Out-of-Sample Forecast Model Averaging with Parameter Instability

Anwen Yin*
Iowa State University
Ames, Iowa

November 9, 2014

Abstract

Forecasting economic variables of interest when structural breaks are possible is highly relevant to many economic applications. As an alternative to model selection by hypothesis testing, model averaging or forecast combination provides possible predictive gains by better managing the model selection risk. This paper extends Hansen's (2009) model averaging theory under uncertainty regarding structural breaks to the out-of-sample forecast setting. We propose 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. We present the theoretical form that the CV penalty term takes in the setting where both parameter instability and conditional heteroscedasticity are present, then derive the sample optimal weights based on the cross-validation criterion. We provide Monte Carlo evidence showing that CV weights outperform competing methods (i.e. Mallows' weights, equal weights, and Schwarz-Bayesian weights) in several simulation designs. Last, we apply the CV weights to forecasting the U.S. and Taiwan quarterly GDP growth rates out-of-sample, and demonstrate their better performance compared with other methods.

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, the 2014 Midwest Econometric Group meeting and the 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 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 [42] [41] 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 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 outperforms 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 [43] for a comprehensive survey of forecast combination.

In an empirical paper studying the U.S. aggregate equity market returns, Rapach, Strauss and Guo [36] 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 [12] 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.

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 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 [19]. 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 [37] [33] 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 [26] 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 [28], 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. 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 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 and out-of-sample forecast comparison and forecast evaluation.

Recently, Hansen has published a series of papers [24] [25] [26] [27] 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 then extends his Mallows' model averaging theory to forecast combination and compares its performance with other related combination methods based on simulated data [25]. 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 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.[4] 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 [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. Following earlier research, in another paper coauthored with Racine [27], 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 asymptotic optimality property is still restricted to random samples. Alternatively, Liu and Okui [30] 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 possible break point. To take the impact of unknown break date into account, in a seminal paper, Andrews [2] proposes a SupW type test statistic for detecting breaks and presents the associated asymptotic distribution for the test statistic. Note that

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

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.[5] Subsequently, Bai [5] [6] and Bai and Perron [7] 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, notably Andrews [4] [3], Hansen [23], Elliott and Muller [20], Rossi [39] and Bunzel and Iglesias [9].

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. See Pesaran and Timmermann [34] and Pesaran, Pick and Pranovich [35].

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. See Giacomini and Rossi [21] [22], Bunzel and Calhoun [8] and Inoue and Kilian [29].

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

3 Econometric Theory

3.1 Model and Estimation

The econometric model used to forecast and its estimation method are closely related to Hansen [26] and Andrews [2].[6] 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.[7] 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 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:

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

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

[7] 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, for example, using Andrews' SupF or SupW test, and then decide which model to choose based on testing results.[8]

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 ($\mathbf{AIC^c}$), Schwarz Bayesian information criterion (**SIC**), Hannan-Quinn (**HQ**) and Mallows' $\mathrm{C_p}$ ($\mathbf{C_p}$). 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 many 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

[8] 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 [2], Bai and Perron [7] and Elliot and Muller [20].

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 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:

$$\widehat{CV}_T(m) = \frac{1}{T}\sum_{t=1}^{T} \tilde{e}_t(m)^2 \tag{4}$$

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 6 implies that we need to run 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 6 as

$$\widehat{CV}_T(m) = \frac{1}{T}\sum_{t=1}^{T} \left(\frac{\hat{e}_t(m)}{1-h_t}\right)^2 \tag{5}$$

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

4. $T^{-1/2}\sum_{t=1}^{[\pi T]} x_te_t \Rightarrow W(\pi)$ *where* $W(\pi)$ *is a* $k\times1$ *Wiener process with symmetric, positive definite long-run covariance matrix* $\Sigma \equiv \lim_{T\to\infty} \text{VAR}(T^{-1/2}\sum_{t=1}^{[\pi T]} x_te_t)$*, for* $0 \leq \pi \leq 1$*. '* $\Rightarrow$ *' denotes the weak convergence of the underlying probability measure as* $T \to \infty$*.*

5. $T^{-1}\sum_{t=1}^{[\pi T]} x_tx_t'$ *converges uniformly to* πQ *for all* $\pi \in [0, 1]$*,* $Q = E(x_tx_t')$ *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 widely adopted in applications of detecting and dating breaks, see Rapach and Wohar [37] and Paye and Timmermann [33].

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

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 3.1. *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, when the conditional heteroscedasticity is absent, we would not expect any significant difference between CV and Cp.

