Out-of-Sample Forecast Model Averaging with Parameter Instability

Anwen Yin*
Iowa State University
Ames, Iowa

October 9, 2014

Abstract

This paper extends Hansen's (2009) optimal weights combining the stable model and the structural break model to the out-of-sample (**OOS**) forecast setting. We propose optimal weights based on the leave-one-out cross-validation criterion (**CV**) to replace Hansen's weights, as CV can be applied more generally compared with Mallows' criterion (**Cp**), especially in that it is robust to heteroscedasticity which is relevant to many economic time series. These weights are optimal in the sense of minimizing the sample cross-validation criterion under this setting. We provide Monte Carlo evidence showing that CV weights outperform several other methods (i.e. Mallows' weights, equal weights, and Schwarz-Bayesian weights) in several designs. Last, we apply our method to forecasting the U.S. and Taiwan quarterly GDP growth rates out-of-sample and demonstrate the improved performance of our method.

Keywords: *Cross-Validation, Conditional Heteroscedasticity, Structural Break, Out-of-Sample, Forecast Evaluation, Forecast Combination*

JEL Classification: C22, C52, C53

*The author is grateful to Richard Ashley, Joydeep Bhattacharya, Helle Bunzel, Gray Calhoun, David Frankel, Ramazan Gencay, Jonathan McFadden, Jarad Niemi, Dan Nordman, T.J. Rakitan, David Rapach, John Schroeter, Yohei Yamamoto, Guofu Zhou and econometrics session participants in the 2014 Missouri Economics Conference, 2014 Midwest Econometric Group meeting and 31st Canadian Econometric Study Group meeting for helpful comments and suggestions. All errors are mine. Contact: 165 Heady Hall, Department of Economics, Iowa State University, Ames, Iowa, 50011. Telephone: (515) 294-3512. Email: anwen@iastate.edu

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 boost 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 forecasting 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 [45] [44] 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 arises when the magnitude of the break is small or the evidence of parameter instability is not convincing. In this case, the pre-test model, 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 break 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 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 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

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

[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 forecasts better than the true DGP evaluated by root mean squared forecast error on average.

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 [47] for a comprehensive survey of forecast combination.

In an empirical paper studying the U.S. aggregate equity market returns, Rapach, Strauss and Guo [39] 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's [13] 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 break.

With these potential benefits mentioned above, 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 perform better than various optimal combination weights proposed by researchers, notably the Granger–Ramanathan combination. A paper attempting to explain this puzzle is written by Elliott [20]. 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 these benefits and drawbacks mentioned above in mind, in this pa-

[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 [40] [36] related to dating structural breaks based on different data frequencies and models.

per, 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 structural breaks 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. Our paper adapts Hansen's Mallows model averaging method [27] to the study of out-of-sample forecasting with breaks. Specifically, we propose model averaging weights derived from the cross–validation information criterion to combine the break model and the stable model.

The cross–validation information criterion is an unbiased estimate of the mean squared forecast error or the expected test error rate in the language of statistical learning [29], 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 model. Under the assumption of conditional homoscedasticity, we show that the cross–validation model averaging weights are asymptotically equivalent to Hansen's weights. A natural extension is to relax this homoscedastic error assumption as it may be too strict for many 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 2 provides related literature review. Section 3 first describes the econometric model and the forecasting problem, then presents theoretical results for the model averaging weights.

Section 4 presents Monte Carlo evidence. Section 5 provides two empirical examples comparing our method with others. Section 6 concludes.

2 Related Literature

This paper relies on the literature related to information criterion-based model selection and averaging, structural breaks testing, heteroscedasticity and autocorrelation consistent/robust (**HAC/HAR**) covariance matrix estimation, and out-of-sample forecast comparison and forecast evaluation.

Recently, Hansen has published a series of papers [25] [26] [27] [28] 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 homoskedasticity and the restriction of weight discretization, model averaging estimators based on Mallows' criterion are asymptotically optimal in the sense of minimizing mean squared error 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 later extends his Mallows' model averaging theory to forecast combination and compares its performance with other related combination methods based on simulated data [26]. 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 optimality to hold, we need the data to be independent and identically distributed. This restriction has made the optimality property less relevant for many empirical applications where the data under study is time series, for example, GDP growth rate, unemployment rate and inflation rate. Even more stringently, Hansen imposes the restriction that the models un-

der consideration are strictly nested[4] in order to ensure optimality. In another paper coauthored with Racine [28], Hansen relaxes 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 optimality is still restricted to random samples and does not allow arrival of new models for consideration. Why cross-validation? Comparing Mallows' criterion with CV, Andrews [1] 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. Liu and Okui [32] propose a heteroscedasticity-robust Mallows' criterion which generalizes Hansen's least squares model averaging optimality results by allowing for conditionally heteroscedastic errors. One major drawback of the model averaging methods discussed above is that none of the optimality results applies to the setting where structural breaks are possible or nonlinear models are present.

