Forecasting and model averaging with structural breaks

by

Anwen Yin

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Economics

Program of Study Committee:

Helle Bunzel, Co-major Professor

Gray Calhoun, Co-major Professor

Joydeep Bharttacharya

David Frankel

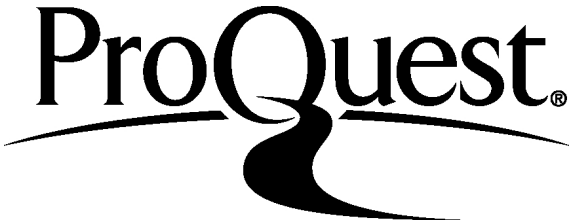
Jarad Niemi

Dan Nordman

Iowa State University

Ames, Iowa

2015



DEDICATION

To my parents and grandparents.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting research and the writing of this dissertation. First and foremost, Helle and Gray for their guidance, patience and support throughout this research and the writing of my dissertation. Their insights and words of encouragement have often inspired me and renewed my hopes for completing my doctoral degree. I would also like to thank my committee members for their efforts and contributions to this work: Dan, David, Jarad and Joydeep. Thank you everyone!

ABSTRACT

This dissertation consists of three papers. Collectively they attempt to investigate on how to better forecast a time series variable when there is uncertainty on the stability of model parameters.

The first chapter applies the newly developed theory of optimal and robust weights to forecasting the U.S. market equity premium in the presence of structural breaks. The empirical results suggest that parameter instability cannot fully explain the weak forecasting performance of most predictors used in related empirical research.

The second chapter introduces a two-stage forecast combination method to forecasting the U.S. market equity premium out-of-sample. In the first stage, for each predictive model, we combine its stable and break cases by using several model averaging methods. Next, we pool all adjusted predictive models together by applying equal weights. The empirical results suggest that this new method can potentially offer substantial predictive gains relative to the simple one-stage overall equal weights method.

The third chapter extends model averaging theory under uncertainty regarding structural breaks to the out-of-sample forecast setting, and proposes new predictive model weights based on the leave-one-out cross-validation criterion (CV), as CV is robust to heteroscedasticity and can be applied generally. It provides Monte Carlo and empirical evidence showing that CV weights outperform several competing methods.

CHAPTER 1. FORECASTING EQUITY PREMIUM WITH STRUCTURAL BREAKS

1.1 Introduction

Recent econometric advances and empirical evidence seem to suggest that the market excess returns are predictable to some degree. Forty years ago this would have been tantamount to an outright rejection of the efficient capital market hypothesis. In fact, the martingale is long considered to be a necessary condition for an efficient asset market, one in which the information contained in past prices is instantly, fully, and perpetually reflected in the asset's current price. If the market is efficient, then it should not be possible to profit by trading on the information contained in the asset's price history, hence the conditional expectation of future price changes, conditional on the price history, cannot be either positive or negative and therefore must be zero. A model associated with the efficient market hypothesis is the random walk model. It assumes that the successive returns are independent, and that the returns are identically distributed over time. Consequently, it implies that the efficient market hypothesis and random walk model combined can fully explain the weak forecasting performance of a wide range of predictors in empirical studies.

However, one of the central tenets of modern financial economics is the necessity of some degree of trade-off between risk and the expected excess returns. In addition, although the martingale hypothesis places a restriction on the expected returns, it does not account for risk in any way. Particularly, if an asset's expected price change is

positive, it may be the reward necessary to attract investors to hold the asset and to bear the associated risk. Therefore, the martingale property may be neither a necessary nor a sufficient condition for rationally determined asset prices. The complex structure of security markets and frictions in the trading process could possibly generate stock return predictability.

Recently, Goyal and Welch (2008) show that the simple historical average model of the U.S. equity market excess returns forecasts future returns better than other models with various predictors suggested by the literature. They argue that the poor out-of-sample performance of linear predictive regressions is a systematic problem, not confined to any decade. They compare predictive regressions with historical average returns and find that historical average returns almost always generate superior return forecasts, so they conclude that "*the profession has yet to find some variable that has meaningful and robust empirical equity premium forecasting power*". Subsequently, in examining the cause of the forecast failure shown in Goyal and Welch (2008), Rapach et al. (2010) argue that model uncertainty and parameter instability impair the forecasting ability. Additionally, Rapach and Wohar (2006) and Paye and Timmermann (2006) have shown empirical evidence of detected structural breaks in equity premium predicative models. But the literature on how to forecast excess returns with detected structural breaks is limited.

In this paper, we attempt to answer two empirical questions. First, if the true data generating process underlying the predictive model indeed has structural breaks, how to forecast excess returns? Second, can structural breaks or parameter instability fully explain the poor out-of-sample performance of those variables evaluated in Goyal and Welch (2008)? For the presence of parameter instability, using monthly data from Goyal and Welch (2008) and the break testing procedure by Bai and Perron (1998), we find that all models except for the one using the stock market variance, do not have significant statistical evidence for breaks. Therefore, parameter instability alone cannot explain the

puzzle of weak out-of-sample predictive power for most variables. Next, for the stock market variance model with estimated breaks, we apply the optimal and robust weights theory proposed by Pesaran et al. (2013) to forecasting the U.S. market equity premium out-of-sample. Our empirical results suggest that the stock market variance does have predictive power in forecasting excess returns. In addition, its predictive ability is present even without assuming parameter instability for the linear predictive model. Our further analysis shows that for the stock market variance, its break model outperforms the stable one.

This paper builds on literature related to out-of-sample forecast evaluation and structural breaks. Researchers, such as Giacomini and Rossi (2009), have provided empirical evidence and suggest that parameter instability or structural break is an important source of forecast failure in macroeconomics and finance. Parameter instability can arise as a result of changes in tastes, technology, institutional arrangements and government policy. If there are breaks in the underlying data generating process and the break sizes are large, predictive models without taking into account this fact tend to forecast poorly out-of-sample. Researchers, such as Inoue and Kilian (2004), Goyal and Welch (2008) and Giacomini and Rossi (2009), have documented this out-of-sample forecast breakdown under parameter instability.

In the modeling of structural breaks, parameters can be assumed to change at discrete time intervals or continuously. With the discrete break model, break dates are estimated and forecasts are typically constructed using the post-break observations. Furthermore, Pesaran and Timmermann (2007) have proposed the optimal window theory to forecast in the presence of breaks. They argue that forecasts from the post-break window may not be mean squared forecast error optimal, as the estimation error could be large due to small post-break sample size. Their optimal estimation window includes pre-break observations which involves a bias-variance trade-off. On the other hand, Pesaran et al. (2013) propose optimal weights in the sense that the resulting forecasts minimize the expected mean

squared forecast error. With known break sizes and dates, their optimal weights follow a step function that allocates constant weights within regimes, but different weights across regimes. Since in practice break dates and sizes are unknown and their estimation could be highly imprecise, Pesaran et al. (2013) also develop weights that are robust to the uncertainty surrounding the break dates and sizes. With the continuously varying parameter model, breaks are assumed to occur at every time instant and observations are down-weighted to take account of the slowly changing nature of the parameters, for example, exponential smoothing.

The remainder of this paper is organized as follows. Section 1.2 reports the break estimation results. Section 1.3 outlines the weighted least squares theory we use to forecast out-of-sample with breaks. Section 1.4 reports empirical results. Section 1.5 concludes.

1.2 Detecting and Dating Structural Breaks

Goyal and Welch (2008) use the stable linear one-step ahead predictive model to evaluate the predictive power of a wide range of variables,[1]

$$y_{t+1} = \bar{y} + \beta x_t + u_{t+1} \tag{1.1}$$

where $t = 1, ..., T$. y_{t+1} is the market excess returns, $\bar{y}$ is the intercept, x_t is the exogenous predictor available at time t to forecast the next period returns y_{t+1} and u_{t+1} is a disturbance term. The un-modeled structural breaks may be the cause why many predictors are week to forecast the excess returns relative to the benchmark which is simply

$$y_{t+1} = \bar{y} + u_{t+1}. \tag{1.2}$$

In this section we will present the break model and outline the method we will use to detect and estimate possible breaks for model (1.1).

[1] They also consider a large linear model which includes all variables.

1.2.1 Break Model

The model subject to m breaks occurring at times $(t_1, t_2, ..., t_m)$ is

$$y_{t+1} = \begin{cases} \overline{y}_1 + \beta_1 x_t + u_{t+1}, & t = 1, ..., t_1 \\ \overline{y}_2 + \beta_2 x_t + u_{t+1}, & t = t_1 + 1, ..., t_2 \\ \vdots & \vdots \\ \overline{y}_m + \beta_m x_t + u_{t+1}, & t = t_{m-1} + 1, ..., t_m \\ \overline{y}_{m+1} + \beta_{m+1} x_t + u_{t+1}, & t = t_m + 1, ..., T \end{cases} \tag{1.3}$$

where y_{t+1} is the one-step ahead market excess returns, x_t is the exogenous predictor available at time t to forecast the next period returns y_{t+1} and u_{t+1} is a disturbance term. The reason for using the discrete, step-function type break model is that some of the potential sources of breaks, such as shifts in economic policy regimes or large macroeconomic shocks, are likely to lead to rather sudden shifts in the parameters of the forecasting model. In addition, we assume that parameter instability only occurs in the regression coefficients $\bar{y}$ and β.

The idea of estimating structural breaks in Bai and Perron (1998) is to find a set of dates which globally minimizes the sum of squared residuals from the least squares regression

$$(\hat{t}_1, \hat{t}_2, ...\hat{t}_M) = \text{argmin} \sum_{i=1}^{m+1} \sum_{s=t_{i-1}+1}^{t_i} [y_{s+1} - \bar{y}_s - \beta_s x_s]^2 \tag{1.4}$$

where i indexes the number of regimes. The regression parameter estimates are the ordinary least squares estimates associated with the m-partition of the data sample. For break identification, a crucial assumption in Bai and Perron (1998) is that there is enough number of observations within each regime. Given the break date estimates, the regression model coefficients, $\left\{\widehat{\beta}_i\right\}_{i=1}^{m+1}$, are the least squares estimates associated with the partition comprised of the estimated break dates.

1.2.2 Data

Our monthly data from January 1871 to December 2011 are obtained from Goyal and Welch (2008). Since not all variables are available for the entire time span, in order to take a comprehensive look at the performance of all predictors, we only consider a subset of the data from May 1937 to December 2011 for our empirical analysis. It is worth mentioning that in this paper we examine more predictive variables than those studied in Paye and Timmermann (2006) and Rapach and Wohar (2006).

The dependent variable, the market equity premium, is the log returns on the S&P 500 index including dividends minus the log returns on the risk-free rate. The predictors are

- Log dividend-price ratio (dp): log of a 12-month moving sum of dividends paid on the S&P 500 index minus the log of stock prices.
- Log dividend yield (dy): log of a 12-month moving sum of dividends minus the log of lagged prices.
- Log earnings-price ratio (ep): log of a 12-month moving sum of earnings on the S&P 500 index minus the log of stock prices.
- Log dividend-payout ratio (de): log of a 12-month moving sum of dividends minus the log of a 12-month moving sum of earnings.
- Stock market variance (svar): monthly sum of squared daily returns on the S&P 500 index.
- Cross sectional premium (csp): the relative valuations of high- and low-beta stocks.
- Book to market ratio (bm): ratio of book value to market value for the Dow Jones Industrial Average.

- Net equity expansion (ntis): ratio of a 12-month moving sum of net equity issues by NYSE-listed stocks to the total end of year market capitalization of NYSE stocks.
- 3-month Treasury bill rate (tbl): interest rate on a three-month secondary market Treasury bill.
- Long term government bond yield (lty): long term government bond yield.
- Term spread (tms): long term yield minus the Treasury bill rate.
- Default premium (dfy): difference between BAA- and AAA-rated corporate bond yields.
- Inflation (infl): inflation is the Consumer Price Index (all urban consumers) from the Bureau of Labor Statistics.

These variables can be put into three categories: stock characteristics variables, such as the dividend price ratio; market micro-structure variables, such as the net equity expansion; and macroeconomic indicators, for example, the inflation rate.

1.2.3 Break Estimation

Our model (1.3) assumes that all regression coefficients are subject to structural breaks, since there is no convincing evidence saying otherwise. Because the total number of breaks is another parameter to estimate, a predictive model with a large number of estimated break dates fully based on equation (1.4) may be overfitted. To correct possible model overfitting, we adopt the approach by Zeileis et al. (2003) to select the number of estimated breaks based on the Bayesian information criterion which penalizes overfitting. The number of breaks associated with the minimum Bayesian information criterion (BIC) value will be selected. If the BIC value achieves its minimum at the point where the total number of breaks is zero, then it favors a stable model with no breaks.

The total number of breaks estimation results for all models are presented in Figure A.1, Figure A.2 and Figure A.3.

We have 14 models in total, 13 univariate regression models plus one historical mean model as benchmark. For each model labeled by its predictor, Figure A.1, Figure A.2 and Figure A.3 report the BIC value and the sum of squared residuals (RSS) as a function of the number of breaks. The RSS is shown in blue colored curve and it is downward-sloping in all figures. This is not surprising because adding one more arbitrary break is analogous to adding one more regressor in a linear model and the RSS will decrease as the result of model overfitting. The black colored BIC curve is the criterion we use in break number selection. By BIC, we can see that only the stock market variance model has evidence of parameter instability with three breaks. But the evidence is not strong enough to rule out the stable model shown in Figure A.1. Both models have approximately the same BIC value, so next we will split the analysis of the stock market variance model into two cases, the break model case and the stable model case. For other models, it is clear from these figures that the stable model is the best choice.

For the break model of stock market variance, the break date estimates are March 1956, September 1974 and November 1985. Note that the second break date, September 1974, corresponds to the timing of the oil shock documented by economists.[2] The last break date may be related to the great moderation.

1.2.4 Full Sample Estimation Results

For all stable models, we simply estimate their parameters by least squares then conduct inference. Separately, for the stock market variance model with breaks, based on previous results, we estimate its parameters for each segment by least squares. Our full sample least squares estimation results for all stable models are presented in Table A.1. The full sample estimation results for the stock market variance model with breaks

[2]Goyal and Welch (2008) pick the year 1974 as the break date without estimation.

are reported in Table A.2. In Table A.1, for each model labeled by its predictor, we report its in-sample R^2 statistic, intercept estimate and predictor coefficient estimate β. Parentheses report the t statistic for each parameter estimate above. In Table A.2, we report all statistics separately for each segment.

For all predictor-based stable models except for the stock market variance model, the in-sample explanatory power of predictors measured by R^2 is very low. Furthermore, most predictor coefficients are insignificant. Our results contradict with studies, such as Giacomini and Rossi (2009) and Goyal and Welch (2008), which conjecture that the insignificant predictive ability of economic variables is likely due to parameter instability. Our results show that most predictors in Goyal and Welch's monthly data are stable in the bivariate predictive model, and the poor forecasting performance of these variables cannot be attributed to un-modeled parameter instability.

For the stock market variance model with three breaks, its R^2 value is higher than any other predictors shown in Table A.1 in all segments. Furthermore, its parameter estimates are significant in all segments. Our results suggest that the stock market variance has predictive power in forecasting excess returns.

Next we will show how to apply the optimal and robust weights to forecasting out-of-sample with breaks.

1.3 Forecast with Parameter Instability

With mounting evidence of parameter instability in many macroeconomic and financial predictive models (see Rapach and Wohar (2006) and Paye and Timmermann (2006)), how to forecast a time series variable of interest with model parameter instability is an important issue. Researchers have proposed various methods to forecast under modeled breaks, and this strand of literature is fast evolving. Here we apply the weighed least squares theory proposed by Pesaran et al. (2013) to forecast in the presence of

breaks. In this section we will outline the construction of optimal weights and robust weights, and examine their empirical performance next. From a forecaster's perspective, the latest break date should be most important to predict the future, so for models with multiple estimated breaks, we only focus on forecasting after the latest break and drive weights accordingly.

1.3.1 Methodology

1.3.1.1 Optimal Weights

The theory supporting the optimal weights assumes that the break dates and sizes are known. Following the notation of related out-of-sample forecast literature, we denote the total sample size $T+1$, and split the sample into two parts: the first R observations for the training sample while the remaining P observations for prediction and forecast evaluation, $R+P=T+1$. In addition, we impose that the break point, τ, falls into the estimation sample, and is bounded far away from both ends, that is, $1 << \tau << R$. We only consider the one-step ahead forecast problem. The predictive model with optimal weights is

$$\widehat{y}_{t+1} = x'_{t+1}\widehat{\beta}_t^{opt} \tag{1.5}$$

The weights used in parameter estimation are optimal in the sense of minimizing the expected mean squared forecast error

$$\mathbf{w} = \arg\min_{\mathbf{w}} E\left[\left(y_{t+1} - x'_{t+1}\widehat{\beta}_t\right)^2\right] \tag{1.6}$$

There are three popular estimation windows in the out-of-sample forecast literature: recursive window, rolling window and fixed window. Under the recursive window, at each point in time, the estimated parameters are updated by adding one more observation starting with sample size R. Under the rolling window, the estimation window

is always fixed at length of R, for example, the first estimate uses data from period 1 to period R, while the second estimate runs from period 2 to period $R+1$. Under the fixed window, parameters are estimated only once using the entire estimation sample R. Mathematically, for the recursive window

$$\widehat{\beta}_t^{opt} = \left(\sum_{s=1}^{t} w_s x_s x_s'\right)^{-1} \left(\sum_{s=1}^{t} w_s x_s y_s\right) \tag{1.7}$$

for the rolling window

$$\widehat{\beta}_t^{opt} = \left(\sum_{s=t-R+1}^{t} w_s x_s x_s'\right)^{-1} \left(\sum_{s=t-R+1}^{t} w_s x_s y_s\right) \tag{1.8}$$

and for the fixed window

$$\widehat{\beta}^{opt} = \left(\sum_{s=1}^{R} w_s x_s x_s'\right)^{-1} \left(\sum_{s=1}^{R} w_s x_s y_s\right) \tag{1.9}$$

where $t = R, ..., R+P-1$.

The optimal weights theory states that observations in each regime will receive different weights for parameter estimation. If there is only one break, then the optimal weights take a simple two-regime form under fixed window, distinct weights across regimes but constant within each regime

$$\begin{cases} w_1 = \frac{1}{R}\frac{1}{\mu+(1-\mu)(1+\mu R\lambda^2\omega^2)} \\ w_2 = \frac{1}{R}\frac{1+\mu R\lambda^2\omega^2}{\mu+(1-\mu)(1+\mu R\lambda^2\omega^2)} \end{cases} \tag{1.10}$$

where τ is the break date, $\mu = \tau/R, \lambda = \frac{\beta_1-\beta_2}{\sigma}, \omega = \frac{1}{\tau}\sum_{s=1}^{\tau} x_s^2$. Optimal weights under recursive window or rolling window take the same form except that we need to update R with the actual sample size in each estimation step. Since we do not know the population value of these parameters, in practice we need to take advantage of our break

detection results earlier to provide sample approximations for the population parameter values of β_1, β_2 and σ. Our ordinary least squares estimates for the βs in the third and fourth segments in table A.2 will serve as proxies for β_1 and β_2. The sample standard deviation from September 1974 to December 2011 will be used to approximate σ.

1.3.1.2 Robust Weights

For optimal weights we have assumed that the dates and the sizes of parameter breaks are known. However, this assumption may not be relevant to real time forecasting. Specifically, the break sizes are difficult to estimate unless a relatively large number of post-break observations is available. So in addition to optimal weights, Pesaran et al. (2013) also propose weights which are robust to the uncertainty of break dates and sizes. In the robust weights theory, break dates and sizes are unknown.

The derivation of robust weights is an extension to deriving optimal weights. To illustrate the main idea of robust weights, we will continue the derivation process from equation (1.10). Rewrite equation (1.10) as

$$\begin{cases} Rw_1 = \frac{1}{\mu+(1-\mu)(1+\mu R\lambda^2\omega^2)} \\ Rw_2 = \frac{1+\mu R\lambda^2\omega^2}{\mu+(1-\mu)(1+\mu R\lambda^2\omega^2)} \end{cases} \tag{1.11}$$

We can reformulate the time profile of the weights as

$$Rw_t\left(\mu,\lambda^2\right) = w_2 + (w_1 - w_2)\, I_{[\tau-t]} \tag{1.12}$$

for $t = 1, 2, ..., R$. Hence,

$$Rw\left(a,\mu,\lambda^2\right) = \frac{\frac{1}{R}+\mu\lambda^2}{\frac{1}{R}+\mu(1-\mu)\lambda^2} - \left(\frac{\mu\lambda^2}{\frac{1}{R}+\mu(1-\mu)\lambda^2}\right) I_{[\mu-a]} \tag{1.13}$$

where $a = t/R \in [0, 1]$.

There is one discrete break in β_i, but now we do not know the exact date of the break, τ. Instead, to derive the robust weights, we can impose a uniform distribution assumption on the break fraction, $\mu \equiv \tau/R \sim U\left[\underline{\mu}, \overline{\mu}\right]$, where $\underline{\mu}$ and $\overline{\mu}$ are some pre-specified lower and upper bounds for the break fraction. $\underline{\mu}$ could take the value of zero while $\overline{\mu}$ can be very close to one. By minimizing the expected mean squared forecast error, the population robust weights can be solved as

$$Rw(a) = \begin{cases} 0 + O(R^{-1}) & \text{for } a < \underline{\mu} \\ \left(\overline{\mu} - \underline{\mu}\right)^{-1} \int_{\underline{\mu}}^{\overline{\mu}} \frac{1}{1-\mu} d\mu - \left(\overline{\mu} - \underline{\mu}\right)^{-1} \int_{a}^{\overline{\mu}} \frac{1}{1-\mu} d\mu + O(R^{-1}) & \text{for } \underline{\mu} \leq a \leq \overline{\mu} \\ \left(\overline{\mu} - \underline{\mu}\right)^{-1} \int_{\underline{\mu}}^{\overline{\mu}} \frac{1}{1-\mu} d\mu + O(R^{-1}) & \text{for } a > \overline{\mu} \end{cases} \tag{1.14}$$

then approximated by

$$w(a) \approx \begin{cases} 0 & \text{for } a < \underline{\mu} \\ \frac{-1}{R\left(\overline{\mu} - \underline{\mu}\right)} \log\left(\frac{1-a}{1-\underline{\mu}}\right) & \text{for } \underline{\mu} \leq a \leq \overline{\mu} \\ \frac{-1}{R\left(\overline{\mu} - \underline{\mu}\right)} \log\left(\frac{1-\overline{\mu}}{1-\underline{\mu}}\right) & \text{for } a > \overline{\mu} \end{cases} \tag{1.15}$$

In the case where $\underline{\mu}$ and $\overline{\mu}$ are close to the end points of 0 and 1, we have

$$w(a) \approx \frac{-\log(1-a)}{R}, a \in [0, \overline{\mu}] \tag{1.16}$$

A discrete time version can be obtained by setting $R\overline{\mu} = R - 1$. Namely,

$$w_t^* = \frac{-\log(1 - t/R)}{R-1}, \text{ for } t = 1, 2, ..., R-1 \tag{1.17}$$

and

$$w_R^* = \frac{-1}{R-1} \log\left(1 - \frac{R-1}{R}\right) = \frac{\log(R)}{R-1} \tag{1.18}$$

Due to approximation error, these weights do not sum to unity, so they need to be re-scaled as

$$w_t = \frac{w_t^*}{\sum_{i=1}^{R} w_i^*}, \text{ for } t = 1, 2, ..., R \tag{1.19}$$

So under fixed window, the sample robust weights take the following form

$$w_t = \begin{cases} \frac{\log(1-s/R)}{\sum_{i=1}^{R-1}\log(1-i/R)-\log(R)}, & s = 1, ..., R-1 \\ \frac{\log(R)}{\log(R)-\sum_{i=1}^{R-1}\log(1-i/R)}, & s = R \end{cases} \tag{1.20}$$

Robust weights under recursive window or rolling window take the same form as in equation (1.20) except that we need to update R with the actual sample size used in each estimation step. With robust weights, the least squares parameter estimates under the fixed window are:

$$\widehat{\beta}^R = \left(\sum_{s=1}^{R} w_s x_s x_s'\right)^{-1} \left(\sum_{s=1}^{R} w_s x_s y_s\right) \tag{1.21}$$

under the rolling window

$$\widehat{\beta}_t^R = \left(\sum_{s=t-R+1}^{t} w_s x_s x_s'\right)^{-1} \left(\sum_{s=t-R+1}^{t} w_s x_s y_s\right) \tag{1.22}$$

and under the recursive window

$$\widehat{\beta}_t^R = \left(\sum_{s=1}^{t} w_s x_s x_s'\right)^{-1} \left(\sum_{s=1}^{t} w_s x_s y_s\right) \tag{1.23}$$

where $t = R, ..., T$.

Note that the robust weights shown in equation (1.20) do not involve break dates and sizes. Comparing robust weights with optimal weights, we can see that robust weights

take different values for different observations, as opposed to constant weights within a structural regime under optimal weights. In our empirical applications, the robust weights are monotonically increasing as time runs toward the end of the sample: the most recent observation receives the highest weight while observations in the distant past receive smaller weights. An example is shown in Figure A.11.

1.4 Out-of-sample Forecast

In the empirical analysis, we reserve the last 36 observations from the monthly data as the evaluation sample, $P = 36$. Theses observations represent the last three years of monthly data from January 2009 to December 2011. For the break model of stock market variance (1.3), the training sample starts with the first observation after the second break date (August 1974) and ends with the observation right before the evaluation sample (December 2008). The justification for our training sample size choice is that the econometric theory for forecasting with more than one break in the coefficient is not fully developed. Furthermore, from a researcher's perspective in empirical analysis, the latest break matters the most. Overall, we have $R = 859$ and $P = 36$ for the stable model of the stock market variance in equation (1.1), while $R = 412$ and $P = 36$ for the structural break model of the stock market variance (1.3). We use the mean squared forecast error (MSFE) to evaluate forecasts and compare results.

1.4.1 Forecast Using the Stable Model of Stock Market Variance

We first examine the out-of-sample performance of model (1.1) for the stock market variance without assuming structural breaks. We use model (1.1) to forecast the last 36 months of the equity premium using all window choices. In addition, we also include forecasts from the historical mean benchmark model (1.2). The results are shown in Figure A.4.

Figure A.4 shows that almost all estimation windows perform at least as well as the benchmark measured by a series of test errors, which supports our in-sample estimation results that the predictive power of stock market variance is significant. It is worth mentioning that forecasting results using annual data in Goyal and Welch (2008) suggest that the regression coefficient for the stock market variance predictor is insignificant, but our results using monthly data state otherwise. This could be due to the fact that we have more observations for parameter estimation using monthly data.

1.4.2 Forecast Using the Break Model of Stock Market Variance

Previously we have shown the forecasting performance of the stable model (1.1). Here we switch to the break model (1.3) and apply the optimal and robust weights to forecasting out-of-sample. In addition, we also consider the post-break window method which only uses observations after the latest break to estimate parameters.

In practice, it is up to the researcher to decide which method to use among optimal weights, robust weights and post-break window. Robust weights involve using observations even before the break date to estimate parameters so it may introduce estimation bias. The post-break window only uses observations after the recent break so it may help reduce estimation bias, but if the post-break window size is small, it may result in a large efficiency loss. Optimal weights assume that the true break dates and sizes are known, but in practice it is almost impossible to estimate them with great precision, especially when either the sample size or the break size is small.

1.4.2.1 Fixed Window

Out-of-sample results for the stock market variance model under fixed window are shown in Figure A.5.

We can see that the stock market variance model performs at least as well as the benchmark over the evaluation sample period measured by a series of test errors. The

robust weights perform especially well towards the end of the evaluation sample period. Comparing weighting methods, our results suggest that the post-break window could be used as an alternative to robust weights if the computation of robust weights is costly.