We know that the information criterion usually consists of two parts, one measuring the in-sample fit while the other penalizing overfitting. Specifically, by proposition 3.1, since CV and Cp are asymptotically equivalent, for the CV criterion, we have

$$E(CV(T)) = E(\hat{\sigma}^2) + 2E(e'Pe) \tag{6}$$

In equation 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, as the population optimal weight w can be obtained by minimizing $E(CV(T))$. 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,[9] so that is how the $\bar{p}$ value proposed by Hansen, where $\bar{p} = \frac{1}{2}(E(SupW) + k)$,[10] enters the break model weight in the following corollary.

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

$$\hat{w} = \frac{(T-2k)(\sum_{t=1}^{T}\tilde{e}_t^2 - \sum_{t=1}^{T}\hat{e}_t^2) - \bar{p}\sum_{t=1}^{T}\hat{e}_t^2}{(T-2k)(\sum_{t=1}^{T}\tilde{e}_t^2 - \sum_{t=1}^{T}\hat{e}_t^2)} \tag{7}$$

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 [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,*

$$e'P(\hat{m})e \xrightarrow{d} \sum_{j=1}^{k}\lambda_j\chi^2(1) + J_0(\xi_\delta) \tag{8}$$

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

[10] $E(SupW)$ is the expectation of the SupW statistic in Andrews [2]. Hansen [26] provides the sample $\bar{p}$ value for a range of the number of regressors based on simulation results.

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 8 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 $Q = E(x_t x_t')$. 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 3.2. *The feasible optimal weight minimizing the sample cross-validation criterion for the break model in the presence of conditional heteroscedasticity takes the form:*

$$\hat{w} = 1 - \frac{\operatorname{tr}\left(\hat{Q}^{-1}\hat{\Sigma}\right) + 2\bar{p} - k}{2\left(\sum_{t=1}^{T}\tilde{e}_t^2 - \sum_{t=1}^{T}\hat{e}_t^2\right)} \tag{9}$$

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,* $\mathrm{tr}(\hat{Q}^{-1}\hat{\Sigma})$ *is the trace of the matrix* $\hat{Q}^{-1}\hat{\Sigma}$*,* $\bar{p}^* = \frac{1}{2}(\mathrm{tr}(\hat{Q}^{-1}\hat{\Sigma}) + 2\bar{p} - k)$*.*

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.

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,

$$y_t = \mu + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \sum_{i=1}^{5} \theta_i x_i + e_t \tag{10a}$$

$$e_t = v_t \sqrt{h_t} \tag{10b}$$

$$h_t = \alpha_0 + \alpha_1 e_{t-1}^2 \tag{10c}$$

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

$$y_t = \mu + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \sum_{i=1}^{2} \theta_i x_i + e_t \tag{11}$$

(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.[11] 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:

$$y_t = \mu + \rho_1 y_{t-1} + \rho_2 y_{t-2} + e_t \tag{12}$$

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, $\tau = 0.3$. 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. 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.[12]

[11] 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.

[12] Bai and Perron [7] assume that the break size is large enough in order to be identified and

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:[13] in case (i) the model to forecast is based on 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 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 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 or test 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, 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, RMSFE or MSFE. See Calhoun [10] [11], McCracken [31] [32], Rossi [40], Clark and McCracken [13] [14] [16], Clark and West [17] and West [44] for more details on out-of-sample forecasting.

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.

[13] 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.

The total sample size, T, is 200. To investigate if the choice of 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 [34] or the robust weight method proposed by Pesaran, Pick and Pranovich [35] could also be considered.[14]

In each case, to evaluate and compare performance, we produce forecasts using four methods:[15] (i) Mallows' model averaging (**Cp**); (ii) CV model averaging (**CV**); (iii) Bayesian model averaging[16] (**SIC**); and (iv) equal weights[17] (**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[18] and compute the relative performance (**Ratio**) for each method, for example, $\mathrm{RMSFE}^{\mathrm{CV}}/\mathrm{RMSFE}^{\mathrm{Equal}}$. If the ratio is less than one, it indicates that

[14] 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 empirical works. 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.

[15] 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 [15] 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.

[16] 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 $w_b = \exp(SIC^b)/(\exp(SIC^b) + \exp(SIC^s))$

[17] Each model receives weight of 0.5.

[18] 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.

