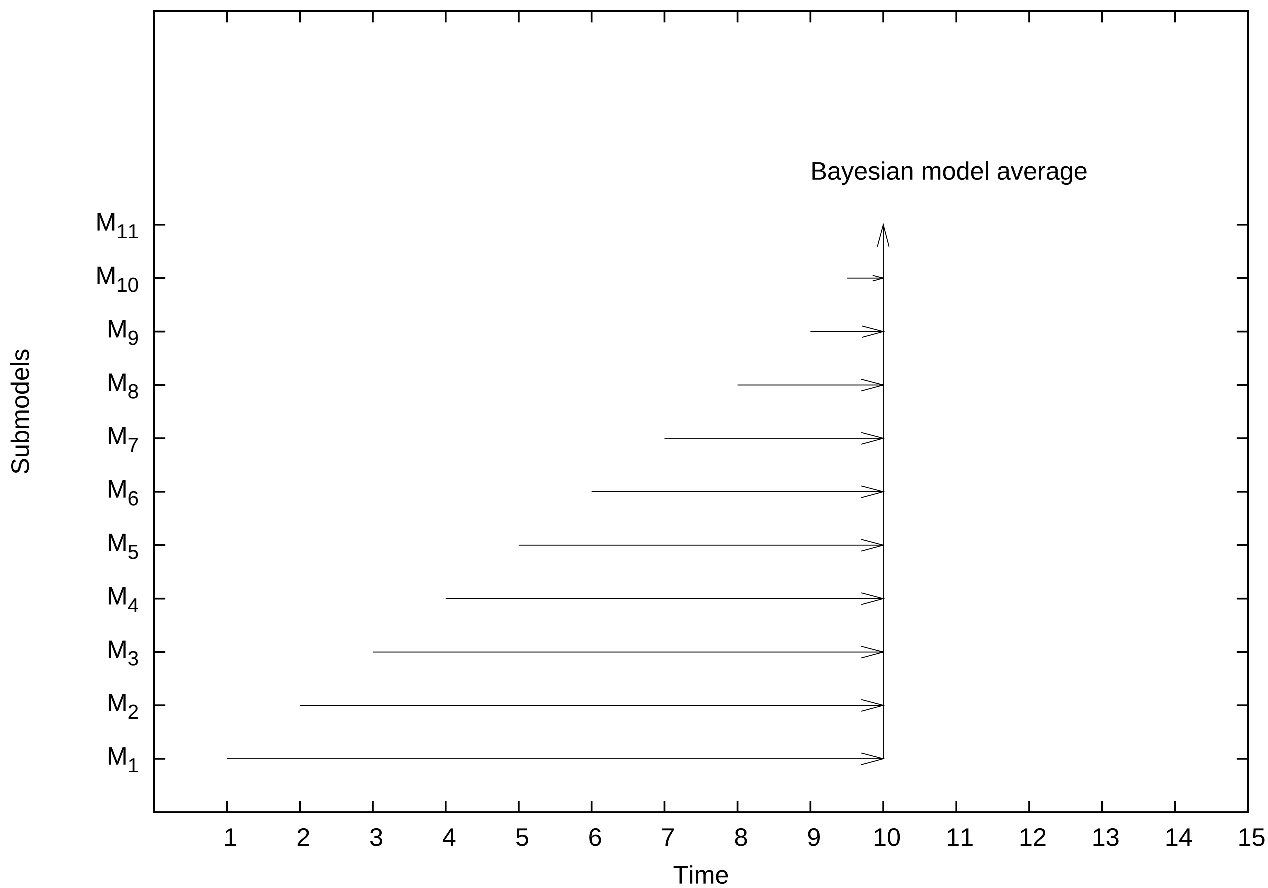
Figure 2: Some Examples of the Distribution From a Two-Component Mixture



This figure displays the density from various configurations of a mixture of two Normal densities. The parameters are $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, p_1)$ and correspond to the submodel in Section 4. Given the parameters the density is

$$\frac{p}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{1}{2\sigma_1^2}(r_t - \mu_1)^2\right) + \frac{1-p}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{1}{2\sigma_2^2}(r_t - \mu_2)^2\right).$$

Figure 3: Individual Submodels and the Bayesian Model 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 (6.6). 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 4: Premium Forecasts through Time, $k = 1$.

