Understanding Models' Forecasting Performance

Barbara Rossi* and Tatevik Sekhposyan†

March 2011 (First version: January 2009)

Abstract

We propose a new methodology to identify the sources of models' forecasting performance. The methodology decomposes the models' forecasting performance into asymptotically uncorrelated components that measure instabilities in the forecasting performance, predictive content, and over-fitting. The empirical application shows the usefulness of the new methodology for understanding the causes of the poor forecasting ability of economic models for exchange rate determination.

Keywords: Forecasting, Forecast Evaluation, Instabilities, Over-fitting, Exchange Rates.

J.E.L. Codes: C22, C52, C53

*Corresponding author at: Department of Economics, Duke University, 213 Social Sciences, P.O. Box 90097, Durham, NC 27708, USA. Tel.: +1-919-660-1801; fax: +1-919-684-8974; e-mail: brossi@econ.duke.edu

†International Economic Analysis Department, Bank of Canada, 234 Wellington Street, Ottawa, ON, K1R 1C8 Canada

1 Introduction

This paper has two objectives. The first objective is to propose a new methodology for understanding why models have different forecasting performance. Is it because the forecasting performance has changed over time? Or is it because there is estimation uncertainty which makes models' in-sample fit not informative about their out-of-sample forecasting ability? We identify three possible sources of models' forecasting performance: predictive content, over-fitting, and time-varying forecasting ability. Predictive content indicates whether the in-sample fit predicts out-of-sample forecasting performance. Over-fitting is a situation in which a model includes irrelevant regressors, which improve the in-sample fit of the model but penalize the model in an out-of-sample forecasting exercise. Time-varying forecasting ability might be caused by changes in the parameters of the models, as well as by unmodeled changes in the stochastic processes generating the variables. Our proposed technique involves a decomposition of the existing measures of forecasting performance into these components to understand the reasons why a model forecasts better than its competitors. We also propose tests for assessing the significance of each component. Thus, our methodology suggests constructive ways of improving models' forecasting ability.

The second objective is to apply the proposed methodology to study the performance of models of exchange rate determination in an out-of-sample forecasting environment. Explaining and forecasting nominal exchange rates with macroeconomic fundamentals has long been a struggle in the international finance literature. We apply our methodology in order to better understand the sources of the poor forecasting ability of the models. We focus on models of exchange rate determination in industrialized countries such as Switzerland, United Kingdom, Canada, Japan, and Germany. Similarly to Bacchetta, van Wincoop and Beutler (2010), we consider economic models of exchange rate determination that involve macroeconomic fundamentals such as oil prices, industrial production indices, unemployment rates, interest rates, and money differentials. The empirical findings are as follows.

Exchange rate forecasts based on the random walk are superior to those of economic models on average over the out-of-sample period. When trying to understand the reasons for the inferior forecasting performance of the macroeconomic fundamentals, we find that lack of predictive content is the major explanation for the lack of short-term forecasting ability of the economic models, whereas instabilities play a role especially for medium term (one-year ahead) forecasts.

Our paper is closely related to the rapidly growing literature on forecasting (see Elliott and Timmermann (2008) for reviews and references). In their seminal works, Diebold and Mariano (1995) and West (1996) proposed tests for comparing the forecasting ability of competing models, inspiring a substantial research agenda that includes West and McCracken (1998), McCracken (2000), Clark and McCracken (2001, 2005, 2006), Clark and West (2006) and Giacomini and White (2006), among others. None of these works, however, analyze the reasons why the best model outperforms its competitors. In a recent paper, Giacomini and Rossi (2009) propose testing whether the relationship between in-sample predictive content and out-of-sample forecasting ability for a given model has worsened over time, and refer to such situations as Forecast Breakdowns. Interestingly, Giacomini and Rossi (2009) show that Forecast Breakdowns may happen because of instabilities and over-fitting. However, while their test detects both instabilities and over-fitting, in practice it cannot identify the exact source of the breakdown. The main objective of our paper is instead to decompose the forecasting performance in components that shed light on the causes of the models' forecast advantages/disadvantages. We study such decomposition in the same framework of Giacomini and Rossi (2010), that allows for time variation in the forecasting performance, and discuss its implementation in fixed rolling window as well as recursive window environments.

The rest of the paper is organized as follows. Section 2 describes the framework and the assumptions, Section 3 discusses the new decomposition proposed in this paper, and Section 4 presents the relevant tests. Section 5 provides a simple Monte Carlo simulation exercise,

whereas Section 6 attempts to analyze the causes of the poor performance of models for exchange rate determination. Section 7 concludes.

2 The Framework and Assumptions

Since the works by Diebold and Mariano (1995), West (1996), and Clark and McCracken (2001), it has become common to compare models according to their forecasting performance in a pseudo out-of-sample forecasting environment. Let $h \geq 1$ denote the (finite) forecast horizon**.** We are interested in evaluating the performance of $h-$steps ahead forecasts for the scalar variable y_t using a vector of predictors x_t, where forecasts are obtained via a direct forecasting method.[1] We assume the researcher has P out-of-sample predictions available, where the first out-of-sample prediction is based on a parameter estimated using data up to time R, the second prediction is based on a parameter estimated using data up to $R+1$, and the last prediction is based on a parameter estimated using data up to $R+P-1=T$, where $R+P+h-1=T+h$ is the size of the available sample.

The researcher is interested in evaluating model(s)' pseudo out-of-sample forecasting performance, and comparing it with a measure of in-sample fit so that the out-of-sample forecasting performance can be ultimately decomposed into components measuring the contribution of time-variation, over-fitting, and predictive content. Let $\{L_{t+h}(.)\}_{t=R}^{T}$ be a sequence of loss functions evaluating $h-$steps ahead out-of-sample forecast errors. This framework is general enough to encompass:

(i) measures of absolute forecasting performance, where $L_{t+h}(.)$ is the forecast error loss of a model;

(ii) measures of relative forecasting performance, where $L_{t+h}(.)$ is the difference of the forecast error losses of two competing models; this includes, for example, the measures of relative forecasting performance considered by Diebold and Mariano (1995) and West (1996);

[1] That is, $h-$steps ahead forecasts are directly obtained by using estimates from the direct regression of the dependent variable on the regressors lagged $h-$periods.

(iii) measures of regression-based predictive ability, where $L_{t+h}(.)$ is the product of the forecast error of a model times possible predictors; this includes, for example, Mincer and Zarnowitz's (1969) measures of forecast efficiency. For an overview and discussion of more general regression-based tests of predictive ability see West and McCracken (1998).

To illustrate, we provide examples of all three measures of predictive ability. Consider an unrestricted model specified as $y_{t+h} = x_t'\alpha + \varepsilon_{t+h}$, and a restricted model: $y_{t+h} = \varepsilon_{t+h}$. Let $\hat{\alpha}_t$ be an estimate of the regression coefficient of the unrestricted model at time t using all the observations available up to time t, i.e. $\hat{\alpha}_t = \left(\sum_{j=1}^{t-h} x_j x_j'\right)^{-1}\left(\sum_{j=1}^{t-h} x_j y_{j+h}\right)$.

(i) Under a quadratic loss function, the measures of absolute forecasting performance for model 1 and 2 would be the squared forecast errors: $L_{t+h}(.) = (y_{t+h} - x_t'\hat{\alpha}_t)^2$ and $L_{t+h}(.) = y_{t+h}^2$ respectively.

(ii) Under the same quadratic loss function, the measure of relative forecasting performance of the models in (i) would be $L_{t+h}(.) = (y_{t+h} - x_t'\hat{\alpha}_t)^2 - y_{t+h}^2$.

(iii) An example of a regression-based predictive ability test is the test of zero mean prediction error, where, for the unrestricted model: $L_{t+h}(.) = y_{t+h} - x_t'\hat{\alpha}_t$.

Throughout this paper, we focus on measures of relative forecasting performance. Let the two competing models be labeled 1 and 2, which could be nested or non-nested. Model 1 is characterized by parameters α and model 2 by parameters γ. We consider two estimation frameworks: a fixed rolling window and an expanding (or recursive) window.

2.1 Fixed Rolling Window Case

In the fixed rolling window case, the model's parameters are estimated using samples of R observations dated $t-R+1, ..., t$, for $t = R, R+1, ..., T$, where $R < \infty$. The parameter estimates for model 1 are obtained by $\widehat{\alpha}_{t,R} = \arg\min_a \sum_{j=t-R+1}^{t} \mathcal{L}_j^{(1)}(a)$, where $\mathcal{L}^{(1)}(.)$ denotes the in-sample loss function for model 1; similarly, the parameters for model 2 are $\widehat{\gamma}_{t,R} = \arg\min_g \sum_{j=t-R+1}^{t} \mathcal{L}_j^{(2)}(g)$. At each point in time t, the estimation will generate a sequence

of R in-sample fitted errors denoted by $\{\eta_{1,j}(\widehat{\alpha}_{t,R}), \eta_{2,j}(\widehat{\gamma}_{t,R})\}_{j=t-R+1}^{t}$; among the R fitted errors, we use the last in-sample fitted errors at time t, $(\eta_{1,t}(\widehat{\alpha}_{t,R}), \eta_{2,t}(\widehat{\gamma}_{t,R}))$, to evaluate the models' in-sample fit at time t, $\mathcal{L}_t^{(1)}(\widehat{\alpha}_{t,R})$ and $\mathcal{L}_t^{(2)}(\widehat{\gamma}_{t,R})$. For example, for the unrestricted model considered previously, $y_{t+h} = x_t'\alpha + \varepsilon_{t+h}$, under a quadratic loss, we have $\widehat{\alpha}_{t,R} = (\sum_{j=t-h-R+1}^{t-h} x_j x_j')^{-1} (\sum_{j=t-h-R+1}^{t-h} x_j y_{j+h})$, for $t = R, R+1, ..., T$. The sequence of in-sample fitted errors at time t is: $\{\eta_{1,j}(\widehat{\alpha}_{t,R})\}_{j=t-R+1}^{t} = \{y_j - x_{j-h}'\widehat{\alpha}_{j,R}\}_{j=t-R+1}^{t}$, of which we use the last in-sample fitted error, $\eta_{1,t}(\widehat{\alpha}_{t,R}) = y_t - x_{t-h}'\widehat{\alpha}_{t,R}$, to evaluate the in-sample loss at time t : $\mathcal{L}_t^{(1)}(\widehat{\alpha}_{t,R}) \equiv \left(y_t - x_{t-h}'\widehat{\alpha}_{t,R}\right)^2$. Thus, as the rolling estimation is performed over the sample for $t = R, R+1, ..., T$, we collect a series of in-sample losses: $\left\{\mathcal{L}_t^{(1)}(\widehat{\alpha}_{t,R}), \mathcal{L}_t^{(2)}(\widehat{\gamma}_{t,R})\right\}_{t=R}^{T}$.