Table 1: Monte Carlo Simulation: Design I

	$P = 30$			$P = 50$		
Break Size	Cp	CV	SIC	Cp	CV	SIC
100	0.6312	0.6298	1.2987	0.6599	0.6585	1.2849
10	0.6644	0.6627	1.2563	0.6871	0.6854	1.2473
5	0.7085	0.7066	1.2063	0.7289	0.7271	1.2005
3	0.7658	0.7636	1.1517	0.7869	0.7850	1.1454
2	0.8330	0.8308	1.0974	0.8500	0.8483	1.0925

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.

the method under consideration ourperforms 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 error design are reported in table 1.[19] 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 break size increases, but CV moves at a slightly faster speed. On the other hand, 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.

Overall, our results imply that when there is ARCH type conditional heteroscedasticity in the data and when the break impact is not strictly dominated

[19] 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	Cp	CV	SIC
100	0.4610	0.2586	1.0717	0.5706	0.3415	1.0649
10	0.7007	0.5681	1.0393	0.6945	0.5419	1.0392
5	0.8422	0.7700	1.0194	0.8699	0.7946	1.0191
3	0.8978	0.8541	1.0111	0.9135	0.8800	1.0126
2	0.9188	0.8778	1.0082	0.9417	0.9320	1.0074

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.

by that of the volatility, the cross-validation weighting method outperforms Mallows' model averaging. Additionally, CV performs better than 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.

4.2 Design II

Simulation results for the second design are reported in table 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 moves at a much faster speed. Bayesian weighting does almost the same as equal weighting, but its performance deteriorates when the break size becomes large like what we have seen in the previous design. The

Table 3: Monte Carlo Simulation: Design III

	$P = 30$			$P = 50$		
Break Size	Cp	CV	SIC	Cp	CV	SIC
100	0.9810	0.9759	1.0011	0.9825	0.9769	1.0011
10	0.9860	0.9789	1.0006	0.9880	0.9822	1.0006
5	0.9919	0.9850	1.0003	0.9933	0.9868	1.0003
3	0.9977	0.9903	1.0000	0.9975	0.9905	1.0001
2	1.0009	0.9940	0.9999	1.0013	0.9952	0.9999

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.

choice of test sample size does not seem to have any significant impact on any weighting methods or pre-test 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 averaging, 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.

4.3 Design III

Simulation results for this Great Moderation type design are reported in table 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 moves at a slightly faster speed. Bayesian weighting does almost the same as equal weighting, but its performance is less sensitive to the break size in this case.

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 larger, the average root mean squared 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.

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 1. We consider the Taiwanese data[20] 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,[21] and it becomes more volatile towards the end of the sample. The U.S. data is obtained from the Bureau of Economic Analysis.[22] The data for Taiwan is

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

[21] 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.

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

from National Statistics.[23]

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 [41] and Stock and Watson [42]. 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 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.