Model averaging under parameter instability relies heavily on the theory of break detection, break dates estimation and inference. Historically, applied econometricians rely on the Chow test to detect 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. In many situations this seems quite unrealistic and requires that econometricians visually examine the time series data to search for a possible break point. In a seminal paper, Andrews [3] has shown the non-standard asymptotic distribution of a class of Sup-type test statistic for detecting breaks and conducting inference when the break point is unknown. This results in a subsequent series of articles related to break detection and optimal testing, notably Andrews [5] [4], Hansen [24], Bai [6] [7], Bai and Perron [8], Elliott and Muller [21], Rossi [42] and Bunzel and Iglesias [10]. Bai and Perron [8] generalize Andrews'

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

test by introducing methods to detect and date multiple change points in a partial structural break linear model, as well as develop relevant asymptotic theory for conducting inference. 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.

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. There are several papers related to forecasting with breaks. See Pesaran and Timmermann [37], Pesaran, Pick and Pranovich [38]. The focus of this string of literature is on the forecasting window choice. Specifically, what portion of the data to use. The selection of forecasting window involves a bias-variance trade-off. Including more data before the estimated break date may help reduce the mean squared forecast error, but doing so could result in more bias in the parameter estimation.

As an alternative to the forecasting model averaging method studied in this paper when parameter instability is possible, several researchers have proposed various in–sample and out–of–sample tests to select a forecasting model which is robust to structural breaks. See Giacomini and Rossi [22] [23], Bunzel and Calhoun [9] and Inoue and Kilian [30].

Since allowing for heteroscedasticity requires the use of HAC or HAR estimators of the covariance matrix, various methods regarding the choice of kernel and bandwidth for estimating covariance matrix can be considered. See Newey and West [35], Andrews [2], Kiefer, Vogelsang and Bunzel [31] and Sun [46].

3 Econometric Theory

3.1 Model and Estimation

The econometric model used to forecast and its estimation method are closely related to Hansen [27] and Andrews [3].[5] 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[6] 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. Parameters are estimated by ordinary least squares. The forecasting model allowing for structural break is:

$$y_t = x_t'\beta_1 I_{[t<m]} + x_t'\beta_2 I_{[t\geq m]} + e_t \tag{1}$$

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 for 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:

$$y_t = x_t'\beta + e_t \tag{2}$$

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

[6] 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.

The traditional pre-test procedure starts with performing a test for structural breaks[7], either by using Andrews' SupF or SupW test, or Bai and Perron's multiple-break test, and then decide to keep the stable or unstable model.

As an alternative to model selection, we can combine these two models by assigning weight w to model 1 and $1-w$ to model 2, where $w \geq 0$. So the combined predictive model is

$$y_t = w\left\{x_t'\beta_1 I_{[t<m]} + x_t'\beta_2 I_{[t\geq m]}\right\} + (1-w)\left\{x_t'\beta\right\} + e_t \tag{3}$$

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.

3.2 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. The quantity measuring in-sample fit are the same for most criteria, but they differ in the degree of penalization. For instance, AIC penalizes each additional parameter by 2 while SIC penalizes overfitting by the logarithm of 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

[7]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 [3], Bai and Perron [8] and Elliot and Muller [21].

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 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 one-step ahead predictive model selection and is shown robust to 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 break.

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

$$CV_T(k) = \frac{1}{T}\sum_{t=1}^{T}\tilde{e}_t(k)^2 \tag{4}$$

where $\tilde{e}_t(k) = y_t - \tilde{\beta}_{-t}(k)'x_t(k)$ are the residuals from the regression with the t^{th} observation dropped and $\tilde{\beta}_{-t}(k) = (\sum_{i\neq t} x_i(k)x_i(k)')^{-1}(\sum_{i\neq t} x_i(k)y_i)$ is the associated vector of parameter estimates. Intuitively, this procedure is trying to estimate the expected test error rate based on the model training data. Though equation 4 implies that we need to run regression T times for given sample size T, luckily, for linear models, we can calculate 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 4 as

$$CV_T(k) = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\hat{e}_t(k)}{1-h_t}\right)^2 \tag{5}$$

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

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 says 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 adopted in applications of detecting and dating breaks, see Rapach and Wohar [40] and Paye and Timmermann [36].

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.

Assumption 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 [19] for more details on advanced asymptotic theory. Assumption 1.6 says that the error term is conditionally heteroscedastic which is less restrictive.