1.4.2.2 Recursive Window

Out-of-sample forecasting results for the stock market variance model under recursive window are shown in Figure A.6.

In this case we see that the robust weights and the post-break window work well over most part of the evaluation sample period. For most forecasts, the efficiency gains are relatively large under either robust weights or post-break window compared with the historical mean model.

1.4.2.3 Rolling Window

Out-of-sample results for the stock market variance model under rolling window are shown in Figure A.7.

Results in this case are similar to those under fixed window. Robust weights forecast better than others at the beginning and towards the end of the sample. Post-break window does well during the middle of the evaluation period.

1.4.3 Comparing the Stable Model with the Break Model

Previously we have shown that the stock market variance has predictive power in forecasting excess returns based on Goyal and Welch's monthly data, and the predictive ability stays regardless of the presence of structural breaks. Since our break detection results presented in section 1.2.3 do not provide a clear guidance on which model to choose, the break model (1.3) or the stable model (1.1) for the stock market variance, a natural extension is to compare the out-of-sample performance between these two models.

Following Goyal and Welch (2008), to construct a graphical device to evaluate the out-of-sample forecasting performance for two competing models, we will create a time series plot of the mean squared forecast error (MSFE) difference between the stable and break model,

$$\Delta\text{MSFE} = \text{MSFE}^{\text{stable}} - \text{MSFE}^{\text{break}} \tag{1.24}$$

We will consider all estimation window choices, namely, recursive window, rolling window and the fixed window. In addition, since we have three weighting choices for the break model, totally we have nine MSFE difference time series plots. These MSFE difference plots are presented in Figure A.8, Figure A.9 and Figure A.10. In each plot, if the curve moves up, it implies that the break model outperforms the stable model during that evaluation period. If the curve moves down, it supports the stable model during that period.

A number of fluctuations can be seen under the recursive window in Figure A.8. All weighting methods show strong support for the break model at the end of the sample, and the MSFE difference curve remains positive for most part of the evaluation period.

Rolling window favors the stable model as shown in Figure A.9. Both optimal weights and robust weights support the stable model at the beginning of the series, and the MSFE difference remains negative for most part of the evaluation period. The post-break window curve is very flat, and it stays close to zero through the entire evaluation period.

For the fixed window shown in Figure A.10, we can see that the robust weights show strong support for the break model at the beginning and towards the end of the evaluation period. Optimal weights and post-break window are flat with the difference remaining positive for most part of the evaluation period.

1.5 Conclusion

Goyal and Welch (2008) have examined the out-of-sample performance of a wide range of predictors suggested by the empirical finance literature in forecasting excess returns using stable linear models. They conclude that most predictors have weak predictive ability. Furthermore, researchers argue that the cause of the failure for these predictors is parameter instability and have provided empirical evidence, see Paye and Timmermann (2006) and Rapach and Wohar (2006). Then the problem is how to forecast out-of-sample with modeled breaks, and how to evaluate forecasts and compare models.

This paper applies the newly developed theory of optimal and robust weights to forecasting the U.S. market equity premium in the presence of structural breaks using Goyal and Welch's data. The weights are optimal in the sense of minimizing the expected mean squared forecast error, or robust to the break dates and size estimation error. Our empirical results suggest that parameter instability cannot fully explain the weak forecasting performance of most predictors considered in Goyal and Welch (2008). We find that out of 13 predictors, only one variable, the stock market variance, has evidence of structural breaks. But the evidence is not strong to rule out the stable model. Our empirical results suggest that the stock market variance has predictive power for market equity premium regardless of the presence of modeled breaks. Comparing the break model with the stable one, our results favor the former in forecasting excess returns.

CHAPTER 2. COMBINING MULTIPLE PREDICTIVE MODELS WITH POSSIBLE STRUCTURAL BREAKS

2.1 Introduction

Forecast combination is receiving growing attention in econometrics and finance. Combining predictive models is a smoothed extension of model selection, and may possibly substantially reduce risk relative to model selection. While a broad consensus is that forecast combination improves forecast accuracy, there is no consensus on how to construct the forecast weights. Particularly, researchers have recognized the usefulness of forecast combination in the presence of model parameter instability, and structural breaks are often mentioned as motivation for combining predictive models. The underlying idea is that models may differ in how they adapt to changes. Thus, when breaks are small, predictive models with stable parameters may outperform models with time-varying parameters. The converse is true in the presence of large breaks happened in the distant past. Since estimating the break dates and sizes precisely is difficult in real time, it is possible that combining forecasts from models with different degrees of adaptability can offer significant gains relative to selecting a single best model. Recent literature on economic forecasting[1] has focused on two particularly appealing methods, equal weights and Bayesian averaging. The equal weights method selects a set of models and then assigns them all equal weight for all forecasts. The Bayesian averaging method produces weights as by-product of Bayesian model averaging. In addition to the aforementioned

[1] See Timmermann (2006) and Rossi (2013).

weights, Hansen (2007) proposes Mallows' model averaging, and has extended the theory to various settings in subsequent research.[2]

In the literature on forecasting the market equity premium, Rapach et al. (2010) and Rapach and Zhou (2013) show that forecast combination can deliver statistically and economically significant out-of-sample gains relative to the historical average returns consistently over time. They argue that model uncertainty and instability seriously impair the predictability of individual model and the empirical explanations for the benefits of forecast combination are that combining forecasts can take advantage of all available information and combining forecasts are linked to the real economy. In their empirical analysis, they report that forecast combination can solve the puzzle presented in Goyal and Welch (2008) that many economic variables have week or no predictive power to forecast the U.S. market excess returns based on linear models. Specifically, Rapach et al. (2010) and Rapach and Zhou (2013) apply combination methods such as equal weights and discounted mean squared forecast error weights to demonstrating its superior out-of-sample performance relative to the historical mean benchmark. But it is not clear in their analysis how the equal weights and the discounted mean squared forecast error weights are related to structural breaks.

This paper introduces a two-stage forecast combination method which explicitly deals with structural breaks. In the first stage, to take into account the uncertainty on parameter instability for each predictive model, we combine its stable and break cases by using one of the four proposed methods, namely, equal weights, discounted mean squared forecast error weights, Schwarz information criterion weights weights and Mallows' weights. Next, we pool all adjusted predictive models obtained from the first stage together by applying equal weights. We recommend using the Mallows' weights in the first stage because it is theoretically justified by Hansen (2009).[3]

[2]See Hansen (2008), Hansen (2009) and Hansen and Racine (2011)

[3]In our empirical analysis, Mallows' weights work the best among all four methods in forecasting excess returns using Goyal and Welch's updated data.

To evaluate our two-stage forecast combination method and to compare results with related literature, we apply the two-stage forecast combination method to forecasting the U.S. market equity premium out-of-sample using an updated comprehensive dataset from Goyal and Welch (2008), and compare our results with those from Rapach et al. (2010) and Rapach and Zhou (2013) in similar studies. It is worth mentioning that in this paper we use all frequencies of data available to thoroughly investigate the empirical performance for all combination methods.[4] Our empirical results suggest that the two-stage forecast combination method, especially the one based on Mallows' weights, can potentially offer substantial forecasting gains relative to a simple one-stage equal weighting method used in Rapach et al. (2010) and Rapach and Zhou (2013) over the same dataset.[5]

This paper builds on an extensive literature on forecast combination and market equity premium prediction. Timmermann (2006) provides a comprehensive survey on forecast combination by analyzing theoretically the factors that determine the advantages from combining forecasts, and discussing several cases related to model misspecification, parameter instability and the role of combinations under asymmetric loss. Hansen (2008) proposes forecast combination based on the Mallows' information criterion which is an asymptotically unbiased estimate of both the in-sample mean squared error and the out-of-sample one-step ahead mean squared forecast error. Clark and McCracken (2010) examines the effectiveness of combining various models of instability in improving VAR forecasts made with real-time data, and considers a wide range of forecast combination methods in their analysis. Elliott (2011a) examines the sizes of the theoretical gains to optimal combination and provides conditions under which averaging and optimal combination are equivalent. Cheng and Hansen (2013) consider forecast combination with factor-augmented regression and investigate forecast combination across models using

[4]Rapach et al. (2010) uses quarterly data only. Rapach and Zhou (2013) uses monthly data only.

[5]Rapach et al. (2010) and Rapach and Zhou (2013) also consider other methods, such as median weighting, trimmed mean weighting and discounted mean squared forecast error weights with different values of the discount factor, the simple one-step overall equal weights perform the best.

weights that minimize the Mallows' and the leave-h-out cross validation criteria. Rapach and Wohar (2006) examine the structural stability of predictive regression models of the U.S. quarterly aggregate real stock returns over the postwar era. They find strong evidence of structural breaks in several bivariate predictive regression models of S&P 500 returns. Goyal and Welch (2008) systematically investigates the in-sample and out-of-sample performance of linear regressions that predict the equity premium with prominent variables suggested by the academic literature. Campbell and Thompson (2008) show that many predictive regressions of equity premium using other financial predictors can beat the historical market average return once some restrictions are imposed on the signs of model parameters and return forecasts. Rapach et al. (2010) recommend combining individual forecasts to predict market equity premium and show that forecast combination offers statistically and economically significant out-of-sample gains relative to the historical average market returns consistently over time. Rapach and Zhou (2013) survey the literature on equity premium forecasting and show strategies, such as economically motivated model restrictions, forecast combination, diffusion indices and regime switching models, can improve forecasting performance by addressing the substantial model uncertainty and parameter instability.

The remainder of this paper is organized as follows. Section 2.2 presents the econometric models, estimation method, out-of-sample forecast procedure and forecast combination methods. Section 2.3 presents the data and our empirical results. Section 2.4 concludes.

2.2 Econometric Model

2.2.1 Bivariate Predictive Model

Following Goyal and Welch (2008) and Rapach et al. (2010), we first consider the stable one-step ahead bivariate predictive model:

$$r_{t+1} = \beta_0^i + \beta_1^i X_{i,t} + e_t \tag{2.1}$$

where r_{t+1} is the one period ahead market equity premium, $X_{i,t}$ is predictor i available at time t to forecast the next period excess returns and e_t is a disturbance term. We generate a series of out-of-sample forecasts of the market equity premium using a recursive estimation window. Specifically, we split the total sample of T observations into two parts, an estimation sample of size R and an evaluation sample of size P, where $R+P=T$. Under the recursive window, at each point in time, the estimated parameters are updated by adding one more observation starting with sample size R. For example, the first out-of-sample forecast of the market equity premium based on predictor $x_{i,t}$ is

$$\hat{r}_{i,\,R+1} = \hat{\beta}^i_{0,\,R} + \hat{\beta}^i_{1,\,R} X_{i,\,R} \tag{2.2}$$

where $\hat{\beta}^i_{0,\,R}$ and $\hat{\beta}^i_{1,\,R}$ are the ordinary least squares estimates of β_0^i and β_1^i, respectively, in equation (2.1) using the first R observations in the sample. Then, the second period out-of-sample forecast is

$$\hat{r}_{i,\,R+2} = \hat{\beta}^i_{0,\,R+1} + \hat{\beta}^i_{1,\,R+1} X_{i,\,R+1} \tag{2.3}$$

where $\hat{\beta}^i_{0,\,R+1}$ and $\hat{\beta}^i_{1,\,R+1}$ are the least squares estimates of β_0^i and β_1^i using the first $R+1$ observations. Proceeding in this manner through the end of the out-of-sample period T, we have recursively produced a sequence of out-of-sample forecasts of size P, $\{\hat{r}_{s+1}\}_{s=1}^{P}$, using predictor $x_{i,t}$. Using predictive model (2.1), we can apply the same procedure to the rest of predictors, $x_{i,t}$, where $i = 1, ..., M$. Since we have 14 predictors available to

forecast excess returns across all data frequencies in our empirical applications presented in section 2.3, $M = 14$.

In additional to the bivariate predictive model (2.1) using various predictors to forecast the market equity premium, we can simply use the historical mean of the excess returns for prediction

$$r_{t+1} = \beta_0^i + e_t. \tag{2.4}$$

We apply the same out-of-sample procedure outlined previously to generate a sequence of forecasts of size P according to model (2.4), we label this series $\{\bar{r}_{s+1}\}_{s=1}^{P}$. In the literature of examining the efficient capital market hypothesis and forecasting stock returns, the historical mean model (2.4) serves as a natural benchmark predictive model to compare with other proposed complex models, see Goyal and Welch (2008), Campbell and Thompson (2008) and Rapach et al. (2010). To compare results with related literature, we continue to use model (2.4) as benchmark in this paper.

2.2.2 Forecast Combination

Since there is no prior information implying that model (2.1) is the true model or the best linear predictor, given that researchers have documented evidence of parameter instability or structural breaks in linear models forecasting stock returns,[6] a natural competing alternative to model (2.1) is a model with breaks in its coefficients

$$r_{t+1} = \beta_{0,t}^i + \beta_{1,t}^i X_{i,t} + e_t. \tag{2.5}$$

Because of sample size concerns and the uncertainty surrounding the quality of the estimates of structural break dates and sizes, we only consider the one break model in this paper,

$$r_{t+1} = \begin{cases} \beta_{0,1}^i + \beta_{1,1}^i X_{i,t} + e_t & \text{if } t < \tau \\ \beta_{0,2}^i + \beta_{1,2}^i X_{i,t} + e_t & \text{if } t \geq \tau \end{cases} \tag{2.6}$$

[6] See Rapach and Wohar (2006) and Paye and Timmermann (2006)

where τ is the time index of the break. The break date τ is restricted to the interval $[\tau_1, \tau_2]$ which is bounded away from the ends of the sample on both sides, $1 < \tau_1 < \tau_2 < R$. Following related literature,[7] we restrict that the break date falls into the middle 70% portion of the estimation sample.

The break date can be estimated by concentration. That is, for a fixed value of τ, we can estimate the piece-wise model parameters by least squares and then calculate the sum of squared errors, $\text{SSE}(\tau) = \hat{e}_t(\tau)$. We apply this procedure to all possible values of τ in the interval $[\tau_1, \tau_2]$, so we have a series of values of the sum of squared errors, $\{\text{SSE}_s(\tau)\}_{s=\tau_1}^{\tau_2}$. Our estimate of the break date, $\hat{\tau}$, would be the value of τ that is the global minimizer of $\{\text{SSE}_s(\tau)\}_{s=\tau_1}^{\tau_2}$.

For a given bivariate predictive model with predictor X_i, we can choose a stable version of model (2.1) or a break version of model (2.6). Next, we are going to present several forecast combination methods to combine model (2.1) and model (2.6) to form a averaged model with predictor X_i. We assign weight w to the break model (2.6) and $1 - w$ to the stable model (2.1), where $w \in [0, 1]$, so the averaged model, MOD_i, is

$$r_{t+1} = w\left\{\beta_{0,t}^i + \beta_{1,t}^i X_{i,t}\right\} + (1-w)\left\{\beta_0^i + \beta_1^i X_{i,t}\right\} + e_t. \tag{2.7}$$

2.2.2.1 Equal Weights

We can equally weight all models without estimating or calculating any additional parameters, so the averaged model (2.7), MOD_i^e, is

$$r_{t+1} = \frac{1}{2}\left\{\beta_{0,t}^i + \beta_{1,t}^i X_{i,t}\right\} + \frac{1}{2}\left\{\beta_0^i + \beta_1^i X_{i,t}\right\} + e_t. \tag{2.8}$$

2.2.2.2 Discounted Mean Squared Forecast Error Weights

Stock and Watson (2003) propose a discounted mean squared forecast error (DMSFE) combination method that computes weights based on the past predictive performance of

[7] See Andrews (1993) and Hansen (2009).

individual models over a holdout out-of-sample period. That is, for model j at time t,

$$w_{j,t}^{d} = \frac{\phi_{j,t}^{-1}}{\sum_{s=1}^{M} \phi_{s,t}^{-1}} \tag{2.9}$$

where

$$\phi_{j,t} = \sum_{l=1}^{t} \theta^{t-l} (r_{l+1} - \hat{r}_{l+1})^2 \tag{2.10}$$

and θ is a discount factor. The DSMFE method assigns greater weight to individual predictive model with lower past mean squared forecast error (MSFE) over the holdout out-of-sample period. When $\theta = 1$, there is no discounting, so all past observations are treated equally when calculating MSFE over the holdout period. If $\theta < 1$, DMSFE allows for greater weights on the more recent observations. The averaged model (2.7), MOD_i^{d}, is

$$r_{t+1} = w_{i,t}^{d} \left\{ \beta_{0,t}^{i} + \beta_{1,t}^{i} X_{i,t} \right\} + (1 - w_{i,t}^{d}) \left\{ \beta_{0}^{i} + \beta_{1}^{i} X_{i,t} \right\} + e_t. \tag{2.11}$$

2.2.2.3 Schwarz Information Criterion Weights

The Schwarz information criterion weight[8] for each predictive model is calculated based on the associated value of the Schwarz information criterion (SIC). For example, at time t, if the the SIC value for the break model (2.6) is $\text{SIC}^b(t)$ and the SIC value for the stable model (2.1) is $\text{SIC}^s(t)$, then the SIC weight for the break model, w_t^s, is

$$w_t^s = \frac{\exp\left(\text{SIC}^b(t)\right)}{\exp\left(\text{SIC}^b(t)\right) + \exp\left(\text{SIC}^s(t)\right)}. \tag{2.12}$$

The averaged model (2.7), MOD_i^{d}, is:

$$r_{t+1} = w_{i,t}^{s} \left\{ \beta_{0,t}^{i} + \beta_{1,t}^{i} X_{i,t} \right\} + (1 - w_{i,t}^{s}) \left\{ \beta_{0}^{i} + \beta_{1}^{i} X_{i,t} \right\} + e_t. \tag{2.13}$$

2.2.2.4 Mallows' Weights

Hansen (2007) proposes an averaging estimator with the weight selected to minimize a Mallows' information criterion, which is an asymptotically unbiased estimate of both

[8] See Timmermann (2006) and Rossi (2013).

the in-sample mean squared error and the out-of-sample one-step ahead mean squared forecast error. Subsequently, Hansen (2009) extends Mallows' model averaging to the structural break case. Specifically, at time period t, the Mallows' weight for the break model (2.6), w_t^m, is

$$w_t^m = \begin{cases} 0 & \text{if } \mathrm{F}_t < \bar{p} \\ 1 - \frac{\bar{p}}{\mathrm{F}_t} & \text{if } \mathrm{F}_t \geq \bar{p} \end{cases} \tag{2.14}$$

where F_t is the standard F-test statistic in Andrews (1993), and $\bar{p}$ is a penalty coefficient whose value depends on the asymptotic distribution of the Andrews' SupF test statistic. Hansen (2009) provides a table of $\bar{p}$ values for various cases.

The averaged model (2.7), $\mathrm{MOD}_i^{\mathrm{m}}$, is

$$r_{t+1} = w_{i,t}^m \left\{\beta_{0,t}^i + \beta_{1,t}^i X_{i,t}\right\} + (1 - w_{i,t}^m) \left\{\beta_0^i + \beta_1^i X_{i,t}\right\} + e_t. \tag{2.15}$$

2.2.2.5 Combining All Predictive Models

Since the seminal work of Bates and Granger (1969), it has been known that combining forecasts across predictive models can produce forecasts that outperform any single individual model. Forecast combination can be viewed as a diversification strategy analogous to portfolio diversification that improves forecasting performance.

For our bivariate predictive model (2.1), since totally we have M predictors available, there are M candidate models to forecast the market equity premium. Once we include break model (2.6), we end up with $2M$ predictive models. Previously we have shown several methods to average the stable and break version of a predictive model with predictor X_i, next, we are going to combine all $2M$ models to form one pooled model.

Specifically, for each bivariate model i with predictor X_i, first, we combine its stable and break cases using the four previously outlined combination methods to get MOD_i^j, $i = 1, ..., M$ and $j \in \{e, d, s, m\}$. So for a given weighting method j, at this point we have M models left from the initial $2M$ models.

Next, we will assign equal weight to all M models for a given method j, that is, $\frac{1}{M}\sum_{i=1}^{M}\text{MOD}_i^j$. This will be our final combined model to forecast,

$$r_{t+1} = \frac{1}{M}\sum_{i=1}^{M}\left\{w_{i,t}^j\left[\beta_{0,t}^i + \beta_{1,t}^i X_{i,t}\right] + (1 - w_{i,t}^j)\left[\beta_0^i + \beta_1^i X_{i,t}\right]\right\} + e_t \tag{2.16}$$

where $j \in \{e, d, s, m\}$.[9]

Intuitively, we have introduced a two-stage weighting procedure. In the first stage, for each predictive model i, we combine its stable and break cases by using one of the four outlined methods, namely, equal weights, DMSFE weights, SIC weights and Mallows' weights. Then, we pool all models together by equal weights.

Note that in the second stage we only consider equal weights to pool all models which have been averaged for parameter instability in the first round. The reason is that in many empirical applications, combining a large number of predictive models by equal weights tend to outperform other complex methods. But at the first stage regarding model parameter instability, complex averaging methods may offer substantial gains.

2.2.3 Forecast Evaluation

A popular metric to evaluate forecasts is the mean squared forecast error (MSFE)

$$\text{MSFE} = \frac{1}{T}\sum_{t=1}^{T}\left(y_t - \hat{y}_t\right)^2. \tag{2.17}$$

Since in this paper we compare forecasting performance of various predictors and combination methods with the historical mean benchmark, and to compare our results with those from related literature, we adopt Campbell and Thompson's out-of-sample R_{OS}^2 statistic to compare methods and models,

$$R_{OS}^2(i) = 100 \times \left(1 - \frac{MSFE^i}{MSFE^0}\right) \tag{2.18}$$

where i indexes the model or method and 0 represents the benchmark. The R_{OS}^2 statistic measures the reduction in MSFE for the predictive model or combination forecast relative

[9] e: equal weights. d: DMSFE weights. s: SIC weights. m: Mallows' weights.

to the historical average model. Thus, if the R^2_{OS} statistic is positive, it indicates better forecasting performance for model i than the benchmark model. The higher the R^2_{OS} value, the better the out-of-sample performance.

In addition, following Goyal and Welch (2008) and Rapach and Zhou (2013), to construct a series of graphical device to evaluate the out-of-sample forecasting performance for the benchmark and the competing models, we adopt the cumulative difference in squared forecast errors (CDSFE) between the historical mean model and the bivariate model or the combined model as another metric,

$$\text{CDSFE}_\text{t} = \sum_{s=1}^{t} (r_{s+1} - \bar{r}_{s+1})^2 - \sum_{s=1}^{t} (r_{s+1} - \hat{r}_{s+1})^2 \tag{2.19}$$

where $\bar{r}_{s+1}$ is the forecast from the benchmark model (2.4) and $\hat{r}_{s+1}$ is the forecast from model (2.1) or model (2.16). At any period t, if the value of CDSFE_t is positive, it implies that the competing model outperforms the benchmark by having a smaller prediction error rate.

Time series plots of the CDSFE curve can be conveniently used to determine if the competing model has a lower MSFE than the historical average benchmark for any period by simply comparing the height of the curve at the beginning and end points of the segment corresponding to the period of evaluation. If the curve is higher at the end of the evaluation period relative to the starting time, then the competing model or method has a lower MSFE than the benchmark during the out-of-sample evaluation period. A model or method which forecasts better than the historical average model will thus have a slope that is positive everywhere during the out-of-sample evaluation period.

2.3 Empirical Results

2.3.1 Data and Out-of-sample Forecast

Our data are from Goyal and Welch (2008). We use the most recently updated data up to the year of 2013. The market equity premium is the log returns on the S&P 500

index including dividends minus the log returns on the risk-free rate. Since we apply forecast combination methods to data of all frequencies, we only keep the following fourteen variables which are available for all data frequencies:

- Log dividend-price ratio (dp): log of a 12-month moving sum of dividends paid on the S&P 500 index minus the log of stock prices.
- Log dividend yield (dy): log of a 12-month moving sum of dividends minus the log of lagged prices.
- Log earnings-price ratio (ep): log of a 12-month moving sum of earnings on the S&P 500 index minus the log of stock prices.
- Log dividend-payout ratio (de): log of a 12-month moving sum of dividends minus the log of a 12-month moving sum of earnings.
- Stock market variance (svar): monthly sum of squared daily returns on the S&P 500 index.
- Book to market ratio (bm): ratio of book value to market value for the Dow Jones Industrial Average.
- Net equity expansion (ntis): ratio of a 12-month moving sum of net equity issues by NYSE-listed stocks to the total end of year market capitalization of NYSE stocks.
- Treasury bill rate (tbl): interest rate on a three-month secondary market Treasury bill.
- Long term yield (lty): long term government bond yield.
- Long term return (ltr): return on long term government bonds.
- Term spread (tms): long term yield minus the Treasury bill rate.

- Default yield spread (dfy): difference between BAA- and AAA-rated corporate bond yields.
- Default return spread (dfr): long term corporate bond returns minus the long term government bond returns.
- Inflation (infl): inflation is the Consumer Price Index (all urban consumers) from the Bureau of Labor Statistics.

As for the sample split choices, to make our results comparable to those of Goyal and Welch (2008), Rapach et al. (2010) and Rapach and Zhou (2013), we adopt the following choices:

- Monthly data: the estimation sample runs from 1927:01 to 1956:12 (R = 360), and the evaluation sample runs from 1957:01 to 2013:12 (P = 684).
- Quarterly data: the estimation sample runs from 1947:1 to 1964:4 (R = 72), and the evaluation sample runs from 1965:1 to 2013:4 (P = 196).
- Yearly data: the estimation sample runs from 1927 to 1964 (R = 38), and the evaluation sample runs from 1965 to 2013 (P = 49).

We use the recursive window to forecast out-of-sample, meaning that the estimation sample always starts from the same beginning period and additional observations are used as they become available. Following this procedure, all model parameters are estimated recursively and out-of-sample forecasts are generated accordingly.[10]

The time series plots of monthly, quarterly and yearly data are presented in Figure B.1, Figure B.2 and Figure B.3, respectively. Additionally, we also present the correlation matrices for all variables across all data frequencies in Figure B.4, Figure B.5 and

[10]Model parameters can also be estimated using a rolling window, which drops earlier observations as additional data become available. Rolling window may be justified by appealing to structural breaks, but our results shown in this paper do not change substantially if we switch to use the rolling window.

Figure B.6. We can see that the dependent variable, the U.S. market equity premium, is weekly correlated with all 14 predictors across all frequencies. This may partially explain the earlier findings in Goyal and Welch (2008) that the simple bivariate predictive model (2.1) forecasts excess returns poorly out-of-sample compared with the historical mean model (2.4). Furthermore, some predictors, such as dividend-price ratio (dp), dividend-yield ratio (dy), earnings-price ratio (ep) and book to market ratio (bm), are highly correlated with each other. This may explain the poor out-of-sample performance of the "kitchen-sink" model which includes all predictors in Goyal and Welch (2008).

2.3.2 Bivariate Model Prediction

To reexamine the empirical results of Goyal and Welch (2008) with updated data, here we use model (2.1) to forecast equity premium out-of-sample for all 14 predictors. Then we present the out-of-sample time series plots of the cumulative difference of squared forecast error (CDSFE) between model (2.1) and the historical benchmark model (2.4) for all predictors. The monthly, quarterly and yearly CDSFE plots are shown in Figure B.7, Figure B.8 and Figure B.9, respectively. This is an informative graphical device that shows an individual model's out-of-sample forecasting performance over time. When the curve in each panel of those figures increases, the predictive model outperforms the historical average model, while the opposite holds when the curve decreases. These plots conveniently illustrate whether an individual model has a lower MSFE than the benchmark over a selected out-of-sample evaluation period. A predictive model that always beats the benchmark for any period will have a curve with a slope that is always positive.