We consider the loss functions $L_{t+h}^{(1)}(\widehat{\alpha}_{t,R})$ and $L_{t+h}^{(2)}(\widehat{\gamma}_{t,R})$ to evaluate out-of-sample predictive ability of direct h-step ahead forecasts for models 1 and 2 made at time t. For example, for the unrestricted model considered previously ($y_{t+h} = x_t'\alpha + \varepsilon_{t+h}$), under a quadratic loss, the out-of-sample multi-step direct forecast loss at time t is: $L_{t+h}^{(1)}(\widehat{\alpha}_{t,R}) \equiv (y_{t+h} - x_t'\widehat{\alpha}_{t,R})^2$. As the rolling estimation is performed over the sample, we collect a series of out-of-sample losses: $\left\{L_{t+h}^{(1)}(\widehat{\alpha}_{t,R}), L_{t+h}^{(2)}(\widehat{\gamma}_{t,R})\right\}_{t=R}^{T}$.

The loss function used for estimation need not necessarily be the same loss function used for forecast evaluation, although in order to ensure a meaningful interpretation of the models' in-sample performance as a proxy for the out-of-sample performance, we require the loss function used for estimation to be the same as the loss used for forecast evaluation. Assumption 4(a) in Section 3 will formalize this requirement.

Let $\theta \equiv (\alpha', \gamma')'$ be the $(p \times 1)$ parameter vector, $\widehat{\theta}_{t,R} \equiv \left(\widehat{\alpha}_{t,R}', \widehat{\gamma}_{t,R}'\right)'$, $L_{t+h}\left(\widehat{\theta}_{t,R}\right) \equiv L_{t+h}^{(1)}(\widehat{\alpha}_{t,R}) - L_{t+h}^{(2)}(\widehat{\gamma}_{t,R})$ and $\mathcal{L}_t\left(\widehat{\theta}_{t,R}\right) \equiv \mathcal{L}_t^{(1)}(\widehat{\alpha}_{t,R}) - \mathcal{L}_t^{(2)}(\widehat{\gamma}_{t,R})$. For notational simplicity, in what follows we drop the dependence on the parameters, and simply use $\widehat{L}_{t+h}$ and $\widehat{\mathcal{L}}_t$ to denote $L_{t+h}(\widehat{\theta}_{t,R})$ and $\mathcal{L}_t(\widehat{\theta}_{t,R})$ respectively.[2]

We make the following Assumption.

[2]Even if we assume that the loss function used for estimation is the same as the loss function used for forecast evaluation, the different notation for the estimated in-sample and out-of-sample losses ($\widehat{L}_t$ and $\widehat{\mathcal{L}}_t$) is necessary to reflect that they are evaluated at parameters estimated at different points in time.

Assumption 1: Let $\widehat{l}_{t+h} \equiv \left(\widehat{L}_{t+h},\ \widehat{\mathcal{L}}_t\widehat{L}_{t+h}\right)'$, $t = R, ..., T$, *and* $Z_{t+h,R} \equiv \widehat{l}_{t+h} - E(\widehat{l}_{t+h})$.

(a) $\Omega_{roll} \equiv \lim_{T\to\infty} Var\left(P^{-1/2}\sum_{t=R}^{T} Z_{t+h,R}\right)$ *is positive definite;*

(b) for some $r > 2$, $|||[Z_{t+h,R}, \widehat{\mathcal{L}}_t]|||_r < \Delta < \infty$;[3]

(c) $\{y_t, x_t\}$ *are mixing with either* $\{\phi\}$ *of size* $-r/2\,(r-1)$ *or* $\{\alpha\}$ *of size* $-r/(r-2)$;

(d) for $k \in [0,1]$, $\lim_{T\to\infty} E\left[W_P(k)\,W_P(k)'\right] = kI_2$, *where* $W_P(k) = \sum_{t=R}^{R+[kP]} \Omega_{roll}^{-1/2} Z_{t+h,R}$;

(e) $P \to \infty$ *as* $T \to \infty$, *whereas* $R < \infty$, $h < \infty$.

Remarks. Assumption 1 is useful for obtaining the limiting distribution of the statistics of interest, and provides the necessary assumptions for the high-level assumptions that guarantee that a Functional Central Limit Theorem holds, as in Giacomini and Rossi (2010). Assumptions 1(a-c) impose moment and mixing conditions to ensure that a Multivariate Invariance Principle holds (Wooldridge and White, 1988). In addition, Assumption 1(d) imposes global covariance stationarity. The assumption allows for the competing models to be either nested or non-nested and to be estimated with general loss functions. This generality has the trade-off of restricting the estimation to a fixed rolling window scheme, and Assumption 1(e) ensures that the parameter estimates are obtained over an asymptotically negligible in-sample fraction of the data (R).[4]

Proposition 1 (Asymptotic Results for the Fixed Rolling Window Case) *For every* $k \in [0,1]$, *under Assumption 1:*

$$\frac{1}{\sqrt{P}} \sum_{t=R}^{R+[kP]} \Omega_{roll}^{-1/2}\left[\widehat{l}_{t+h} - E(\widehat{l}_{t+h})\right] \Rightarrow \mathcal{W}(k), \tag{1}$$

where $\mathcal{W}(.)$ *is a* (2×1) *standard vector Brownian Motion.*

[3]Hereafter, ||.|| denotes the Euclidean norm.

[4]Note that researchers might also be interested in a rolling window estimation case where the size of the rolling window is "large" relative to the sample size. In this case, researchers may apply the Clark and West's (2006, 2007) test, and obtain results similar to those above. However, using Clark and West's test would have the drawback of eliminating the over-fitting component, which is the focus of this paper.

Comment. A consistent estimate of Ω_{roll} can be obtained by

$$\widehat{\Omega}_{roll} = \sum_{i=-q(P)+1}^{q(P)-1} (1 - |i/q(P)|) P^{-1} \sum_{t=R}^{T} \widehat{l}_{t+h}^{d} \widehat{l}_{t+h}^{d\prime}, \tag{2}$$

where $\widehat{l}_{t+h}^{d} \equiv \widehat{l}_{t+h} - P^{-1} \sum_{t=R}^{T} \widehat{l}_{t+h}$ and $q(P)$ is a bandwidth that grows with P (Newey and West, 1987).

2.2 Expanding Window Case

In the expanding (or recursive) window case, the forecasting environment is the same as in Section 2.1, with the following exception. The parameter estimates for model 1 are obtained by $\widehat{\alpha}_{t,R} = \arg\min_a \sum_{j=1}^{t} \mathcal{L}_j^{(1)}(a)$; similarly, the parameters for model 2 are $\widehat{\gamma}_{t,R} = \arg\min_g \sum_{j=1}^{t} \mathcal{L}_j^{(2)}(g)$. Accordingly, the first prediction is based on a parameter vector estimated using data from 1 to R, the second on a parameter vector estimated using data from 1 to $R+1$, ..., and the last on a parameter vector estimated using data from 1 to $R+P-1=T$. Let $\widehat{\theta}_{t,R}$ denote the estimate of the parameter θ based on data from period t and earlier. For example, for the unrestricted model considered previously, $y_{t+h} = x_t'\alpha + \varepsilon_{t+h}$, we have: $\widehat{\alpha}_{t,R} = (\sum_{j=1}^{t-h} x_j x_j')^{-1} (\sum_{j=1}^{t-h} x_j y_{j+h})$, for $t = R, R+1, ..., T$, $L_{t+h}^{(1)}(\widehat{\alpha}_{t,R}) \equiv (y_{t+h} - x_t'\widehat{\alpha}_{t,R})^2$ and $\mathcal{L}_t^{(1)}(\widehat{\alpha}_{t,R}) \equiv \left(y_t - x_{t-h}'\widehat{\alpha}_{t,R}\right)^2$. Finally, let $L_{t+h} \equiv L_{t+h}(\theta^*)$, $\mathcal{L}_t \equiv \mathcal{L}_t(\theta^*)$, where θ^* is the pseudo-true parameter value.

We make the following Assumption.

Assumption 2: Let $\widehat{l}_{t+h} \equiv \left(\widehat{L}_{t+h}, \widehat{\mathcal{L}}_t \widehat{L}_{t+h}\right)'$, $l_{t+h} \equiv (L_{t+h}, \mathcal{L}_t L_{t+h})'$.

(a) $R, P \to \infty$ as $T \to \infty$, and $\lim_{T\to\infty} (P/R) = \pi \in [0, \infty)$, $h < \infty$;

(b) In some open neighborhood N *around* θ^*, *and with probability one,* $L_{t+h}(\theta), \mathcal{L}_t(\theta)$ *are measurable and twice continuously differentiable with respect to* θ. *In addition, there is a constant* $\overline{K} < \infty$ *such that for all* t, $sup_{\theta\in N} |\partial^2 l_t(\theta) / \partial\theta\partial\theta'| < M_t$ *for which* $E(M_t) < \overline{K}$.

(c) $\widehat{\theta}_{t,R}$ *satisfies* $\widehat{\theta}_{t,R} - \theta^* = J_t H_t$, *where* J_t *is* $(p \times q)$, H_t *is* $(q \times 1)$, $J_t \underset{as}{\to} J$, *with* J *of*

rank p; $H_t = t^{-1}\sum_{s=1}^{t} h_s$ *for a* $(q \times 1)$ *orthogonality condition vector* $h_s \equiv h_s(\theta^*)$; $E(h_s) = 0$.

(d) Let $D_{t+h} \equiv \frac{\partial l_{t+h}(\theta)}{\partial \theta}|_{\theta=\theta^*}$, $D \equiv E(D_{t+h})$ and $\xi_{t+h} \equiv \left[vec(D_{t+h})', l'_{t+h}, h'_t, \mathcal{L}_t\right]'$. *Then: (i) For some* $d > 1$, $sup_t E||\xi_{t+h}||_{4d} < \infty$. *(ii)* $\xi_{t+h} - E(\xi_{t+h})$ *is strong mixing, with mixing coefficients of size* $-3d/(d-1)$. *(iii)* $\xi_{t+h} - E(\xi_{t+h})$ *is covariance stationary. (iv) Let* $\Gamma_{ll}(j) = E(l_{t+h} - E(l_{t+h}))(l_{t+h-j} - E(l_{t+h}))'$, $\Gamma_{lh}(j) = E(l_{t+h} - E(l_{t+h}))(h_{t-j} - E(h_t))'$, $\Gamma_{hh}(j) = E(h_t - E(h_t))(h_{t-j} - E(h_t))'$, $S_{ll} = \sum_{j=-\infty}^{\infty}\Gamma_{ll}(j)$, $S_{lh} = \sum_{j=-\infty}^{\infty}\Gamma_{lh}(j)$, $S_{hh} = \sum_{j=-\infty}^{\infty}\Gamma_{hh}(j)$, $S = \begin{pmatrix} S_{ll} & S_{lh}J' \\ JS'_{lh} & JS_{hh}J' \end{pmatrix}$. *Then* S_{ll} *is positive definite.*

Assumption 2(a) allows both R and P to grow as the total sample size grows; this is a special feature of the expanding window case.[5] Assumption 2(b) ensures that the relevant losses are well approximated by smooth quadratic functions in the neighborhood of the parameter vector, as in West (1996). Assumption 2(c) is specific to the recursive window estimation procedure, and requires that the parameter estimates be obtained over recursive windows of data.[6] Assumption 2(d,i-iv) imposes moment conditions that ensure the validity of the Central Limit Theorem, as well as covariance stationarity for technical convenience. As in West (1996), positive definiteness of S_{ll} rules out the nested model case.[7] Thus, for nested model comparisons, we recommend the rolling window scheme discussed in Section 2.1.