To obtain model weights, first, we need to show what the cross-validation cri-

terion looks like under the above assumptions. We know that the information criterion usually consists of two parts, one measuring in-sample fit while the other penalizing overfitting. For model averaging, the penalty term is crucial in determining the optimal weights. The proofs of all theoretical results are provided in the appendix.

Proposition 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,* $E(CV(T)) \xrightarrow{p} E(Cp(T))$.

The intuition for this result is that since the cross-validation criterion is robust to heteroscedasticity compared with Mallows' criterion, so when conditional heteroscedasticity is absent, we would not expect any significant difference between CV and Cp. With this result in hand, we can obtain the feasible sample optimal weight for the break model:

Corollary 3.1. *The feasible sample CV weight for the break model is:*

$$\hat{w} = \frac{(T-2k)(\sum_{t=1}^{T}\tilde{e}_t^2 - \sum_{t=1}^{T}\hat{e}_t^2) - \bar{p}\sum_{t=1}^{T}\hat{e}_t^2}{(T-2k)(\sum_{t=1}^{T}\tilde{e}_t^2 - \sum_{t=1}^{T}\hat{e}_t^2)} \tag{6}$$

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.*

By proposition 3.1, Hansen's weights [27] still apply in this case. Following Hansen, the population penalty term in the cross-validation criterion involves a distribution which is a function of the true data generating process. To obtain the feasible sample optimal weights, the sample CV penalty term is approximated by

averaging two extreme cases [8], so that is how Hansen's $\bar{p}$ value enters the formula for the break model weight.

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 [1], our next proposition establishes the asymptotic distribution of the CV penalty term in the presence of conditional heteroscedasticity.

Proposition 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 the Brownian bridge,*

$$e'P(\hat{m})e \xrightarrow{d} \sum_{j=1}^{k} \lambda_j \chi^2(1) + J_0(\xi_\delta) \tag{7}$$

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.

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 3.2. 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 7 reflexes the impact of conditional heteroscedasticity while the second term deals with 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 $Q = E(x_t x_t')$.

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

Empirically, Σ can be estimated by HAC estimators and Q can be consistently estimated by its sample analogue $\frac{1}{T}\sum_{t=1}^{T} x_t x_t'$. The feasible sample optimal weight for the break model in this case is:

Corollary 3.2. *The feasible sample optimal weight for the break model in the presence of conditional heteroscedasticity takes the form:*

$$\hat{w} = 1 - \frac{\operatorname{tr}\left(\hat{Q}^{-1}\hat{\Sigma}\right) + 2\bar{p} - k}{2\left(\sum_{t=1}^{T}\tilde{e}_t^2 - \sum_{t=1}^{T}\hat{e}_t^2\right)} \tag{8}$$

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* $\hat{Q}^{-1}\hat{\Sigma}$*,* $\bar{p}^* = \frac{1}{2}(\operatorname{tr}(\hat{Q}^{-1}\hat{\Sigma}) + 2\bar{p} - k)$.

Again, $\bar{p}$ comes from averaging two extreme cases approximating the infeasible expectation of the population penalty term.

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

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 use three different designs of the true data generating process: (i). an AR(2) process plus five exogenous predictors with ARCH(1) errors,

$$y_t = \mu + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \sum_{i=1}^{5}\theta_i x_i + e_t \tag{9a}$$

$$e_t = v_t\sqrt{h_t} \tag{9b}$$

$$h_t = \alpha_0 + \alpha_1 e_{t-1}^2 \tag{9c}$$

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

$$y_t = \mu + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \sum_{i=1}^{2} \theta_i x_i + e_t \tag{10}$$

(iii). an AR(2) process with a break in the variance of the error term. We consider this design to study the forecasting performance of CV model averaging in the Great Moderation type environment. The break date of the error term variance is not identical to that of the conditional mean[9]. We allow for this break date difference hoping to better approximate the environment forecasters face in practice. Mathematically, the data generating process for design (iii) is the following:

$$y_t = \mu + \rho_1 y_{t-1} + \rho_2 y_{t-2} + e_t \tag{11}$$

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 happening at the $30\%th$ observation of the training sample R, that is, $\tau = 0.3$. We let the structural break of the parameters take the multiplicative form, that is, if the pre-break coefficient is β, the post-break parameter becomes $\delta\beta$, where δ controls 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 $x_5 \sim \text{Binomial}(1, 0.02)$. The parameter values for all data generating processes are $\mu = 2, \rho_1 = 0.4, \rho_2 = 0.2, \theta_1 = 0.8, \theta_2 = -0.4, \theta_3 = 2, \theta_4 = -3.5, \theta_5 = 10,$

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

$\alpha_0 = 1, \alpha_1 = 0.4$. These values are chosen to satisfy the stationarity and ARCH error regularity assumptions. Note that in our simulation, the post-break coefficient values become smaller than their pre-break counterparts. 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, to ensure regime-wise stationarity, δ should not take values greater than 1.1 if we prefer larger post-break parameter value[10].