From Figure B.7, Figure B.8 and Figure B.9, we can see that the historical mean model still outperforms most predictors even with updated data. Furthermore, there is no models or predictors which consistently outperform the historical mean over the evaluation period for all data frequencies. These figures also suggest that dividend-price

ratio (dp), dividend-yield ratio (dy), earnings-price ratio (ep) and book to market ratio (bm) tend to outperform the benchmark over the most part of the evaluation period. This is not surprising because in the previous section we have shown that these variables are highly positively correlated with each other, but they are weakly correlated with the market premium. Overall, whether these variables have statistically significant predictive power is questionable in linear predictive models, so it is still difficult to identify individual predictors that reliably beat the benchmark in predicting excess returns.

2.3.3 Forecast Excess Returns Using Combined Model

Rapach et al. (2010) and Rapach and Zhou (2013) conclude that forecast combination appears successful for out-of-sample premium prediction because it can substantially reduce forecast variance and include information from various economic predictors. They also suggest that the usefulness of forecast combination ultimately stems from the highly uncertain, complex and constantly evolving data-generating process underlying the expected market excess returns. They find that combining forecasts, especially using equal weights averaging all models, outperform the historical average by statistically and economically meaningful margins, and more consistently than a range of individual models suggested by the literature.

Our main empirical contribution is to prove that the two-stage forecast combination method outlined in section 2.2.2.5, can substantially improve the out-of-sample forecast performance using the same dataset studied by Rapach et al. (2010) and Rapach and Zhou (2013). The out-of-sample performance are evaluated and compared using both the CDSFE plots and the Campbell and Thompson (2008) R^2_{OS}. The results hold consistently for all data frequencies.

Before reporting our empirical results on two-stage forecast combination, we start with presenting the out-of-sample forecast correlation matrices for all bivariate models for monthly, quarterly and yearly data in Figure B.10, Figure B.11 and Figure B.12,

respectively. From these figures we can see that forecasts form some models are correlated to some degree, for example, forecasts from the dividend-price ratio model (dp) and the long term yield model (lty) are negatively correlated. This graphical device help us confirm that our choice of equal weights for the second stage combination outlined in section 2.2.2.5 is more desirable than optimization based complex weights, for example, the least squares weights.[11]

In additional to the four first-step combination methods outlined in section 2.2.2, namely, equal weights, DMSFE weights, SIC weights and Mallows' weights, we also consider combining all 14 stable models[12] (model (2.1)) and all 14 break models (model (2.6)) in the following analysis. They are labeled "stable" and "break", respectively, in our subsequent figures and tables. So there are six combined models available.

The solid lines in Figure B.13, Figure B.14 and Figure B.15 plot the differences between the cumulative squared forecast error for the historical mean model and the cumulative squared forecast error for the combined model of all six methods for monthly, quarterly and annual data, respectively. In contrast to previous figures for the bivariate model, the CDSFE curves in Figure B.13, Figure B.14 and Figure B.15 are predominantly positive, indicating that for this particular dataset and forecast problem, forecast combination delivers substantial and consistent gains compared with individual predictive model. Note that the combined models work particularly well for the monthly data as the slope of the CDSFE curve is almost positive for the entire out-of-sample evaluation period for all methods. While for the annual data, it looks like the forecast performance of the combined model somehow deteriorates towards the end of the evaluation sample period.

Furthermore, we are interested in comparing the six averaging methods used in our combined models. This can be done by comparing their associated R^2_{OS} statistics. Ta-

[11] See Timmermann (2006).

[12] This is equivalent to the "Mean" combination method used in Rapach et al. (2010), and the "POOL-AVG" method in Rapach and Zhou (2013).

ble B.1 reports the out-of-sample R^2_{OS} statistics for all combination methods for all data frequencies. R^2_{OS} measures the percent reduction in mean squared forecast error (MSFE) for the combination methods (2.16) given in the first row of Table B.1 relative to the historical average benchmark forecast by model (2.4). Cp stands for Mallows' weights. DMSFE is the discounted mean squared forecast error weights with discount factor $\theta = 1$. The column titled "Break" shows equal weights for the break version of all bivariate predictive models. The column titled "Stable" shows equal weights for the stable version of all bivariate predictive models. We apply the two-stage forecast combination procedure to the first four columns, meaning that in the first stage, for each bivariate predictive model, we use Cp, DMSFE, Equal or SIC weights to average its stable and break cases, then we apply equal weights to all 14 break-adjusted models. For the last two columns, we simply equally weight all 14 break or stable bivariate predictive models.

From Table B.1 we can see that forecast combination offers the largest gains relative to the benchmark for monthly data. More than 10% reduction in mean squared forecast error can be achieved by using combined models. Among the four methods dealing with the uncertainty on model parameter instability, Mallows' weights (Cp) perform the best for all data frequencies. In addition, Mallows' weights outperform the "mean" or "POOL-AVG" method[13] used in Rapach et al. (2010) and Rapach and Zhou (2013) in studying the quarterly and monthly data. It is surprising to see that Mallows' weights outperform "POOL-AVG" by more than 1% reduction in MSFE for quarterly data.[14] We conclude that using Mallows' weights to control for the uncertainty on model parameter instability in the first stage of forecast combination may offer substantial out-of-sample gains relative to a simple overall equal weights strategy.

The last two columns of Table B.1 suggest that structural breaks or parameter instability may be one of the reasons why individual predictive model (2.1) fails to beat the

[13]They are equivalent to the "Stable" column shown in Table B.1.

[14]In Rapach et al. (2010), "mean" or "POOL-AVG" combination method offers the largest gains relative to the benchmark among all models and methods considered for the quarterly evaluation sample starting from 1965:Q1.

historical mean model (2.4) in forecasting U.S. market equity premium out-of-sample as shown in Goyal and Welch (2008) and Rapach et al. (2010) for monthly and quarterly data. But it is not clear whether we can attribute the failure of individual predictive models to structural breaks for yearly data because our empirical results show that stable models offer almost 2% more reduction in MSFE than combining break models.

Since the main goal of this paper is to propose a two-stage forecast combination method, and to compare and evaluate model averaging methods using a popular dataset studied by other prominent researchers, we do not address the efficient capital market hypothesis problem in this paper. We propose a new forecast combination method which explicitly deals with structural breaks in linear predictive models and prove its effectiveness in forecasting the U.S. market excess returns out-of-sample. Specifically, we demonstrate the outstanding performance of Mallows' weights in averaging models with possible breaks via our empirical applications.

2.4 Conclusion

This paper has extended the forecast combination methods to predict the U.S. market equity premium out-of-sample. With the strong uncertainty on model parameter instability, we introduce a two-stage forecast combination method: in the first stage, for each predictive model, we combine its stable and break cases by using one of the four outlined methods, namely, equal weights, discounted mean squared forecast error weights, Schwarz information criterion weights weights and Mallows' weights. Next, we pool all adjusted predictive models obtained from the first stage together by applying equal weights. We apply our two-stage forecast combination method to forecasting the U.S. market equity premium out-of-sample using an updated comprehensive dataset from Goyal and Welch (2008), and compare our results with those from Rapach et al. (2010) and Rapach and Zhou (2013) in similar studies. Our empirical results using Goyal and

Welch's data suggest that the two-stage forecast combination method, especially the one based on Mallows' weights proposed by Hansen (2009), can potentially offer substantial forecast gains relative to a simple equal weighting method used in Rapach et al. (2010) and Rapach and Zhou (2013) over the same dataset. To compare empirical results with related literature, we use the out-of-sample R^2 statistic proposed by Campbell and Thompson (2008) to evaluate forecasts.

Our theory is confined to the context of linear predictive models. While it would be greatly desirable to extend the analysis to include other types of model. Another unexplored issue is inference. At this point it is not clear how to rigorously test whether or not the mean squared forecast error difference is statistically significant for combined models with explicitly modeled structural breaks. This is a challenging topic and quite important for future investigation.

CHAPTER 3. OUT-OF-SAMPLE FORECAST MODEL AVERAGING WITH PARAMETER INSTABILITY

3.1 Introduction

Forecast combination or model averaging has been a useful tool employed by econometricians and industry forecasters in studying many macroeconomic and financial time series, for example, GDP growth rate, unemployment rate, inflation rate and stock market returns. Combination methods such as Granger–Ramanathan, Bates–Granger, Bayesian model averaging, least squares combination, discounted mean square forecast error weights, time–varying combination and survey forecasts combination have been developed for forecasting under various settings.

There are several reasons explaining the popularity of forecast combination or model averaging in empirical research. First, it is highly possible that a single forecasting model is misspecified due to information constraints. For example, predictors that potentially could help improve forecasting performance are not included in the underlying model, so combining forecasts or averaging models may help the forecaster better manage the risk induced in the model selection process and take advantage of all available information. Even in a stationary world, the true data generating process may be a highly complicated nonlinear function of lags of infinite order and variables which are difficult to measure precisely in practice, consequently, most linear forecasting models proposed by researchers can only be viewed as local approximations for the best linear predictor. It is hard to believe that one predictive model strictly outperforms all other models at

all points in time, rather, the best forecasting model may change over time. Due to small sample size for some variables of interest and imperfect information, it is difficult to track the best model based on past forecasting performance. Therefore, combining models can be taken as a practical way to make forecasts robust to misspecification bias, especially when forecasts from various sources are not highly correlated. For example, if the bias is idiosyncratic in each individual model, then combining forecasts from all candidate models may help average out this bias.

Second, a forecasting model's adaptability to parameter instability or structural breaks may not be constant across time. Drastic government policy changes or financial institution reform may bring about structural breaks in the time series variable of interest. An example worth mentioning here is the Great Moderation. Many researchers, such as Stock and Watson (2003) and Stock (2004), agree that there is a structural break in the volatility of the U.S. GDP growth rate around mid-1980s as the series becomes less volatile since then. Other developed countries, such as Canada and Germany, have seen the same pattern starting around the same period.[1] Depending on the magnitude and the frequency of the break process, forecasters may prefer a non-stationary model in which all or some of the parameters have changed around the estimated break dates to a stable model where all parameters are assumed constant, but problems arise when the magnitude of the break is small or the evidence of parameter instability is not convincing. In this case, the pre-test model, that is, the single forecasting model selected based on hypothesis testing or information criteria, may not be the best choice for prediction if we assess and compare its performance with other candidates according to mean squared forecast error (**MSFE**). Why? On one hand, the estimation or dating of structural breaks can be very imprecise. On the other hand, the quality of the break dates estimates depends not only on the break size measured by some metric, but also on whether

[1] Arguments explaining this phenomenon include technology progress or innovation, monetary policy change and financial system reform, etc.

the impact of the break is dominated by the volatility of the process.[2] Additionally, for some time series variable of interest, we may reach different conclusions if we study the same variable with a different data frequency. For instance, researchers have conducted research on the stock market returns based on various frequency choices, daily, monthly, quarterly or yearly. For the structural break analysis, it is hard to confirm or prove that the estimated structural break dates from all frequencies coincide.[3] Given this model selection uncertainty, forecast combination may offer diversification gains that make it attractive to average the break and stationary models, rather than relying on a pre-test model. See Timmermann (2006) for a comprehensive survey of forecast combination.

In an empirical paper studying the U.S. aggregate equity market returns, Rapach et al. (2010) argue that forecast combination is a powerful tool against structural breaks in predicting excess stock returns. For given sample split choices, according to Campbell and Thompson (2008) out-of-sample R^2 statistic, they show that forecasts generated by pooling all fifteen models are more accurate than those obtained from any single forecasting model or the large kitchen-sink model. But they do not provide detailed econometric theory explaining why forecast combination methods, such as equal weight and discounted mean squared forecast error weight used in their paper, may help deal with structural breaks.

In spite of these aforementioned possible benefits, a puzzle associated with forecast combination is that in many empirical applications, equally weighted forecast schemes, i.e., each candidate model receives weight one divided by the total number of models, tend to outperform various optimal combination weights proposed by researchers, notably the Granger–Ramanathan combination. A paper attempting to explain this puzzle

[2]We have conducted simulation for this case. Our simulation results indicate that, even if there is a break in the conditional mean of the DGP, as long as the magnitude of the break is strictly dominated by the variance of the error term, it turns out that the stable version of the DGP outperforms the true DGP evaluated by root mean squared forecast error on average.

[3]For example, the estimated break date based on monthly data does not fall into the same year if estimated using yearly data. There are several empirical papers (Rapach and Wohar (2006) Paye and Timmermann (2006)) related to dating structural breaks based on different data frequencies and models.

is written by Elliott (2011b). Elliott argues that if the variance of the unforecastable component of the variable is large, the gains from optimal forecast combination will be strictly dominated by the unpredictable component. Additionally, the noise introduced by estimating various optimal combination weights, especially when the number of weights is large, further reduces combination gains.

Having all the benefits and drawbacks mentioned above in mind, in this paper, we focus on the situation where forecasts are generated by two competing models and study if we can come up with model averaging weights possibly superior to others in terms of better managing structural breaks and conditional heteroscedasticity. These two competing models share the same regressors, but one has a structural break in the conditional mean while the other is stable. This framework applies to situations in which: (i) Researchers or forecasters cannot find convincing evidence supporting structural breaks; (ii) The model is not correctly specified. Specifically, we propose model averaging weights derived from the cross–validation information criterion to combine the break model and the stable model in the out-of-sample forecast setting.

The cross–validation information criterion is an unbiased estimate of the mean squared forecast error or the expected test error rate, so naturally, it is appropriate to apply CV to the out-of-sample forecasting and forecast evaluation analysis. Studies have shown that the cross–validation criterion outperforms various other criteria in model selection under conditional heteroscedasticity, notably in determining the order of ARMA models. Under the assumption of conditional homoscedasticity, we show that the cross–validation criterion is asymptotically equivalent to Mallows' Cp criterion, so the asymptotic optimality properties associated with Mallows' weights carry over to the cross-validation weights. A natural extension is to relax this homoscedastic error assumption as it may be too strict for relevant empirical applications. Our main contribution is to derive the cross–validation model averaging weights under conditional heteroscedasticity with breaks, and to show that CV weights are the correct weights minimizing the expected

mean squared forecast error in this situation. Monte Carlo evidence and empirical examples are provided to support our results.

The remainder of this paper is organized as follows: Section 3.2 provides a review of the related literature. Section 3.3 first describes the econometric model and the forecasting problem, then presents theoretical results for the model averaging weights. Section 3.4 presents Monte Carlo evidence. Section 3.5 provides two empirical examples comparing our method with others. Section 3.6 concludes.

3.2 Related Literature

This paper relies on the literature related to the information criterion-based model selection and averaging, structural breaks testing and out-of-sample forecast comparison and forecast evaluation.

Recently, Hansen has published a series of papers[4] which help develop relevant econometric theory for the use of model averaging under various situations, and has pushed the forecast combination theory to a new level. He establishes that under the assumption of conditional homoscedasticity and the restriction of weight discretization, model averaging estimators based on Mallows' criterion are asymptotically optimal in the sense of minimizing the expected mean squared error (**MSE**) while controlling omitted variable bias. The reason for using Mallows' criterion is because it is an asymptotically unbiased estimator of the in-sample MSE or one-step ahead out-of-sample MSFE compared with other criteria, such as Akaike information criterion (**AIC**) or Schwarz-Bayesian information criterion (**SIC**). Hansen (2008) then extends his Mallows' model averaging theory to forecast combination and compares its performance with other related combination methods based on simulated data. He shows that Mallows' criterion is an approximately unbiased estimator of MSFE even for a stationary time series, but the optimality results do not apply. In order for the asymptotic optimality results to hold, we need the data of

[4]See Hansen (2007), Hansen (2008), Hansen (2009) and Hansen and Racine (2011).

interest to be independent and identically distributed. Unfortunately, this restriction of i.i.d. data has made the optimality property less relevant to many empirical applications where the data under study is time series, for example, GDP growth rate, stock returns, inflation rate and currency market volatility. Even more stringently, Hansen imposes the restriction that the models under consideration are strictly nested in order to ensure optimality.[5] Having these restrictions mentioned above, it is natural to replace Mallows' Cp with a criterion which can be applied more generally. Comparing Mallows' Cp with the cross-validation criterion, Andrews (1991) demonstrates that Mallows' criterion is no longer optimal in model selection if allowing for conditional heteroscedasticity, and CV is the only feasible criterion among popular candidates that are asymptotically optimal under general conditions. Following earlier research, Hansen and Racine (2011) relax the assumption of conditional homoscedasticity and nested linear models to show model averaging optimality by replacing the Mallows' criterion with the cross-validation criterion, but the asymptotic optimality property is still restricted to random samples. Alternatively, Liu and Okui (2012) propose a heteroscedasticity-robust Mallows' criterion which generalizes Hansen's least squares model averaging optimality results by allowing for conditionally heteroscedastic errors.

To make model averaging more appealing to empirical applications, it is natural to extend the optimal weighting theory to the structural breaks setting, so bringing leading research on dating and estimating breaks to model combination is desirable. Historically, applied econometricians rely on the Chow test to test for structural breaks, but the use of Chow's test assumes that the researcher knows the exact date of the structural break, if it indeed happens. If the researcher or policy maker has superior information set on possible break dates, or events potentially leading to parameter instability, conducting inference by Chow test seems reasonable. Otherwise, this assumption seems quite unrealistic and requires that econometricians visually examine the time series data to search for a

[5]Hansen considers a sequence of nested MA models.

possible break point. To take the impact of unknown break date into account, in a seminal paper, Andrews (1993) proposes a SupW type test statistic for detecting breaks and presents the associated asymptotic distribution for the test statistic. Note that Andrews' paper does not explicitly show how to estimate the break date and its consistency, but it implies that the break date can be estimated by concentration.[6] Subsequently, Bai (1997), Bai (1999) and Bai and Perron (1998) have a series of articles on rigorous break date estimation and testing, and have extended the econometric theory to multiple breaks and partial breaks. Bai and Perron's computational procedure for detecting breaks is adopted in many empirical works related to macroeconomic and financial time series since it is reasonable to think that there could be multiple structural breaks, for example, the U.S. equity markets have experienced institutional change and several financial crises since the early twentieth century. Additionally, there is research on optimal testing in the structural change setting, see Andrews and Ploberger (1994), Andrews (2003), Hansen (2000), Elliott and Muller (2006) and Rossi (2005).

For the prediction problem, from the perspective of a forecaster, testing for structural breaks is not the end. How to better predict the future and evaluate forecasts is of great importance to econometricians working on economic forecasting. Theory on forecasting with breaks is still evolving as new methods are proposed and evaluated. One specific research topic is the selection of the optimal data window to estimate the predictive model. The choice of window involves a bias-variance trade-off: For a given break date estimate, including more data before the estimated date may help reduce the mean squared forecast error, but doing so could result in more bias in the parameter estimation.[7] As an alternative to model averaging when parameter instability is possible, researchers have proposed various in–sample and out–of–sample tests to select a predictive model which is robust to structural breaks.[8]

[6]The date that leads to the largest reduction of the sum of squared errors relative to the no break benchmark.

[7]See Pesaran and Timmermann (2007) and Pesaran et al. (2011).

[8]See Giacomini and Rossi (2010), Bunzel and Calhoun (2012) and Inoue and Kilian (2004).

3.3 Econometric Theory

3.3.1 Model and Estimation

The econometric model used to forecast and its estimation method are closely related to Hansen (2009) and Andrews (1993).[9] The model we are interested in is a linear time series regression with a possible structural break in the conditional mean. The observations we have are time series $\{y_t, x_t\}$ for $t = 1, ..., T$, where y_t is the scalar dependent variable and x_t is a $k \times 1$ vector of related predictors and possibly lagged values of y_t, k is the total number of regressors or predictors included.[10] Parameters are estimated by ordinary least squares. The forecasting model allowing for structural break is:

$$y_t = x_t'\beta_1 I_{[t<m]} + x_t'\beta_2 I_{[t\geq m]} + e_t \tag{3.1}$$

where $I_{[\bullet]}$ is an indicator function, m is the time index of the break and $E(e_t|x_t) = 0$. The break date is restricted to the interval $[m_1, m_2]$ which is bounded away from the ends of the sample on both sides, $1 < m_1 < m_2 < T$. In practice, a popular choice is to use the middle 70% portion of the sample. We assume that all information relevant to forecasting is included in the regressors x_t, and the source of model misspecification comes solely from the uncertainty about parameter stability. This is in contrast to many applied econometric models where model misspecification bias comes from the wrong choice of regressors but the parameters are assumed stable.

We can also use a stable linear model to forecast:

$$y_t = x_t'\beta + e_t \tag{3.2}$$

The traditional pre-test procedure starts with performing a test for structural breaks, for example, using Andrews' SupF or SupW test, and then decide which model to choose

[9] Andrews considers GMM as the primary estimation method.

[10] Since we are interested in forecasting, y_t can be thought of as the variable to be predicted for the next period using currently available information x_t.

based on testing results.[11] As an alternative to model selection, we can combine these two models by assigning weight w to model 3.1 and $1-w$ to model 3.2, where $w \geq 0$. So the combined predictive model is

$$y_t = w\left\{x_t'\beta_1 I_{[t<m]} + x_t'\beta_2 I_{[t\geq m]}\right\} + (1-w)\left\{x_t'\beta\right\} + e_t \tag{3.3}$$

With the forecasting model ready, next, we are going to present the cross-validation criterion in detail which is crucial in determining the optimal weight w in equation 3.3.

3.3.2 Cross-Validation Criterion

There are several popular information criteria for model selection: for example, Akaike information criterion (**AIC**), corrected AIC (**AIC$^{\mathbf{c}}$**), Schwarz-Bayesian information criterion (**SIC**), Hannan-Quinn (**HQ**) and Mallows' C_p (**C_p**). Most criteria have two components in their formulas: the first part measures model fit while the second penalizes overfitting. Many information criteria share the same component measuring the in-sample fit, but they differ in the degree of overfitting penalization. For instance, AIC penalizes each additional parameter by 2 while SIC penalizes overfitting by the logarithm of the sample size, so SIC tends to select a more parsimonious model than AIC if the sample size is large.

For the forecasting analysis, what we care about is the test error rate assessing the model predictive ability, not the training error rate produced in the model estimation stage, so selecting a information criterion which gives a good estimate of the expected test error rate is crucial. Cross-validation is such a criterion. Specifically, we focus on the use of the leave-one-out cross-validation for this paper, though other CV variants, such as K–fold cross-validation, may be considered. Cross-validation is computationally simple for the one-step ahead predictive model selection and is shown robust to

[11] This can be done in various ways. One is to treat various possible number of breaks as different models, then select one according to some information criterion, e.g., AIC, SIC or Mallow's. Another way is hypothesis testing, following the relevant testing procedures outlined in Andrews (1993), Bai and Perron (1998) and Elliott and Muller (2006).

conditional heteroscedasticity in the econometrics and statistics literature. For forecast combination, researchers have applied CV to the quadratic programming based model averaging analysis, but its setting does not include structural change.

The sample leave-one-out cross-validation criterion can be computed by the following procedure:

$$\widehat{CV_T}(m) = \frac{1}{T}\sum_{t=1}^{T}\tilde{e}_t(m)^2 \tag{3.4}$$

where $\tilde{e}_t(m) = y_t - \tilde{\beta}_{-t}(m)'x_t(m)$ are the residuals from the regression with the t^{th} observation dropped and $\tilde{\beta}_{-t}(m) = (\sum_{i\neq t} x_i(m)x_i(m)')^{-1}(\sum_{i\neq t} x_i(m)y_i)$ is the associated vector of parameter estimates. Intuitively, this procedure is trying to estimate the expected test error rate based on the training data. Though equation 3.4 implies that we need to run the regression T times for given sample size T, fortunately, for linear regression models, we can calculate the sample CV value by running regression only once. Formally, the leave-one-out cross-validation residuals can be computed from the full sample least squares residuals, $\tilde{e}_t = \frac{\hat{e}_t}{1-h_t}$, where $h_t = x_t'(X_t'X_t)^{-1}x_t$ is the leverage associated with observation t, $\hat{e}_t$ is the full sample least squares residual and $\tilde{e}_t$ is the cross-validation residual. So we can rewrite equation 3.4 as

$$\widehat{CV_T}(m) = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\hat{e}_t(m)}{1-h_t}\right)^2 \tag{3.5}$$

In the next section we are going to show how model averaging weights are derived from the cross-validation criterion.

3.3.3 Cross-Validation Weights

We start this section by listing relevant assumptions needed for our results.

Assumption 1. *Suppose the following holds:*

1. *The true data generating process satisfies the linear process* $y_t = x_t'\beta_t + e_t$, $t = 1, ..., T, \beta_t \in \mathbb{R}^k$, *where* $\beta_t = \beta + T^{-1/2}\eta(t/T)\delta\sigma_t$. $\eta(\bullet)$ *is a* $\mathbb{R}^k$ *valued Riemann integrable function on* $[0, 1]$ *and* $\delta \in \mathbb{R}\backslash\{0\}$ *is a scalar indexing the magnitude of parameter variation,* σ_t *is the standard deviation of the error term at period* t.

2. $\{(x_t', e_t)\}$ *is* α*-mixing of size* $-r/(r-2), r > 2$ *or* ϕ*-mixing of size* $-r/(2r-2), r \geq 2$.

3. $E(x_t e_t) = 0, \forall t$, *and the process* $\{x_t e_t\}$ *is uniformly* L_r*-bounded, i.e.,* $||x_t e_t||_r < B$, *where* B *is a constant and* $B < \infty$.

4. $T^{-1/2}\sum_{t=1}^{[\pi T]} x_t e_t \Rightarrow W(\pi)$ *where* $W(\pi)$ *is a* $k \times 1$ *Wiener process with symmetric, positive definite long-run covariance matrix* $\Sigma \equiv \lim_{T\to\infty} \text{VAR}(T^{-1/2}\sum_{t=1}^{[\pi T]} x_t e_t)$, *for* $0 \leq \pi \leq 1$. '$\Rightarrow$' *denotes the weak convergence of the underlying probability measure as* $T \to \infty$.

5. $T^{-1}\sum_{t=1}^{[\pi T]} x_t x_t'$ *converges uniformly to* πQ *for all* $\pi \in [0, 1]$, $Q = E(x_t x_t')$ *and all eigenvalues of* Q *are uniformly bounded away from zero.* $[\pi T]$ *denotes the integer part of the product* πT.

6. $E(e_t|x_t) = 0$; $E(e_t^2|x_t) = \sigma_t^2$.

Assumption 1.1 states that the true data generating process for y_t takes a general parameter variation form and structural break occurs in all parameters. In each period, the change of the true parameter value is of small magnitude so that the asymptotic distributions are asymptotically continuous. Additionally, the parameter variation is proportional to the unconditional standard deviation of the error term, so the impact of parameter instability will not be dominated by that of the volatility. This type of data

generating process is quite general, as it includes several commonly used models, for example, the single break model with the absolute change of parameter values positive in one period while zero in others.

In practice, if there is no clear guidance or information on which subset of parameters are unstable *a priori*, it is natural to assume that all parameters are subject to break. This full-break in the conditional mean assumption is less restrictive, so empirically it is widely adopted in applications of detecting and dating breaks, see Rapach and Wohar (2006) and Paye and Timmermann (2006).

Notice that our predictive model outlined earlier only allows for one possible break in the conditional mean, so it is highly possible that the forecasting model, either the pre-test model or the averaged model, is misspecified. We make this assumption allowing for the gap between the true data generating process and the forecasting model primarily for two reasons. First, in practice the true data generating process is almost always unknown to researchers, as it may be a complicated process possibly involving past values of infinite order. In addition, the true dynamics and parameter stability are very difficult to capture by models based on limited information. Second, for the prediction problem, the goal is not to come up with a highly complex model to fit the training data as closely as possible measured in terms of the learning error rate. Instead, forecasters pay more attention to the test error rate. By reducing the complexity of the predictive model, we hope our model to be more adaptive to environment change in the future.

Assumptions 1.2 – 1.5 ensure that we can apply all relevant mixing laws of large numbers, functional central limit theorem or Donsker's invariance principle when proving our results. See Davidson (1994) for more details on advanced asymptotic theory. Assumption 1.6 states that the error term is conditionally heteroscedastic.