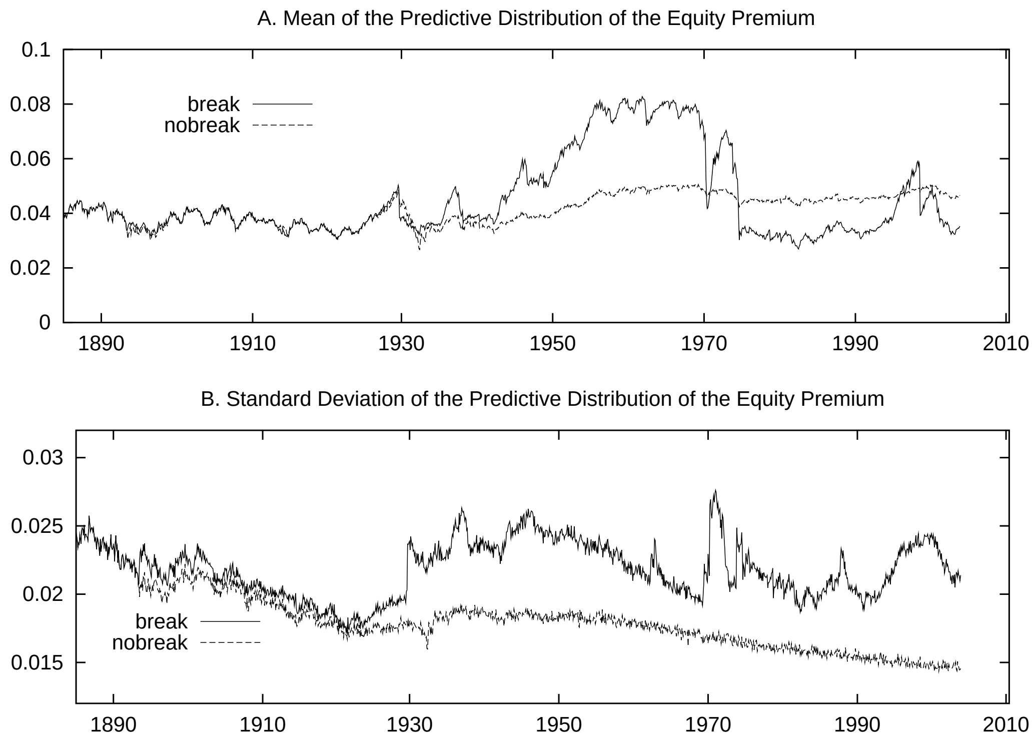
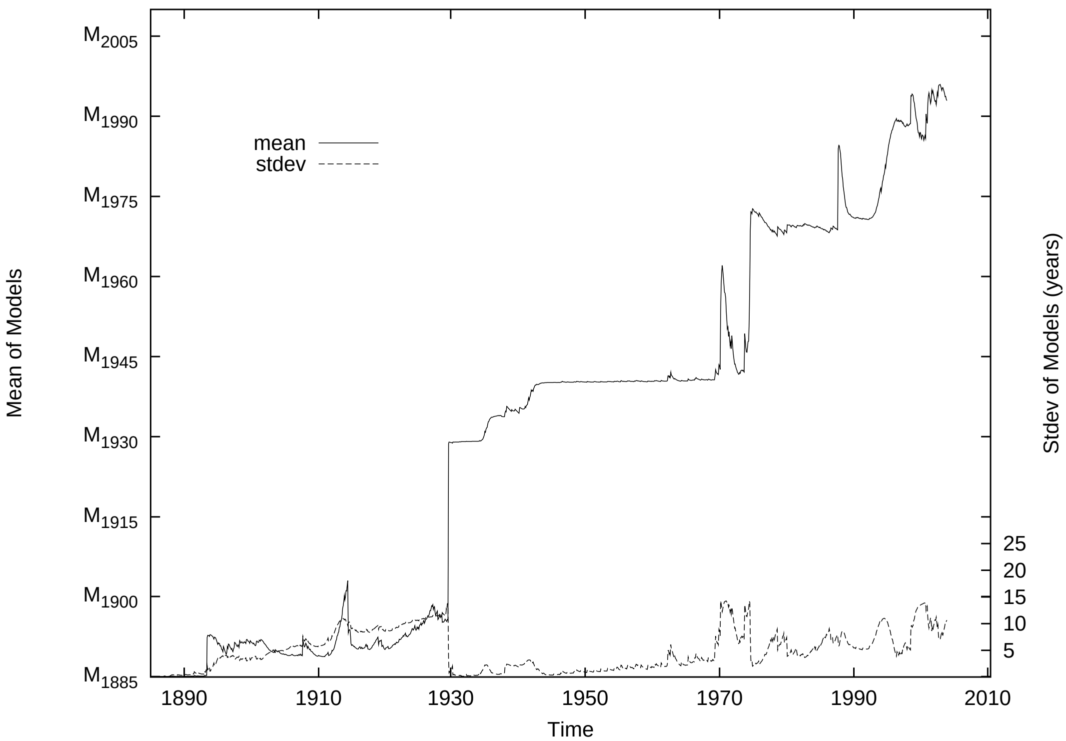


Figure A displays the out-of-sample forecasts (predictive mean) of the equity premium period by period for both the structural break model and the no break alternative. Figure B displays the corresponding standard deviation of the predictive distribution of the equity premium.

Figure 5: Posterior Mean and Standard Deviation of the Distribution of Submodels



This figure displays the posterior mean and the standard deviation of the distribution of submodels at each point in time. The moments are calculated from (6.7) for each observations $t = 1885:2 - 2003:12$, based on data up to and including t. The moments are

$$\text{mean}_t = \sum_{i=1885}^{t} iP(M_i|I_t); \quad \text{stdev}_t = \sqrt{\sum_{i=1885}^{t} i^2 P(M_i|I_t) - \text{mean}_t^2}$$

Submodels are indexed by the calendar time when they begin. The mean of the distribution of submodels is displayed on the vertical axis.