To make our simulation design close to relevant empirical studies and to capture the difficulty many researchers face in finding the best approximating models, the model used to forecast differs from the true data generating process[11]: in case (i) the model to forecast is based those five exogenous predictors in the DGP, $y_t = \mu + \sum_{i=1}^{5} \theta_i x_i + e_t$; 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 the AR(1) with intercept, $y_t = \mu + \rho_1 y_{t-1} + e_t$.

For each case, we evaluate the out-of-sample (OOS) forecasting performance by comparing the 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. Out-of-sample forecast is 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 sample of size P. Under the recursive window, at each point in time, the estimated parameter is 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

[10]Bai and Perron [8] 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 as in the example is not large enough to identify especially when the data is highly volatile as shown in our simulation work.

[11]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 or practitioners tend to select a model of small order.

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 evaluate our forecasts according to some loss function, for example, RMSFE or MSFE. See Calhoun [11] [12], McCracken [33] [34], Rossi [43], Clark and McCracken [14] [15] [17], Clark and West [18] and West [48] for more details on out-of-sample forecasting.

The total sample size, T, is 200[12]. In our pseudo one-step ahead out-of-sample forecasting simulation, we reserve the first 170 or 150 ($R = 170$ or $R = 150$) observations as the training sample and the rest as the prediction sample ($P = 30$ or $P = 50$). For the break model, we use the post-break window method to forecast out-of-sample. Other techniques, such as the optimal window method proposed by Pesaran and Timmermann [37] or the robust weight method proposed by Pesaran, Pick and Pranovich [38] could also be considered. For simplicity, we only apply the post-break window method in this paper [13].

In each case, to evaluate and compare performance, we produce forecasts using six methods [14]: (i) Mallows' model averaging (**Cp**); (ii) CV model averaging (**CV**); (iii) Bayesian model averaging[15] (**SIC**); (iv) stable model (**Stable**); (v) break model (**Break**); and (vi) equal weights[16] (**Equal**). We evaluate their forecast performance by root mean-square forecast error (**RMSFE**). For ease of

[12]This sample size is chosen to be relevant to most macroeconomic time series.

[13]Currently, researchers are sill working on developing theory and methods related to forecasting with breaks, and we are not aware of any dominant method that performs well in most situations faced by practitioners. 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.

[14]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 [16] 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.

[15]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 beak model is $w_b = \exp(SIC^b)/(\exp(SIC^b) + \exp(SIC^s))$

[16]Each model receives weight of 0.5.

Table 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 $y_t = \mu + \sum_{i=1}^{5} \theta_i x_i + e_t$. P is the evaluation sample size, total sample size is 200, break fraction relative to the training sample is $\tau = 0.3$, 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 break. Break: model with a full structural break.

comparison, we pick the equal weight method as the benchmark[17] 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 better performance than equal weights. The smaller the ratio is, the better the forecasting performance is for given sample split.

4.1 Design I

Simulation results for the ARCH[18] error design are presented in table 1. We can see from the table that CV forecasts better than Cp across all considered break sizes and prediction sample sizes. Both of CV and Cp's relative RMSFE decreases monotonically as the break size increases, but CV decreases at a slightly faster speed. Bayesian weighting does slightly worse than the equal weight method, but its performance deteriorates when the break size becomes large as it fails to capture the fact that the evidence supporting break is becoming stronger. It should not be a surprise that the break model does well since structural break indeed happens

[17] The reason to pick equal weights as the benchmark is because of the aforementioned forecast combination puzzle: equally weighted forecasts tend to perform better than other complicated methods in many applications. The puzzle is generally discussed in the forecasting literature without allowing breaks. In this study we try to examine whether it dominates our method when facing structural breaks.

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

Table 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 $\tau = 0.3$, 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 break. Break: model with a full structural break.

in the DGP, but it performs slightly worse than CV because it does not take the volatility into account.

Our results indicate that when there is ARCH type conditional heteroscedasticity in the data and when the break is not strictly dominated by the volatility, CV forecasts better than Mallows' model averaging. Additionally, CV forecasts better than equal weight so the forecast combination puzzle does not apply in this design. Bayesian model averaging is approximately equivalent to equal weight, but it is less sensitive to the size of break. Compared with CV, Bayesian weighting method does not put more weight on the proper model even the break size increases significantly.