Because the cross-validation criterion estimates the expected test error rate, or the expected mean squared forecast error rate, the optimal weights should be those minimizing the cross-validation criterion, which can be interpreted as weights minimizing the

expected test error rate. To obtain model weights, first, we need to show what the cross-validation criterion looks like under our assumptions. We start with a proposition on the cross-validation criterion form when the error term is conditionally homoscedastic. The proofs of all theoretical results are provided in the appendix.

Proposition 3.3.1. *If Assumption 1 holds but* $E(e_t^2|x_t) = \sigma^2$*, the leave-one-out cross–validation criterion is asymptotically equivalent to Mallows' criterion, that is,* $E(CV(T)) \xrightarrow{p} E(Cp(T))$.

This proposition states that since conditional homoscedasticity is a special case of conditional heteroscedasticity, all the asymptotic optimality results from Mallows' criterion carry over to the cross-validation criterion when the errors are conditionally homoscedastic.

We know that the information criterion usually consists of two parts: one measures the in-sample fit while the other penalizes overfitting. Specifically, by proposition 3.3.1, since CV and Cp are asymptotically equivalent, for the CV criterion, we have

$$E(CV(T)) = E(\hat{\sigma}^2) + 2E(e'Pe) \tag{3.6}$$

In equation 3.6, $\hat{\sigma}^2$ measures the in-sample fit, $2E(e'Pe)$ is the population penalty term where e is the vector of the errors and P is the projection matrix. The penalty term, $2E(e'Pe)$, is crucial in determining the optimal weights for the averaged model 3.3, as the population optimal weight w can be obtained by minimizing $E(CV(T))$. If we can find a sample analogue of $E(CV(T))$, the optimal weights can be obtained by minimizing the sample criterion. Because the population penalty term $2E(e'Pe)$ depends on the true data generating process, it cannot be consistently estimated in practice. To obtain the feasible sample CV criterion and the associated sample optimal weight $\hat{w}$, following Hansen's approach, the value of $2E(e'Pe)$ can be approximated by averaging two extreme

cases,[12] so that is how the $\bar{p}$ value proposed by Hansen, where $\bar{p} = \frac{1}{2}(E(SupW) + k)$,[13] enters the break model weight in the following corollary.

Corollary 3.3.1. *With conditionally homoscedastic errors, the feasible sample optimal CV weight for the break model is:*

$$\hat{w} = \frac{(T-2k)(\sum_{t=1}^{T}\tilde{e}_t^2 - \sum_{t=1}^{T}\hat{e}_t^2) - \bar{p}\sum_{t=1}^{T}\hat{e}_t^2}{(T-2k)(\sum_{t=1}^{T}\tilde{e}_t^2 - \sum_{t=1}^{T}\hat{e}_t^2)} \tag{3.7}$$

if $(T-2k)(\sum_{t=1}^{T}\tilde{e}_t^2 - \sum_{t=1}^{T}\hat{e}_t^2)(\sum_{t=1}^{T}\hat{e}_t^2)^{-1} \geq \bar{p}$ *while* $\hat{w} = 0$ *otherwise.* T *is the sample size,* k *is the number of regressors,* $\hat{e}_t s$ *are the ordinary least squares residuals from the break model,* $\tilde{e}_t s$ *are residuals from the stable model,* $\bar{p}$ *is the penalty coefficient whose value depends on the asymptotic distribution of the SupW test statistic.*

The sample optimal weight $\hat{w}$ is obtained by minimizing the sample CV criterion for the weighted model.

It is widely known in the model selection literature that the CV criterion is superior to Mallows' and other information criteria because of its robustness to heteroscedasticity Andrews (1991), our next proposition establishes the asymptotic distribution of the CV penalty term in the presence of conditional heteroscedasticity.

Proposition 3.3.2. *If Assumption 1 holds, then the penalty term in the cross-validation criterion converges in distribution to a weighted sum of independent* χ^2 *distribution with degree of freedom one, plus a term whose distribution is a function of a Brownian bridge,*

$$e'P(\hat{m})e \xrightarrow{d} \sum_{j=1}^{k}\lambda_j\chi^2(1) + J_0(\xi_\delta) \tag{3.8}$$

where $\lambda_j s$ *are the eigenvalues of the matrix* $Q^{-1}\Sigma$, Σ *is the long-run variance of* $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}X_t e_t$, $Q = E(x_t x_t')$ *and* $J_0(\xi_\delta)$ *is the asymptotic distribution of the Sup-Wald type statistic under the true data generating process.*

[12] One is that the break size is extremely large while in the other case the break size is 0.

[13] $E(SupW)$ is the expectation of the SupW statistic in Andrews (1993). Hansen (2009) provides a table of the sample $\bar{p}$ value for a range of the number of regressors based on simulation results.

Comparing this result with Hansen's, we can see that the distribution under conditional homoscedasticity is just a special case of what is shown in proposition 3.3.2. That is, the weights for the χ^2 random variables are identical and they take the value of one, which results in a χ^2 distribution with degrees of freedom equal to the total number of regressors. In our results, λ_js can take different values which capture the impact brought to the weight by allowing for conditional heteroscedasticity. Intuitively, the first term on the right-hand-side of equation 3.8 reflexes the impact of conditional heteroscedasticity while the second term deals with the structural break.

The expectation of $\sum_{j=1}^{k} \lambda_j \chi^2(1)$ is simply $\sum_{j=1}^{k} \lambda_j$ which is the trace of the matrix $Q^{-1}\Sigma$, where Σ is the long-run variance of $\frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_t e_t$ and $Q = E(x_t x_t')$. Empirically, Σ can be estimated by HAC estimators and Q can be consistently estimated by its sample analogue $\frac{1}{T} \sum_{t=1}^{T} x_t x_t'$.

Again, the penalty term of the CV criterion depends on the true data generating process as reflected in the $J_0(\xi_\delta)$ term, it cannot be consistently estimated in practice. To obtain the feasible sample CV criterion, following earlier approach we can approximate $J_0(\xi_\delta)$ by averaging two extreme cases utilizing Hansen's $\bar{p}$ value. The feasible sample optimal weight $\hat{w}$ for the break model can be obtained by minimizing the sample CV criterion associated with the averaged model.

Corollary 3.3.2. *The feasible optimal weight minimizing the sample cross-validation criterion for the break model in the presence of conditional heteroscedasticity takes the form:*

$$\hat{w} = 1 - \frac{\operatorname{tr}\left(\hat{Q}^{-1}\hat{\Sigma}\right) + 2\bar{p} - k}{2\left(\sum_{t=1}^{T} \tilde{e}_t^2 - \sum_{t=1}^{T} \hat{e}_t^2\right)} \tag{3.9}$$

if $(\sum_{t=1}^{T} \tilde{e}_t^2 - \sum_{t=1}^{T} \hat{e}_t^2) \geq \bar{p}^*$ *while* $\hat{w} = 0$ *otherwise.* $\hat{e}_t$*s are the OLS residuals from the break model and* $\tilde{e}_t$*s are residuals from the stable model,* $\operatorname{tr}(\hat{Q}^{-1}\hat{\Sigma})$ *is the trace of the matrix* $\hat{Q}^{-1}\hat{\Sigma}$*,* $\bar{p}^* = \frac{1}{2}(\operatorname{tr}(\hat{Q}^{-1}\hat{\Sigma}) + 2\bar{p} - k)$.

In the next section, through several designs we are going to assess the sample performance of CV weights comparing with Cp weights and other related methods in controlled simulations.

3.4 Simulation Results

Here we are going to evaluate the forecast performance of CV model averaging through controlled numerical simulation. Specifically, we are going to consider three different designs of the true data generating process: (i) an AR(2) process plus five exogenous predictors with ARCH(1) errors,

$$y_t = \mu + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \sum_{i=1}^{5} \theta_i x_i + e_t \tag{3.10a}$$

$$e_t = v_t \sqrt{h_t} \tag{3.10b}$$

$$h_t = \alpha_0 + \alpha_1 e_{t-1}^2 \tag{3.10c}$$

(ii) an AR(2) process plus two exogenous predictors with heteroscedastic errors drawing from the Normal distribution $N(0, y_{t-1}^2)$

$$y_t = \mu + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \sum_{i=1}^{2} \theta_i x_i + e_t \tag{3.11}$$

(iii) an AR(2) process with a single break in the variance of the error term. Although our theory does not explicitly address the volatility break situation, we consider this design to investigate and compare the predictive performance of CV model averaging with other related methods in the Great Moderation type environment. In this design, the break date of the error term variance is not identical to that of the conditional mean.[14] We allow for this break date difference hoping to better approximate the environment forecasters face in practice.

[14] In this simulation, we set the break fraction of the error term variance at 0.5 relative to the training sample, while the break fraction for the conditional mean is set at 0.3 relative to the training sample.

Mathematically, the data generating process for design (iii) is the following:

$$y_t = \mu + \rho_1 y_{t-1} + \rho_2 y_{t-2} + e_t \tag{3.12}$$

where

$$e_t \sim \begin{cases} N(0, \sigma^2) & t \in [1, \tau_v] \\ N(0, \frac{1}{4}\sigma^2) & t \in [\tau_v + 1, R] \end{cases}$$

In all three designs there is a one-time structural break in all coefficients of the conditional mean occurring at the $30\%th$ observation of the training sample R, that is, $\pi = 0.3$. We let the structural break take the multiplicative form, that is, if the pre-break coefficient is β, then the post-break value becomes $\delta\beta$, where δ is a tuning parameter controlling for the break size. For the ARCH process, v_ts are drawn independently and identically from the standard normal distribution. Other predictors are drawn i.i.d. as the following: $x_1 \sim \mathrm{N}(0, 4)$, $x_2 \sim \mathrm{U}[-2, 2]$, $x_3 \sim \mathrm{N}(0, 16)$, $x_4 \sim \mathrm{t}(5)$ and $x_5 \sim \mathrm{Binomial}(1, 0.02)$. The parameter values for all data generating processes listed above are: $\mu = 2, \rho_1 = 0.4, \rho_2 = 0.2$, $\theta_1 = 0.8, \theta_2 = -0.4, \theta_3 = 2, \theta_4 = -3.5, \theta_5 = 10$, $\alpha_0 = 1, \alpha_1 = 0.4$. These values are chosen to satisfy the stationarity and ARCH error regularity restrictions. It is worth mentioning that, in our simulations, the post-break parameter values of interest become smaller than their pre-break counterparts ($\delta < 1$). This choice of break direction provides us with more freedom in controlling the break size, for example, if the true data generating process is an intercept-free AR(1) model with pre-break parameter value 0.9, δ cannot take values greater than 1.1 if stationarity is to be maintained.[15]

After presenting the data generating processes, next, to capture the model selection uncertainty researchers face in choosing the best local approximating models, the forecasting model in each design differs from the true data generating process:[16] in

[15] Bai and Perron (1998) assume that the break size is large enough in order to be identified and estimated. Though we have not found any leading metric measuring the break size, break size of 1.1 mentioned in the example is not large enough for identification purpose, especially when the data is highly volatile as those generated in our simulations.

[16] The difference of the AR order between the DGP and the forecasting model captures the fact that in practice, it is hard to fully capture the dynamics by selecting the 'true' order. By the principle of parsimony, researchers tend to select a model of small order.

case (i) the model to forecast is based on some exogenous predictors in the DGP, $y_t = \mu + \sum_{i=1}^{4} \theta_i x_i + e_t$; in case (ii), again the model to forecast does not involve the AR component, $y_t = \mu + \sum_{i=1}^{2} \theta_i x_i + e_t$; in case (iii), the model to forecast is AR(1) with intercept, $y_t = \mu + \rho_1 y_{t-1} + e_t$.

In each design, for a given weighting method, we evaluate its out-of-sample (OOS) performance by comparing the average root mean squared forecast error divided by that of the equal weights method. Recursive window is used to generate OOS forecasts as it mimics the practice that forecasters update their forecast when new data become available, so weights are also constructed recursively. Specifically, out-of-sample forecasts are constructed by the following steps: First, we split the time series sample into two parts, the prediction or training sample of size R and the evaluation or test sample of size P. Under the recursive window, at each point in time, the estimated parameters are updated by adding one more observation starting with sample size R. For example, $\beta_t = (\sum_{s=1}^{t-1} x_s x_s')^{-1} \sum_{s=1}^{t-1} x_s y_{s+1}, \beta_{t+1} = (\sum_{s=1}^{t} x_s x_s')^{-1} \sum_{s=1}^{t} x_s y_{s+1}$. By this procedure, we estimate parameters recursively, and then generate a sequence of forecasts of size P based on these estimated parameters. We can compare this sequence of forecasts with those reserved data in the evaluation sample, and assess the quality of our forecasts according to some loss function, for example, MSFE. See Calhoun (2013), Calhoun (2014), McCracken (2000), McCracken (2007), Rossi (2013), Clark and McCracken (2001), Clark and McCracken (2005), Clark and McCracken (2013), Clark and West (2007) and West (2006) for more details on out-of-sample forecasting.

The total sample size, $T = R+P$, is 200. To investigate if the choice of the evaluation sample size has an impact on forecasting results, in our pseudo one-step ahead out-of-sample forecasting simulations, we reserve the first 170 and 150 ($R = 170$ and $R = 150$) observations as the training sample and the rest as the prediction sample ($P = 30$ and $P = 50$) in two separate experiments for each design. For the break model, we use the post-break window method to forecast out-of-sample as it is simple to implement and

does not involve the estimation of additional parameters. Other techniques, such as the optimal window method proposed by Pesaran and Timmermann (2007) or the robust weight method proposed by Pesaran et al. (2011) could also be considered.[17]

In each case, to evaluate and compare performance, we generate forecasts using six methods:[18] (i) Mallows' model averaging (**Cp**); (ii) CV model averaging (**CV**); (iii) Bayesian model averaging[19] (**SIC**); (iv) stable model (**Stable**); (v) break model (**Break**); and (vi) equal weights[20] (**Equal**). We assess their predictive performance by root mean squared forecast error (**RMSFE**). For ease of comparison, we pick the equal weight method as the benchmark[21] and compute the relative performance (**Ratio**) for each method, for example, $\text{RMSFE}^{\text{CV}}/\text{RMSFE}^{\text{Equal}}$. If the ratio is less than one, it indicates that the method under consideration outperforms the benchmark. The smaller the ratio is, the better the forecasting performance is for a given sample split.

3.4.1 Design I

Simulation results for the ARCH error design are reported in table C.1.[22] We can see from the table that CV outperforms Cp across all considered break sizes and test sample sizes. Both of CV and Cp's relative RMSFE decrease monotonically as the

[17] Currently, researchers are still working on developing theory and methods related to forecasting with breaks, and we are not aware of any dominant method that works well in most empirical applications. The simulation conducted by Pesaran and Timmermann suggests that there is little gain from complicated methods. The simple rule, to forecast using the data after the detected break, seems to work as well as anything else.

[18] Methods such as Bates-Granger combination, Granger-Ramanathan combination and common factor combination are not considered in our simulation. In a related paper, Clark and McCracken (2011) conclude that "*...it is clear that the simplest forms of model averaging—such as those that use equal weights across all models—consistently perform among the best methods...forecasts based on OLS-type combination and factor-based combination rank among the worst*". So we only compare our method with either closely related or empirically proven effective methods.

[19] We call this method "Bayesian" not in a strict sense: the Bayesian weight for each model is calculated based on the value of the Schwarz-Bayesian information criterion, i.e., the weight for the break model is $w_b = \exp(SIC^b)/(\exp(SIC^b) + \exp(SIC^s))$

[20] Each model receives weight of 0.5.

[21] The reason to pick equal weights as the benchmark is because of the aforementioned forecast combination puzzle: equally weighted forecasts tend to outperform other complex methods in empirical works. Here we would like to examine whether it dominates our method when facing structural breaks.

[22] Our results also hold in the GARCH error case.

break size increases, but CV's relative RMSFE decreases slightly faster. On the other hand, Bayesian weighting does slightly worse than the equal weights method, and its performance deteriorates when the break size becomes large as it fails to capture the fact that the evidence supporting the break is becoming stronger.

For the non-averaged models, it is not surprising that the break model does well because structural break indeed happens in the DGP, and its relative RMSFE decreases as the break size becomes large.

Overall, our results imply that when there is ARCH type conditional heteroscedasticity in the data and when the break impact is not strictly dominated by that of the volatility, the cross-validation weighting method outperforms Mallows' model averaging. Additionally, CV outperforms equal weights so the forecast combination puzzle does not apply in this design. Bayesian model averaging is approximately equivalent to equal weighting, but it is less sensitive to the change of break size. Compared with CV, Bayesian criterion weighting does not put more weight on the proper model even when the break size becomes large.

3.4.2 Design II

Simulation results for the second design are reported in table C.2. Here we can see that CV outperforms Cp across all break sizes and test sample sizes considered. Both of their relative RMSFE decrease monotonically as the break size increases, but now the RMSFE of CV decreases faster. Bayesian weighting does almost the same as equal weighting, but its performance deteriorates when the break size becomes large as we have seen in the previous design. The choice of the test sample size does not seem to have any significant impact on any weighting methods or non-averaged models.

Overall, our results indicate that when there is "wild" type heteroscedasticity in the data as modeled in the DGP and when the break impact is not strictly dominated by that of the volatility, the cross-validation weighting outperforms Mallows' model averag-

ing, especially when the break size is large. Additionally, CV outperforms equal weights so the forecast combination puzzle does not apply in this design. Bayesian model averaging is approximately equivalent to equal weights. Again, compared with CV, Bayesian weighting method does not put more weight on the proper model when the break size increases.

3.4.3 Design III

Simulation results for this Great Moderation type design are reported in table C.3. The general pattern shown in the previous two designs remains in this case. CV outperforms Cp across all considered break sizes and prediction sample sizes. Both of their relative RMSFE decrease monotonically as the break size increases, but the relative RMSFE of CV decreases faster. Bayesian weighting does almost the same as equal weighting, but its performance is less sensitive to the break size in this case.

3.4.4 Summary

We have compared the statistical performance of CV weights with other competing methods, such as Mallows' Cp weights, equal weights and Bayesian information criterion weights, in three simulation designs. All the experiments show that CV weights outperform the rest in the presence of structural breaks and heteroscedasticity. As the break size becomes large, the average root mean squared forecast error associated with either CV or Cp weights decreases monotonically, but CV's error tends to decrease faster in some cases. Additionally, the forecast combination puzzle does not apply in any of these experiments for our CV weights.

3.5 Empirical Application

In this section we are going to apply our CV model averaging method and other related methods to forecasting the quarterly GDP growth rate for the U.S. and Taiwan.

We plot these two series separately in figure C.1. We consider the Taiwanese data[23] because it has some interesting features compared with the U.S. data, for example, the Taiwanese data seems to have a break in the mean around the early 1990s,[24] and it becomes more volatile towards the end of the sample. The U.S. data is obtained from the Bureau of Economic Analysis.[25] The data for Taiwan is from National Statistics.[26]

For the U.S. series, we can see that the growth rate becomes less volatile toward the end of the sample. This pattern is the so called Great Moderation phenomenon, see Stock (2004) and Stock and Watson (2003). On the prediction of U.S. GDP growth, Stock and Watson argue that the forecasting relationship is time-varying and combination forecasts reliably improve upon the AR benchmark. They claim:

> From the perspective of forecasting methods, this evidence of sporadic predictive content poses the challenge of developing methods that provide reliable forecasts in the face of time-varying relations...the finding that averaging individually unreliable forecasts produces a reliable combination forecast is not readily explained by the standard theory of forecast combination, which relies on information pooling in a stationary environment...fully articulated statistical or economic models consistent with this observation could help to produce combination forecasts with even lower MSFEs.

Motivated by these remarks, we will demonstrate that our theory based CV model averaging method outperforms Mallows' weight, Bayesian weight, and most importantly, the equal weighting method in terms of smaller root mean squared forecast error.

[23]The data length for Taiwan is shorter than that of the U.S. because Taiwan officially starts its post-war modernization in the early 1950s.

[24]This may be explained by the fact that Taiwan started drastic political reform around this period, moving from an authoritarian central government to a modern democracy.

[25]http://www.bea.gov/

[26]National Statistics is the Taiwanese government agency commissioned with producing statistics to help better understand Taiwan, its population, resources, economy, society, and culture. See http://eng.stat.gov.tw/

3.5.1 Forecast U.S. GDP Growth

Here we apply our method to forecasting the U.S. quarterly GDP growth rate[27] out-of-sample and compare its performance with others. We have quarterly data running from 1960:Q1 to 2012:Q1, 209 observations in total. The variable we are interested in predicting is the U.S. quarterly GDP growth rate. Predictors considered are the quarterly change of U.S. 3-month treasury rate (ΔSR), the quarterly change of U.S. 10-year treasury rate (ΔLR) and the quarterly change of default premium (ΔDP).[28]

Because we do not know the "true" model, or the "true" predictors to include, five candidate models are considered. For each candidate, we are going to combine the break version and stable version of the model using CV weights and other competing weights, then forecast out-of-sample and calculate the root mean squared forecast errors. From small to large the candidate models considered are:

$$\Delta\text{GDP}_t = \beta_0 + \beta_1 \Delta\text{GDP}_{t-1} + \epsilon_t \tag{3.13a}$$

$$\Delta\text{GDP}_t = \beta_0 + \beta_1 \Delta\text{GDP}_{t-1} + \beta_2 \Delta\text{GDP}_{t-2} + \epsilon_t \tag{3.13b}$$

$$\Delta\text{GDP}_t = \beta_0 + \beta_1 \Delta\text{GDP}_{t-1} + \beta_2 \Delta\text{SR}_{t-1} + \epsilon_t \tag{3.13c}$$

$$\Delta\text{GDP}_t = \beta_0 + \beta_1 \Delta\text{GDP}_{t-1} + \beta_2 \Delta\text{SR}_{t-1} + \beta_3 \Delta\text{LR}_{t-1} + \epsilon_t \tag{3.13d}$$

$$\Delta\text{GDP}_t = \beta_0 + \beta_1 \Delta\text{GDP}_{t-1} + \beta_2 \Delta\text{SR}_{t-1} + \beta_3 \Delta\text{LR}_{t-1} + \beta_4 \Delta\text{DP}_{t-1} + \epsilon_t \tag{3.13e}$$

Consistent with what is done in the simulation section, for each model we apply the recursive window to forecast out-of-sample. To investigate the impact of the test sample size, for each model, we vary the evaluation sample size from 20 to 50 with increments of 5, then calculate the RMSFE for each weighting method for a given test sample size.

Forecast results from all models are reported in table C.4. For each model, the column shows the OOS performance for a given weighting method. The rows report results for

[27] The data used for this application are from Bruce Hansen's website:http://www.ssc.wisc.edu/~bhansen/cbc/.

[28] The default premium is calculated by the difference between the AAA bond rate and BAA bond rate.

different evaluation sample sizes. For the entries in the table, following our Monte Carlo simulation, we select the equal weighting method as the benchmark and normalize all OOS forecasting performance around one. If the value of the relative RMSFE for a given method is below one, it implies that the method under consideration outperforms the benchmark.

We can see that in all five models approximating the DGP, CV outperforms SIC, Cp and equal weights under recursive window across all evaluation sample sizes. Additionally, CV is the only method exceeding the benchmark regardless of the test sample size and the base predictive model choice. The forecast gains of CV relative to the benchmark range from about 1% to 6% across evaluation sample sizes and models. As for Mallows' weights, in four out of five models, their performance gets close to the benchmark as the test sample size increases, so this may suggest that Mallows' weights are more sensitive to the test sample size compared with CV. Last, for the SIC weights, their performance is almost identical to the benchmark, and is quite stable across all models and test sample sizes, though in some cases SIC weights marginally outperform the benchmark.

3.5.2 Forecast Taiwan GDP Growth

For the Taiwanese series, it demonstrates two interesting features in the figure. First, it looks like that the Taiwanese average growth rate has dropped toward the end of the sample. This may be explained by the economic growth theory that during the early period of modernization or industrialization, a country tends to experience high economic growth rate. But as time goes, the growth rate approaches to the low equilibrium rate. Second, it seems like that the series becomes more volatile toward the end of the sample compared with the U.S. data. This phenomenon contrasts with many other developed counties which exhibit the similar Great Moderation pattern shown in the U.S. data, for example, Canada and Germany.

We have quarterly data running from 1962:Q1 to 2013:Q4, 208 observations in total.

The variable we are interested in forecasting is the Taiwanese quarterly GDP growth rate. Since we do not have any exogenous predictors available, we only consider two AR predictive models of different order, namely, the AR(1) model and the AR(2) model, and combine the break version and stable version of each model using various weighting methods. Out-of-sample forecast results from these two models are reported in table C.5. Again, we keep the general setting outlined in the previous application: For each model, we generate a sequence of scaled RMSFE by varying the evaluation sample size P, from 20 to 50, with increments of 5; Equal weighting is the benchmark; All entries in the table are RMSFE divided by that of the benchmark.

For the AR(1) model, we can see from table C.5 that all weighting methods perform roughly the same as the benchmark, though CV leads the rest marginally. For the AR(2) model, both CV and Cp outperform the benchmark, but CV leads Cp across all the test sample sizes considered. Overall, both applications demonstrate the superior performance of CV weights compared with related methods.

3.6 Conclusion

We are interested in answering a basic question of how to forecast a time series variable of interest when there is uncertainty about parameter instability. Specifically, which model should be selected for prediction: the break model or the stable one? If the uncertainty is strong and we decide to combine these two predictive models, what is the optimal rule in terms of some information criterion about assigning weights? Built upon Hansen's Mallows' model averaging method, we propose using the cross-validation criterion to combine predictive models.

In many empirical applications related to macroeconomic or financial time series, researchers usually cannot avoid explicitly dealing with heteroscedasticity for analysis and prediction, so assuming conditional homoscedasticity in the model averaging theory

may seem restrictive. To adapt Hansen's weights to the out-of-sample forecast setting, we need to relax the conditional homoscedasticity assumption and adjust weights accordingly. In the literature of model selection, the cross-validation criterion is shown to be robust to heteroscedasticity unlike other information criteria, such as AIC, BIC and Mallows', so it is natural to replace Cp with CV and then derive the new optimal weights.

Researchers have found that in many applications, equally weighted forecasts outperform other complex combination methods. This so called forecast combination puzzle has cast doubt on the use of complicated model averaging methods, so comparing a new method with the equal weights method becomes necessary for validation. Both CV and Cp weights are easy to compute and do not rely on weight estimation as in the Granger-Ramanathan forecast combination. This feature should be appealing to practitioners and professional forecasters because simplicity may help reduce the excess noise introduced by applying complex weighting methods. This may help explain why our cross-validation weights exceed equal weighting as shown in simulations and in empirical examples on forecasting U.S. and Taiwan quarterly GDP growth rates out-of-sample.

APPENDIX A. FORECASTING EQUITY PREMIUM WITH STRUCTURAL BREAKS

TABLES AND FIGURES

Table A.1 Estimation Results for Stable Predictive Models

	R^2	Intercept	β
Historical Mean		0.0027	
		(1.743)	
Dividend-yield	0.0027	0.0206	0.0053
		(1.773)	(1.556)
Dividend-payout	0.0005	0.0050	0.0033
		(1.290)	(0.642)
Dividend-price	0.0025	-0.0146	-0.0051
		(-1.257)	(-1.502)
Earnings-price	0.0043	-0.0157	-0.0068
		(-1.650)	(-1.959)
Stock Market Variance	0.080	0.0096	-3.1314
		(5.717)	(-8.830)
Cross Sectional Premium	0.0066	0.0021	1.5988
		(1.253)	(2.289)
Book-to-market	0.0043	0.0098	-0.0123
		(2.486)	(-1.958)
Net Equity Expansion	0.0020	0.0045	-0.1179
		(2.187)	(-1.330)
3-month Treasury Bill	0.0049	0.0068	-0.1038
		(2.736)	(-2.107)
Long term Yield	0.0025	0.0072	-0.0800
		(2.120)	(-1.487)
Term Spread	0.0035	-0.0008	0.2097
		(-0.330)	(1.776)
Default Premium	0.0002	0.0039	0.1190
		(1.130)	(0.390)
Inflation	0.0029	0.0043	-0.5339
		(2.338)	(-1.599)

Note: The stable model is $y_{t+1} = \bar{y} + \beta x_t + u_{t+1}$, where $t = 1, ..., T$. y_{t+1} is the market excess returns, $\bar{y}$ is the intercept, x_t is the exogenous predictor available at time t to forecast the next period returns y_{t+1} and u_{t+1} is a disturbance term. The historical mean model is $y_{t+1} = \bar{y} + u_{t+1}$. Each predictive model is labeled with its predictor except for the historical mean model. For each model, we report its in-sample R^2 statistic, intercept estimate and predictor coefficient estimate β. Parentheses report the t statistic for each parameter estimate above. Monthly data rums from 1937:05 to 2011:12.