Figure 6: Submodel Probabilties over Time, $k = 1$

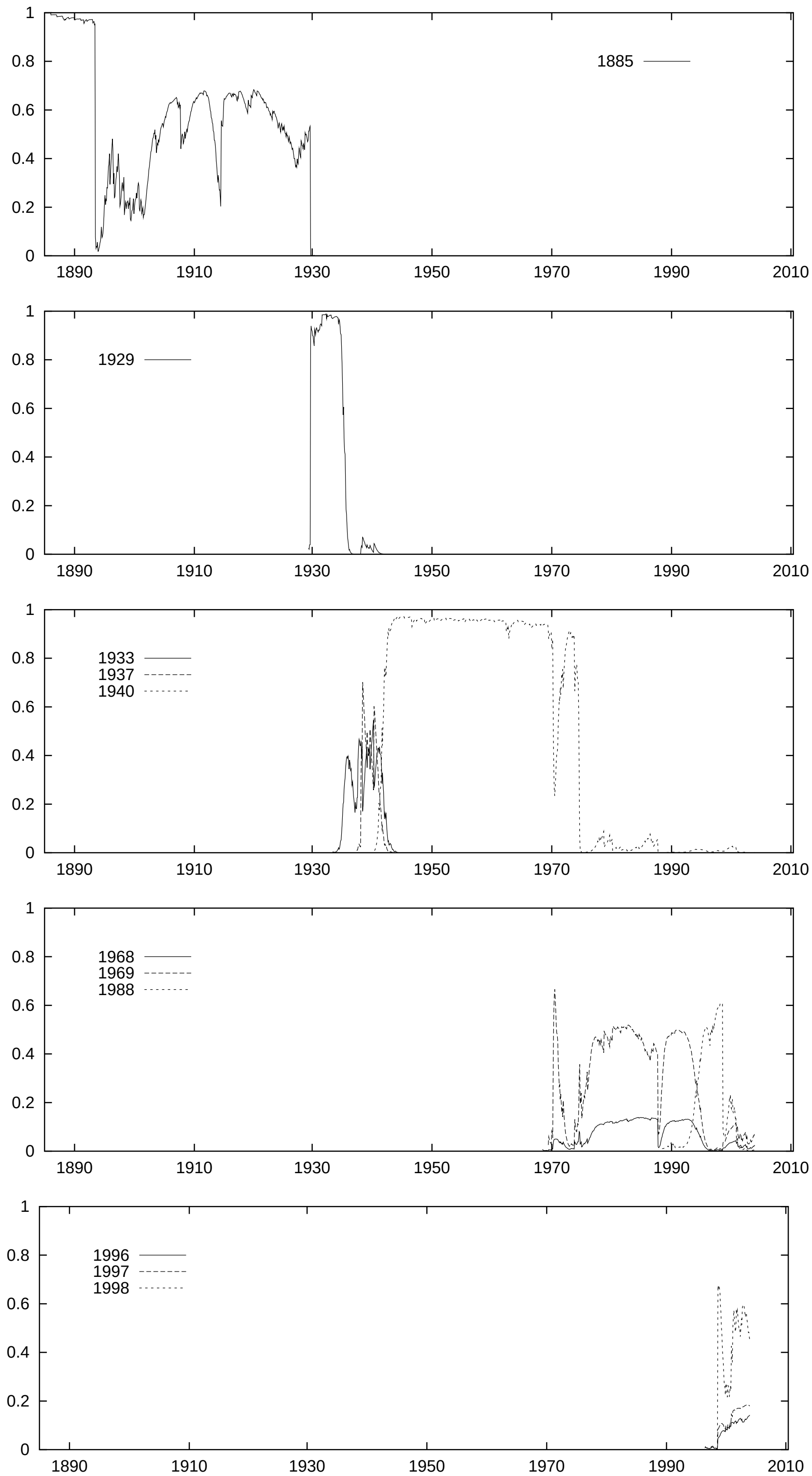
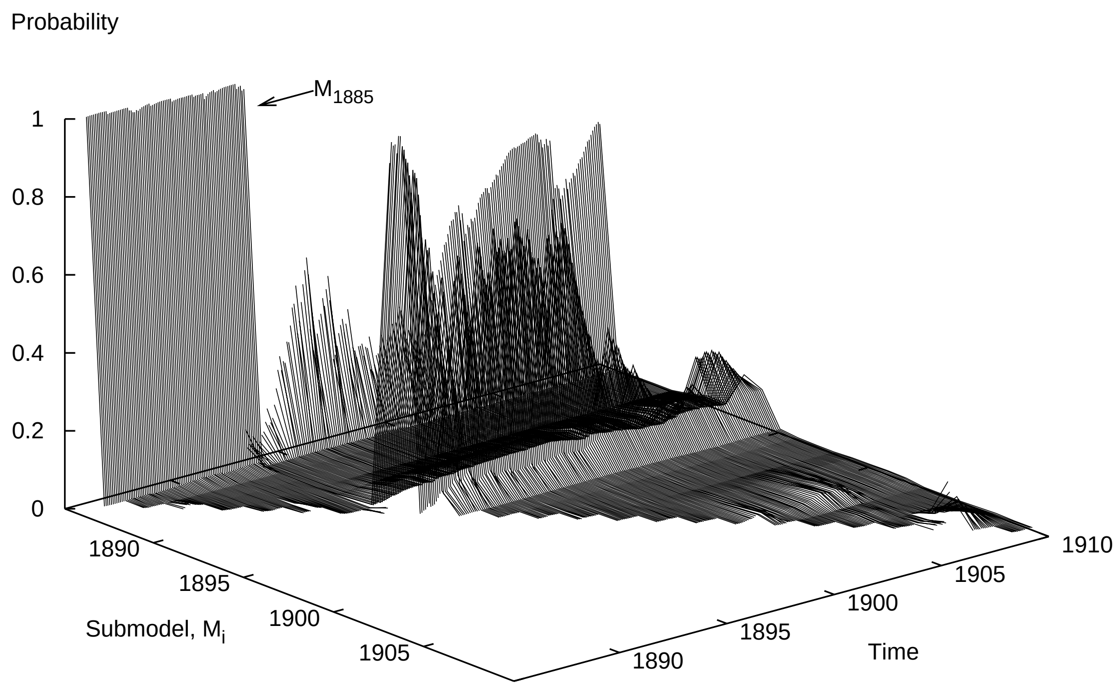


Figure 7: Submodel Probabilities through Time, 1885:2-1910:1

A. k=1



B. k=2

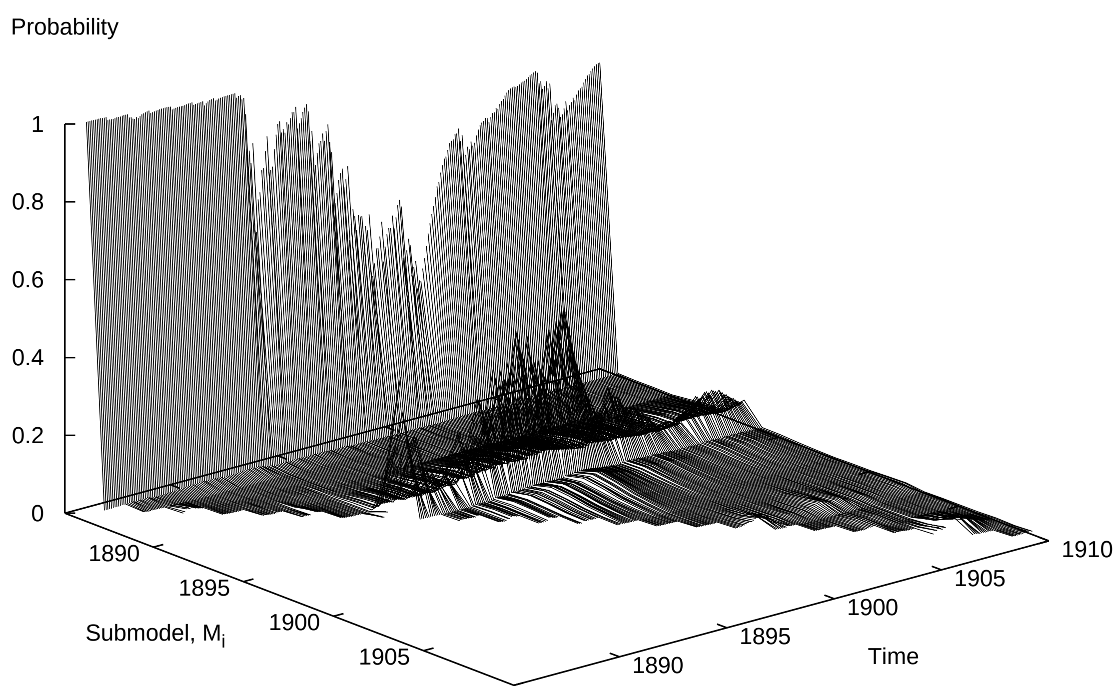
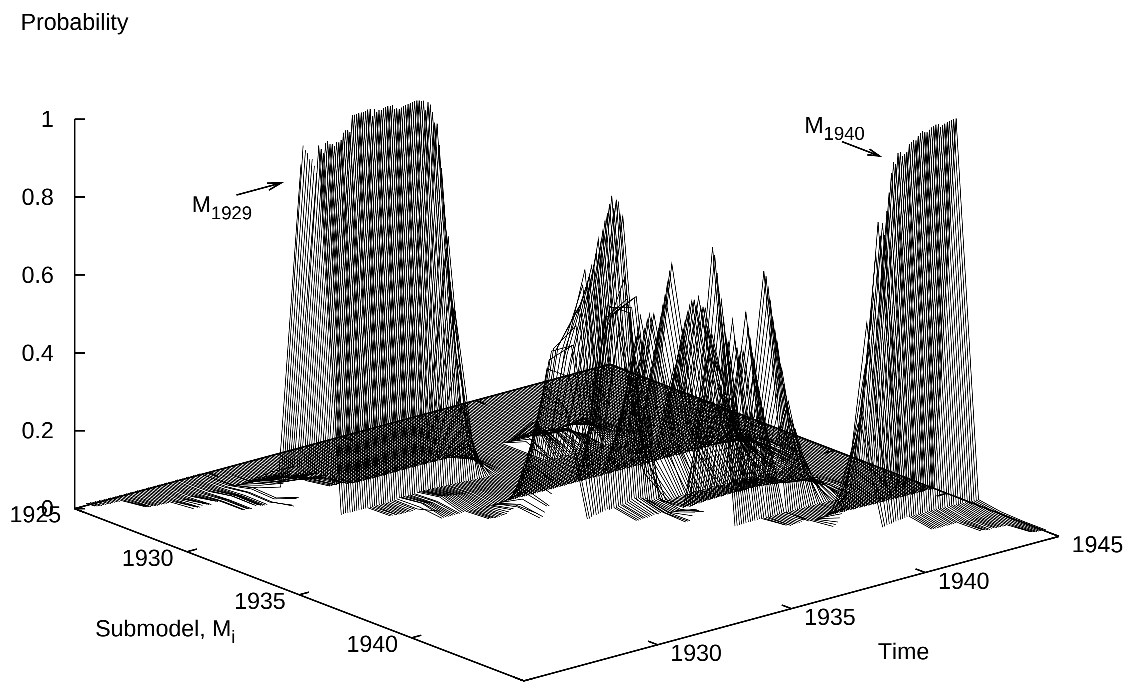