It is worth mentioning that in our design, the post-break coefficients become smaller than their pre-break counterparts by various degrees. This procedure is adopted to ensure that piece-wise stationarity is maintained under structural break. [19]

4.2 Design II

Simulation results for the second design are shown in table 2. In this case we can see from the table that CV forecasts much better than Cp across all break sizes and prediction sample sizes. Both of their relative RMSFE decreases monotonically as the break size increases, but in this design the RMSFE of CV decreases at a much faster speed. Bayesian weighting does almost the same as equal weight, but its performance deteriorates when the break size becomes large as shown in the previous design. The pattern of our results holds for both cases of the out-of-sample size.

Our results indicate that when there is this "wild" type heteroscedasticity in the data as modeled in the DGP and when the break is not strictly dominated by the volatility, CV forecasts significantly better than Mallows' model averaging, especially when the break size is large. Additionally, CV forecasts better than equal weight so the forecast combination puzzle does not apply in this design. Bayesian model averaging is approximately equivalent to equal weight. Again, compared with CV, Bayesian weighting method does not put more weight on the proper model when the break size increases.

4.3 Design III

Simulation results for this Great Moderation type design are shown in table 3. The general pattern shown in the previous two designs remains in this case. CV forecasts better than Cp across all listed break sizes and prediction sample sizes. Both of their relative RMSFE decreases monotonically as the break size increases, but the relative RMSFE of CV decreases at a slightly faster speed. Bayesian weighting does almost the same as equal weight, but its performance is less sensitive to the break size in this case. The pattern of our results holds for both cases

[19]The simulation results stay the same if we reverse the break size direction by starting with smaller pre-break parameter values.

Table 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 $\tau = 0.3$, 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 break. Break: model with a full structural break.

of the out-of-sample size.

In this design our results indicate that when the data has the Great Moderation appearance as shown in the U.S. GDP growth data, and when the break is not strictly dominated by the volatility, CV forecasts better than Mallows' model averaging. Additionally, the forecast combination puzzle does not apply here. This motivates us to apply our method to forecasting real U.S. GDP growth rate and to comparing its performance with other related methods.

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. These two series are shown in figure 1. The U.S. data is obtained from the Bureau of Economic Analysis[20]. The data for Taiwan is from National Statistics[21]. The U.S. series is shown in the blue colored solid line while the Taiwan series is denoted in red colored dash line. Note that the data for Taiwan is shorter than that of the the U.S. This is because Taiwan officially starts its modernization with American aid in the early 1950s.

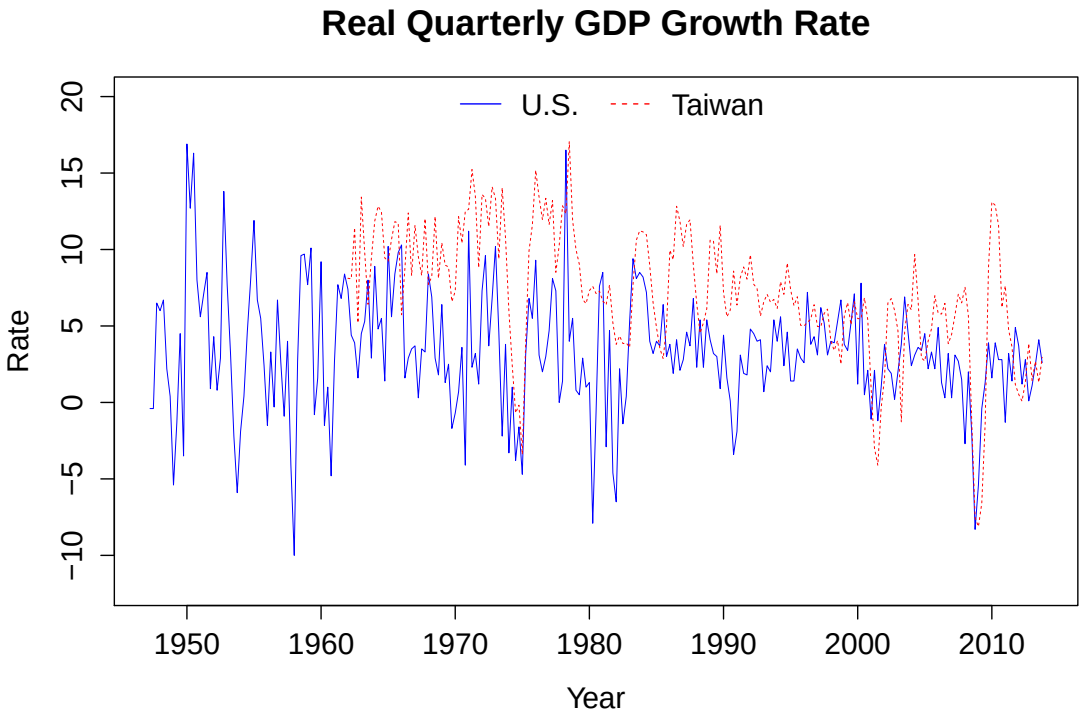
[20] http://www.bea.gov/

[21] http://eng.stat.gov.tw/

For the U.S. series, we can see that the growth rate becomes less volatile toward the end of the sample. This is the so called Great Moderation phenomenon, see Stock [44] and Stock and Watson [45]. Stock and Watson argue that the forecasting relationship for the GDP growth 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.*" In the next section, we demonstrate that our theory based CV model averaging method performs better than Mallows' weight, Bayesian weight, and most importantly, the equal weight method in terms of smaller root MSFE.