There are two important cases where the asymptotic theory for the recursive window case simplifies. These are cases in which the estimation uncertainty vanishes asymptotically. A first case is when $\pi = 0$, which implies that estimation uncertainty is irrelevant since there is an arbitrary large number of observations used for estimating the models' parameters, R, relative to the number used to estimate $E(l_{t+h})$, P. The second case is when $D = 0$. A

[5]This also implies that the probability limit of the parameter estimates are constant, at least under the null hypothesis.

[6]In our example: $J_t = (\frac{1}{t}\sum_{j=1}^{t-h} x_j x'_j)^{-1}$ and $H_t = \frac{1}{t}\sum_{j=1}^{t-h} x_j y_{j+h}$.

[7]The extension of the expanding window setup to nested models would require using non-normal distributions for which the Functional Central Limit Theorem cannot be easily applied.

leading case that ensures $D = 0$ is the quadratic loss (i.e. the Mean Squared Forecast Error) and i.i.d. errors. Proposition 2 below considers the case of vanishing estimation uncertainty. Proposition 5 in Appendix A provides results for the general case in which either $\pi \neq 0$ or $D \neq 0$. Both propositions are formally proved in Appendix B.

Proposition 2 (Asymptotic Results for the Recursive Window Case) *For every* $k \in [0,1]$, *under Assumption 2, if* $\pi = 0$ *or* $D = 0$ *then*

$$\frac{1}{\sqrt{P}} \sum_{t=R}^{R+[kP]} \Omega_{rec}^{-1/2} \left[\widehat{l}_{t+h} - E\left(l_{t+h}\right) \right] \Rightarrow \mathcal{W}\left(k\right),$$

where $\Omega_{rec} = S_{ll}$, *and* $\mathcal{W}\left(.\right)$ *is a* (2×1) *standard vector Brownian Motion.*

Comment. Upon mild strengthening of the assumptions on h_t as in Andrews (1991), a consistent estimate of Ω_{rec} can be obtained by

$$\widehat{\Omega}_{rec} = \sum_{i=-q(P)+1}^{q(P)-1} (1 - |i/q(P)|) P^{-1} \sum_{t=R}^{T} \widehat{l}_{t+h}^{d} \widehat{l}_{t+h}^{d\prime}, \tag{3}$$

where $\widehat{l}_{t+h}^{d} \equiv \widehat{l}_{t+h} - P^{-1} \sum_{t=R}^{T} \widehat{l}_{t+h}$ and $q(P)$ is a bandwidth that grows with P (Newey and West, 1987).

3 Understanding the Sources of Models' Forecasting Performance: A Decomposition

Existing forecast comparison tests, such as Diebold and Mariano (1995) and West (1996), inform the researcher only about which model forecasts best, and do not shed any light on why that is the case. Our main objective, instead, is to decompose the sources of the out-of-sample forecasting performance into uncorrelated components that have meaningful

economic interpretation, and might provide constructive insights to improve models' forecasts. The out-of-sample forecasting performance of competing models can be attributed to model instability, over-fitting, and predictive content. Below we elaborate on each of these components in more detail.

We measure time variation in models' relative forecasting performance by averaging relative predictive ability over rolling windows of size m, as in Giacomini and Rossi (2010), where $m < P$ satisfies assumption 3 below.

Assumption 3: $\lim_{T\to\infty} (m/P) \to \mu \in (0, \infty)$ *as* $m, P \to \infty$.

We define predictive content as the correlation between the in-sample and out-of-sample measures of fit. When the correlation is small, the in-sample measures of fit have no predictive content for the out-of-sample and vice versa. An interesting case occurs when the correlation is strong, but negative. In this case the in-sample predictive content is strong yet misleading for the out-of-sample. We define over-fitting as a situation in which a model fits well in-sample but loses predictive ability out-of-sample; that is, where in-sample measures of fit fail to be informative regarding the out-of-sample predictive content.

To capture predictive content and over-fitting, we consider the following regression:

$$\widehat{L}_{t+h} = \beta \widehat{\mathcal{L}}_t + u_{t+h} \quad \text{for} \quad t = R, R+1, ..., T. \tag{4}$$

Let $\widehat{\beta} \equiv \left(\frac{1}{P} \sum_{t=R}^{T} \widehat{\mathcal{L}}_t^2 \right)^{-1} \left(\frac{1}{P} \sum_{t=R}^{T} \widehat{\mathcal{L}}_t \widehat{L}_{t+h} \right)$ denote the OLS estimate of β in regression (4), $\widehat{\beta}\widehat{\mathcal{L}}_t$ and $\widehat{u}_{t+h}$ denote the corresponding fitted values and regression errors. Note that $\widehat{L}_{t+h} = \widehat{\beta}\widehat{\mathcal{L}}_t + \widehat{u}_{t+h}$. In addition, regression (4) does *not* include a constant, so that the error term measures the average out-of-sample losses not explained by in-sample performance. Then, the average Mean Square Forecast Error (MSFE) can be decomposed as:

$$\frac{1}{P} \sum_{t=R}^{T} \widehat{L}_{t+h} = B_P + U_P, \tag{5}$$

where $B_P \equiv \widehat{\beta}\left(\frac{1}{P}\sum_{t=R}^{T}\widehat{\mathcal{L}}_t\right)$ and $U_P \equiv \frac{1}{P}\sum_{t=R}^{T}\widehat{u}_{t+h}$. B_P can be interpreted as the component that was predictable on the basis of the in-sample relative fit of the models (predictive content), whereas U_P is the component that was unexpected (over-fitting).

The following example provides more details on the interpretation of the components measuring predictive content and over-fitting.

Example: Let the true data generating process (DGP) be $y_{t+h} = \alpha + \varepsilon_{t+h}$, where $\varepsilon_{t+h} \sim iidN(0, \sigma^2)$. We compare the forecasts of y_{t+h} from two nested models' made at time t based on parameter estimates obtained via the fixed rolling window approach. The first (unrestricted) model includes a constant only, so that its forecasts are $\widehat{\alpha}_{t,R} = \frac{1}{R}\sum_{j=t-h-R+1}^{t-h} y_{j+h}$, $t = R, R+1, ..., T$, and the second (restricted) model sets the constant to be zero, so that its forecast is zero. Consider the (quadratic) forecast error loss difference, $\widehat{L}_{t+h} \equiv L_{t+h}^{(1)}(\widehat{\alpha}_{t,R}) - L_{t+h}^{(2)}(0) \equiv (y_{t+h} - \widehat{\alpha}_{t,R})^2 - y_{t+h}^2$, and the (quadratic) in-sample loss difference $\widehat{\mathcal{L}}_t \equiv \mathcal{L}_t^{(1)}(\widehat{\alpha}_{t,R}) - \mathcal{L}_t^{(2)}(0) \equiv (y_t - \widehat{\alpha}_{t,R})^2 - y_t^2$.

Let $\beta \equiv E\left(\widehat{L}_{t+h}\widehat{\mathcal{L}}_t\right)/E\left(\widehat{\mathcal{L}}_t^{\,2}\right)$. It can be shown that

$$\beta = (\alpha^4 + 4\sigma^2\alpha^2 + (4\sigma^2 + 2\sigma^2\alpha^2)/R)^{-1}(\alpha^4 - 3\sigma^2/R^2).^8 \tag{6}$$

When the models are nested, in small samples $E(\hat{\mathcal{L}}_t) = -(\alpha^2 + \sigma^2/R) < 0$, as the in-sample fit of the larger model is always better than that of the small one. Consequently, $E(B_P) = \beta E(\hat{\mathcal{L}}_t) = 0$ only when $\beta = 0$. The calculations show that the numerator for β has two distinct components: the first, α^4, is an outcome of the mis-specification in model 2; the other, $3\sigma^2/R^2$, changes with the sample size and "captures" estimation uncertainty in model

[8] $E(\widehat{L}_{t+h}\widehat{\mathcal{L}}_t) = E((-2\epsilon_{t+h}(\sum_{j=t-h-R+1}^{t-h}\epsilon_{j+h})/R + (\sum_{j=t-h-R+1}^{t-h}\epsilon_{j+h})^2/R^2 - \alpha^2 - 2\alpha\epsilon_{t+h})$ $(-2\epsilon_t(\sum_{j=t-h-R+1}^{t-h}\epsilon_{j+h})/R + (\sum_{j=t-h-R+1}^{t-h}\epsilon_{j+h})^2/R^2 - \alpha^2 - 2\alpha\epsilon_t)) = \alpha^4 - 3\sigma^4/R^2$, where the derivation of the last part relies on the assumptions of normality and iid for the error terms in that $E(\epsilon_t) = 0$, $E(\epsilon_t^2) = \sigma^2$, $E(\epsilon_t^3) = 0$, $E(\epsilon_t^4) = 3\sigma^4$ and, when $j > 0$, then $E(\epsilon_{t+j}^2\epsilon_t^2) = \sigma^4$, $E(\epsilon_{t+j}\epsilon_t) = 0$. In addition, it is useful to note that $E((\sum_{j=1}^{t}\epsilon_j)^4) = tE(\epsilon_t^4) + t(t-1)(\sigma^2)^2 + 4\sum_{j=1}^{t-1}(\sigma^2)^2 = tE(\epsilon_t^4) + 3t(t-1)\sigma^4 = 3t^2\sigma^4$, $E((\sum_{j=1}^{t}\epsilon_j)^3) = 0$, and $E((\epsilon_t\sum_{j=1}^{t}\epsilon_j)^3) = E(\epsilon_t^4) + 3(t-1)\sigma^4$. The expression for the denominator can be derived similarly.

1. When the two components are equal to each other, the in-sample loss differences have no predictive content for the out-of-sample. When the mis-specification component dominates, then the in-sample loss differences provide information content for the out-of-sample. On the other hand, when β is negative, though the in-sample fit has predictive content for the out-of-sample, it is misleading in that it is driven primarily by the estimation uncertainty. For any given value of β, $E(B_P) = \beta E(\hat{\mathcal{L}}_t) = -\beta(\alpha^2 + \sigma^2/R)$, where β is defined in eq. (6).