Figure 8: Submodel Probabilities through Time, 1925:1-1945:1

A. k=1



B. k=2

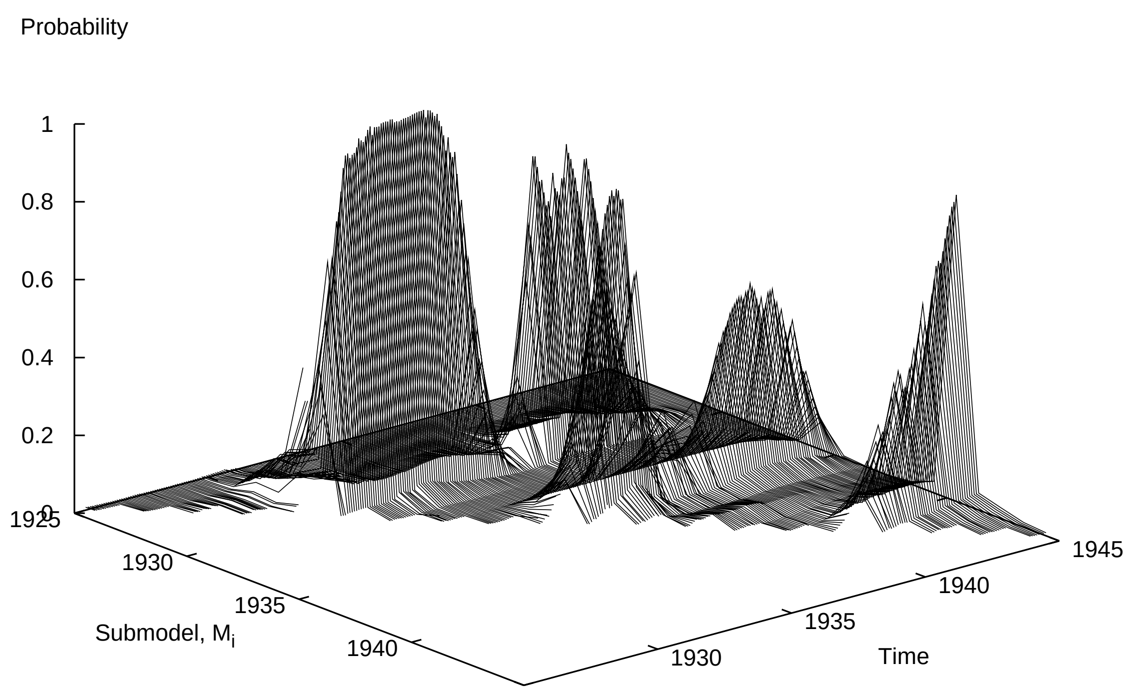
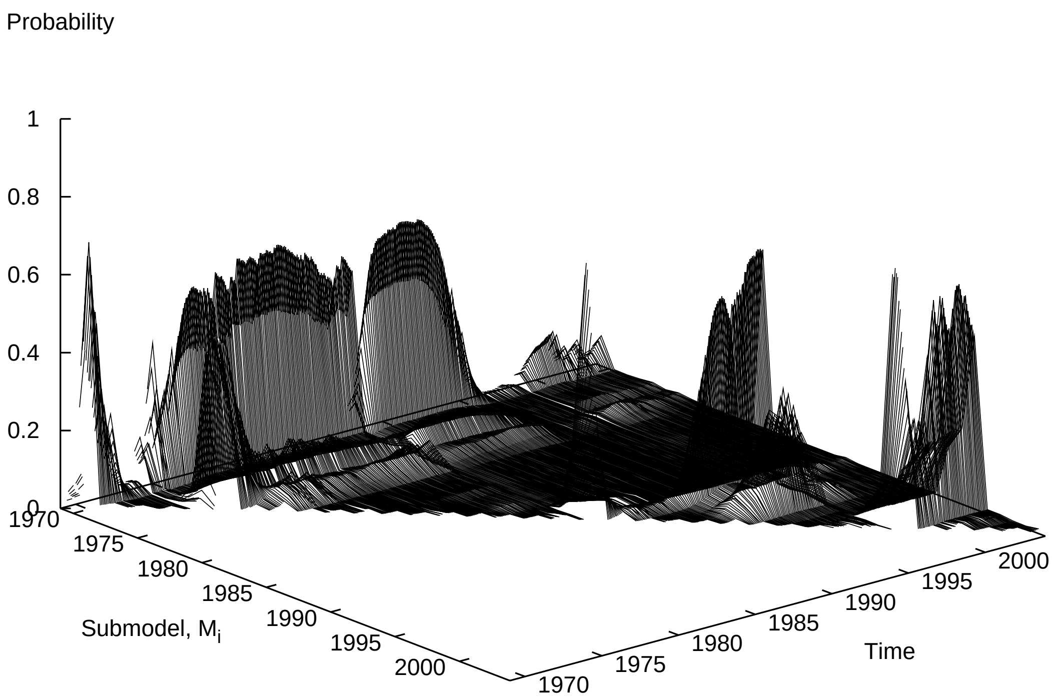