Table A.2 Estimation Results for the Stock Market Variance Model with Three Breaks

	R^2	Intercept	β
Segment 1: 1937:05 - 1956:02	0.073	0.015	-3.692
		(3.602)	(-4.189)
Segment 2: 1956:03 - 1974:08	0.215	0.013	-13.032
		(4.308)	(-7.766)
Segment 3: 1974:09 - 1985:10	0.121	-0.017	11.930
		(-2.873)	(4.263)
Segment 4: 1985:11 - 2011:12	0.172	0.013	-3.193
		(5.052)	(-8.033)

Note: For the stock market variance model with three breaks, we report its in-sample R^2 statistic, intercept estimate and predictor coefficient estimate β for each regime. Parentheses report the t statistic for each parameter estimate above.

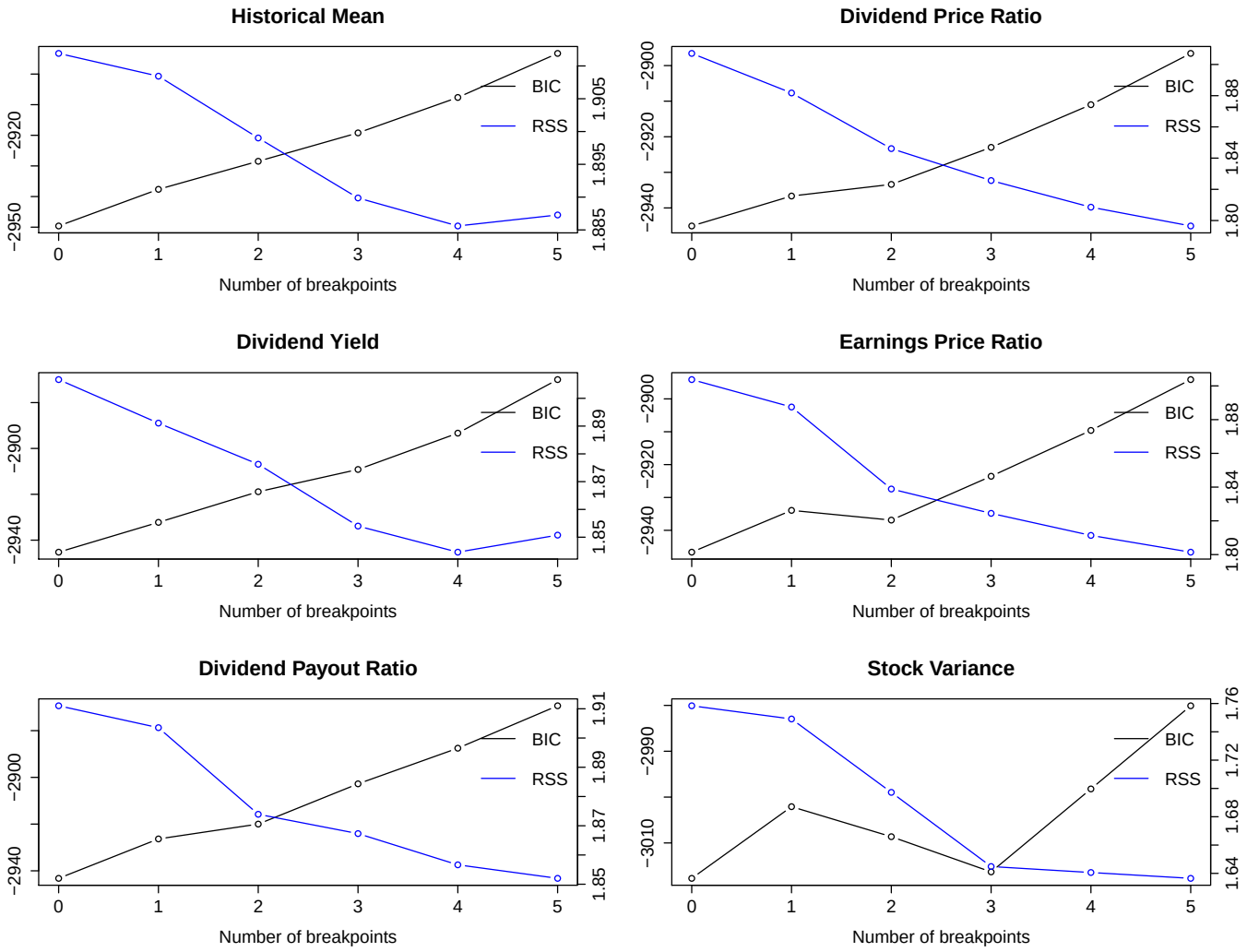


Figure A.1 Break Estimation Results for Historical Mean, Dividend-price Ratio, Dividend Yield, Earnings-price Ratio, Dividend-payout Ratio and Stock Market Variance

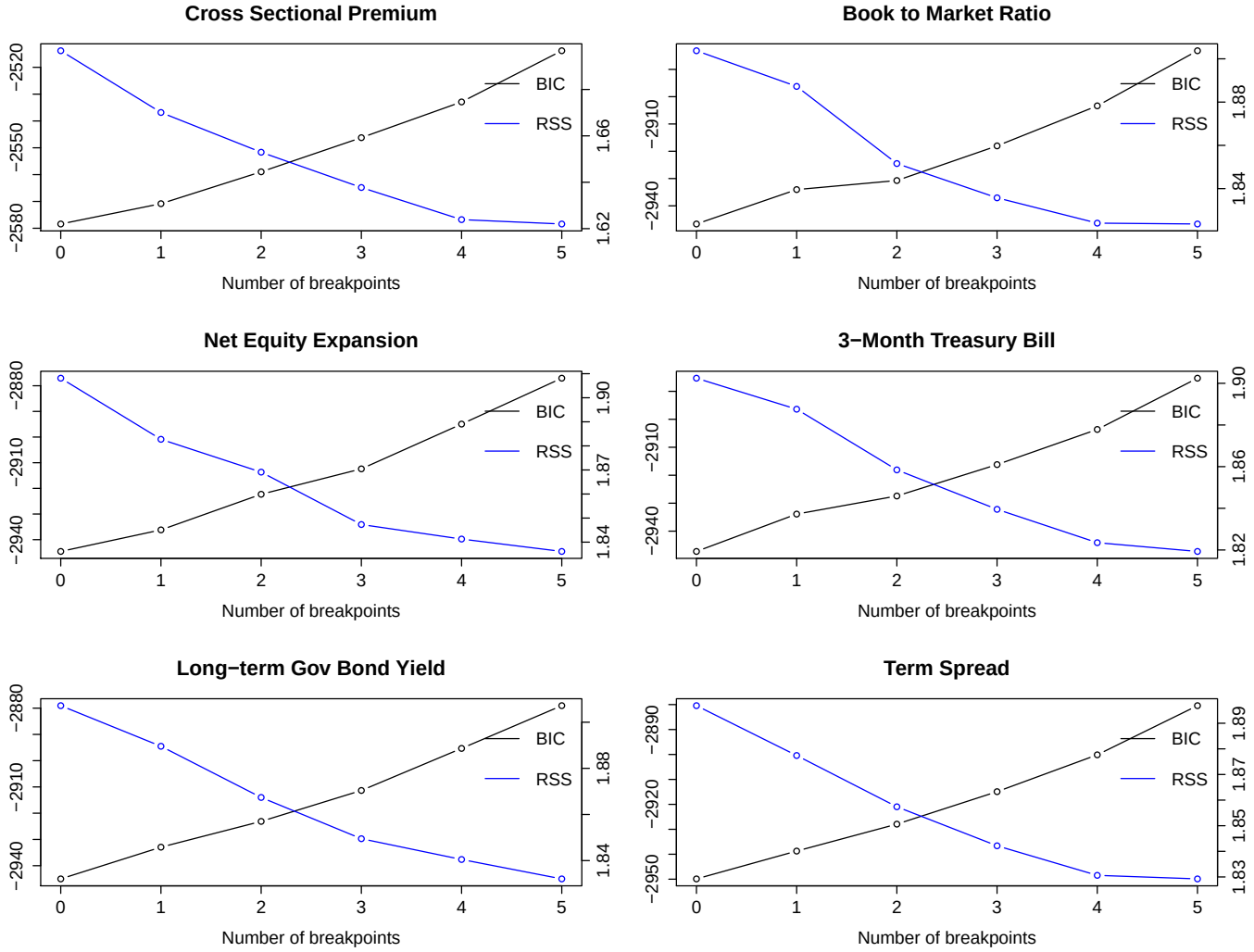


Figure A.2 Break Estimation Results for Cross Sectional Premium, Book-to-market Ratio, Net Equity Expansion, Treasury Bill, Long Term Yield and Term Spread

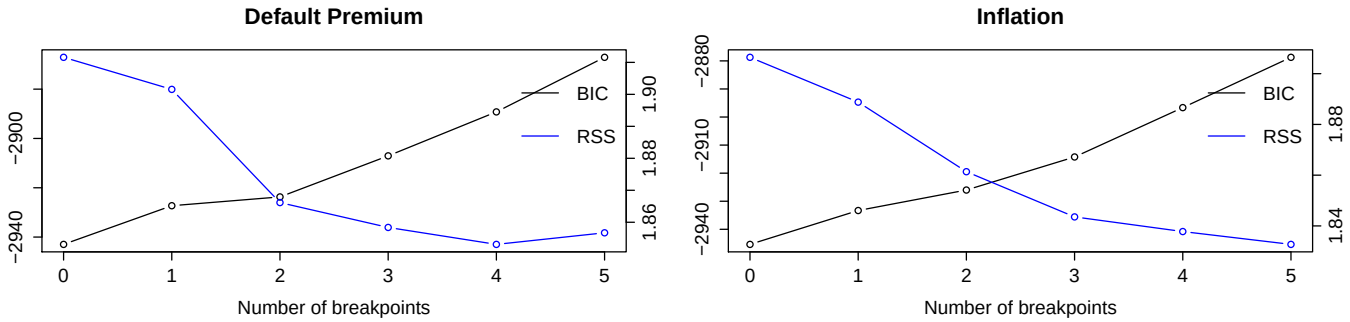


Figure A.3 Break Estimation Results for Default Premium and Inflation

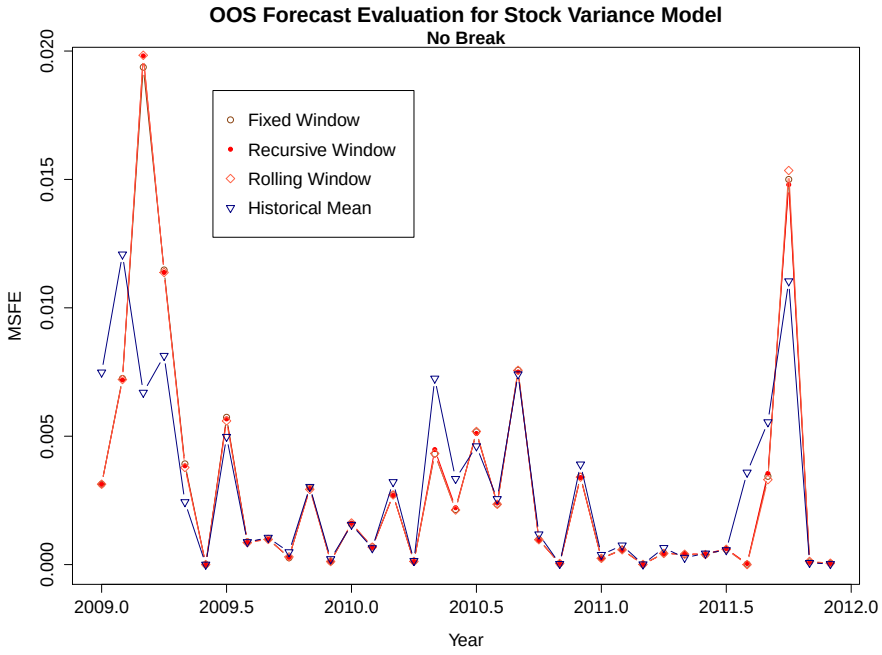


Figure A.4 Out-of-Sample Forecast Evaluation for the Stable Model

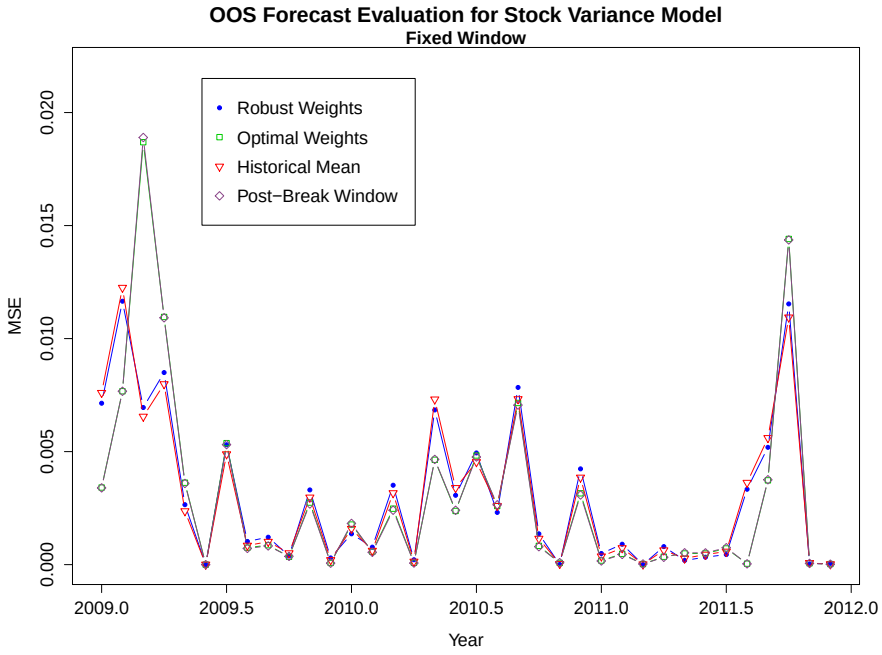


Figure A.5 Out-of-Sample Forecast Evaluation for the Break Model under Fixed Window

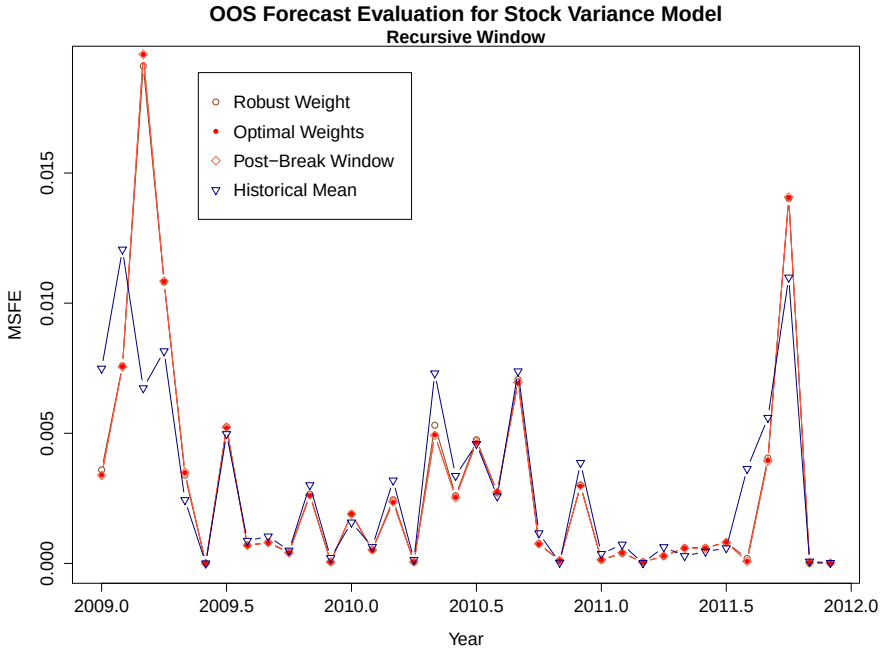


Figure A.6 Out-of-Sample Forecast Evaluation for the Break Model under Recursive Window

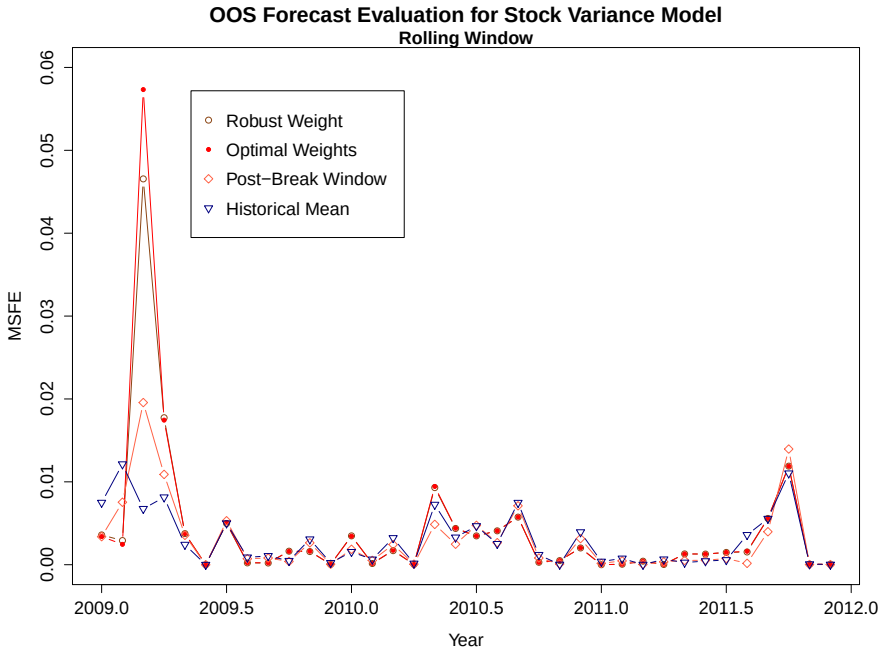


Figure A.7 Out-of-Sample Forecast Evaluation for the Break Model under Rolling Window

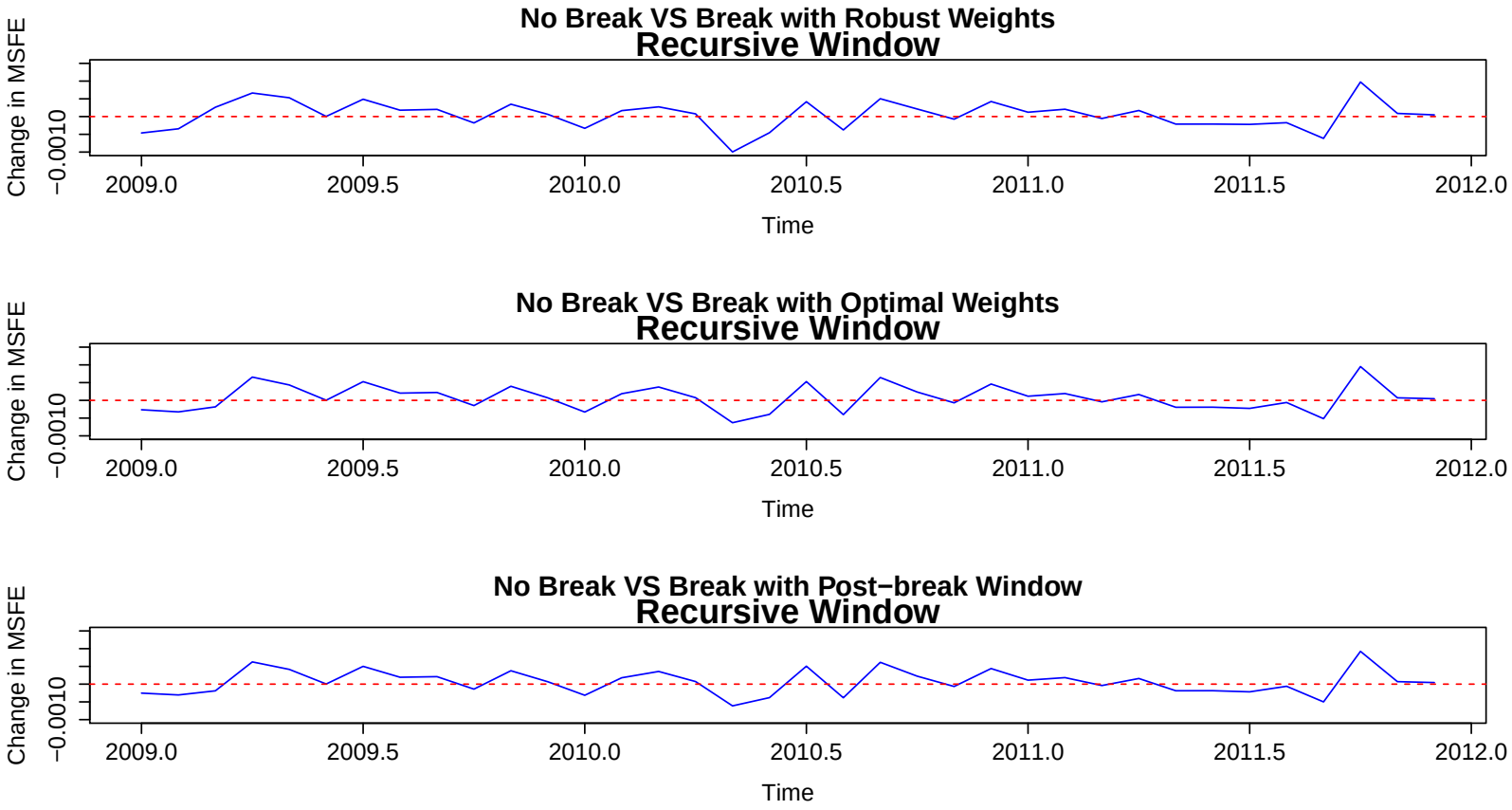


Figure A.8 Recursive window out-of-sample forecast comparison between the break Model and stable model

Figure A.9 Rolling window out-of-sample forecast comparison between the break Model and stable model

Figure A.10 Fixed window out-of-sample forecast comparison between the break Model and stable model

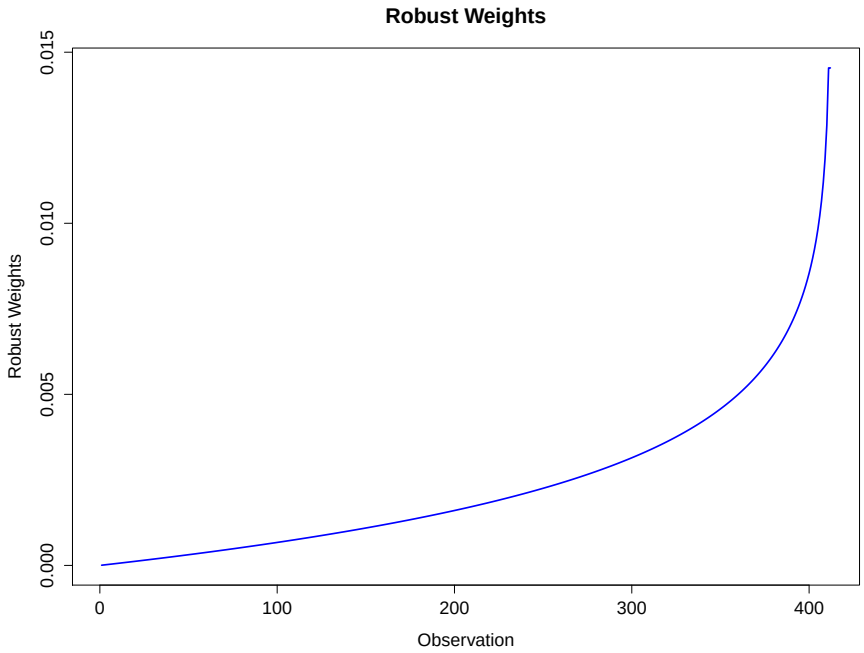


Figure A.11 Robust Weights Example

APPENDIX B. COMBINING MULTIPLE PREDICTIVE MODELS WITH POSSIBLE STRUCTURAL BREAKS

TABLES AND FIGURES

Table B.1 U.S. Market Equity Premium Out-of-Sample R^2_{OS} Statistics for Combining Methods

	Cp	DMSFE	Equal	SIC	Break	Stable
Monthly Data	10.484	10.441	10.406	10.406	10.635	10.143
Quarterly Data	6.171	6.008	5.835	5.826	6.214	5.071
Yearly Data	3.608	3.459	3.206	3.157	2.199	3.897

Note: R^2_{OS} is the Campbell and Thompson (2008) out-of-sample R^2 statistic, which measures the percent reduction in mean squared forecast error (MSFE) for the combination methods given in the first row of the table relative to the historical average benchmark forecast. Cp: Mallows' weights. DMSFE: discounted mean squared forecast error weights with discount factor $\theta = 1$. Equal: equal weights. SIC: Schwarz Information Criterion Weights. Break: equal weights for the break version of all bivariate predictive models. Stable: equal weights for the stable version of all bivariate predictive models. We apply the two-stage forecast combination procedure to the first four columns, meaning that in the first stage, for each bivariate predictive model, we use Cp, DMSFE, Equal or SIC weights to average its stable and break cases, then we apply equal weights to all 14 break-adjusted models. For the last two columns, we simply equally weight all 14 break or stable bivariate predictive models. Monthly data: the estimation sample runs from 1927:01 to 1956:12, and the evaluation sample runs from 1957:01 to 2013:12. Quarterly data: the estimation sample runs from 1947:1 to 1964:4, and the evaluation sample runs from 1965:1 to 2013:4. Yearly data: the estimation sample runs from 1927 to 1964, and the evaluation sample runs from 1965 to 2013.