By construction, $E(U_P) = E(\hat{L}_{t+h}) - E(B_P) = (\sigma^2/R - \alpha^2) - E(B_P)$. Similar to the case of B_P, the component designed to measure over-fitting is affected by both mis-specification and estimation uncertainty. One should note that for $\beta > 0$, the mis-specification component affects both $E(B_P)$ and $E(U_P)$ in a similar direction, while the estimation uncertainty moves them in opposite directions. Estimation uncertainty penalizes the predictive content B_P and makes the unexplained component U_P larger.[9]

To ensure a meaningful interpretation of models' in-sample performance as a proxy for the out-of-sample performance, we assume that the loss used for in-sample fit evaluation is the same as that used for out-of-sample forecast evaluation. This assumption is formalized in Assumption 4(a). Furthermore, our proposed decomposition depends on the estimation procedure. Parameters estimated in expanding windows converge to their pseudo-true values so that, in the limit, expected out-of-sample performance is measured by $\overline{L}_{t+h} \equiv E(L_{t+h})$ and expected in-sample performance is measured by $\overline{\mathcal{L}}_t \equiv E(\mathcal{L}_t)$; we also define $\overline{\mathcal{L}L}_{t,t+h} \equiv E(\mathcal{L}_t L_{t+h})$. However, for parameters estimated in fixed rolling windows the estimation uncertainty remains asymptotically relevant, so that expected out-of-sample performance is measured by $\overline{L}_{t+h} \equiv E(\widehat{L}_{t+h})$, and the expected in-sample performance is measured by $\overline{\mathcal{L}}_t \equiv E(\widehat{\mathcal{L}}_t)$, and $\overline{\mathcal{L}L}_{t,t+h} \equiv E(\widehat{\mathcal{L}}_t \widehat{L}_{t+h})$. We focus on the relevant case where the models' have different in-sample performance. This assumption is formalized in Assumption 4(b).

[9]Note that $E(\hat{L}_t) = \sigma^2/R - \alpha^2$, whereas $E(\hat{\mathcal{L}}_t) = -\sigma^2/R - \alpha^2$. Thus, the same estimation uncertainty component σ^2/R penalizes model 2 in-sample and at the same time improves model 2's performance out-of-sample (relative to model 1). This result was shown by Hansen (2008), who explores its implications for information criteria. This paper differs by Hansen (2008) in a fundamental way by focusing on a decomposition of the out-of-sample forecasting ability into separate and economically meaningful components.

The assumption is always satisfied in small-samples for nested models, and it is also trivially satisfied for measures of absolute performance.

Assumption 4: (a) For every t, $L_t(\theta) = \mathcal{L}_t(\theta)$*; (b)* $\lim_{T\to\infty} \frac{1}{P} \sum_{t=R}^{T} \overline{\mathcal{L}}_t \neq 0$.

Let $\lambda \in [\mu, 1]$. For $\tau = [\lambda P]$ we propose to decompose the out-of-sample loss function differences $\left\{\widehat{L}_{t+h}\right\}_{t=R}^{T}$ calculated in rolling windows of size m into their difference relative to the average loss, $A_{\tau,P}$, an average forecast error loss expected on the basis of the in-sample performance, B_P, and an average unexpected forecast error loss, U_P. We thus define:

$$\begin{aligned} A_{\tau,P} &= \frac{1}{m} \sum_{t=R+\tau-m}^{R+\tau-1} \widehat{L}_{t+h} - \frac{1}{P} \sum_{t=R}^{T} \widehat{L}_{t+h}, \\ B_P &= \left(\frac{1}{P} \sum_{t=R}^{T} \widehat{\mathcal{L}}_t \right) \widehat{\beta}, \\ U_P &= \frac{1}{P} \sum_{t=R}^{T} \widehat{u}_{t+h}, \end{aligned}$$

where B_P, U_P are estimated from regression (4).

The following assumption states the hypotheses that we are interested in:

Assumption 5: Let $\overline{A}_{\tau,P} \equiv E(A_{\tau,P})$, $\overline{B}_P \equiv \beta \overline{\mathcal{L}}_t$, *and* $\overline{U}_P \equiv \overline{L}_{t+h} - \beta \overline{\mathcal{L}}_t$. *The following null hypotheses hold:*

$$H_{0,A} : \overline{A}_{\tau,P} = 0 \text{ for all } \tau = m, m+1, ...P, \tag{7}$$

$$H_{0,B} : \overline{B}_P = 0, \tag{8}$$

$$H_{0,U} : \overline{U}_P = 0. \tag{9}$$

Proposition 3 provides the decomposition for both rolling and recursive window estimation schemes.

Proposition 3 (The Decomposition) *Let either (a) [Fixed Rolling Window Estimation] Assumptions 1,3,4 hold; or (b) [Recursive Window Estimation] Assumptions 2,3,4 hold and*

either $D = 0$ *or* $\pi = 0$. *Then, for* $\lambda \in [\mu, 1]$ *and* $\tau = [\lambda P]$*:*

$$\frac{1}{m}\sum_{t=R+\tau-m}^{R+\tau-1}[\widehat{L}_{t+h} - \overline{L}_{t+h}] = \left(A_{\tau,P} - \overline{A}_{\tau,P}\right) + \left(B_P - \overline{B}_P\right) + \left(U_P - \overline{U}_P\right). \tag{10}$$

Under Assumption 5, $A_{\tau,P}, B_P, U_P$ *are asymptotically uncorrelated, and provide a decomposition of the out-of-sample measure of forecasting performance,* $\frac{1}{m}\sum_{t=R+\tau-m}^{R+\tau-1}[\widehat{L}_{t+h} - \overline{L}_{t+h}]$.

Appendix B proves that $A_{\tau,P}$, B_P and U_P are asymptotically uncorrelated. This implies that (10) provides a decomposition of rolling averages of out-of-sample losses into a component that reflects the extent of instabilities in the relative forecasting performance, $A_{\tau,P}$, a component that reflects how much of the average out-of-sample forecasting ability was predictable on the basis of the in-sample fit, B_P, and how much it was unexpected, U_P. In essence, $B_P + U_P$ is the average forecasting performance over the out-of-sample period considered by Diebold and Mariano (1995) and West (1996), among others.

To summarize, $A_{\tau,P}$ measures the presence of time variation in the models' performance *relative* to their average performance. In the presence of no time variation in the expected relative forecasting performance, $\overline{A}_{\tau,P}$ should equal zero. When instead the sign of $A_{\tau,P}$ changes, the out-of-sample predictive ability swings from favoring one model to favoring the other model. B_P measures the models' out-of-sample relative forecasting ability reflected in the in-sample relative performance. When B_P has the same sign as $\frac{1}{P}\sum_{t=R}^{T}\widehat{L}_{t+h}$, this suggests that in-sample losses have predictive content for out-of-sample losses. When they have the opposite sign, there is predictive content, although it is misleading because the out-of-sample performance will be the opposite of what is expected on the basis of in-sample information. U_P measures models' out-of-sample relative forecasting ability not reflected by in-sample fit, which is our definition of over-fitting.

Similar results hold for the expanding window estimation scheme in the more general case where either $\pi \neq 0$ or $D \neq 0$. Proposition 6, discussed in Appendix A and proved in Appendix B, demonstrates that the only difference is that, in the more general case,

the variance changes as a deterministic function of the point in time in which forecasts are considered, which requires a decomposition normalized by the variances.

4 Statistical Tests

This section describes how to test the statistical significance of the three components in decompositions (10) and (20). Let $\sigma_A^2 \equiv \lim_{T\to\infty} Var(P^{-1/2}\sum_{t=R}^{T}\widehat{L}_{t+h})$, $\sigma_B^2 \equiv \lim_{T\to\infty} Var\left(P^{1/2}B_P\right)$, $\sigma_U^2 \equiv \lim_{T\to\infty} Var\left(P^{1/2}U_P\right)$, and $\widehat{\sigma}_A^2$, $\widehat{\sigma}_B^2$, $\widehat{\sigma}_U^2$ be consistent estimates of σ_A^2, σ_B^2 and σ_U^2 (such as described in Proposition 4). Also, let $\widehat{\Omega}_{(i,j),roll}$ and $\widehat{\Omega}_{(i,j),rec}$ denote the (i-th,j-th) element of $\widehat{\Omega}_{roll}$ and $\widehat{\Omega}_{rec}$, for $\widehat{\Omega}_{roll}$ defined in eq.(2) and for $\widehat{\Omega}_{rec}$ defined in eq.(3).

Proposition 4 provides test statistics for evaluating the significance of the three components in decomposition (10).

Proposition 4 (Significance Tests) *Let either: (a) [Fixed Rolling Window Estimation] Assumptions 1,3,4 hold, and* $\widehat{\sigma}_A^2 = \widehat{\Omega}_{(1,1),roll}$, $\widehat{\sigma}_B^2 = \Phi_P^2\widehat{\Omega}_{(2,2),roll}$*; or (b) [Recursive Window Estimation] Assumptions 2,3,4 hold, and* $\widehat{\sigma}_A^2 = \widehat{\Omega}_{(1,1),rec}$, $\widehat{\sigma}_B^2 = \Phi_P^2\widehat{\Omega}_{(2,2),rec}$*, and either* $\pi = 0$ *or* $D = 0$.

In addition, let $A_{\tau,P}, B_P, U_P$ *be defined as in Proposition 3, and the tests be defined as:*

$$\begin{aligned}\Gamma_P^{(A)} &\equiv \sup_{\tau=m,\ldots,P}|\sqrt{P}\widehat{\sigma}_A^{-1}A_{\tau,P}|,\\ \Gamma_P^{(B)} &\equiv \sqrt{P}\widehat{\sigma}_B^{-1}B_P,\\ \Gamma_P^{(U)} &\equiv \sqrt{P}\widehat{\sigma}_U^{-1}U_P,\end{aligned}$$

where $\Phi_P \equiv \left(\frac{1}{P}\sum_{t=R}^{T}\widehat{\mathcal{L}}_t\right)\left(\frac{1}{P}\sum_{t=R}^{T}\widehat{\mathcal{L}}_t^2\right)^{-1}$*, and* $\widehat{\sigma}_U^2 = \widehat{\sigma}_A^2 - \widehat{\sigma}_B^2$*. Then:*
(i) Under $H_{0,A}$,

$$\sqrt{P}\widehat{\sigma}_A^{-1}A_{\tau,P} \Rightarrow \frac{1}{\mu}\left[\mathcal{W}_1(\lambda) - \mathcal{W}_1(\lambda-\mu)\right] - \mathcal{W}_1(1), \tag{11}$$

for $\lambda \in [\mu, 1]$ *and* $\tau = [\lambda P]$*, where* $\mathcal{W}_1(\cdot)$ *is a standard univariate Brownian motion. The*

critical values for significance level α *are* $\pm\, k_\alpha$*, where* k_α *solves:*

$$\Pr\left\{\sup_{\lambda\in[\mu,1]}\left|\left[\mathcal{W}_1(\lambda)-\mathcal{W}_1(\lambda-\mu)\right]/\mu-\mathcal{W}_1(1)\right|>k_\alpha\right\}=\alpha. \tag{12}$$

Table 1 reports the critical values k_α *for typical values of* α*.*

(ii) Under $H_{0,B}:\Gamma_P^{(B)}\Rightarrow N(0,1)$*; under* $H_{0,U}:\Gamma_P^{(U)}\Rightarrow N(0,1)$*.*

The null hypothesis of no time variation ($H_{0,A}$) is rejected when $\Gamma_P^{(A)} > k_\alpha$; the null hypothesis of no predictive content ($H_{0,B}$) is rejected when $|\Gamma_P^{(B)}| > z_{\alpha/2}$, where $z_{\alpha/2}$ is the $(\alpha/2)$-th percentile of a standard normal; the null hypothesis of no over-fitting ($H_{0,U}$) is rejected when $|\Gamma_P^{(U)}| > z_{\alpha/2}$. Appendix B provides the formal proof of Proposition 4. In addition, Proposition 7 in Appendix A provides the generalization of the recursive estimation case to either $\pi \neq 0$ or $D \neq 0$.