Figure 9: Submodel Probabilities through Time, 1970:1-2003:12

A. k=1



B. k=2

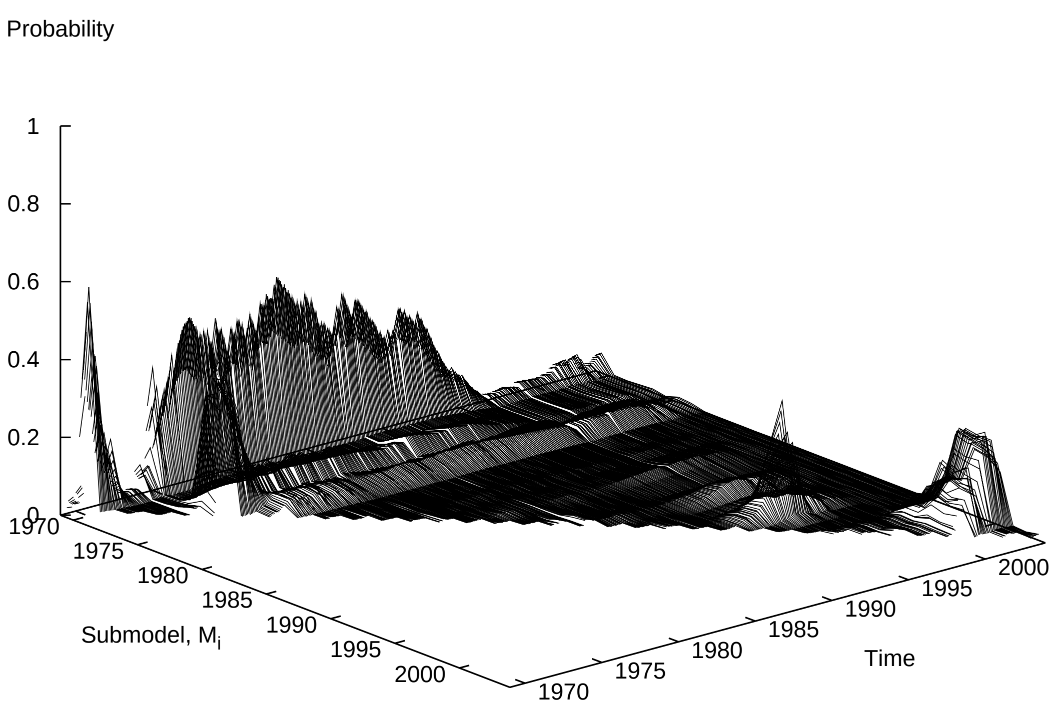
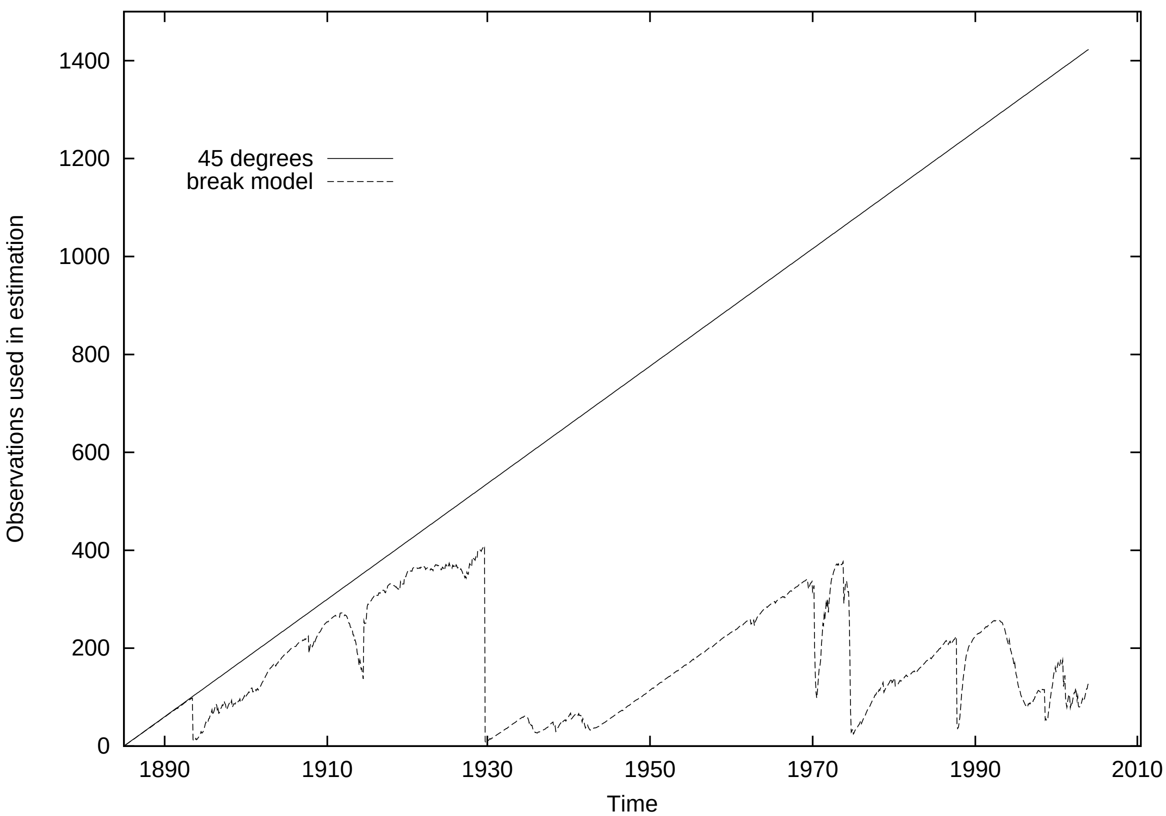


Figure 10: Mean useful Observations, $k = 1$



This figure shows the mean useful observations MUO_t defined as

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

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. If there are no structural breaks then MUO_t would follow the 45 degree line.