Figure B.1 Monthly Data Time Series Plots

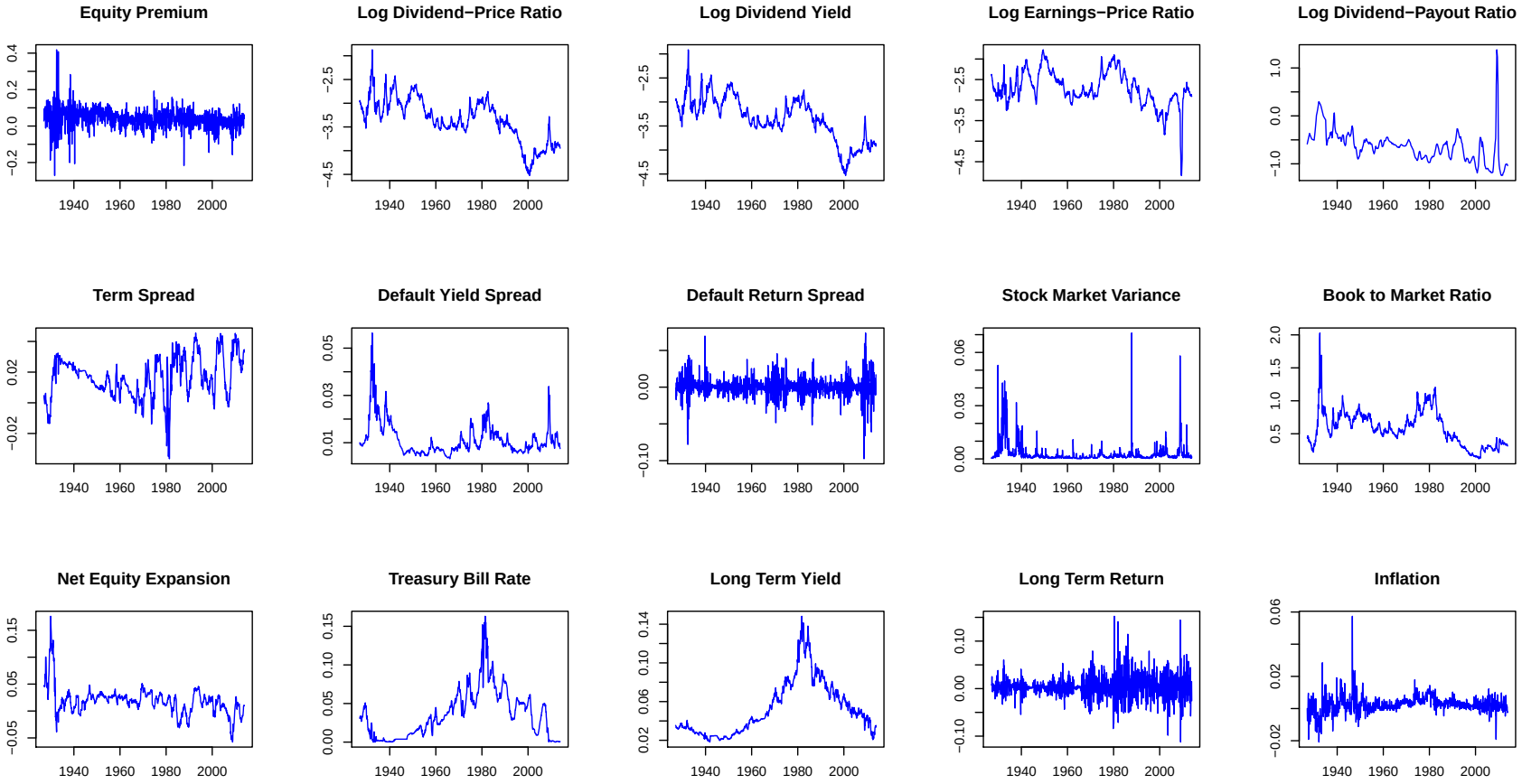


Figure B.2 Quarterly Data Time Series Plots

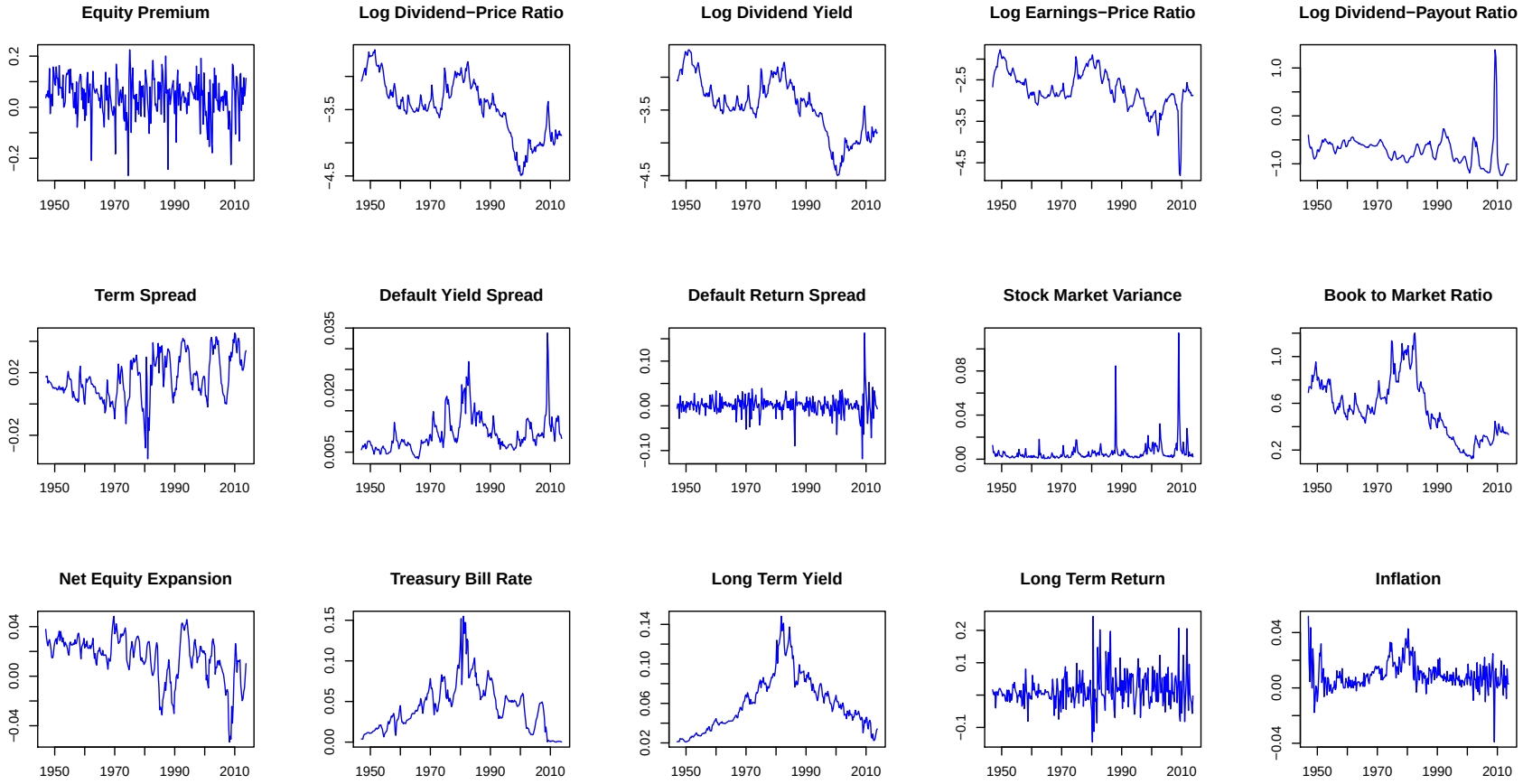


Figure B.3 Annual Data Time Series Plots

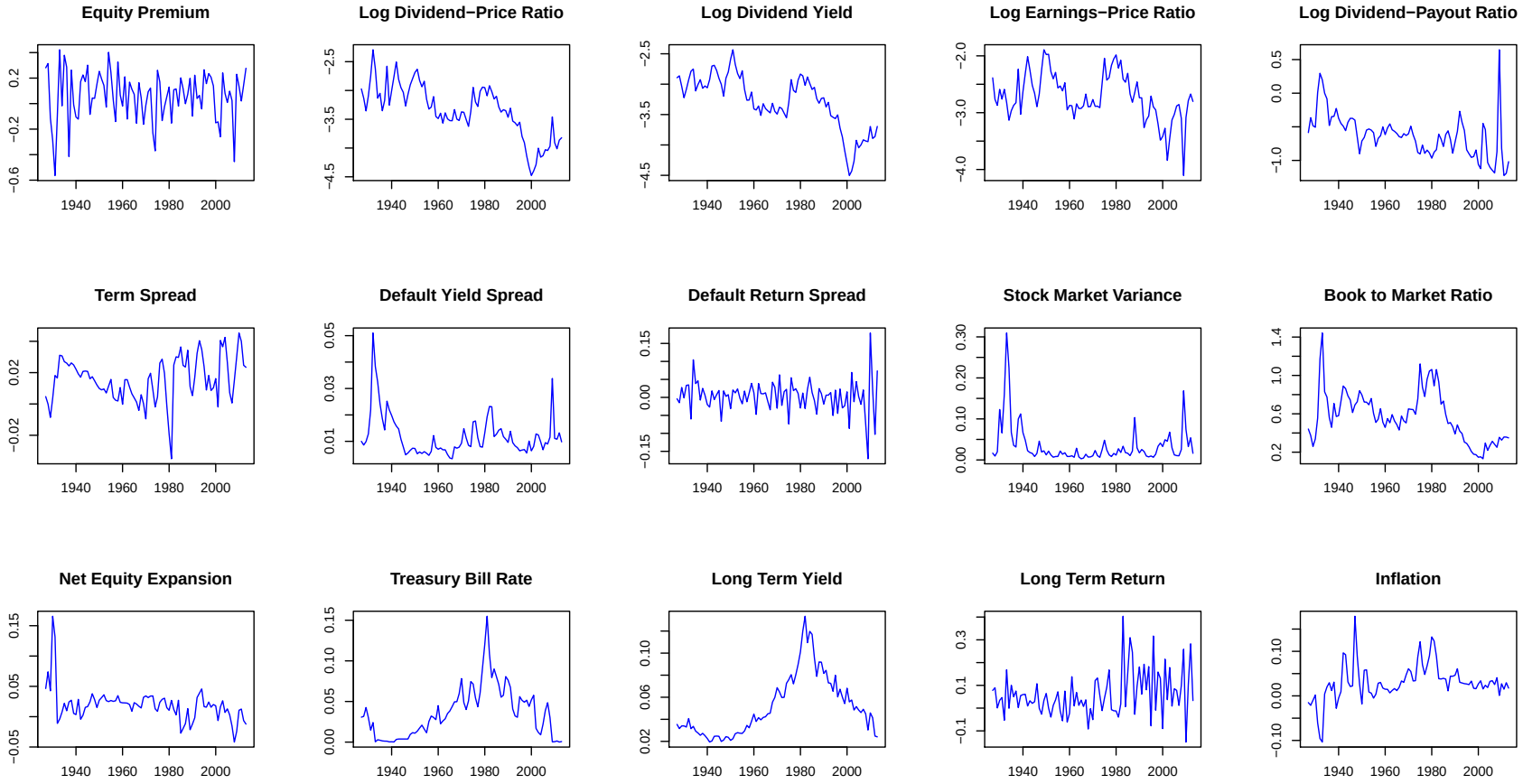


Figure B.4 Monthly Data Variable Correlation Matrix

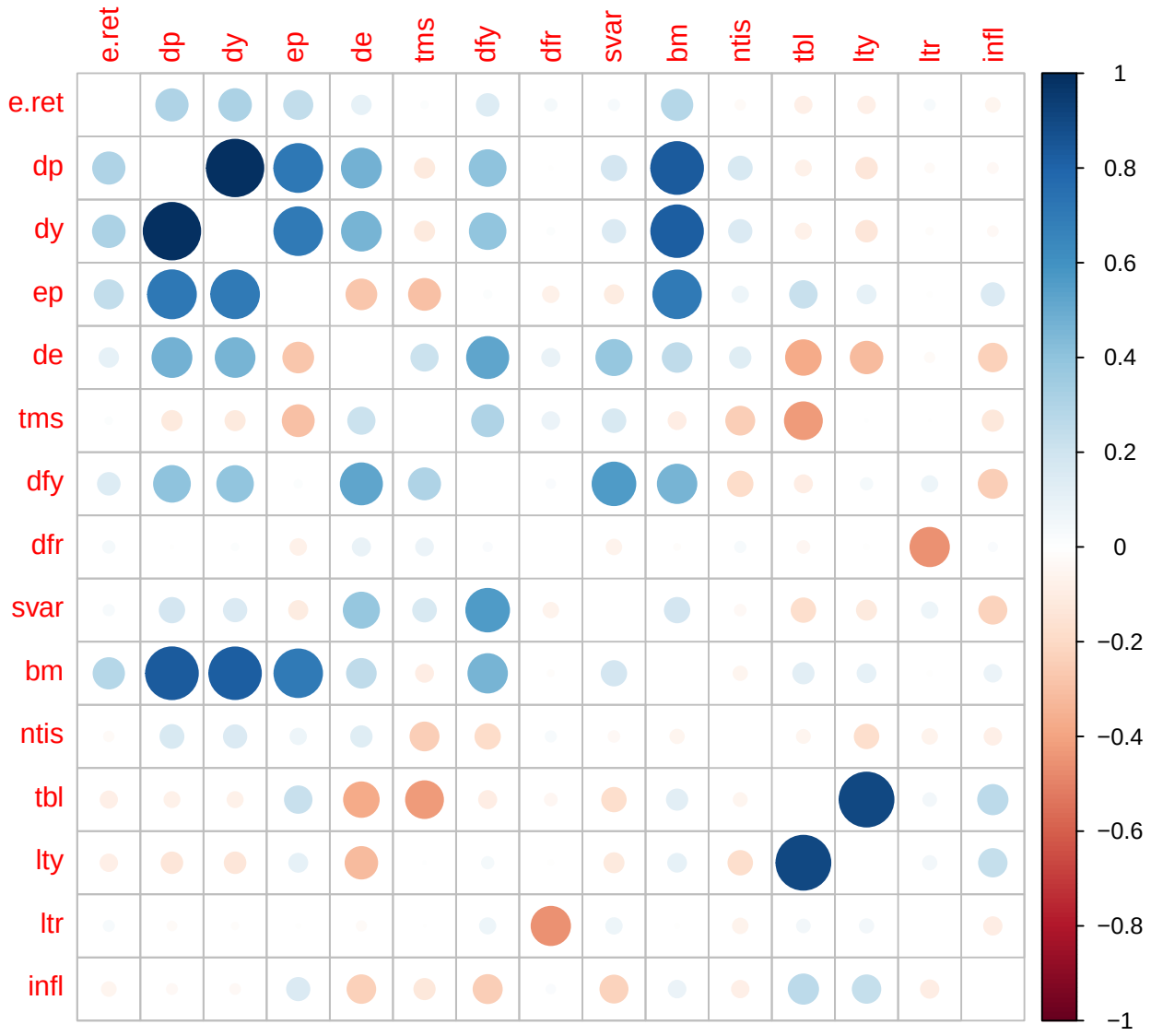


Figure B.5 Quarterly Data Variable Correlation Matrix

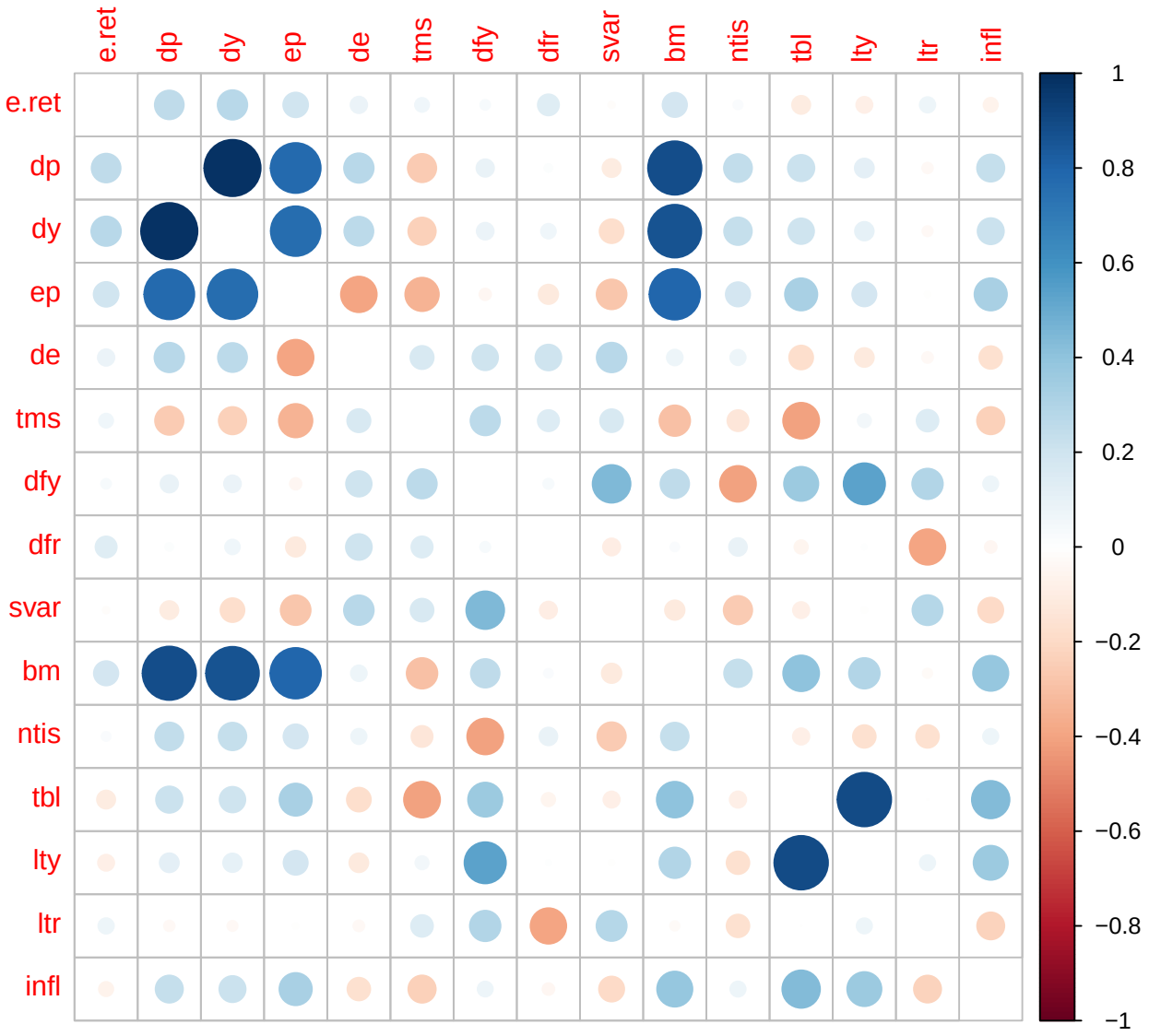


Figure B.6 Annual Data Variable Correlation Matrix

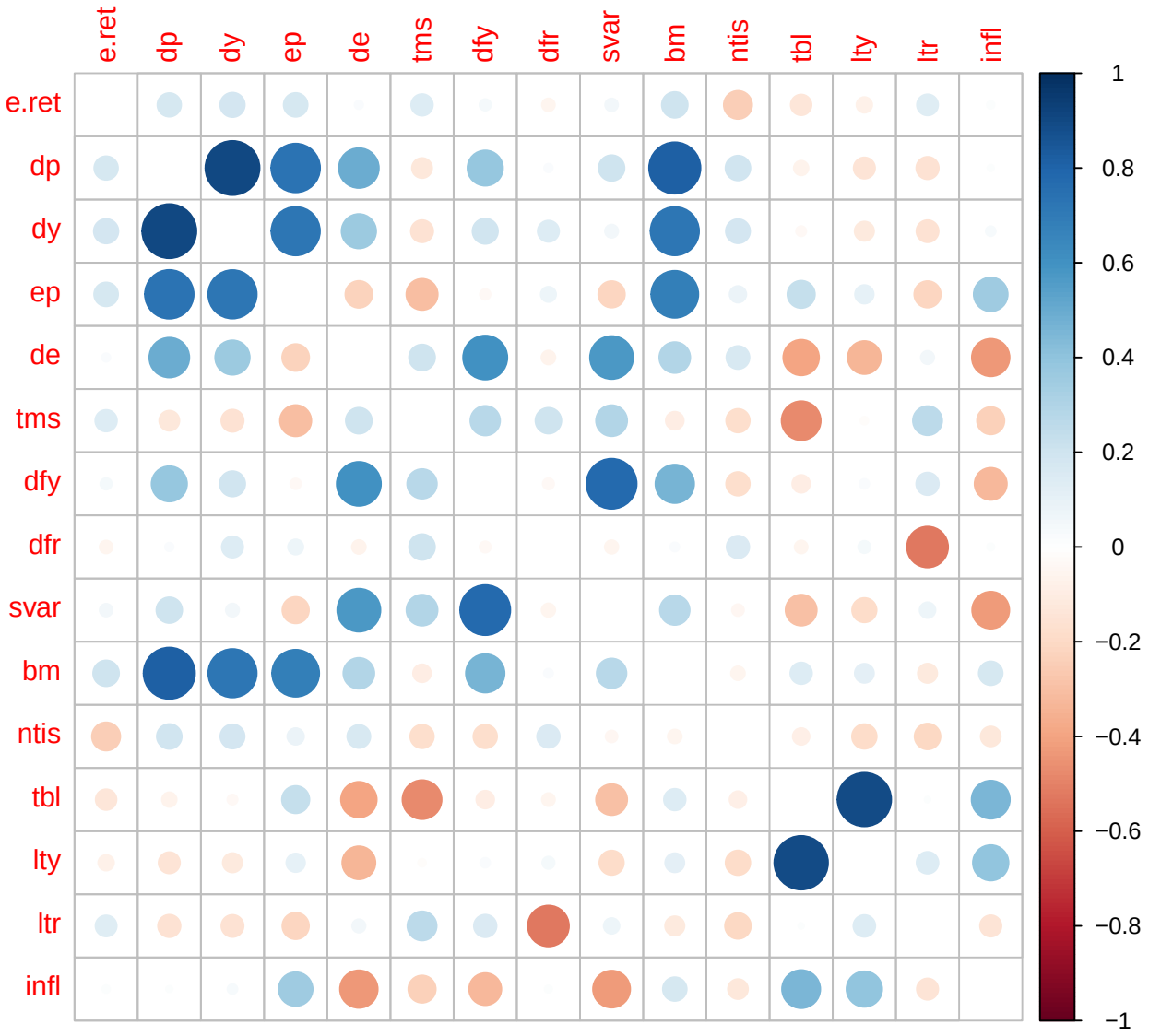


Figure B.7 Cumulative Difference in Squared Forecast Error (CDSFE): Individual Model, Monthly Data

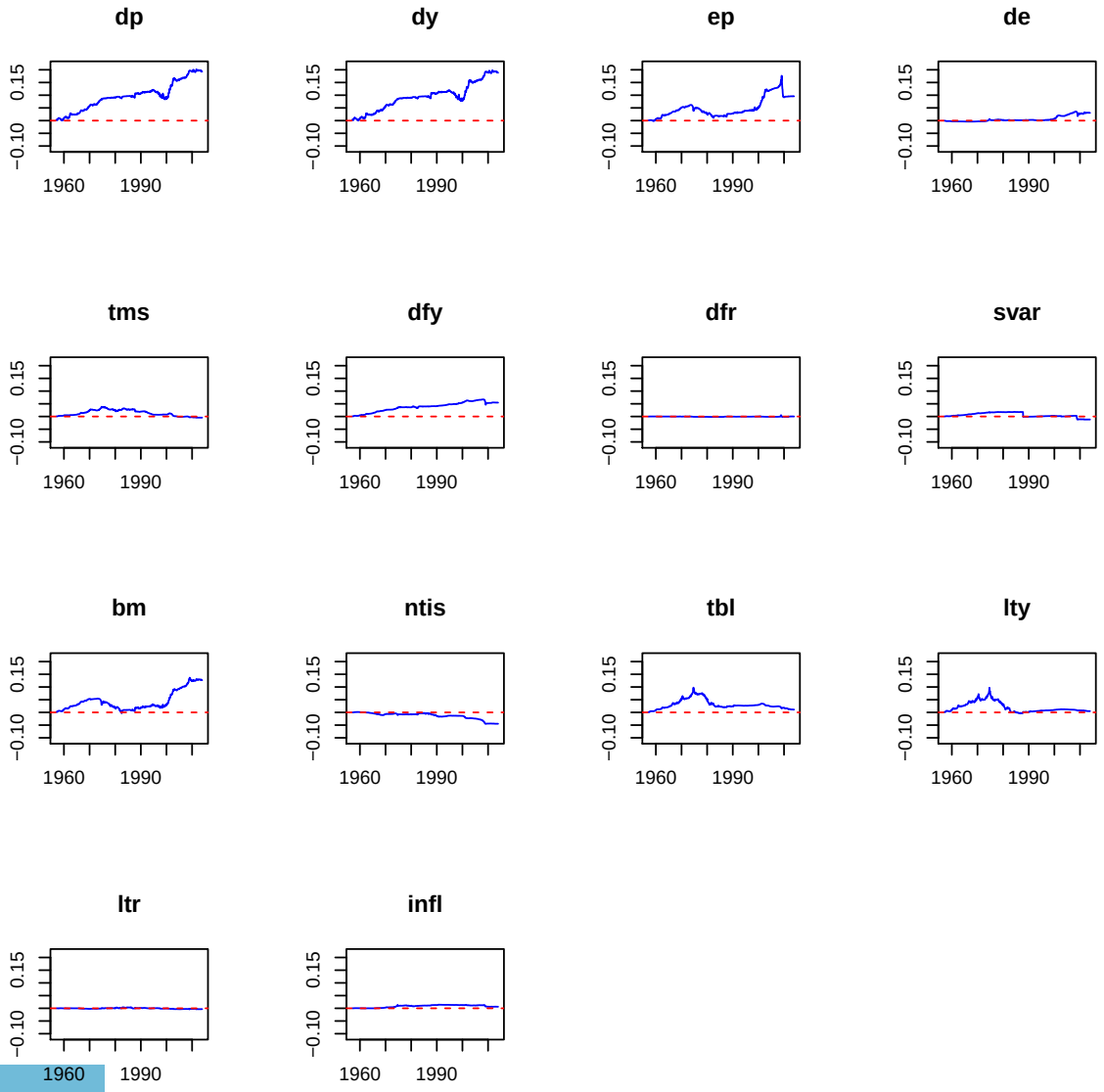


Figure B.8 Cumulative Difference in Squared Forecast Error (CDSFE): Individual Model, Quarterly Data

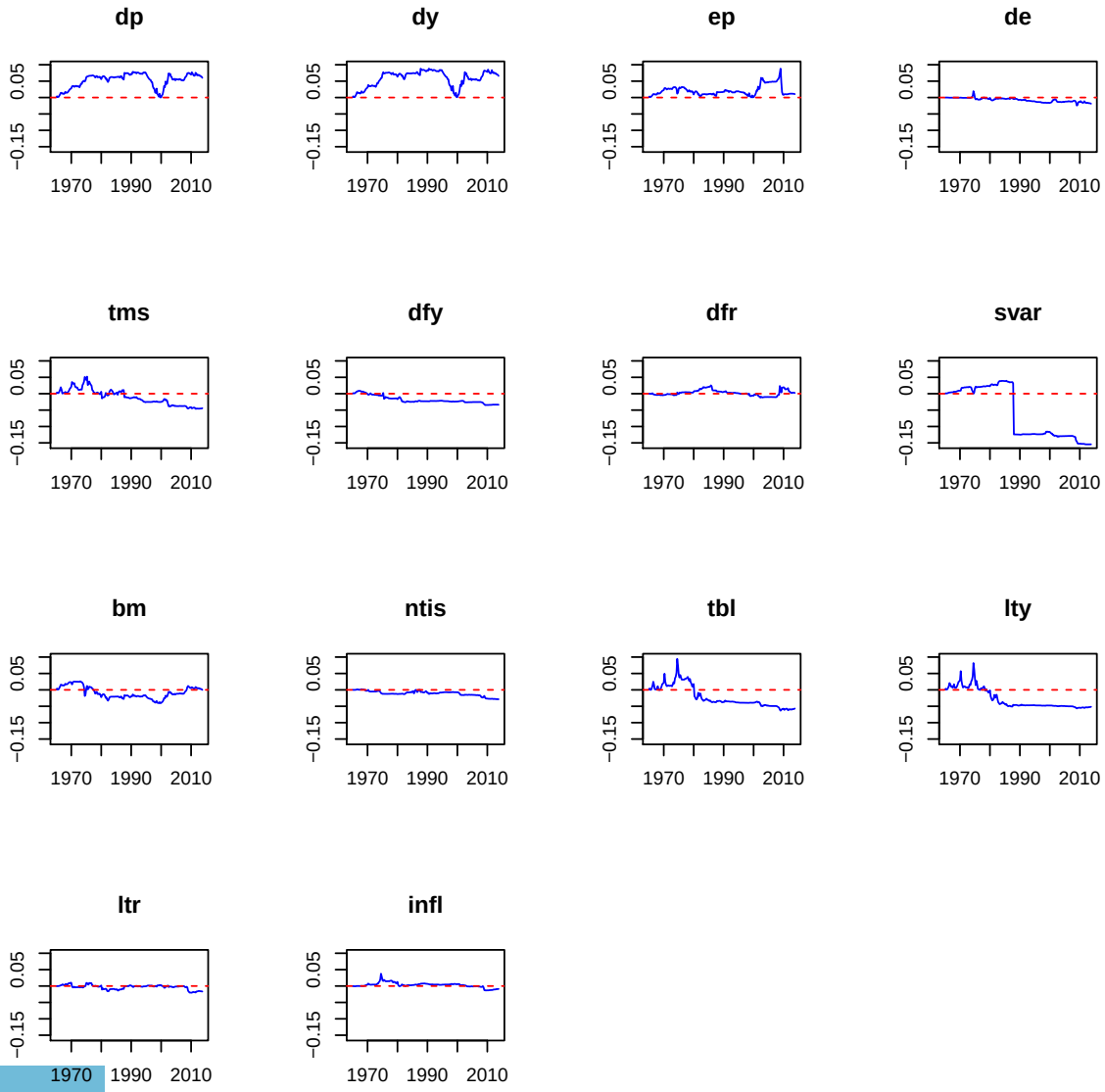


Figure B.9 Cumulative Difference in Squared Forecast Error (CDSFE): Individual Model, Annual Data