5 Monte Carlo Analysis

The objective of this section is twofold. First, we evaluate the performance of the proposed method in small samples; second, we examine the role of the $A_{\tau,P}$, B_P, and U_P components in the proposed decomposition. The Monte Carlo analysis focuses on the rolling window scheme used in the empirical application. The number of Monte Carlo replications is 5,000.

We consider two Data Generating Processes (DGP): the first is the simple example discussed in Section 3; the second is tailored to match the empirical properties of the Canadian exchange rate and money differential data considered in Section 6. Let

$$y_{t+h}=\alpha_t x_t+e_{t+h},\quad t=1,2,...,T, \tag{13}$$

where $h=1$ and either: (i) (IID) $e_t\sim iid\ N(0,\sigma_\varepsilon^2)$, $\sigma_\varepsilon^2=1$, $x_t=1$, or (ii) (Serial Correlation) $e_t=\rho e_{t-1}+\varepsilon_t$, $\rho=-0.0073$, $\varepsilon_t\sim iid\ N(0,\sigma_\varepsilon^2)$, $\sigma_\varepsilon^2=2.6363$, $x_t=b_1+\sum_{s=2}^{6}b_s x_{t-s+1}+v_t$,

$v_t \sim iid\ N(0,1)$ independent of ε_t and $b_1 = 0.1409$, $b_2 = -0.1158$, $b_3 = 0.1059$, $b_4 = 0.0957$, $b_5 = 0.0089$, $b_6 = 0.1412$.

We compare the following two nested models' forecasts for y_{t+h}:

$$\begin{aligned} \text{Model 1 forecast} &: \ \widehat{\alpha}_t x_t \\ \text{Model 2 forecast} &: \ 0, \end{aligned}$$

where model 1 is estimated by OLS in rolling windows, $\widehat{\alpha}_t = \left(\Sigma_{j=t-h-R+1}^{t-h} x_j y_{j+h}\right)\left(\Sigma_{j=t-h-R+1}^{t-h} x_j^2\right)^{-1}$ for $t = R, ..., T$.

We consider the following cases. The first DGP (DGP 1) is used to evaluate the size properties of our procedures in small samples. We let $\alpha_t = \alpha$, where: (i) for the IID case, $\alpha = \sigma_e R^{-1/2}$ satisfies $H_{0,A}$, $\alpha = \sigma_e (3R^{-2})^{1/4}$ satisfies $H_{0,B}$, and $\alpha = \frac{1}{R}\sigma_e \left(\sqrt{-3R + R^2 + 1} + 1\right)^{1/2}$ satisfies $H_{0,U}$;[10] (ii) for the correlated case, the values of $\alpha \in \mathbb{R}$ satisfying the null hypotheses were calculated via Monte Carlo approximations, and $\sigma_e^2 = \sigma_\varepsilon^2/(1-\rho^2)$.[11] A second DGP (DGP 2) evaluates the power of our procedures in the presence of time variation in the parameters: we let $\alpha_t = \alpha + b \cdot \cos(1.5\pi t/T) \cdot (1 - t/T)$, where $b = \{0, 0.1, ..., 1\}$. A third DGP (DGP 3) evaluates the power of our tests against stronger predictive content. We let $\alpha_t = \alpha + b$, for $b = \{0, 0.1, ..., 1\}$. Finally, a fourth DGP (DGP 4) evaluates our procedures against over-fitting. We let Model 1 include $(p-1)$ redundant regressors and its forecast for y_{t+h} is specified as: $\widehat{\alpha}_t x_t + \sum_{s=1}^{p-1} \widehat{\gamma}_{s+1,t} x_{s,t}$, where $x_{1,t}, ..., x_{p-1,t}$ are $(p-1)$ independent standard

[10] The values of α satisfying the null hypotheses can be derived from the Example in Section 3.

[11] In detail, let $\delta_t = \left(\Sigma_{j=t-h-R+1}^{t-h} x_j^2\right)^{-1}\left(\Sigma_{j=t-h-R+1}^{t-h} x_j \varepsilon_{j+h}\right)^2$. Then, $\alpha = E\left(\delta_t^2\right)$ is the value that satisfies $H_{0,A}$ when ρ is small, and the expectation is approximated via Monte Carlo simulations. Similarly, $H_{0,B}$ sets $\alpha = -\frac{1}{2A}\left(B - \sqrt{-4AD + B^2}\right)$, where $A = E\left(x_t^2 x_{t-h}^2\right)$, $B = 2\left[E\left(\delta_t e_t x_{t-h} x_t^2\right) - E\left(\delta_t^2 x_t^2 x_{t-h}^2\right)\right]$, $C = \left[-2E\left(\delta_t^2 e_t x_{t-h} x_t^2\right)\right] = 0$, $D = E\left(\delta_t^4 x_t^2 x_{t-h}^2\right) - 2E\left(\delta_t^3 e_t x_{t-h} x_t^2\right)$. $H_{0,U}$ instead sets $\alpha \in \mathbb{R}$ s.t. $G\alpha^6 + H\alpha^4 + L\alpha^2 + M = 0$, where $G = \left\{E\left(x_{t-h}^2\right) E\left(x_t^2 x_{t-h}^2\right) - E\left(x_t^2\right) E\left(x_{t-h}^4\right)\right\}$, $H = 2E\left(x_{t-h}^2\right)\left[E\left(\delta_t e_t x_{t-h} x_t^2\right) - E\left(\delta_t^2 x_t^2 x_{t-h}^2\right)\right] - 2E\left(x_t^2\right)\left[2E\left(x_{t-h}^3 e_t \delta_t\right) - E\left(x_{t-h}^4 \delta_t^2\right) + 2E\left(e_t^2 x_{t-h}^2\right)\right] + E\left(x_t^2 x_{t-h}^2\right)\left[2E\left(e_t x_{t-h}\delta_t\right) - E\left(x_{t-h}^2 \delta_t^2\right)\right] + Ex_{t-h}^4 E\left(\delta_t^2 x_t^2\right)$, $L = [2E\left(\delta_t e_t x_{t-h} x_t^2\right) - 2E\left(\delta_t^2 x_t^2 x_{t-h}^2\right)][2E\left(e_t x_{t-h}\delta_t\right) - E\left(x_{t-h}^2 \delta_t^2\right)] + E\left(x_{t-h}^2\right)\left[E\left(\delta_t^4 x_t^2 x_{t-h}^2\right) - 2E\left(\delta_t^3 e_t x_{t-h} x_t^2\right) - 4Ee_t^2 x_{t-h}^2 \delta_t^2 + 4Ex_{t-h}^3 e_t \delta_t^3 - Ex_{t-h}^4 \delta_t^4\right] + E\left(\delta_t^2 x_t^2\right)\left[4Ex_{t-h}^3 e_t \delta_t - 2Ex_{t-h}^4 \delta_t^2 + 4Ee_t^2 x_{t-h}^2\right]$, and $M = [E\left(\delta_t^2 x_t^2\right)][4Ee_t^2 x_{t-h}^2 \delta_t^2 - 4Ex_{t-h}^3 e_t \delta_t^3 + Ex_{t-h}^4 \delta_t^4] + [E\left(\delta_t^4 x_t^2 x_{t-h}^2\right) - 2E\left(\delta_t^3 e_t x_{t-h} x_t^2\right)][2E\left(e_t x_{t-h}\delta_t\right) - E\left(x_{t-h}^2 \delta_t^2\right)]$. We used 5,000 replications.

normal random variables, whereas the true DGP remains (13) where $\alpha_t = \alpha$, and α takes the same values as in DGP1.

The results of the Monte Carlo simulations are reported in Tables 2-5. In all tables, Panel A reports results for the IID case and Panel B reports results for the serial correlation case. Asymptotic variances are estimated with a Newey West's (1987) procedure, eq. (2), with $q(P) = 1$ in the IID case and $q(P) = 2$ in the Serial Correlation case.

First, we evaluate the small sample properties of our procedure. Table 2 reports empirical rejection frequencies for DGP 1 for the tests described in Proposition 4 considering a variety of out-of-sample (P) and estimation window (R) sizes. The tests have nominal level equal to 0.05, and $m = 100$. Panel A reports results for the IID case whereas Panel B reports results for the Serial Correlation case. Therefore, $A_{\tau,P}$, B_P, and U_P should all be statistically insignificantly different from zero. Table 2 shows indeed that the rejection frequencies of our tests are close to the nominal level, even in small samples, although serial correlation introduces mild size distortions.

Second, we study the significance of each component in DGP 2-5. DGP 2 allows the parameter to change over time in a way that, as b increases, instabilities become more important. Table 3 shows that the $\Gamma_P^{(A)}$ test has power against instabilities. The $\Gamma_P^{(B)}$ and $\Gamma_P^{(U)}$ tests are not designed to detect instabilities, and therefore their empirical rejection rates are close to nominal size.

DGP 3 is a situation in which the parameters are constant, and the information content in Model 1 becomes progressively better as b increases (the model's performance is equal to its competitor in expectation when $b = 0$). Since the parameters are constant and there are no other instabilities, the $A_{\tau,P}$ component should not be significantly different from zero, whereas the B_P and U_P components should become different from zero when b is sufficiently different from zero. These predictions are supported by the Monte Carlo results in Table 4.

DGP 4 is a situation in which Model 1 includes an increasing number of irrelevant regressors (p). Table 5 shows that, as p increases, the estimation uncertainty caused by the

increase in the number of parameters starts penalizing Model 1: its out-of-sample performance relative to its in-sample predictive content worsens, and B_P, U_P become significantly different from zero. On the other hand, $A_{\tau,P}$ is not significantly different from zero, as there is no time-variation in the models' relative loss differentials.[12]

6 The relationship between fundamentals and exchange rate fluctuations

This section analyzes the link between macroeconomic fundamentals and nominal exchange rate fluctuations using the new tools proposed in this paper. Explaining and forecasting nominal exchange rates has long been a struggle in the international finance literature. Since Meese and Rogoff (1983a,b) first established that the random walk generates the best exchange rate forecasts, the literature has yet to find an economic model that can consistently produce good in-sample fit and outperform a random walk in out-of-sample forecasting, at least at short- to medium-horizons (see e.g. Engel, Mark and West, 2008). In their papers, Meese and Rogoff (1983a,b) conjectured that sampling error, model mis-specification and instabilities can be possible explanations for the poor forecasting performance of the economic models. We therefore apply the methodology presented in Section 2 to better understand *why* the economic models' performance is poor.