Figure 11: Premium Forecasts through Time, $k = 2$.

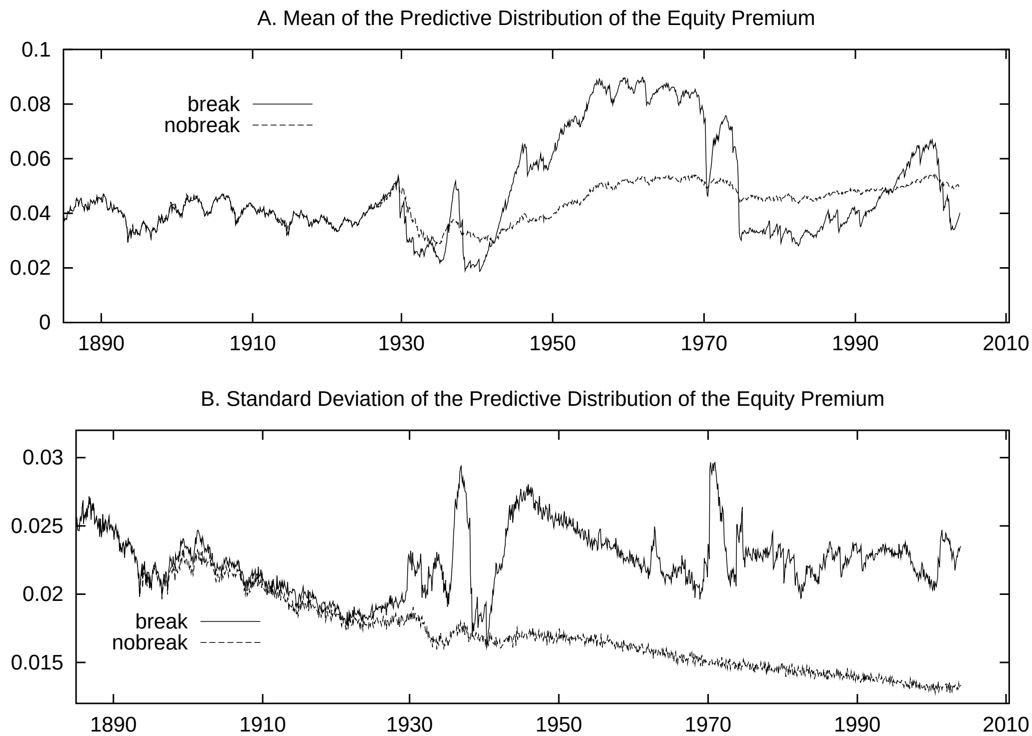
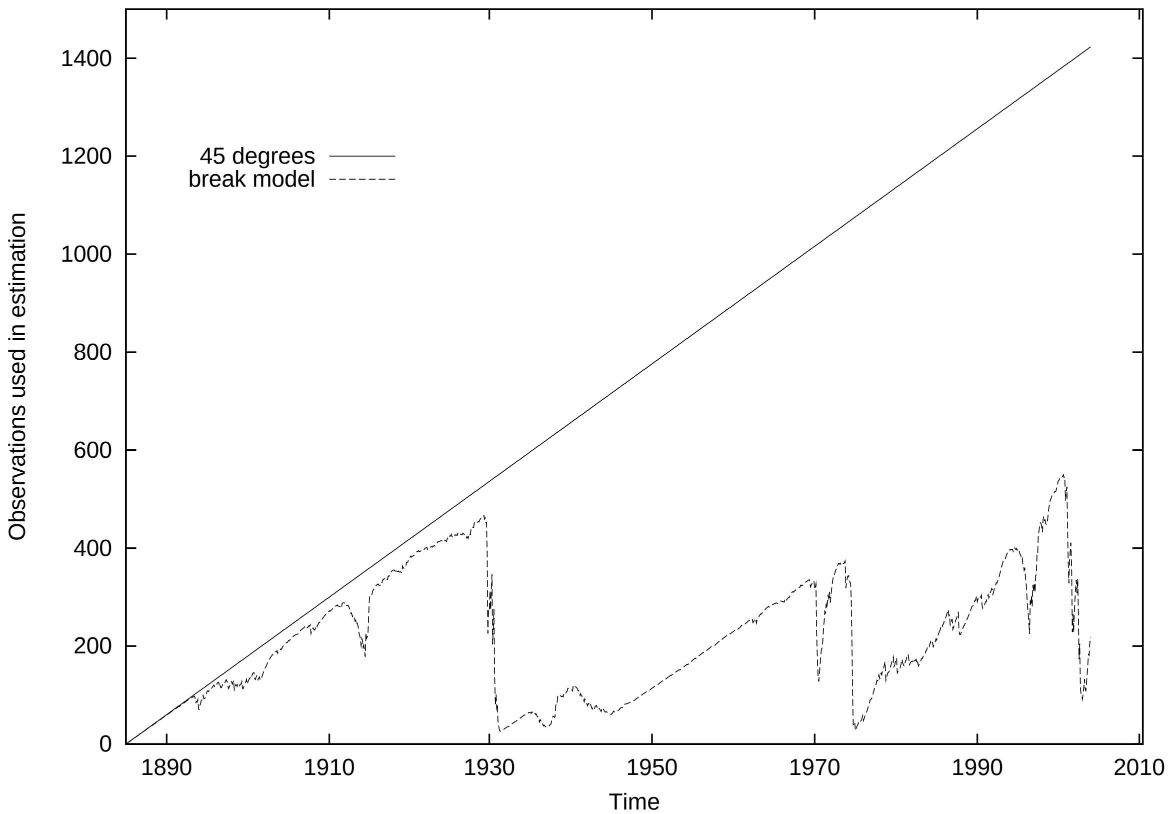


Figure A displays the out-of-sample forecasts (predictive mean) of the equity premium period by period for both the structural break model and the no break alternative. Figure B displays the corresponding standard deviation of the predictive distribution of the equity premium.

Figure 12: Mean useful Observations, $k = 2$.