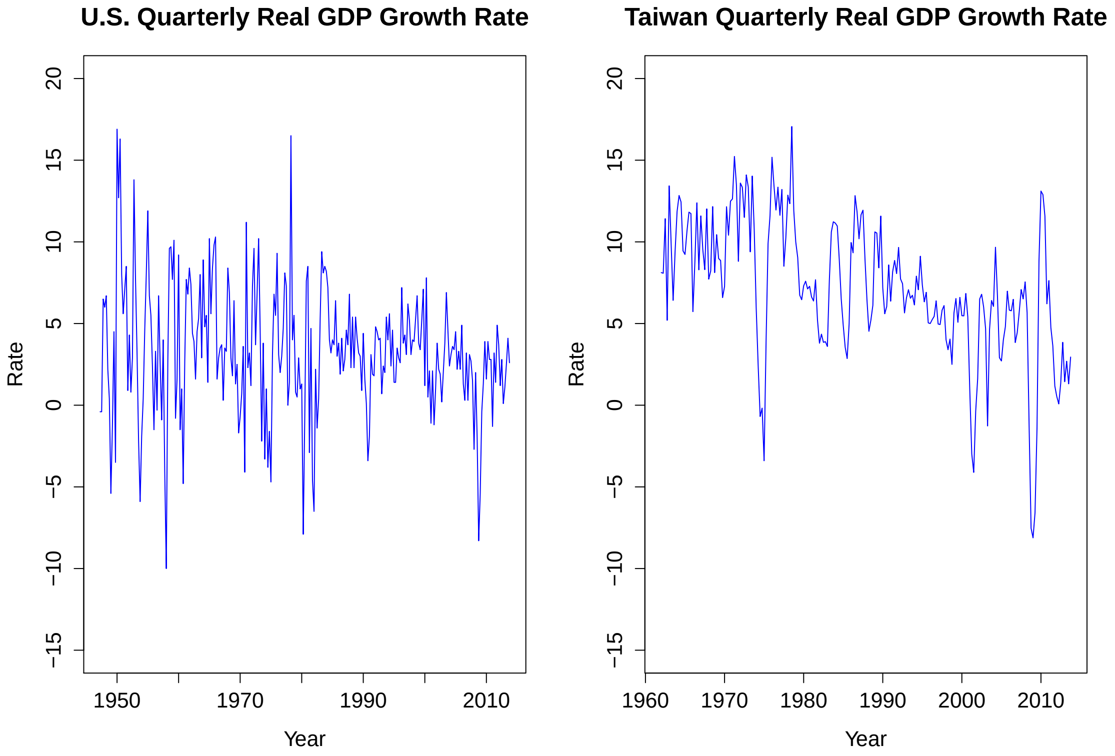
5.1 Forecast U.S. GDP Growth

Here we apply our method to forecasting the U.S. quarterly GDP growth rate[24] out-of-sample and compare its performance with others. We have quarterly data

[23] 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/

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

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


U.S. Quarterly Real GDP Growth Rate

Taiwan Quarterly Real GDP Growth Rate

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).[25]

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:

$$\Delta \mathrm{GDP}_t = \beta_0 + \beta_1 \Delta \mathrm{GDP}_{t-1} + \epsilon_t \tag{13a}$$

$$\Delta \mathrm{GDP}_t = \beta_0 + \beta_1 \Delta \mathrm{GDP}_{t-1} + \beta_2 \Delta \mathrm{GDP}_{t-2} + \epsilon_t \tag{13b}$$

$$\Delta \mathrm{GDP}_t = \beta_0 + \beta_1 \Delta \mathrm{GDP}_{t-1} + \beta_2 \Delta \mathrm{SR}_{t-1} + \epsilon_t \tag{13c}$$

$$\Delta \mathrm{GDP}_t = \beta_0 + \beta_1 \Delta \mathrm{GDP}_{t-1} + \beta_2 \Delta \mathrm{SR}_{t-1} + \beta_3 \Delta \mathrm{LR}_{t-1} + \epsilon_t \tag{13d}$$

$$\Delta \mathrm{GDP}_t = \beta_0 + \beta_1 \Delta \mathrm{GDP}_{t-1} + \beta_2 \Delta \mathrm{SR}_{t-1} + \beta_3 \Delta \mathrm{LR}_{t-1} + \beta_4 \Delta \mathrm{DP}_{t-1} + \epsilon_t \tag{13e}$$

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 4. For each model, the column shows the OOS performance for a given weighting method. The rows report results for different evaluation sample sizes. For the entries in the table, following our Monte Carlo simulation, we select the equal weighting method as

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

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

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 beating the benchmark regardless of the test sample size and base 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 moves close to the beach mark 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 they marginally outperform the benchmark.

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 lower 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. These features in the Taiwanese data motivate us to apply model averaging methods to this new environment and compare and evaluate these methods.

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 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 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 test sample sizes. Overall, both applications demonstrate the superior performance of CV weights compared with related 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 selected for prediction: the break model or the stable one? If 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 forecasting models.

In many empirical applications related to macroeconomic or financial time series, researchers usually can not 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 than 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 exceed other complex combination methods. This so called forecast combination puzzle has cast doubt on the use of complicated model averaging methods, so comparing the 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 method exceeds equal weighting as shown in simulations and in empirical examples on 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 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, we can expand the above equation as

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

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 3.1. Since CV is asymptotically equivalent to Mallows' Cp, following Hansen's [26] proof, 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 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 3.2. The proof of this proposition is adapted from Hansen [26]. 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

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

and

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

so

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

By continuous mapping theorem we have

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

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

Putting together 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. Combined with Hansen's [26] theorem 1 without assuming conditional homoscedasticity or Andrews' [2] 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 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 [38]), 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 3.2. 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 $E(\sum_{j=1}^{k} \lambda_j \chi^2(1)) = \sum_{j=1}^{k} \lambda_j$. 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 $E(J_0(\xi_\delta)) \approx \frac{1}{2}(\mathrm{tr}(\hat{Q}^{-1}\hat{\Sigma}) + 2\bar{p} - k) \equiv \bar{p}^*$. Then by the same procedure in the proof of corollary 3.2, the sample CV criterion is

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

The sample optimal CV weight is the value in $[0,1]$ that minimizes $\widehat{\mathrm{CV}}(w)$, 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. Tests for parameter instability and structural change with unknown change point. *Econometrica*, 61(04):821–856, 1993.

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

[4] 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.

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

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

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

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

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

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

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

[12] 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.

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

[14] 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[29] 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.

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

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

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

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

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

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

[36] 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.

[37] 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.

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

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

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

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

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

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

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