We reconsider the forecasting relationship between the exchange rates and economic fundamentals in a multivariate regression with growth rate differentials of the following country-specific variables relative to their US counterparts: money supply (M_t), the industrial production index (IPI_t), the unemployment rate (UR_t), and the lagged interest rate (R_{t-1}), in addition to the growth rate of oil prices (whose level is denoted by OP_t). In addition, we separately consider the predictive ability of the commodity price index (CP) growth rate for the Canadian exchange rate. We focus on monthly data from 1975:9 to 2008:9 for

[12]Unreported results show that the magnitude of the correlation of the three components is very small.

a few industrial countries, such as Switzerland, United Kingdom, Canada, Japan, and Germany. The data are collected from the IFS, OECD, Datastream, as well as country-specific sources; see Appendix C for details.

We compare one-step ahead forecasts for the following models. Let y_t denote the growth rate of the exchange rate at time t, $y_t = [\ln(S_t/S_{t-1})]$. The "economic" model is specified as:

$$\text{Model 1}: \; y_t = \alpha x_t + \epsilon_{1,t}, \tag{14}$$

where x_t is the vector containing $\ln(M_t/M_{t-1})$, $\ln(IPI_t/IPI_{t-1})$, $\ln(UR_t/UR_{t-1})$, $\ln(R_{t-1}/R_{t-2})$, and $\ln(OP_t/OP_{t-1})$, and $\epsilon_{1,t}$ is the error term. In the case of commodity prices, x_t contains only the growth rate of the commodity price index. The benchmark is a simple random walk:

$$\text{Model 2} \; : y_t = \varepsilon_{2,t}, \tag{15}$$

where $\varepsilon_{2,t}$ is the error term.

First, we follow Bacchetta et al. (2010) and show how the relative forecasting performance of the competing models is affected by the choice of the rolling estimation window size R. Let $R = 40, ..., 196$. Given that our total sample size is fixed ($T = 396$) and the estimation window size varies, the out-of-sample period $P(R) = T + h - R$ is not constant throughout the exercise. Let $\underline{P} = \min P(R) = 200$ denote the minimum common out-of-sample period across the various estimation window sizes. One could proceed in two ways. One way is to average the first $\underline{P}$ out-of-sample forecast losses, $\frac{1}{\underline{P}} \sum_{t=R}^{R+\underline{P}} \widehat{L}_{t+h}$. Alternatively, one can average the last $\underline{P}$ out-of-sample forecast losses, $\frac{1}{\underline{P}} \sum_{t=T-\underline{P}+1}^{T} \widehat{L}_{t+h}$. The difference is that in the latter case the out-of-sample forecasting period is the same despite the estimation window size differences, which allows for a fair comparison of the models over the same forecasting period.

Figures 1 and 2 consider the average out-of-sample forecasting performance of the economic model in eq. (14) relative to the benchmark, eq. (15). The horizontal axes report R.

The vertical axes report the ratio of the mean square forecast errors (MSFEs). Figure 1 depicts $MSFE^{Model}/MSFE^{RW}$, where $MSFE^{Model} = \frac{1}{\underline{P}} \sum_{t=R}^{R+\underline{P}} \epsilon_{1,t+h}^2(\hat{\alpha}_{t,R})$, and $MSFE^{RW} = \frac{1}{\underline{P}} \sum_{t=R}^{R+\underline{P}} \varepsilon_{2,t+h}^2$, whereas Figure 2 depicts the ratio of $MSFE^{Model} = \frac{1}{\underline{P}} \sum_{t=T-\underline{P}+1}^{T} \epsilon_{1,t+h}^2(\hat{\alpha}_{t,R})$ relative to $MSFE^{RW} = \frac{1}{\underline{P}} \sum_{t=T-\underline{P}+1}^{T} \varepsilon_{2,t+h}^2$. If the MSFE ratio is greater than one, then the economic model is performing worse than the benchmark on average. The figures show that the forecasting performance of the economic model is inferior to that of the random walk for all the countries except Canada. However, as the estimation window size increases, the forecasting performance of the models improves. The degree of improvement deteriorates when the models are compared over the same out-of-sample period, as shown in Figure 2.[13] For example, in the case of Japan, the model's average out-of-sample forecasting performance is similar to that of the random walk starting at $R = 110$ when compared over the same out-of-sample forecast periods, while otherwise (as shown in Figure 1), its performance becomes similar only for $R = 200$. Figure 1 reflects Bacchetta, van Wincoop and Beutler's (2010) findings that the poor performance of economic models is mostly attributed to over-fitting, as opposed to parameter instability. Figure 2 uncovers instead that the choice of the window is not crucial, so that over-fitting is a concern only when the window is too small.

Next, we decompose the difference between the MSFEs of the two models (14) and (15) calculated over rolling windows of size $m = 100$ into the components $A_{\tau,P}$, B_P, and U_P, as described in the Section 3. Negative MSFE differences imply that the economic model (14) is better than the benchmark model (15).[14] In addition to the relative forecasting performance of one-step ahead forecasts, we also consider one-year-ahead forecasts in a direct multistep forecasting exercise. More specifically, we consider the following "economic" and benchmark models:

$$\text{Model 1 - multistep: } y_{t+h} = \alpha(L)x_t + \epsilon_{1,t+h} \tag{16}$$

[13] Except for Germany, whose time series is heavily affected by the adoption of the Euro.

[14] The size of the window is chosen to strike a balance between the size of the out-of-sample period (P) and the total sample size in the databases of the various countries. To make the results comparable across countries, we keep the size of the window constant across countries and set it equal to $m = 100$, which results in a sequence of 186 out-of-sample rolling loss differentials for each country.

where y_{t+h} is the h period ahead rate of growth of the exchange rate at time t, defined by $y_{t+h} = [\ln(S_{t+h}/S_t)]/h$, and x_t is as defined previously. $\alpha(L)$ is a lag polynomial, such that $\alpha(L)x_t = \sum_{j=1}^{p} \alpha_j x_{t-j+1}$, where p is selected recursively by BIC, and $\epsilon_{1,t+h}$ is the h-step ahead error term. The benchmark model is the random walk:

$$\text{Model 2 - multistep: } y_{t+h} = \varepsilon_{2,t+h}, \tag{17}$$

where $\varepsilon_{2,t+h}$ is the h-step ahead error term. We focus on one-year ahead forecasts by setting $h = 12$ months.

Figure 3 plots the estimated values of $A_{\tau,P}$, B_P and U_P for the decomposition in Proposition (3) for one-step ahead forecasts (eqs. 14, 15) and Figure 4 plots the same decomposition for multi-step ahead forecasts (eqs. 16, 17). In addition, the first column of Table 6 reports the test statistics for assessing the significance of each of the three components, $\Gamma_P^{(A)}, \Gamma_P^{(B)}$ and $\Gamma_P^{(U)}$, as well as the Diebold and Mariano (1995) and West (1996) test, labeled "DMW".

Overall, according to the DMW test, there is almost no evidence that economic models forecast exchange rates significantly better than the random walk benchmark. This is the well-known "Meese and Rogoff puzzle". It is interesting, however, to look at our decomposition to understand the causes of the poor forecasting ability of the economic models. Figures 3 and 4 show empirical evidence of time variation in $A_{\tau,P}$, signaling possible instability in the relative forecasting performance of the models. Table 6 shows that such instability is statistically insignificant for one-month ahead forecasts across the countries. Instabilities, however, become statistically significant for one-year ahead forecasts in some countries. The figures uncover that the economic model was forecasting significantly better than the benchmark in the late 2000s for Canada and in the early 1990s for the U.K. For the one-month ahead forecasts, the B_P component is mostly positive, except for the case of Germany. In addition, the test statistic indicates that the component is statistically significant for Switzerland, Canada, Japan and Germany. We conclude that the lack of out-of-sample predictive ability

is related to lack of in-sample predictive content. For Germany, even though the predictive content is statistically significant, it is misleading (since $B_P < 0$). For the U.K. instead, the bad forecasting performance of the economic model is attributed to over-fit. As we move to one-year ahead forecasts, the evidence of predictive content becomes weaker. Interestingly, for Canada, there is statistical evidence in favor of predictive content when we forecast with the large economic model. We conclude that lack of predictive content is the major explanation for the lack of one-month ahead forecasting ability of the economic models, whereas time variation is mainly responsible for the lack of one-year ahead forecasting ability.

7 Conclusions

This paper proposes a new decomposition of measures of relative out-of-sample predictive ability into uncorrelated components related to instabilities, in-sample fit, and over-fitting. In addition, the paper provides tests for assessing the significance of each of the components. The methods proposed in this paper have the advantage of identifying the sources of a model's superior forecasting performance and might provide valuable information for improving forecasting models.

Acknowledgments

We thank the editors, two anonymous referees, and participants of the 2008 Midwest Econometrics Group, the EMSG at Duke University, and the 2010 NBER-NSF Time Series Conference for comments. Barbara Rossi gratefully acknowledges support by NSF grant 0647627. The views expressed in this paper are those of the authors. No responsibility should be attributed to the Bank of Canada.

Appendix A. Additional Theoretical Results

This Appendix contains theoretical propositions that extend the recursive window estimation results to situations where $\pi \neq 0$ or $D \neq 0$.

Proposition 5 (Asymptotic Results for the General Recursive Window Case) *For every $k \in [0,1]$, under Assumption 2:*
(a) If $\pi = 0$ then

$$\frac{1}{\sqrt{P}} \sum_{t=R}^{R+[kP]} \Omega_{rec}^{-1/2} \left[\widehat{l}_{t+h} - E\left(l_{t+h}\right) \right] \Rightarrow \mathcal{W}\left(k\right),$$

where $\Omega_{rec} = S_{ll}$; and
(b) If S is p.d. then

$$\frac{1}{\sqrt{P}} \sum_{t=R}^{R+[kP]} \Omega\left(k\right)_{rec}^{-1/2} \left[\widehat{l}_{t+h} - E\left(l_{t+h}\right) \right] \Rightarrow \mathcal{W}\left(k\right),$$

where

$$\Omega\left(k\right)_{rec} \equiv \begin{pmatrix} I & D \end{pmatrix} \begin{pmatrix} S_{ll} & \Upsilon S_{lh} J' \\ \Upsilon J S_{lh}' & 2\Upsilon J S_{hh} J' \end{pmatrix} \begin{pmatrix} I \\ D' \end{pmatrix}, \tag{18}$$

and $\mathcal{W}\left(.\right)$ is a (2×1) standard vector Brownian Motion, $\Upsilon \equiv 1 - \frac{\ln(1+k\pi)}{k\pi}$, $k\pi \in [0,\pi]$.