For the Taiwan series, we can see that it has two interesting features. First, it looks like that the average growth rate has dropped toward the end of the sample. This could be explained by the economic growth theory in Macroeconomics that during the early period of modernization or industrialization, a country tends to have high GDP growth rate. But as time goes, the growth rate approaches to the lower equilibrium rate. Second, it seems like 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 as they exhibit the similar Great Moderation pattern shown in the U.S. data, for example, Canada and Germany. This motivates us to include Taiwan as an additional empirical example to evaluate our method.

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



5.1 Forecast U.S. GDP Growth

Here we apply our method to forecasting U.S. GDP growth rate[22] out-of-sample and compare its performance with others. We have quarterly data from 1960:1 to 2012:1, 209 total observations. The variable we are interested in forecasting is the U.S. quarterly GDP growth rate. Predictors included in some models 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)[23].

[22]The data we use for this application are from Bruce Hansen's website:http://www.ssc.wisc.edu/~bhansen/cbc/. Hansen has several exogenous predictors included in his data in addition to the GDP growth data.

[23]In the database, these predictors are given in levels, but we have converted them to rates of change, so the actual sample size in all models except for the AR(1) is 207. The default premium is calculated by the difference between the AAA bond rate and BAA bond rate.

We consider five models, from small to large they are:

$$\Delta \text{GDP}_t = \beta_0 + \beta_1 \Delta \text{GDP}_{t-1} + \epsilon_t \tag{12a}$$

$$\Delta \text{GDP}_t = \beta_0 + \beta_1 \Delta \text{GDP}_{t-1} + \beta_2 \Delta \text{GDP}_{t-2} + \epsilon_t \tag{12b}$$

$$\Delta \text{GDP}_t = \beta_0 + \beta_1 \Delta \text{GDP}_{t-1} + \beta_2 \Delta \text{SR}_{t-1} + \epsilon_t \tag{12c}$$

$$\Delta \text{GDP}_t = \beta_0 + \beta_1 \Delta \text{GDP}_{t-1} + \beta_2 \Delta \text{SR}_{t-1} + \beta_3 \Delta \text{LR}_{t-1} + \epsilon_t \tag{12d}$$

$$\Delta \text{GDP}_t = \beta_0 + \beta_1 \Delta \text{GDP}_{t-1} + \beta_2 \Delta \text{SR}_{t-1} + \beta_3 \Delta \text{LR}_{t-1} + \beta_4 \Delta \text{DP}_{t-1} + \epsilon_t \tag{12e}$$

Since in reality we do not know the "true" econometric model specification, we consider the above five candidate models. For each model, we examine the performance of various model averaging methods. Consistent with what is done in the previous simulation section, for each model we use the recursive window to forecast out-of-sample. To better examine the OOS performance, in each case we generate a sequence of RMSFE by varying the evaluation sample size, from 20 to 50, with increments of 5. Forecast results from all five models are presented in table 4. For simplicity and ease of comparison, following our Monte Carlo simulation we pick the equal weight method as the benchmark and normalize all OOS forecasting performance around 1. If the value of relative RMSFE is below 1, then the OOS forecasts perform better than that of the equal weight method.

We can see that in all five models approximating the DGP, CV forecasts better than SIC, Cp and equal weight under recursive window across all evaluation sample sizes. Additionally, CV is the only method which beats equal weight. The forecast gains of CV relative to equal weight range from about 1% to 6% across evaluation sample sizes and models. CV solves the forecast combination puzzle in this application.

Table 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

5.2 Forecast Taiwan GDP Growth

Here we apply our method to forecasting Taiwan GDP growth rate out-of-sample and compare its performance with others. We have quarterly data from 1962:1 to 2013:4, 208 total observations. The variable we are interested in forecasting is the Taiwan quarterly GDP growth rate. Since we do not have any exogenous predictors available, we only consider two AR forecasting models, namely, the AR(1) model and the AR(2) model. We continue the general setting outlined in the previous application. For each model, we generate a sequence of RMSFE by varying the evaluation sample size, from 20 to 50, with increments of 5. Forecast results from all five models are shown in table 5. For simplicity and ease of comparison, following our Monte Carlo simulation we pick the equal weight method as the benchmark and normalize all OOS forecasting performance around 1. If the value of relative RMSFE is below 1, then the OOS forecasts perform better than that of the equal weight method.