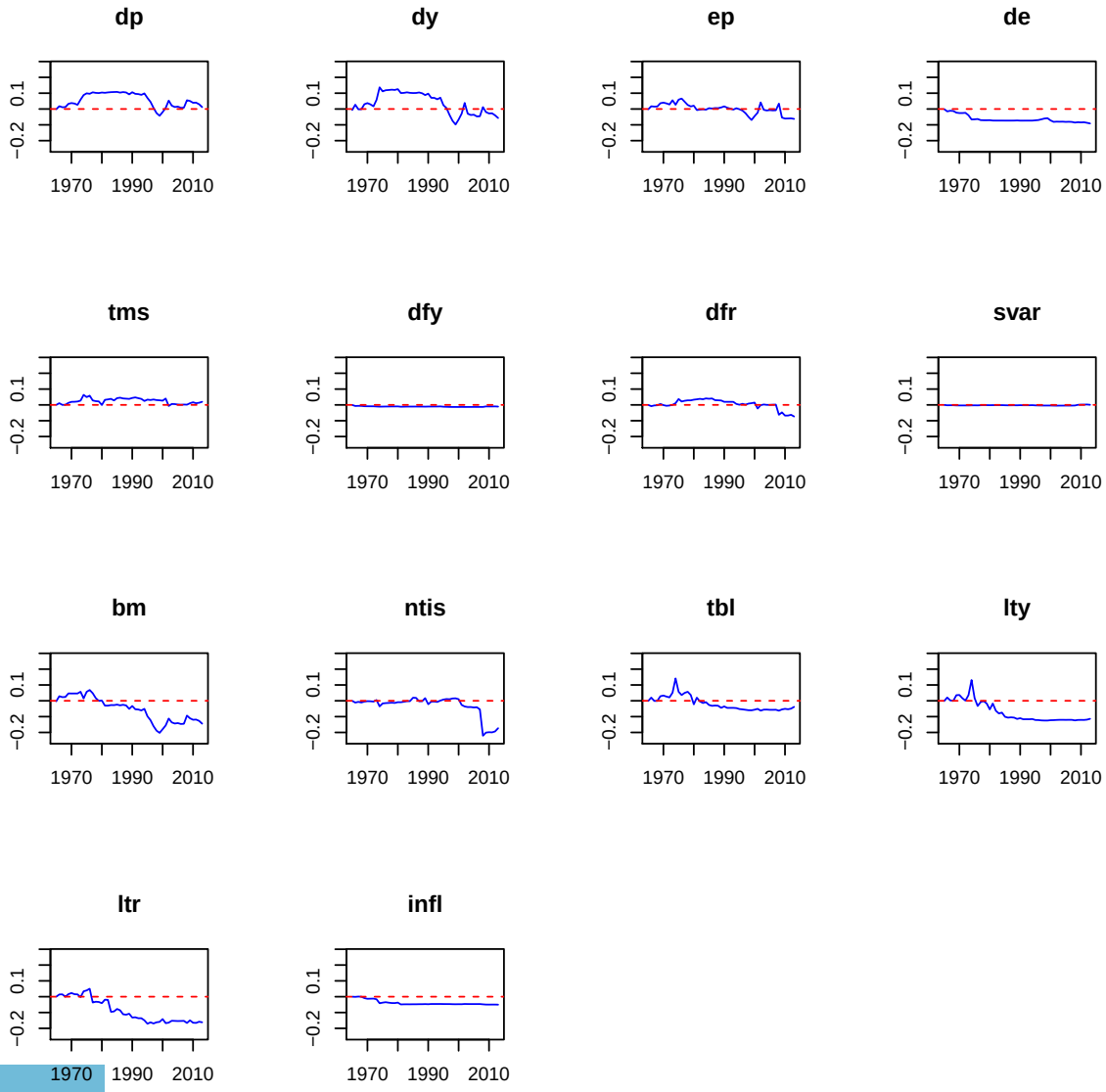


Figure B.10 Monthly Data: Model Out-of-Sample Forecasts Correlation Matrix

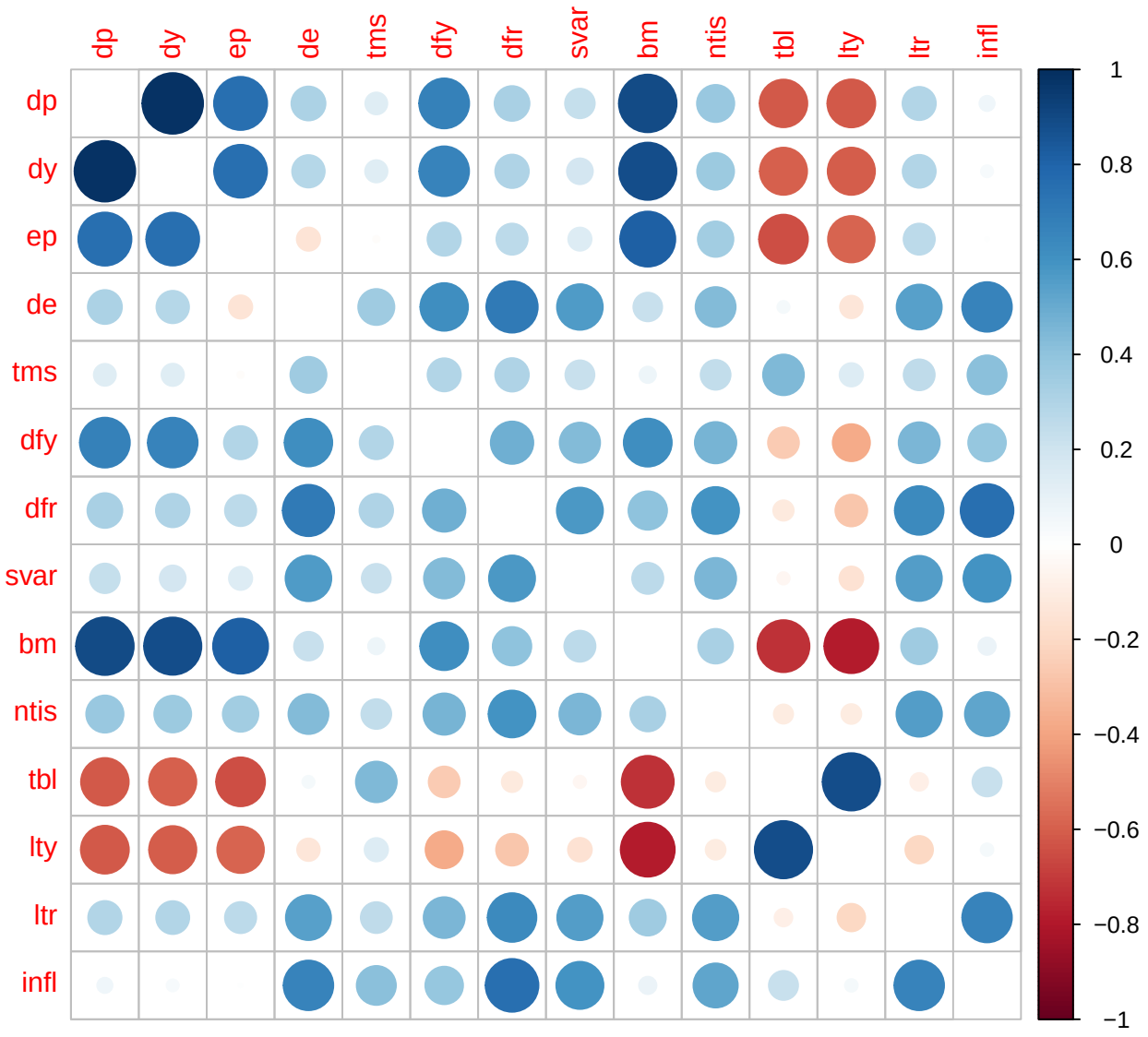


Figure B.11 Quarterly Data: Model Out-of-Sample Forecasts Correlation Matrix

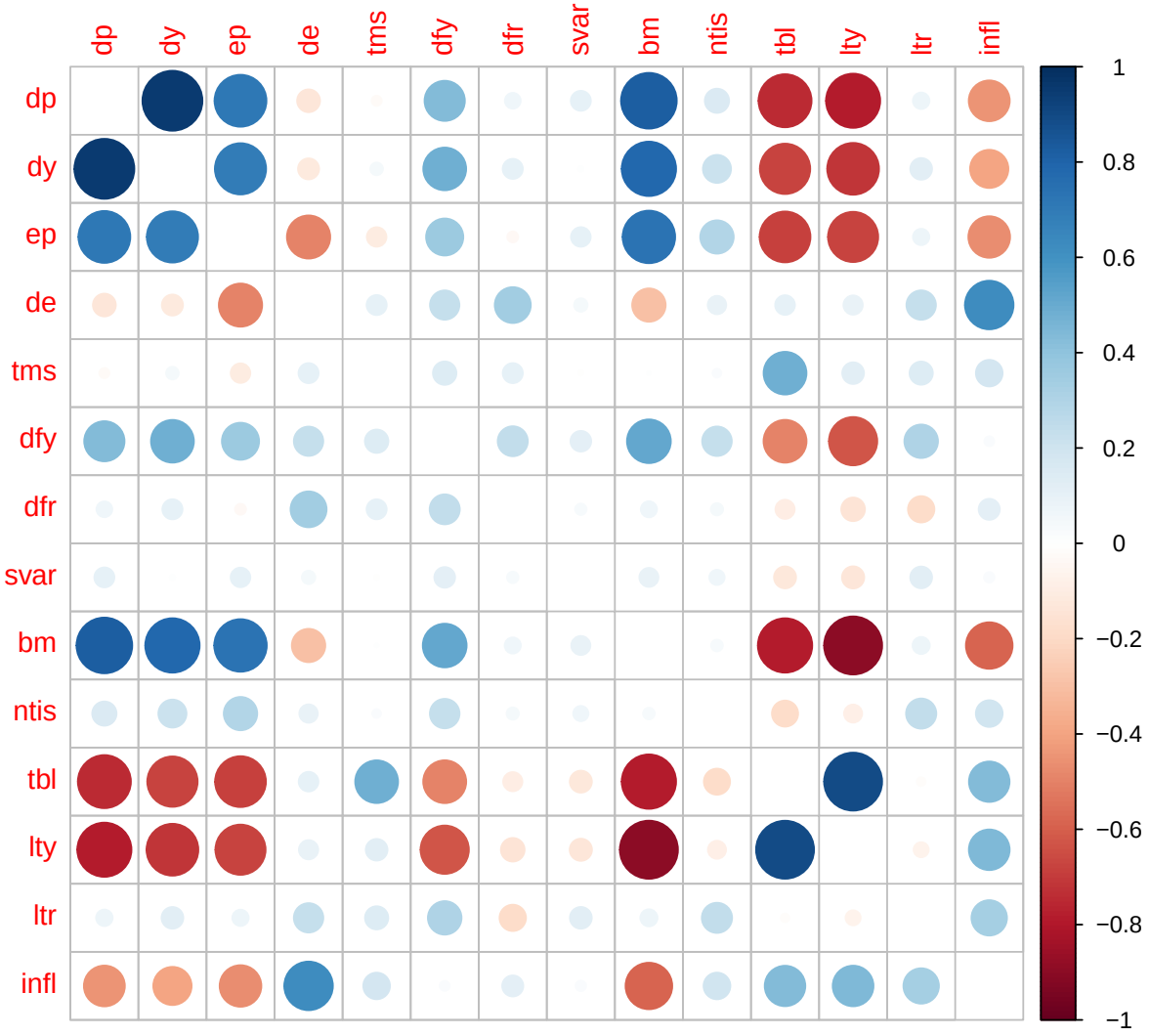


Figure B.12 Annual Data: Model Out-of-Sample Forecasts Correlation Matrix

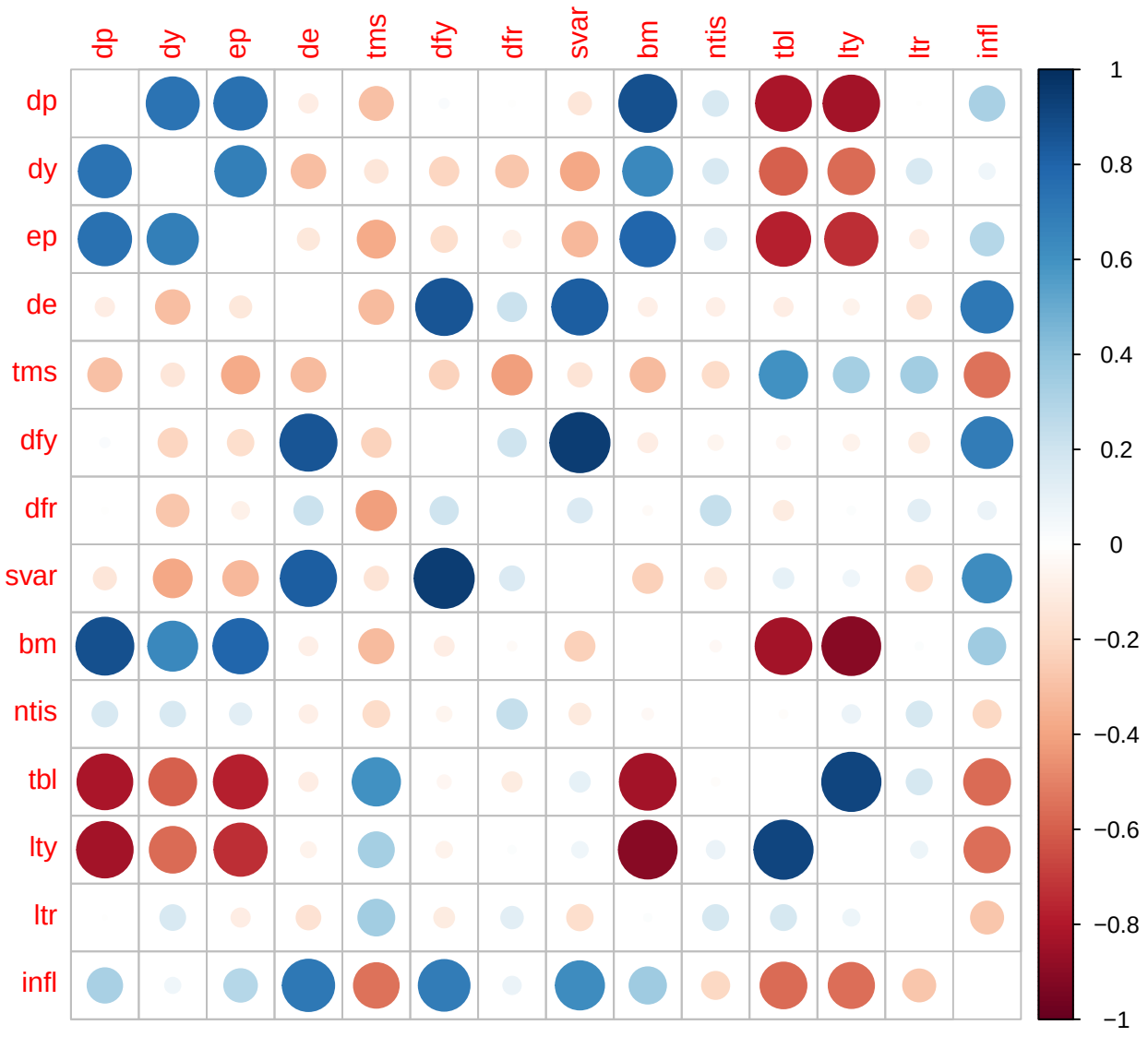


Figure B.13 Cumulative Difference in Squared Forecast Error (CDSFE): Combined Model, Monthly Data

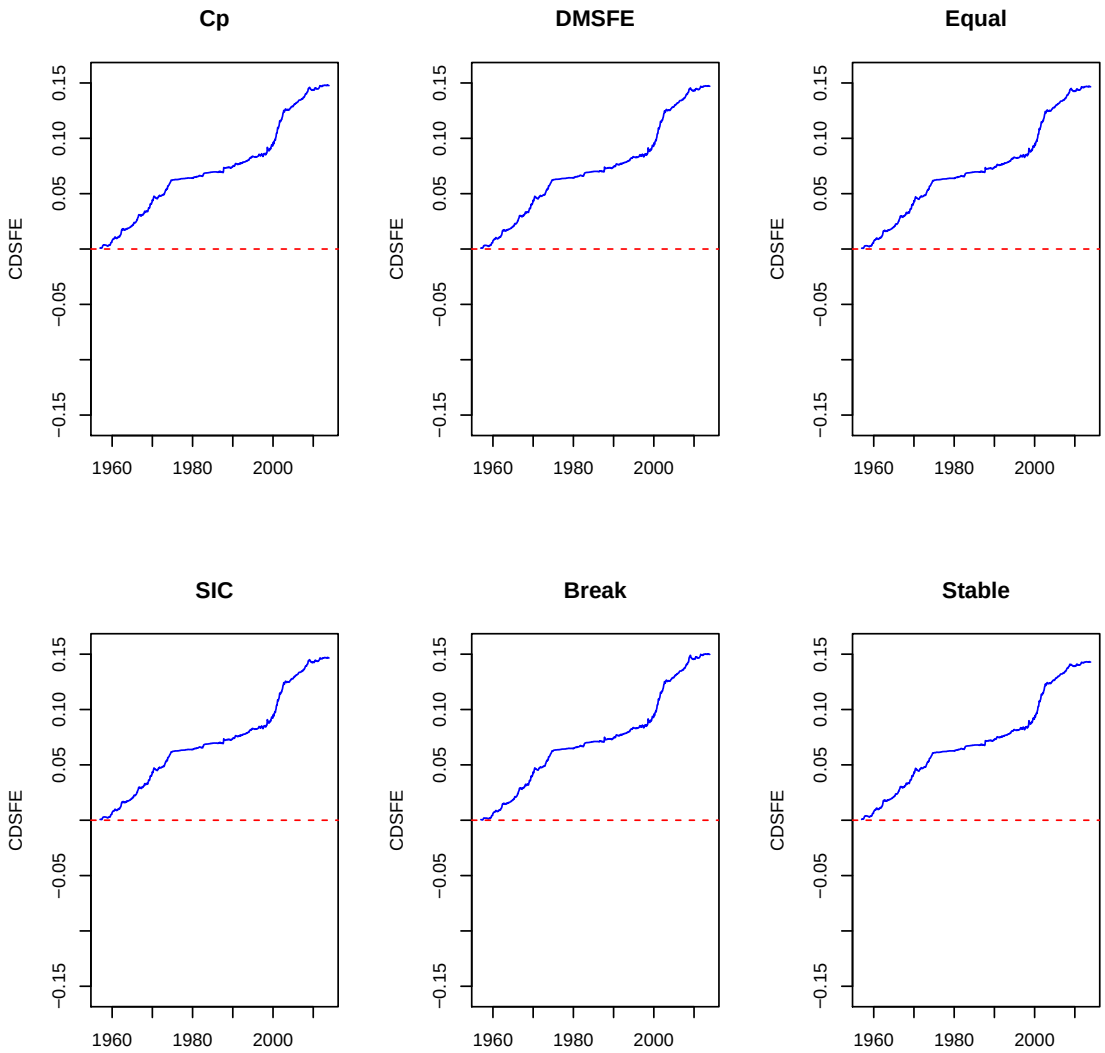


Figure B.14 Cumulative Difference in Squared Forecast Error (CDSFE): Combined Model, Quarterly Data

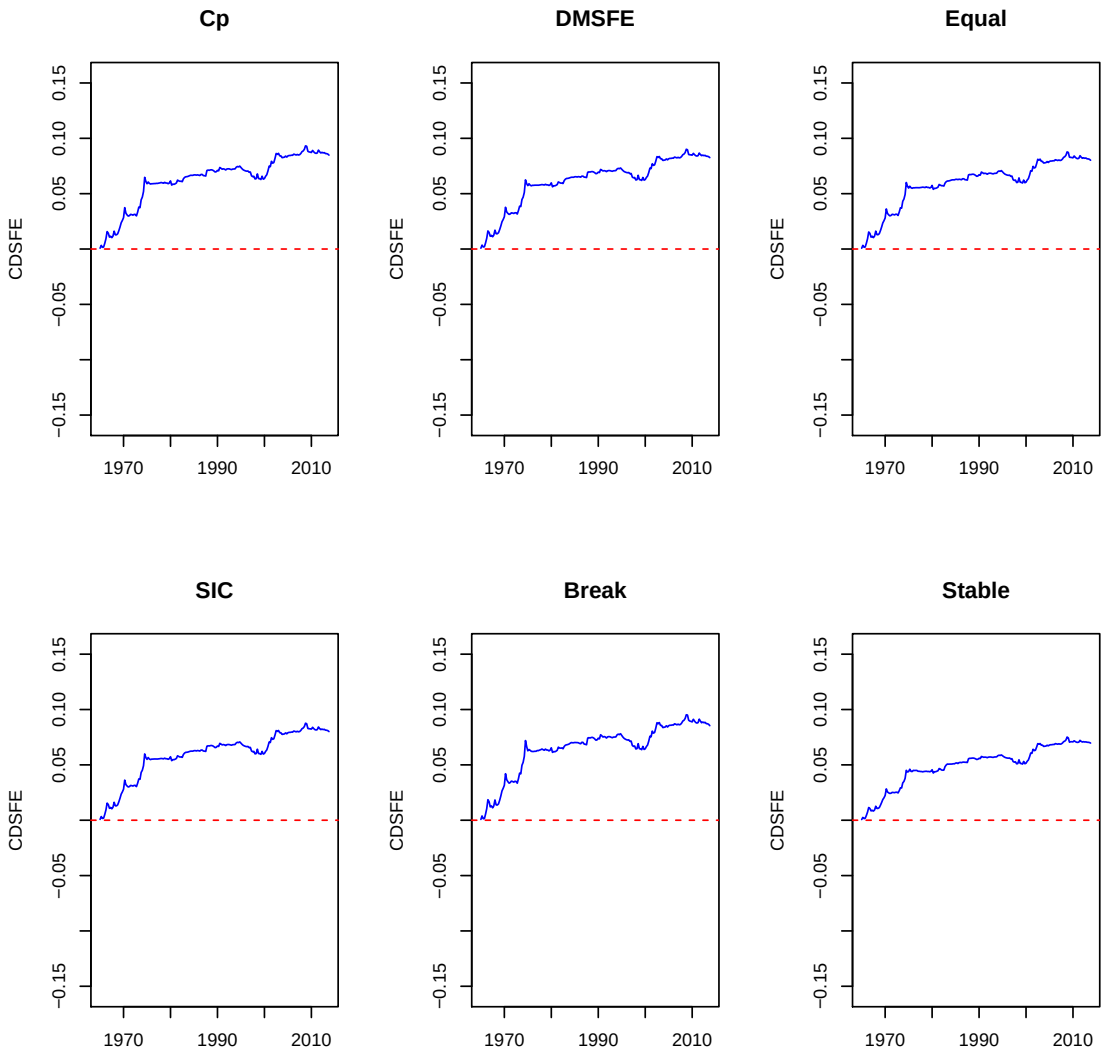
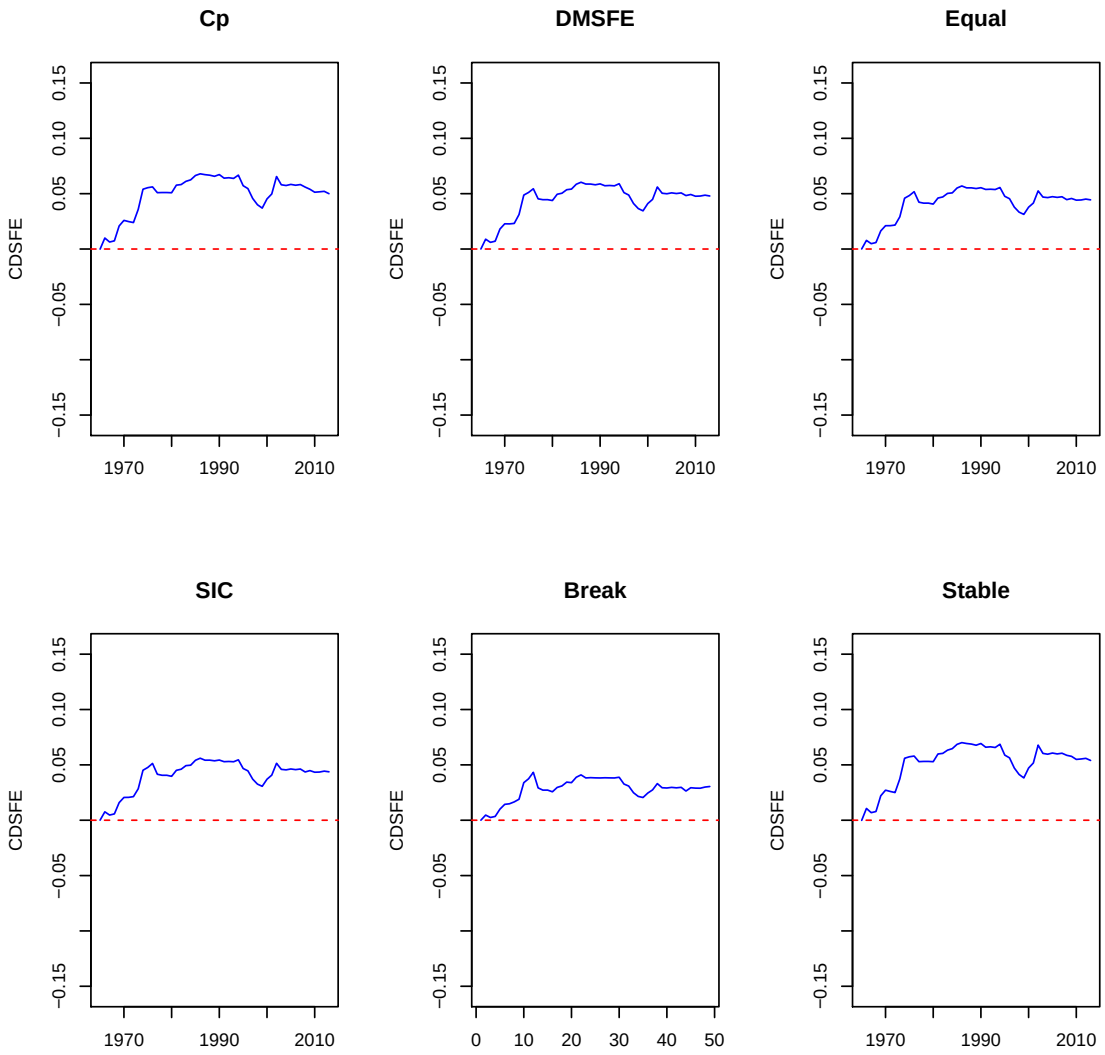


Figure B.15 Cumulative Difference in Squared Forecast Error (CDSFE): Combined Model, Annual Data



APPENDIX C. OUT-OF-SAMPLE FORECAST MODEL AVERAGING WITH PARAMETER INSTABILITY

Tables, Figures and Proofs

Table C.1 Monte Carlo Simulation: Design I

	$P = 30$					$P = 50$				
Break Size	Cp	CV	SIC	Stable	Break	Cp	CV	SIC	Stable	Break
100	0.6312	0.6298	1.2987	1.6557	0.6297	0.6599	0.6585	1.2849	1.6220	0.6584
10	0.6644	0.6627	1.2563	1.6148	0.6627	0.6871	0.6854	1.2473	1.5874	0.6853
5	0.7085	0.7066	1.2063	1.5605	0.7065	0.7289	0.7271	1.2005	1.5335	0.7270
3	0.7658	0.7636	1.1517	1.4782	0.7636	0.7869	0.7850	1.1454	1.4489	0.7850
2	0.8330	0.8308	1.0974	1.3734	0.8308	0.8500	0.8483	1.0925	1.3471	0.8483

Notes: The DGP is $y_t = \mu + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \sum_{i=1}^{5} \theta_i x_i + e_t, e_t = v_t\sqrt{h_t}, h_t = \alpha_0 + \alpha_1 e_{t-1}^2$ and the forecasting model is $y_t = \mu + \sum_{i=1}^{4} \theta_i x_i + e_t$. P is the evaluation sample size, total sample size is 200, break fraction relative to the training sample is $\pi = 0.3$, OOS forecasts are generated by the recursive window, 5000 times replication. Equal weight is chosen as the benchmark and the numbers in the table represent the RMSFE ratio between each individual method and equal weight. Cp: Mallows' weights. CV: cross-validation weights. SIC: Schwarz-Bayesian weights. Stable: model without structural breaks. Break: model with a full structural break.