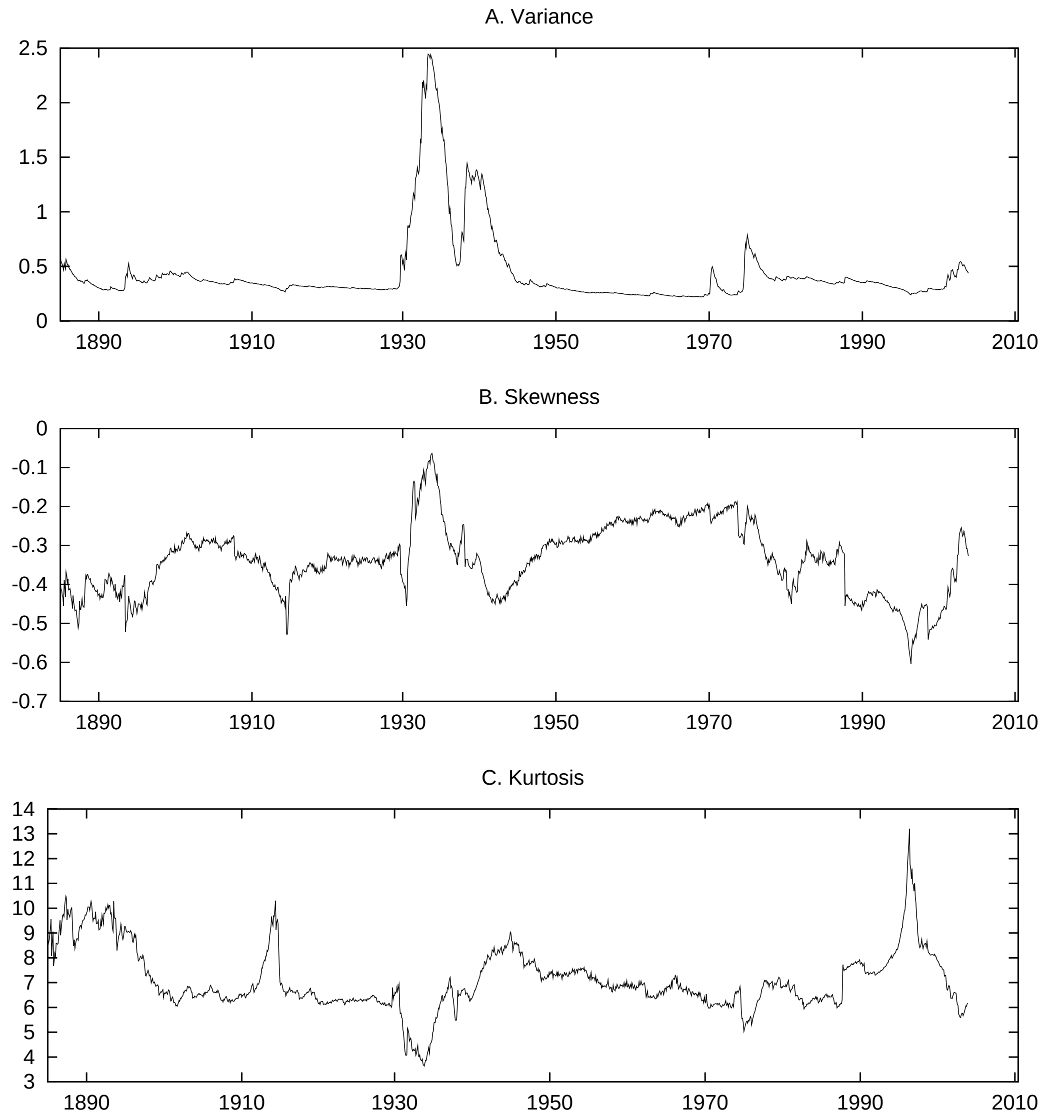


This figure shows the mean useful observations MUO_t defined as

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

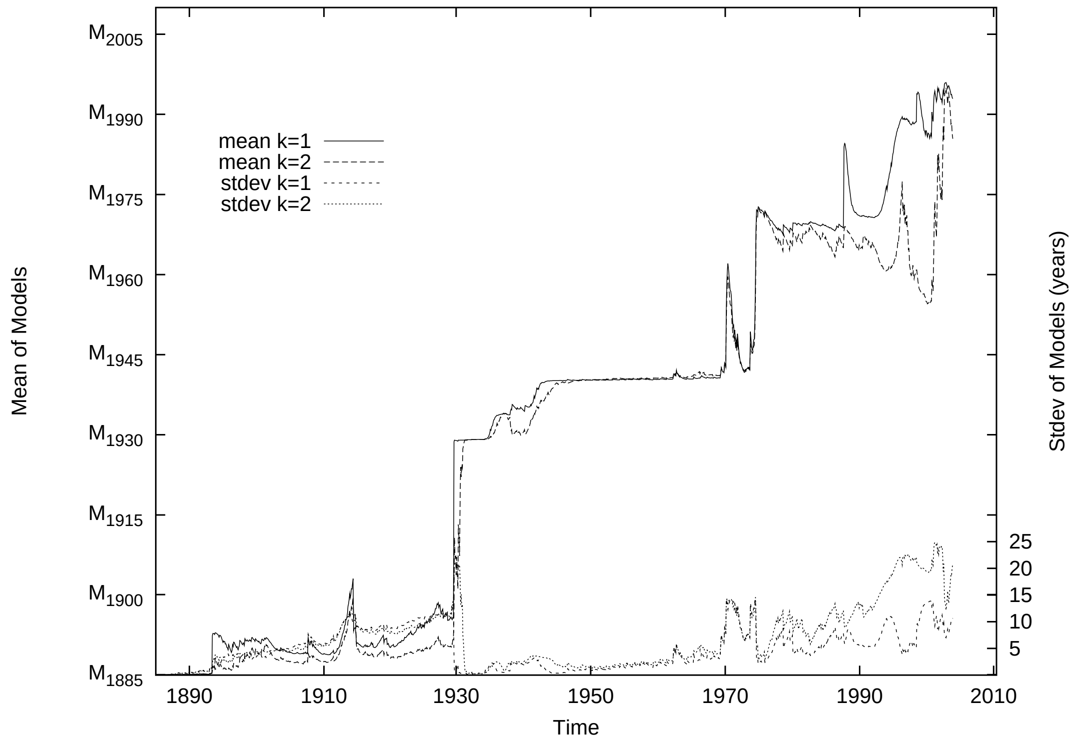
which is the expected number of useful observation for model estimation at each point in time. $p(M_i|I_t)$ is the posterior submodel probability for M_i given the information set I_t. If there are no structural breaks then MUO_t would follow the 45 degree line.