From table 5, we can see that now the Mallows' weights do better than the equal weight in both models and across all evaluation sample sizes. For the AR(1) model, all three methods perform roughly the same but CV has the smallest RMSFE ratio across all evaluation sample sizes. For the AR(2) model, CV performs significantly better than the other two methods.

Table 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.

6 Conclusion

We are interested in answering a basic question of how to forecast a time series variable of interest when there is uncertainly about parameter instability. Specifically, which model should be used for forecasting: the break model or the stable one? If uncertainty is strong and we decide to combine forecasts from two 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 forecasting models. In the literature of model selection, CV is shown to be robust to heteroscedasticity than other information criteria, such as, AIC, BIC and Mallows'. Without assuming conditional heteroscedasticity, we show that CV model averaging is approximately equivalent to Mallows' model averaging. But in many empirical applications related to macroeconomic time series or financial time series, researchers usually can not avoid explicitly dealing with heteroscedasticity for analysis and forecast. This motivates our generalization of the model averaging method by allowing for conditional heteroscedasticity.

Researchers have found that in many applications, equally weighted forecasts perform better than other complex combination methods. This forecast combination puzzle has cast double on the use of complicated model averaging methods. Both CV and Cp weights are easy to compute and do not rely on direct weight estimation as in the Granger-Ramanathan forecast combination. This feature should be appealing to practitioners and professional forecasters because simplicity can help reduce the excess noise introduced by using complex weighting methods. This may help explain why our methods forecast better than equal weight in both controlled simulation and in forecasting U.S. and Taiwan quarterly GDP growth rates out-of-sample.

A Proof

Proof of Proposition 3.1. From the cross-validation criterion, for linear models we have the well-known result that

$$\frac{1}{T}\sum_{i=1}^{T}\tilde{e_t}^2 = \frac{1}{T}\sum_{i=1}^{T}\frac{\tilde{e_t}^2}{(1-h_t)^2}$$

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

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

Under regularity conditions listed in Assumption 1, we have $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$, and 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 3.1. Since CV is asymptotically equivalent to Mallows' Cp, following Hansen's [27] proof, write the sample CV criterion for the weighted model as a function of the break model weight w,

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

where $\bar{p}$ proposed by Hansen is used to approximate the infeasible expected value of the population penalty term. The CV weight is the value in $[0, 1]$ that minimizes $\mathrm{CV}(w)$, 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 3.2. The proof of this proposition is adapted from Hansen [27]. 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 the meat part, $X^*(m)'X^*(m)$, 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 continuous mapping theorem we have

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

For the bread part, $X^*(m)'e = X(m) - X(X'X)^{-1}X(m)'X(m))'e$, 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

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

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

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

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

where $W(1)$ is the 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 the Brownian vector at time π.

Putting together 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. Combined with Hansen's [27] theorem 1 without assuming conditional homoscedasticity or Andrews' [3] theorem 4, we have $\frac{1}{T} e' P^*(m) e \Rightarrow J_0(\xi_\delta)$.

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 relevant 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

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

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

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 [41]), we have

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

where λ_js are the eigenvalues of the matrix $Q^{-1}\Sigma$, $\chi^2(1)$ is a random variable having the χ^2 distribution with degree of freedom one.

Collecting all 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 3.2. From proposition 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 $E(\sum_{j=1}^{k} \lambda_j \chi^2(1)) = \sum_{j=1}^{k} \lambda_j$, applying Hansen's technique, approximate the value of $E(J_0(\xi_\delta))$ by averaging two extreme cases, so $E(J_0(\xi_\delta)) \approx \frac{1}{2}(\mathrm{tr}(\hat{Q}^{-1}\hat{\Sigma}) +$

$2\bar{p} - k) \equiv \bar{p}^*$. Then by the same procedure in the proof of corollary 3.2,

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

The CV weight is the value in $[0, 1]$ that minimizes $\mathrm{CV}(w)$, so

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

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. □

References

[1] Donald W.K Andrews. Asymptotic optimality of generalized cl,cross-validation, and generalized cross-validation in regression with heteroskedastic errors. *Journal of Econometrics*, 47:359–377, 1991.

[2] Donald W.K Andrews. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59(03):817–858, 1991.

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

[4] Donald W.K Andrews. End-of-sample instability tests. *Econometrica*, 71(06):1661–1694, 2003.

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

[6] Jushan Bai. Estimating multiple breaks one at a time. *Econometric Theory*, 13:315–352, 1997.

[7] Jushan Bai. Likelihood ratio tests for multiple structural changes. *Journal of Econometrics*, pages 299–323, 1999.