Table C.2 Monte Carlo Simulation: Design II

	$P = 30$					$P = 50$				
Break Size	Cp	CV	SIC	Stable	Break	Cp	CV	SIC	Stable	Break
100	0.4610	0.2586	1.0717	1.9477	0.2586	0.5706	0.3415	1.0649	1.8951	0.3415
10	0.7007	0.5681	1.0393	1.6830	0.5683	0.6945	0.5419	1.0392	1.6930	0.5421
5	0.8422	0.7700	1.0194	1.4212	0.7701	0.8699	0.7946	1.0191	1.3916	0.7948
3	0.8978	0.8541	1.0111	1.2809	0.8543	0.9135	0.8800	1.0126	1.2551	0.8803
2	0.9188	0.8778	1.0082	1.2352	0.8781	0.9417	0.9320	1.0074	1.1578	0.9323

Notes: The DGP is $y_t = \mu + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \sum_{i=1}^{2} \theta_i x_i + e_t, e_t \sim \mathrm{N}(0, y_{t-1}^2)$ and the forecasting model is $y_t = \mu + \sum_{i=1}^{2} \theta_i x_i + e_t$. P is the evaluation sample size, total sample size is 200, break fraction relative to the training sample is $\pi = 0.3$, OOS forecasts are generated by the recursive window, 5000 times replication. Equal weight is chosen as the benchmark and the numbers in the table represent the RMSFE ratio between each individual method and equal weight. Cp: Mallows' weights. CV: cross-validation weights. SIC: Schwarz-Bayesian weights. Stable: model without structural breaks. Break: model with a full structural break.

Table C.3 Monte Carlo Simulation: Design III

	$P = 30$					$P = 50$				
Break Size	Cp	CV	SIC	Stable	Break	Cp	CV	SIC	Stable	Break
100	0.9810	0.9759	1.0011	1.0839	0.9760	0.9825	0.9769	1.0011	1.0789	0.9770
10	0.9860	0.9789	1.0006	1.0716	0.9790	0.9880	0.9822	1.0006	1.0656	0.9823
5	0.9919	0.9850	1.0003	1.0583	0.9852	0.9933	0.9868	1.0003	1.0534	0.9870
3	0.9977	0.9903	1.0000	1.0455	0.9906	0.9975	0.9905	1.0001	1.0428	0.9908
2	1.0009	0.9940	0.9999	1.0347	0.9944	1.0013	0.9952	0.9999	1.0316	0.9958

Notes: The DGP is $y_t = \mu + \rho_1 y_{t-1} + \rho_2 y_{t-2} + e_t$, $e_t \sim N(0, \sigma^2) \quad t \in [1, \tau_v]$ and $e_t \sim N(0, \frac{1}{4}\sigma^2) \quad t \in [\tau_v + 1, R]$, $\tau_v = 0.5R$, the forecasting model is $y_t = \mu + \rho_1 y_{t-1} + e_t$. P is the evaluation sample size, total sample size is 200, break fraction relative to the training sample is $\pi = 0.3$, OOS forecasts are generated by the recursive window, 5000 times replication. Equal weight is chosen as the benchmark and the numbers in the table represent the RMSFE ratio between each individual method and equal weight. Cp: Mallows' weights. CV: cross-validation weights. SIC: Schwarz-Bayesian weights. Stable: model without structural breaks. Break: model with a full structural break.

Table C.4 U.S. Quarterly GDP Growth Rate Forecast Comparison

	Model a			Model b			Model c			Model d			Model e		
	Cp	CV	SIC	Cp	CV	SIC	Cp	CV	SIC	Cp	CV	SIC	Cp	CV	SIC
P = 20	1.044	0.967	0.999	1.031	0.983	0.999	1.017	0.987	0.999	1.038	0.970	0.998	1.043	0.960	0.997
P = 25	1.038	0.968	0.999	1.021	0.984	0.999	1.036	0.976	0.999	1.038	0.969	0.998	1.017	0.967	0.998
P = 30	1.022	0.977	0.999	1.022	0.983	0.999	1.007	0.996	1.000	1.013	0.991	0.998	1.032	0.975	0.998
P = 35	1.020	0.980	1.000	1.036	0.996	0.999	1.022	0.983	0.999	1.024	0.983	0.999	1.034	0.973	0.998
P = 40	1.022	0.979	0.999	1.012	0.987	1.000	1.024	0.982	0.999	1.025	0.982	0.999	1.033	0.974	0.998
P = 45	1.024	0.978	1.000	1.014	0.986	1.000	1.025	0.982	0.999	1.026	0.981	0.999	1.037	0.974	0.998
P = 50	1.021	0.987	1.000	1.011	0.989	1.000	1.027	0.984	0.999	1.023	0.987	0.999	1.022	0.988	0.999

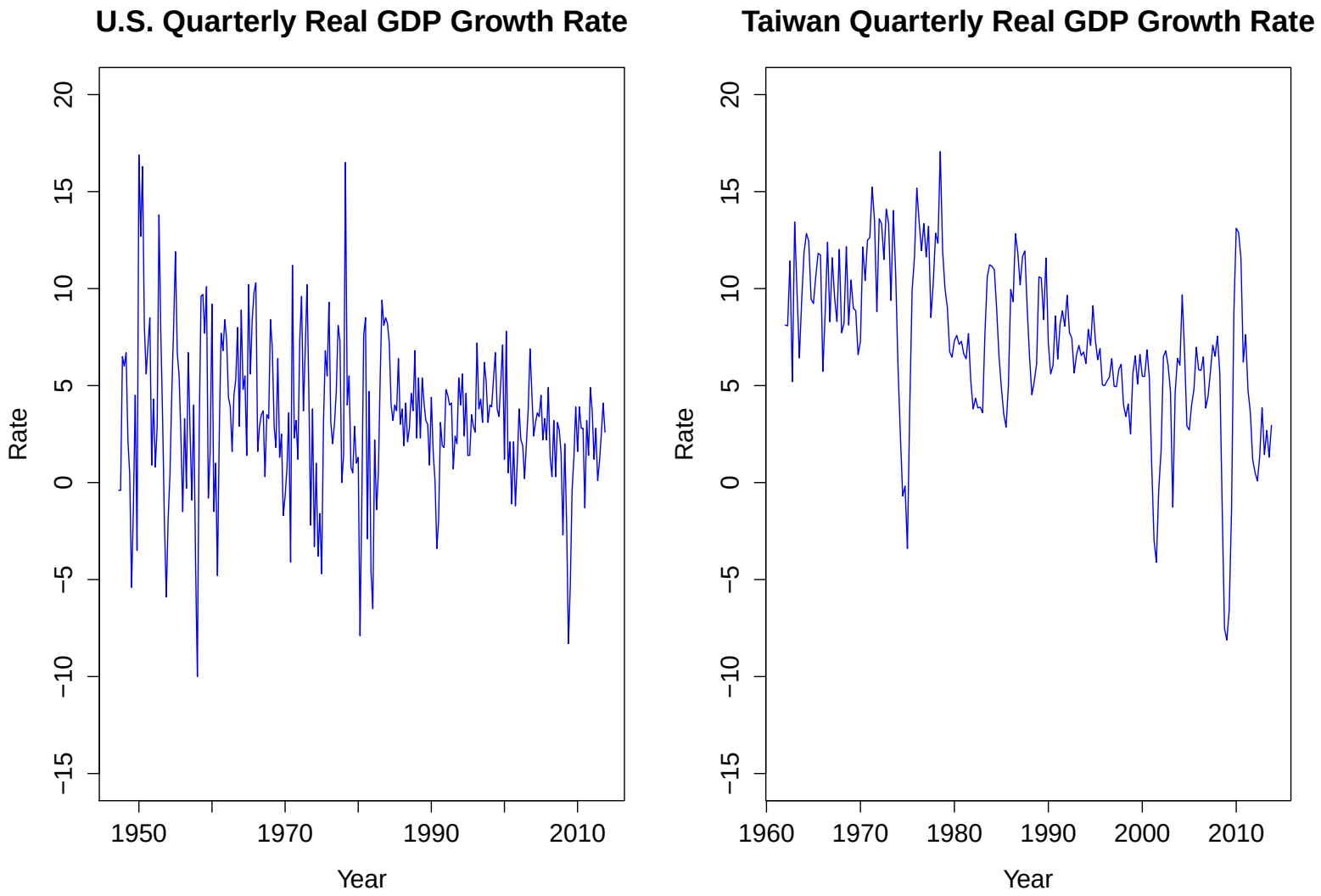
Notes: Quarterly data from 1960:1 to 2012:1. P is the evaluation sample size. Equal weight is chosen as the benchmark and the numbers in the table represent the RMSFE ratio between each individual method and equal weight. Smaller number indicates better forecasting performance. Cp: Mallows' weights. CV: cross-validation weights. SIC: Schwarz-Bayesian weights.
Model a: AR(1)
Model b: AR(2)
Model c: AR(1) + SR
Model d: AR(1) + SR + LR
Model e: AR(1) + SR + LR + DP

Table C.5 Taiwan Quarterly GDP Growth Rate Forecast Comparison

	Model AR(1)			Model AR(2)		
	Cp	CV	SIC	Cp	CV	SIC
P = 20	0.991	0.947	0.999	0.968	0.944	1.000
P = 25	0.998	0.994	1.000	0.972	0.942	1.000
P = 30	0.998	0.995	1.000	0.973	0.943	1.000
P = 35	0.999	0.995	1.000	0.974	0.945	1.000
P = 40	0.998	0.993	1.000	0.976	0.948	1.000
P = 45	0.998	0.993	1.000	0.982	0.961	1.000
P = 50	0.997	0.996	1.000	0.984	0.962	1.000

Notes: Quarterly data from 1962:1 to 2013:4. P is the evaluation sample size. Equal weight is chosen as the benchmark and the numbers in the table represent the RMSFE ratio between each individual method and equal weight. Smaller number indicates better forecasting performance. Cp: Mallows' weights. CV: cross-validation weights. SIC: Schwarz-Bayesian weights.

Figure C.1 U.S. and Taiwan Quarterly GDP Growth Rate



Proof of Proposition 3.3.1. From the cross-validation criterion, for linear regression models we have the well-known result that

$$\frac{1}{T}\sum_{i=1}^{T}\tilde{e}_t^{\,2} = \frac{1}{T}\sum_{i=1}^{T}\frac{\hat{e}_t^{\,2}}{(1-h_t)^2}$$

where $h_t = x_t'(X'X)^{-1}x_t$ is the leverage associated with observation t. Applying Taylor expansion to h_t around 0, we can expand the above equation as

$$\begin{aligned}\frac{1}{T}\sum_{i=1}^{T}\tilde{e}_t^{\,2} &= \frac{1}{T}\sum_{i=1}^{T}\frac{\hat{e}_t^{\,2}}{(1-h_t)^2}\\ &\approx \frac{1}{T}\sum_{i=1}^{T}\hat{e}_t^{\,2} + \frac{2}{T}\sum_{i=1}^{T}\hat{e}_t^{\,2}h_t\\ &= \hat{\sigma}^2 + \frac{2}{T}\sum_{i=1}^{T}\hat{e}_t^{\,2}x_t'(X'X)^{-1}x_t\end{aligned}$$

Under regularity conditions listed in Assumption 1, we have $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$. For the penalty term, $\frac{1}{T}\sum_{i=1}^{T}\hat{e}_t^{\,2}x_t'(X'X)^{-1}x_t \xrightarrow{p} E(e'Pe)$. Putting these two parts together, we can see that CV is asymptotically equivalent to Mallows' Cp under our assumptions except for conditionally homoscedastic errors. □

Proof of Corollary 3.3.1. Since CV is asymptotically equivalent to Mallows' Cp, following proof in Hansen (2009), write the sample CV criterion ($\widehat{\mathrm{CV}}(w)$) for the weighted model as a function of the break model weight w,

$$\widehat{\mathrm{CV}}(w) = (w\hat{e} + (1-w)\tilde{e})'(w\hat{e} + (1-w)\tilde{e}) + 2(T-2k)^{-1}(k + w\bar{p})\hat{e}'\hat{e}$$

where $\bar{p}$ proposed by Hansen is used to approximate the infeasible expected value of the population penalty term. The sample optimal CV weight $\hat{w}$ is the value in $[0,1]$ that minimizes $\widehat{\mathrm{CV}}(w)$, so

$$\hat{w} = \frac{(T-2k)(\sum_{t=1}^{T}\tilde{e}_t^2 - \sum_{t=1}^{T}\hat{e}_t^2) - \bar{p}\sum_{t=1}^{T}\hat{e}_t^2}{(T-2k)(\sum_{t=1}^{T}\tilde{e}_t^2 - \sum_{t=1}^{T}\hat{e}_t^2)}$$

if $(T-2k)(\sum_{t=1}^{T}\tilde{e}_t^2 - \sum_{t=1}^{T}\hat{e}_t^2)(\sum_{t=1}^{T}\hat{e}_t^2)^{-1} \geq \bar{p}$ while $\hat{w} = 0$ otherwise. □

Proof of Proposition 3.3.2. The proof of this proposition is adapted from Hansen (2009). By projection arguments, $P(m) = P + P^*(m)$, where

$$P = X(X'X)^{-1}X',$$

$$P^*(m) = X^*(m)(X^*(m)'X^*(m))^{-1}X^*(m)',$$

$X^*(m) = X(m) - PX(m) = X(m) - X(X'X)^{-1}X'X(m) = X(m) - X(X'X)^{-1}X(m)'X(m)$,

and $X(m)$ is the matrix of stacked regressors $x_t (t < m)$. The cross-validation penalty term can be expanded as:

$$\begin{aligned} e'P(m)e &= e'Pe + e'P^*(m)e \\ &= e'Pe + e'X^*(m)(X^*(m)'X^*(m))^{-1}X^*(m)'e \end{aligned}$$

We start by showing the asymptotic distribution of the second term on the right-hand-side of the above equation, $e'P^*(m)e = e'X^*(m)(X^*(m)'X^*(m))^{-1}X^*(m)'e$. For this term, $X^*(m)'X^*(m)$, we have

$$\begin{aligned} X^*(m)'X^*(m) &= (X(m) - X(X'X)^{-1}X(m)'X(m))'(X(m) - X(X'X)^{-1}X(m)'X(m)) \\ &= X(m)'X(m) - X(m)'X(X'X)^{-1}X(m)'X(m) \\ &\quad -X(m)'X(m)(X'X)^{-1}X'X(m) \\ &\quad +X(m)'X(m)(X'X)^{-1}X(m)'X(m) \\ &= X(m)'X(m) - X(m)'X(X'X)^{-1}X(m)'X(m) \end{aligned}$$

From our assumptions and $\frac{m}{T} \to \pi$, by laws of large numbers, we have

$$\frac{1}{T}X(m)'X(m) \xrightarrow{P} \pi Q$$

and

$$\frac{1}{T}X(m)'X(X'X)^{-1}X(m)'X(m) \xrightarrow{P} \pi Q Q^{-1} \pi Q$$

so

$$\frac{1}{T}X^*(m)'X^*(m) \xrightarrow{P} \pi(1-\pi)Q$$

By the continuous mapping theorem we have

$$(\frac{1}{T}X^*(m)'X^*(m))^{-1} \xrightarrow{P} (\pi(1-\pi))^{-1}Q^{-1}$$

For this term, $X^*(m)'e = (X(m) - X(X'X)^{-1}X(m)'X(m))'e$, we can show

$$\begin{aligned}(X(m) - X(X'X)^{-1}X(m)'X(m))'e &= X(m)'e - X(m)'X(m)(X'X)^{-1}X'e \\ &= \sum_{t=1}^{[T\pi]} x_t e_t - \sum_{t=1}^{[T\pi]} x_t x_t' \left(\sum_{t=1}^{T} x_t x_t'\right)^{-1} \left(\sum_{t=1}^{T} x_t e_t\right)\end{aligned}$$

Next, applying laws of large numbers and the mixing functional central limit theorem, we have

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{[T\pi]} x_t e_t \Rightarrow W(\pi)$$

$$\frac{1}{T}\sum_{t=1}^{[T\pi]} x_t x_t' \xrightarrow{P} \pi Q$$

$$\left(\frac{1}{T}\sum_{t=1}^{T} x_t x_t'\right)^{-1} \xrightarrow{P} Q^{-1}$$

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} x_t e_t \Rightarrow W(1)$$

where $W(1)$ is a Brownian motion vector with covariance matrix $\Sigma \equiv \lim_{n\to\infty} \text{VAR}(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_i e_i)$, and $W(\pi)$ is a Brownian vector indexed at time π.

Putting together the results obtained above, we have

$$\frac{1}{\sqrt{T}}X^*(m)'e \Rightarrow W(\pi) - \pi W(1)$$

Then we have

$$\frac{1}{T}e'P^*(m)e \Rightarrow \frac{1}{\pi(1-\pi)}(W(\pi) - \pi W(1))'Q^{-1}(W(\pi) - \pi W(1)) = \frac{\mathbf{B}(\pi)'\mathbf{B}(\pi)}{\pi(1-\pi)}$$

where $\mathbf{B}(\pi)$ is a Brownian bridge. Since the break date m needs to be estimated, combined with Andrews' Andrews (1993) theorem 4, we have $\frac{1}{T}e'P^*(\hat{m})e \Rightarrow J_0(\xi_\delta)$.

For the first component in the penalty term, $e'Pe$, we have

$$e'Pe = (\frac{1}{\sqrt{T}}\sum_{t=1}^{T} x_t e_t)'(\frac{1}{T}\sum_{t=1}^{T} x_t x_t')^{-1}(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} x_t e_t)$$

Again, applying laws of large numbers and central limit theorem,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} x_t e_t \Rightarrow W(1)$$

$$\frac{1}{T}\sum_{t=1}^{T} x_t x_t' \xrightarrow{p} Q$$

so

$$e'Pe \xrightarrow{p} \Psi' Q^{-1}\Psi$$

where $\Psi \sim N(0, \Sigma)$.

Σ is symmetric and positive definite, Q^{-1} is of the same rank of Σ, applying results of the distribution of quadratic forms (see section 5.4 of Ravishanker and Dipak (2001)), we have

$$e'Pe \xrightarrow{d} \sum_{j=1}^{k} \lambda_j \chi^2(1)$$

Collecting all the results shown above, we have

$$e'P(\hat{m})e \xrightarrow{d} \sum_{j=1}^{k} \lambda_j \chi^2(1) + J_0(\xi_\delta)$$

□

Proof of Corollary 3.3.2. From proposition 3.3.2, take expectation of the CV penalty term,

$$E(e'P(\hat{m})e) = E(\sum_{j=1}^{k} \lambda_j \chi^2(1)) + E(J_0(\xi_\delta))$$

We have $E(\sum_{j=1}^{k}\lambda_j\chi^2(1)) = \sum_{j=1}^{k}\lambda_j$. For $E(J_0(\xi_\delta))$, because it depends on the true data generating process which is unknown in practice, following Hansen's approach, we can approximate the value of $E(J_0(\xi_\delta))$ by averaging two extreme cases, so $E(J_0(\xi_\delta)) \approx \frac{1}{2}(\text{tr}(\hat{Q}^{-1}\hat{\Sigma}) + 2\bar{p} - k) \equiv \bar{p}^*$. Then by the same procedure in the proof of corollary 3.3.1, the sample CV criterion is

$$\widehat{\text{CV}}(w) = (w\hat{e} + (1-w)\tilde{e})'(w\hat{e} + (1-w)\tilde{e}) + 2(\text{tr}(\hat{Q}^{-1}\hat{\Sigma}) + w\bar{p}^*)$$

The sample optimal CV weight $\hat{w}$ is the value in $[0,1]$ that minimizes $\widehat{\text{CV}}(w)$, so

$$\hat{w} = 1 - \frac{\text{tr}\left(\hat{Q}^{-1}\hat{\Sigma}\right) + 2\bar{p} - k}{2\left(\sum_{t=1}^{T}\tilde{e}_t^2 - \sum_{t=1}^{T}\hat{e}_t^2\right)}$$

if $(\sum_{t=1}^{T}\tilde{e}_t^2 - \sum_{t=1}^{T}\hat{e}_t^2) \geq \bar{p}^*$ while $\hat{w} = 0$ otherwise. □

BIBLIOGRAPHY

Andrews, D. W. (1991). Asymptotic optimality of generalized cl, cross-validation, and generalized cross-validation in regression with heteroskedastic errors. *Journal of Econometrics*, 47:359–377.

Andrews, D. W. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica*, 61(04):821–856.

Andrews, D. W. (2003). End-of-sample instability tests. *Econometrica*, 71(06):1661–1694.

Andrews, D. W. and Ploberger, W. (1994). Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica*, 62(06):1383–1414.

Bai, J. (1997). Estimating multiple breaks one at a time. *Econometric Theory*, 13:315–352.

Bai, J. (1999). Likelihood ratio tests for multiple structural changes. *Journal of Econometrics*, pages 299–323.

Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(01):47–78.

Bates, J. and Granger, C. (1969). The combination of forecasts. *Operational Research Quarterly*, 20:451–468.

Bunzel, H. and Calhoun, G. (2012). Cross-validation as a tool for inference under instability.

Calhoun, G. (2013). An asymptotically normal out-of-sample test of equal predictive accuracy for nested models.

Calhoun, G. (2014). Out-of-sample comparisons of overfit models.

Campbell, J. Y. and Thompson, S. B. (2008). Predicting excess stock returns out of sample: can anything beat the historical average? *Review of Financial Studies*, 21(04):1509–1531.

Cheng, X. and Hansen, B. E. (2013). Forecasting with factor-augmented regression: a frequentist model averaging approach.

Clark, T. E. and McCracken, M. W. (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, 105:85–110.

Clark, T. E. and McCracken, M. W. (2005). The power of tests of predictive ability in the presence of structural breaks. *Journal of Econometrics*, 124:1–31.

Clark, T. E. and McCracken, M. W. (2010). Averaging forecasts from vars with uncertain instabilities. *Journal of Applied Econometrics*, 25(01):5–29.

Clark, T. E. and McCracken, M. W. (2011). Averaging forecasts from vars with uncertain instabilities.

Clark, T. E. and McCracken, M. W. (2013). Advances in forecast evaluation. *Handbook of Economic Forecasting*, 2:1107–1201.

Clark, T. E. and West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138:291–311.

Davidson, J. (1994). *Stochastic Limit Theory: An Introduction for Econometricians*. Oxford University Press.

Elliott, G. (2011a). Averaging and the optimal combination of forecasts.

Elliott, G. (2011b). Forecast combination when outcomes are difficult to predict.

Elliott, G. and Muller, U. K. (2006). Efficient tests for general persistent time variation in regression coefficients. *Review of Economics Studies*, 73:907–940.

Giacomini, R. and Rossi, B. (2009). Detecting and predicting forecast breakdowns. *Review of Economic Studies*, 76(02):669–705.

Giacomini, R. and Rossi, B. (2010). Model comparisons in unstable environments.

Goyal, A. and Welch, I. (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, 21(04):1455–1508.

Hansen, B. E. (2000). Testing for structural change in conditional models. *Journal of Econometrics*, 97:93–115.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica*, 75(04):1175–1189.

Hansen, B. E. (2008). Least-squares forecast averaging. *Journal of Econometrics*, 146:342–350.

Hansen, B. E. (2009). Averaging estimators for regressions with a possible structural break. *Econometric Theory*, 25(06):1498–1514.

Hansen, B. E. and Racine, J. S. (2011). Jackknife model averaging. *Journal of Econometrics*.

Inoue, A. and Kilian, L. (2004). In-sample or out-of-sample tests of predictability: which one should we use? *Econometric Review*, 23(04):371–402.

Liu, Q. and Okui, R. (2012). Heteroskedasticity-robust cp model averaging.

McCracken, M. W. (2000). Robust out-of-sample inference. *Journal of Econometrics*, 99:195–223.

McCracken, M. W. (2007). Asymptotics for out of sample tests of granger causality. *Journal of Econometrics*, 140:719–752.

Paye, B. and Timmermann, A. (2006). Instability of return prediction models? *Journal of Empirical Finance*, 13(03):274–315.

Pesaran, M., Pick, A., and Pranovich, M. (2011). Optimal forecasts in the presence of structural breaks.

Pesaran, M., Pick, A., and Pranovich, M. (2013). Optimal forecasts in the presence of structural breaks. *Journal of Econometrics*, 177(02):134–152.

Pesaran, M. H. and Timmermann, A. (2007). Selection of estimation window in the presence of breaks. *Journal of Econometrics*, 137:134–161.

Rapach, D., Strauss, J., and Zhou, G. (2010). Out-of-sample equity premium prediction: combination forecasts and links to the real economy. *Review of Financial Studies*, 23:821–862.

Rapach, D. E. and Wohar, M. E. (2006). Structural breaks and predictive regression models of aggregate u.s. stock returns. *Journal of Financial Econometrics*, 4(02):238–274.

Rapach, D. E. and Zhou, G. (2013). Forecasting stock returns. *Handbook of Economic Forecasting*, 2:328–383.

Ravishanker, N. and Dipak, K. (2001). *A First Course in Linear Model Theory*. Chapman and Hall–CRC.

Rossi, B. (2005). Optimal tests for nested model selection with underlying parameter instability. *Econometric Theory*, 21:962–990.

Rossi, B. (2013). Advances in forecasting under instability. *Handbook of Economic Forecasting*, 2:1203–1324.

Stock, J. H. (2004). Structural stability and models of the business cycle. *De Economist*, 152:197–209.

Stock, J. H. and Watson, M. W. (2003). Forecasting output and inflation: the role of asset prices. *Journal of Economic Literature*, 41:788–829.

Timmermann, A. (2006). Forecast combinations. *Handbook of Economic Forecasting*, 1:135–196.

West, K. D. (2006). Forecast evaluation. *Handbook of Economic Forecasting*, 1:99–134.

Zeileis, A., Kleiber, W., Kramer, W., and Hornik, K. (2003). Testing and dating of structural changes in practice. *Computational Statistics and Data Analysis*, 44:109–123.