Comment to Proposition 5. Upon mild strengthening of the assumptions on h_t as in Andrews (1991), a consistent estimate of $\Omega\left(k\right)_{rec}$ can be obtained by

$$\begin{aligned} \widehat{\Omega}\left(k\right)_{rec} &= \begin{pmatrix} I & \widehat{D} \end{pmatrix} \begin{pmatrix} \widehat{S}_{ll} & \Upsilon \widehat{S}_{lh} \widehat{J}' \\ \Upsilon \widehat{J} \widehat{S}_{lh}' & 2\Upsilon \widehat{J} \widehat{S}_{hh} \widehat{J}' \end{pmatrix} \begin{pmatrix} I \\ \widehat{D}\ ' \end{pmatrix} \qquad (19) \\ \begin{pmatrix} \widehat{S}_{ll} & \widehat{S}_{lh} \\ \widehat{S}_{hl} & \widehat{S}_{hh} \end{pmatrix} &= \sum_{i=-q(P)+1}^{q(P)-1} \left(1 - |i/q(P)|\right) P^{-1} \sum_{t=R}^{T} \left(\varsigma_{t+h} - P^{-1} \sum_{t=R}^{T} \varsigma_{t+h} \right)^2, \end{aligned}$$

where $\varsigma_{t+h} = (\widehat{l}_{t+h}, h_t')'$, $\widehat{D} = P^{-1} \sum_{t=R}^{T} \frac{\partial l_{t+h}}{\partial \theta}|_{\theta=\widehat{\theta}_{t,R}}$, $\widehat{J} = J_T$, and $q(P)$ is a bandwidth that grows with P (Newey and West, 1987).

Proposition 6 (The Decomposition: General Expanding Window Estimation) *Let Assumptions 2,3,4 hold, $\lambda \in [\mu, 1]$, $\Omega\left(\lambda\right)_{(i,j),rec}^{-1/2}$ denote the (i-th,j-th) element of $\Omega\left(\lambda\right)_{rec}^{-1/2}$,*

defined in eq. 19, and for $\tau = [\lambda P]$*:*

$$\frac{1}{m}[\sum_{t=R}^{R+\tau-1} \Omega(\lambda)^{-1/2}_{(1,1),rec} \widehat{L}_{t+h} - \sum_{t=R}^{R+\tau-m-1} \Omega(\lambda-\mu)^{-1/2}_{(1,1),rec} \widehat{L}_{t+h}] = \left[\widetilde{A}_{\tau,P} - E(\widetilde{A}_{\tau,P})\right] + \left[\widetilde{B}_P - E(\widetilde{B}_P)\right] + \left[\widetilde{U}_P - E(\widetilde{U}_P)\right], \text{ for } \tau = m, m+1, ..., P, \quad (20)$$

where

$$\widetilde{A}_{\tau,P} \equiv \frac{1}{m}[\sum_{t=R}^{R+\tau-1} \Omega(\lambda)^{-1/2}_{(1,1),rec} \widehat{L}_{t+h} - \sum_{t=R}^{R+\tau-m-1} \Omega(\lambda-\mu)^{-1/2}_{(1,1),rec} \widehat{L}_{t+h}] - \frac{1}{P}\sum_{t=R}^{T} \Omega(1)^{-1/2}_{(1,1),rec} \widehat{L}_{t+h},$$

$\widetilde{B}_P \equiv \Omega(1)^{-1/2}_{(1,1),rec} B_P$*, and* $\widetilde{U}_P \equiv \Omega(1)^{-1/2}_{(1,1),rec} U_P$*. Under Assumption 5, where* $\overline{A}_{\tau,P} \equiv E(\widetilde{A}_{\tau,P})$*,* $\widetilde{A}_{\tau,P}, \widetilde{B}_P, \widetilde{U}_P$ *are asymptotically uncorrelated, and provide a decomposition of the rolling average (standardized) out-of-sample measure of forecasting performance, eq. (20).*
Comment. Note that, when $D = 0$*, so that estimation uncertainty is not relevant, (20) is the same as (10) because* $\Omega(\lambda)_{rec}$ *does not depend on* λ*.*

Proposition 7 (Significance Tests: Expanding Window Estimation) *Let Assumptions 2,3,4 hold,* $\widehat{\sigma}^2_A = \widehat{\Omega}(1)_{(1,1),rec}$*,* $\widehat{\sigma}^2_{A,\lambda} = \widehat{\Omega}(\lambda)_{(1,1),rec}$*,* $\widehat{\sigma}^2_B = \Phi^2_P \widehat{\Omega}(1)_{(2,2),rec}$*, for* $\widehat{\Omega}_{rec}$ *defined in eq.(19),* $\widehat{\sigma}^2_U = \widehat{\sigma}^2_A - \widehat{\sigma}^2_B$*,* $\widetilde{A}_{\tau,P} \equiv \frac{1}{m}[\sum_{t=R}^{R+\tau-1} \widehat{\sigma}^{-1}_{A,\lambda} \widehat{L}_{t+h} - \sum_{t=R}^{R+\tau-m-1} \widehat{\sigma}^{-1}_{A,\lambda-\mu} \widehat{L}_{t+h}] - \frac{1}{P}\sum_{t=R}^{T} \widehat{\sigma}^{-1}_A \widehat{L}_{t+h}$*,* $\widetilde{B}_P \equiv \widehat{\sigma}^{-1}_A B_P$*,* $\widetilde{U}_P \equiv \widehat{\sigma}^{-1}_A U_P$*, and the tests be defined as:*

$$\begin{aligned} \Gamma^{(A)}_P &\equiv \sup_{\tau=m,...,P} |\widetilde{A}_{\tau,P}|, \\ \Gamma^{(B)}_P &\equiv \sqrt{P}(\widehat{\sigma}_A/\widehat{\sigma}_B)\widetilde{B}_P \\ \Gamma^{(U)}_P &\equiv \sqrt{P}(\widehat{\sigma}_A/\widehat{\sigma}_U)\widetilde{U}_P. \end{aligned}$$

Then: (i) Under $H_{0,A}$*, where* $\overline{A}_{\tau,P} \equiv E(\widetilde{A}_{\tau,P}) : \Gamma^{(A)}_P \Rightarrow \sup_{\lambda\in[\mu,1]} \left|\frac{1}{\mu}[\mathcal{W}_1(\lambda) - \mathcal{W}_1(\lambda-\mu)] - \mathcal{W}_1(1)\right|$*, where* $\tau = [\lambda P]$*,* $m = [\mu P]$ *and* $\mathcal{W}_1(\cdot)$ *is a standard univariate Brownian motion. The critical values for significance level* α *are* $\pm k_\alpha$*, where* k_α *solves (12). Table 1 reports the critical values* k_α*.*
(ii) Under $H_{0,B} : \Gamma^{(B)}_P \Rightarrow N(0,1)$*; under* $H_{0,U} : \Gamma^{(U)}_P \Rightarrow N(0,1)$*.*

Appendix B. Proofs

This Appendix contains the proofs of the theoretical results in the paper.

Lemma 1 (The Mean Value Expansion) *Under Assumption 2, we have*

$$\frac{1}{\sqrt{P}}\sum_{t=R}^{R+[kP]}\widehat{l}_{t+h}=\frac{1}{\sqrt{P}}\sum_{t=R}^{R+[kP]}l_{t+h}+D\frac{1}{\sqrt{P}}\sum_{t=R}^{R+[kP]}JH_t+o_p\left(1\right). \tag{21}$$

Proof of Lemma 1. Consider the following mean value expansion of $\widehat{l}_{t+h}=\widehat{l}_{t+h}(\widehat{\theta}_{t,R})$ around θ^*:

$$\widehat{l}_{t+h}\left(\widehat{\theta}_{t,R}\right)=l_{t+h}+D_{t+h}\left(\widehat{\theta}_{t,R}-\theta^*\right)+r_{t+h}, \tag{22}$$

where the i-th element of r_{t+h} is: $0.5\left(\widehat{\theta}_{t,R}-\theta^*\right)'\left(\frac{\partial^2 l_{i,t+h}(\widetilde{\theta}_{t,R})}{\partial\theta\partial\theta'}\right)\left(\widehat{\theta}_{t,R}-\theta^*\right)$ and $\widetilde{\theta}_{t,R}$ is an intermediate point between $\widehat{\theta}_{t,R}$ and θ^*. From (22) we have

$$\frac{1}{\sqrt{P}}\sum_{t=R}^{R+[kP]}\widehat{l}_{t+h}=\frac{1}{\sqrt{P}}\sum_{t=R}^{R+[kP]}l_{t+h}+\frac{1}{\sqrt{P}}\sum_{t=R}^{R+[kP]}D_{t+h}\left(\widehat{\theta}_{t,R}-\theta^*\right)+\frac{1}{\sqrt{P}}\sum_{t=R}^{R+[kP]}r_{t+h}.$$

Eq. (21) follows from:

(a) $\frac{1}{\sqrt{P}}\sum_{t=R}^{R+[kP]}r_{t+h}=o_p\left(1\right)$ and

(b) $\frac{1}{\sqrt{P}}\sum_{t=R}^{R+[kP]}D_{t+h}\left(\widehat{\theta}_{t,R}-\theta^*\right)=D\frac{1}{\sqrt{P}}\sum_{t=R}^{R+[kP]}JH_t.$

(a) follows from Assumption 2(b) and Equation 4.1(b) in West (1996, p. 1081). To prove (b), note that by Assumption 2(c)

$$\frac{1}{\sqrt{P}}\sum_{t=R}^{R+[kP]}D_{t+h}\left(\widehat{\theta}_{t,R}-\theta^*\right)=\frac{1}{\sqrt{P}}\sum_{t=R}^{R+[kP]}D_{t+h}J_tH_t=\frac{1}{\sqrt{P}}\sum_{t=R}^{R+[kP]}DJH_t+$$
$$\frac{1}{\sqrt{P}}\sum_{t=R}^{R+[kP]}\left(D_{t+h}-D\right)JH_t+\frac{1}{\sqrt{P}}\sum_{t=R}^{R+[kP]}D\left(J_t-J\right)H_t+\frac{1}{\sqrt{P}}\sum_{t=R}^{R+[kP]}\left(D_{t+h}-D\right)\left(J_t-J\right)H_t.$$

We have that, in the last equality: (i) the second term is $o_p\left(1\right)$ as $\frac{1}{\sqrt{P}}\sum_{t=R}^{R+[kP]}\left(D_{t+h}-D\right)JH_t$ $=\frac{\sqrt{[kP]}}{\sqrt{P}}(\frac{1}{\sqrt{[kP]}}\sum_{t=R}^{R+[kP]}D_{t+h}-D)H_t\underset{p}{\rightarrow}0$ by Assumption 2(d), $\frac{\sqrt{[kP]}}{\sqrt{P}}=O\left(1\right)$ and Lemma A4(a) in West (1996). (ii) Similar arguments show that the third and fourth terms are $o_p\left(1\right)$ by Lemma A4(b,c) in West (1996). ■