[8] Jushan Bai and Pierre Perron. Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(01):47–78, 1998.

[9] Helle Bunzel and Gray Calhoun. Cross-validation as a tool for inference under instability. 2012.

[10] Helle Bunzel and Emma M Iglesias. Testing for breaks using alternating observations. 2007.

[11] Gray Calhoun. An asymptotically normal out-of-sample test of equal predictive accuracy for nested models. 2013.

[12] Gray Calhoun. Out-of-sample comparisons of overfit models. 2014.

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

[14] Todd E Clark and Michael W McCracken. Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, 105:85–110, 2001.

[15] Todd E Clark and Michael W McCracken. The power of tests of predictive ability in the presence of structural breaks. *Journal of Econometrics*, 124:1–31, 2005.

[16] Todd E Clark and Michael W McCracken. Averaging forecasts from vars with unvertain instabilities. 2011.

[17] Todd E Clark and Michael W McCracken. Advances in forecast evaluation. *Handbook of Economic Forecasting*, 2:1107–1201, 2013.

[18] Todd E Clark and Kenneth D West. Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138:291–311, 2007.

[19] James Davidson. *Stochastic Limit Theory: An Introduction for Econometricians*. Oxford University Press, 1994.

[20] Graham Elliott. Forecast combination when outcomes are difficult to predict. 2011.

[21] Graham Elliott and Ulrich K Muller. Efficient tests for general persistent time variation in regression coefficients. *Review of Economics Studies*, 73:907–940, 2006.

[22] Raffaella Giacomini and Barbara Rossi. Forecast comparisons in unstable environments. 2008.

[23] Raffaella Giacomini and Barbara Rossi. Model comparisons in unstable environments. 2010.

[24] Bruce E Hansen. Testing for structural change in conditional models. *Journal of Econometrics*, 97:93–115, 2000.

[25] Bruce E Hansen. Least squares model averaging. *Econometrica*, 75(04):1175–1189, 2007.

[26] Bruce E Hansen. Least-squares forecast averaging. *Journal of Econometrics*, 146:342–350, 2008.

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

[28] Bruce E Hansen and Jeffrey S Racine. Jackknife model averaging. *Journal of Econometrics*, 2011.

[29] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2 edition, 2009.

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

[31] Nicholas M Kiefer, Timothy J Volgosang, and Helle Bunzel. Simple robust testing of regression hypothesis. *Econometrica*, 68(03):695–714, 2000.

[32] Qingfeng Liu and Ryo Okui. Heteroskedasticity-robust cp model averaging. 2012.

[33] Michael W McCracken. Robust out-of-sample inference. *Journal of Econometrics*, 99:195–223, 2000.

[34] Michael W McCracken. Asymptotics for out of sample tests of granger causality. *Journal of Econometrics*, 140:719–752, 2007.

[35] Whitney K Newey and Kenneth D West. A simple positive semi-definite heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(03):703–708, 1987.

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

[37] M. H Pesaran and Allan Timmermann. Selection of estimation window in the presence of breaks. *Journal of Econometrics*, 137:134–161, 2007.

[38] M.H Pesaran, A Pick, and M Pranovich. Optimal forecasts in the presence of structural breaks. 2011.

[39] David Rapach, Jack Strauss, and Guofu Zhou. Out-of-sample equity premium prediction: combination forecasts and links to the real economy. *Review of Financial Studies*, 23:821–862, 2010.

[40] David E Rapach and Mark E Wohar. Strictural breaks and predictive regression models of aggregate u.s. stock returns. *Journal of Financial Econometrics*, 4(02):238–274, 2006.

[41] Nalini Ravishanker and K.Dey Dipak. *A First Course in Linear Model Theory*. Chapman and Hall–CRC, 2001.

[42] Barbara Rossi. Optimal tests for nested model selection with underlying parameter instability. *Econometric Theory*, 21:962–990, 2005.

[43] Barbara Rossi. Advances in forecasting under instability. *Handbook of Economic Forecasting*, 2:1203–1324, 2013.

[44] James H Stock. Structural stability and models of the business cycle. *De Economist*, 152:197–209, 2004.

[45] James H Stock and Mark W Watson. Forecasting output and inflation: the role of asset prices. *Journal of Economic Literature*, 41:788–829, 2003.

[46] Yixiao Sun. Let's fix it: fixed-b asymptotics versus small-b asymptotics in heteroscedasticity and autocorrelation robust inference. 2010.

[47] Allan Timmermann. Forecast combinations. *Handbook of Economic Forecasting*, 1:135–196, 2006.

[48] Kenneth D West. Forecast evaluation. *Handbook of Economic Forecasting*, 1:99–134, 2006.