Figure 13: Higher-Order Moments of Excess Returns through Time



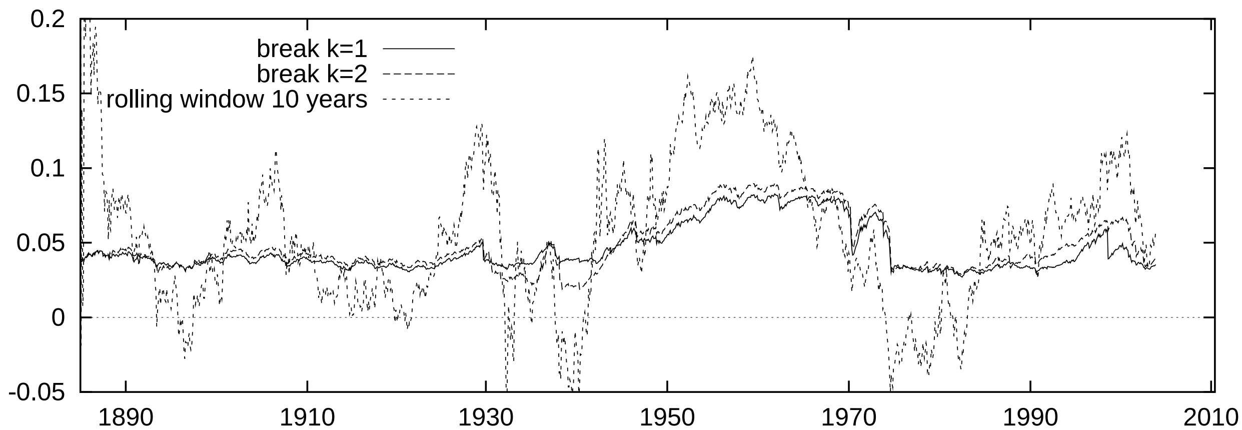
Displayed are the posterior means of the moments of the excess return distribution as inferred from the structural break model, $k = 2$. Each moment is estimated using only information in I_t at each point in time. The moments in (4.6)-(4.8) are computed for each Gibbs draw from the posterior distribution for each of the submodels M_i. The submodel specific moments are averaged using (6.8). 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 14: Comparison of Posterior Mean and Standard Deviation of the Distribution of Submodels



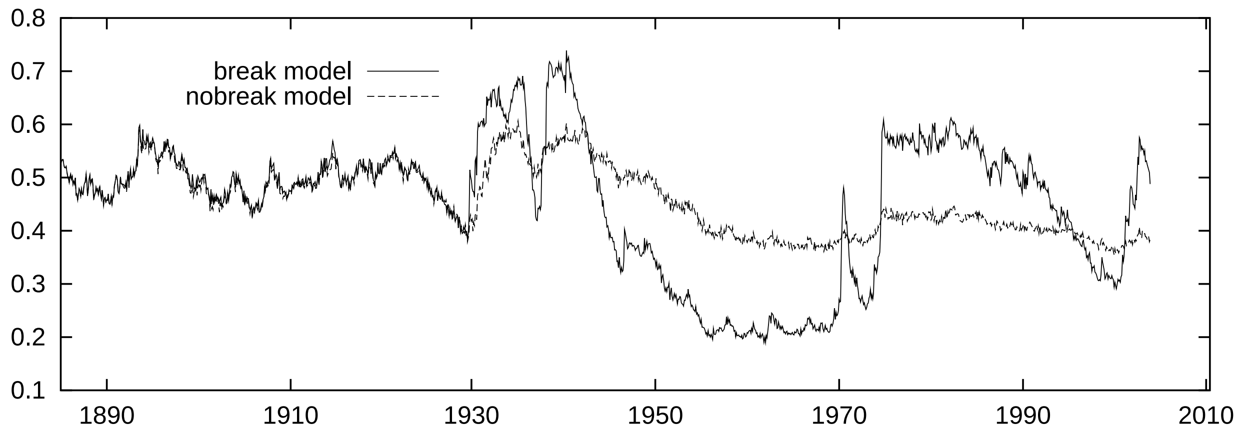
This figure compares the posterior mean and standard deviation of the distribution of submodels for $k = 1$, and 2 specifications. See the notes to Figure 5.

Figure 15: Comparison of Premium Forecasts



This figure compares the forecasts (predictive mean) of the equity premium from the structural break model with 1 and 2 components, along with the sample average that uses a rolling window of 10 years of data. The sample average at time t is defined as $\frac{1}{120}\sum_{i=1}^{120} r_{t-i+1}$.

Figure 16: Implications of Structural Breaks for a Pension Liability. $k = 2$



This figure compares the expected investment required today to receive $1 twenty years in the future. This is calculated as $E_t[1/(1+\gamma)^{20}]$ for both the break and no-break models at each point in time based on the most recent data available. The expectation is taken with respect to the predictive distribution of the equity premium γ, assuming a riskfree rate of 0.

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.

Bernardo, J. M., and A. F. M. Smith (2000): *Bayesian Theory*. Wiley.

Campbell, J. Y., and S. B. Thompson (2004): "Predicting the Equity Premium Out of Sample: Can Anything Beat the Historical Average?," manuscript, Harvard University.

Chib, S. (1998): "Estimation and Comparison of Multiple Change Point Models," *Journal of Econometrics*, 86, 221–241.

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

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 (2004): "Stare Down the Barrel and Center the Crosshairs: Targeting the Ex Ante Equity Premium," manuscript, York University.

Fama, E. F., and K. R. French (1992): "The Cross-Section of Expected Stock Returns," *Journal of Finance*, 47, 427–465.

——— (2002): "The Equity Premium," *Journal of Finance*, 57(2), 637–659.

Gelfand, A., and D. Dey (1994): "Bayesian Model Choice: Asymptotics and Exact Calculations," *Journal of The Royal Statistical Society*, B, 56, 501–514.

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.

——— (2003): “Contemporary Bayesian Econometrics and Statistics,” Department of Economics and Statistics, University of Iowa.

Goyal, A., and I. Welch (2004): “A Comprehensive Look at the Empirical Performance of Equity Premium Prediction,” manuscript, Brown University.

Graham, J. R., and C. R. Harvey (2005): “The long-run equity risk premium,” Fuqua School of Business, Duke University.

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.

Kandel, S., and R. Stambaugh (1996): “On the predictability of stock returns: an asset allocation perspective,” *Journal of Finance*, 51, 385–424.

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.

Maheu, J. M., and S. Gordon (2005): “Learning, Forecasting and Structural Breaks,” manuscript, Department of Economics, University of Toronto.

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.

Pettenuzzo, D., and A. Timmermann (2004): “Optimal Asset Allocation under Structural Breaks,” Department of Economics, University of California, San Diego.

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.

Ross, S. (1976): "The Arbitrage Theory of Capital Asset Pricing," *Journal of Economic Theory*, 13, 341–360.

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.