Lemma 2 (Joint Asymptotic Variance of l_{t+h} and H_t) *For every $k \in [0,1]$,*

$$\lim_{T\to\infty} Var\left(\begin{bmatrix} \frac{1}{\sqrt{P}} \sum_{t=R}^{R+[kP]} l_{t+h} \\ \frac{1}{\sqrt{P}} \sum_{t=R}^{R+[kP]} J_t H_t \end{bmatrix}\right) = k \begin{pmatrix} S_{ll} & 0 \\ 0 & 0 \end{pmatrix} \text{ if } \pi = 0, \text{ and}$$

$$\lim_{T\to\infty} Var\left(\begin{bmatrix} \frac{1}{\sqrt{P}} \sum_{t=R}^{R+[kP]} l_{t+h} \\ \frac{1}{\sqrt{P}} \sum_{t=R}^{R+[kP]} J_t H_t \end{bmatrix}\right) = k \begin{pmatrix} S_{ll} & \Upsilon S_{lh} J' \\ \Upsilon J S_{lh}' & 2\Upsilon J S_{hh} J' \end{pmatrix} \text{ if } \pi > 0.$$

Proof of Lemma 2. We have,

$$(i) \lim_{T\to\infty} Var\left(\frac{1}{\sqrt{P}} \sum_{t=R}^{R+[kP]} J\ H_t\right) = \begin{cases} 0 \text{ if } \pi = 0 \\ k(2\Upsilon J S_{hh} J') \text{ if } \pi > 0, \end{cases}$$

where $\pi = 0$ case follows from Lemma A5 in West (1996). The result for $\pi > 0$ case follows from $\lim_{T\to\infty} Var\left(\frac{1}{\sqrt{[kP]}} \sum_{t=R}^{R+[kP]} H_t\right) = 2\left[1 - (k\pi)^{-1} \ln(1+k\pi)\right] S_{hh} = 2\Upsilon S_{hh}$, together with West (1996, Lemmas A2(b), A5) with P being replaced by $R + [kP]$ and $[kP]/P \underset{T\to\infty}{\to} k$.

Using similar arguments,

$$(ii)\ Var\left(\frac{1}{\sqrt{P}} \sum_{t=R}^{R+[kP]} l_{t+h}\right) = \frac{[kP]}{P} Var\left(\frac{1}{\sqrt{[kP]}} \sum_{t=R}^{R+[kP]} l_{t+h}\right) \underset{T\to\infty}{\to} kS_{ll};$$

$$(iii)\quad Cov\left(\frac{1}{\sqrt{P}} \sum_{t=R}^{R+[kP]} l_{t+h}, \frac{1}{\sqrt{P}} \sum_{t=R}^{R+[kP]} H_t\right) = \frac{[kP]}{P} Cov\left(\frac{1}{\sqrt{[kP]}} \sum_{t=R}^{R+[kP]} l_{t+h}, \frac{1}{\sqrt{[kP]}} \sum_{t=R}^{R+[kP]} H_t\right)$$
$$\underset{T\to\infty}{\to} \begin{cases} 0 \text{ if } \pi = 0 \text{ by West (1994)} \\ k\Upsilon S_{lh} \text{ if } \pi > 0 \text{ by Lemma A5 in West (1994)} \end{cases}$$

■

Lemma 3 (Asymptotic Variance of $\widehat{l}_{t+h}$) *For $k \in [0,1]$, and $\Omega(k)_{rec}$ defined in eq. (19):*

$$(b) \lim_{T\to\infty} Var\left(\frac{1}{\sqrt{P}}\sum_{t=R}^{R+[kP]}\widehat{l}_{t+h}\right) = kS_{ll} \text{ if } \pi = 0, \text{ and}$$

$$(b) \lim_{T\to\infty} Var\left(\frac{1}{\sqrt{P}}\sum_{t=R}^{R+[kP]}\widehat{l}_{t+h}\right) = k\Omega(k)_{rec} \text{ if } \pi > 0.$$

Proof of Lemma 3. From Lemma 1, we have

$$\frac{1}{\sqrt{P}}\sum_{t=R}^{R+[kP]}\left[\widehat{l}_{t+h} - E(l_{t+h})\right] = \frac{1}{\sqrt{P}}\sum_{t=R}^{R+[kP]}\left[l_{t+h} - E(l_{t+h})\right] + D\frac{1}{\sqrt{P}}\sum_{t=R}^{R+[kP]} J\ H_t\ + o_p(1).$$

That the variance is as indicated above follows from Lemma 2 and $\lim_{T\to\infty} Var\left(\frac{1}{\sqrt{P}}\sum_{t=R}^{R+[kP]}\widehat{l}_{t+h}\right)$ $= \begin{pmatrix} I & D \end{pmatrix} \lim_{T\to\infty} Var\begin{pmatrix} \frac{1}{\sqrt{P}}\sum_{t=R}^{R+[kP]} l_{t+h} \\ \frac{1}{\sqrt{P}}\sum_{t=R}^{R+[kP]} JH_t \end{pmatrix}\begin{pmatrix} I \\ D' \end{pmatrix}$, which equals $k\Omega(k)_{rec}$ if $\pi > 0$, and equals kS_{ll} if $\pi = 0$. ■

Proof of Proposition 1. Since $Z_{t+h,R}$ is a measurable function of $\widehat{\theta}_{t,R}$, which includes only a finite (R) number of lags (leads) of $\{y_t, x_t\}$, and $\{y_t, x_t\}$ are mixing, $Z_{t+h,R}$ is also mixing, of the same size as $\{y_t, x_t\}$. Then $W_P \Rightarrow \mathcal{W}$ by Corollary 4.2 in Wooldridge and White (1988). ■

Proof of Propositions 2 and 5. Let $m_{t+h}(k) = \Omega(k)_{rec}^{-1/2}\left[\widehat{l}_{t+h} - E(l_{t+h})\right]$ if $\pi > 0$ (where $\Omega(k)_{rec}$ is positive definite since S is positive definite), and $m_{t+h}(k) = S_{ll}^{-1/2}[\widehat{l}_{t+h} - E(l_{t+h})]$ if $\pi = 0$ or $D = 0$. That $\frac{1}{\sqrt{P}}\sum_{t=R}^{R+[kP]} m_{t+h}(k) = \frac{1}{\sqrt{P}}\sum_{s=1}^{[kP]} m_{s+R+h}(k)$ satisfies Assumption D.3 in Wooldridge and White (1989) follows from Lemma 3. The limiting variance also follows from Lemma 3, and is full rank by Assumption 2(d). Weak convergence of the standardized partial sum process then follows from Corollary 4.2 in Wooldridge and White (1989). The convergence can be converted to uniform convergence as follows: first, an argument similar to that in Andrews (1993, p. 849, lines 4-18) ensures that Assumption (i) in Lemma A4 in Andrews (1993) holds; then, Corollary 3.1 in Wooldridge and White (1989) can be used to show that $m_{t+h}(k)$ satisfies assumption (ii) in Lemma A4 in Andrews (1993). Uniform convergence then follows by Lemma A4 in Andrews (1993). ■

Proof of Proposition 3. Let $\mathcal{W}(.) = [\mathcal{W}_1(.), \mathcal{W}_2(.)]'$ denote a two-dimensional vector of independent standard univariate Brownian Motions.
(a) For the fixed rolling window estimation, let $\Omega_{(i,j),roll}$ denote the i-th row and j-th column element of Ω_{roll}, and let $\mathcal{B}(.) \equiv \Omega_{roll}^{1/2}\mathcal{W}(.)$. Note that

$$\frac{1}{m}\sum_{t=R+\tau-m}^{R+\tau-1}\widehat{L}_{t+h} = \left(\frac{1}{m}\sum_{t=R+\tau-m}^{R+\tau-1}\widehat{L}_{t+h} - \frac{1}{P}\sum_{t=R}^{T}\widehat{L}_{t+h}\right) + \frac{1}{P}\sum_{t=R}^{T}\widehat{L}_{t+h}, \text{ for } \tau = m, .., P,$$

where, from Proposition 1 and Assumptions 3,5: $\frac{1}{\sqrt{P}}\sum_{t=R}^{T}(\widehat{L}_{t+h} - \overline{L}_{t+h}) \Rightarrow \mathcal{B}_1(1)$, and

$$\sqrt{P}A_{\tau,P} = \frac{P}{m}\frac{1}{\sqrt{P}}\sum_{t=R+\tau-m}^{R+\tau-1}(\widehat{L}_{t+h}-\overline{L}_{t+h}) - \frac{1}{\sqrt{P}}\sum_{t=R}^{T}(\widehat{L}_{t+h}-\overline{L}_{t+h}) \Rightarrow \frac{1}{\mu}\left[\mathcal{B}_1(\lambda) - \mathcal{B}_1(\lambda-\mu)\right] - \mathcal{B}_1(1),$$

for μ defined in Assumption 3. By construction, B_P and U_P are asymptotically uncorrelated under either $H_{0,B}$ or $H_{0,U}$,[15] and $\frac{1}{P}\sum_{t=R}^{T}\widehat{L}_{t+h} = B_P + U_P$.
Note that $B_P = \widehat{\beta}\frac{1}{P}\sum_{t=R}^{T}\widehat{\mathcal{L}}_t = \left(\widehat{\beta}-\beta\right)\frac{1}{P}\sum_{t=R}^{T}\widehat{\mathcal{L}}_t + \beta\frac{1}{P}\sum_{t=R}^{T}\widehat{\mathcal{L}}_t$. Thus, under $H_{0,B}$:

$$\begin{aligned} B_P - \overline{B}_P &= \left(\widehat{\beta}-\beta\right)\frac{1}{P}\sum_{t=R}^{T}\widehat{\mathcal{L}}_t + \beta\left(\frac{1}{P}\sum_{t=R}^{T}\widehat{\mathcal{L}}_t\right) \\ &= \left(\frac{1}{P}\sum_{t=R}^{T}\widehat{\mathcal{L}}_t^2\right)^{-1}\left(\frac{1}{P}\sum_{t=R}^{T}\widehat{\mathcal{L}}_t\left[\widehat{L}_{t+h} - \beta\widehat{\mathcal{L}}_t\right]\right)\left(\frac{1}{P}\sum_{t=R}^{T}\widehat{\mathcal{L}}_t\right) + \beta\left(\frac{1}{P}\sum_{t=R}^{T}\widehat{\mathcal{L}}_t\right) \\ &= \Phi_P\left(\frac{1}{P}\sum_{t=R}^{T}\widehat{\mathcal{L}}_t\widehat{L}_{t+h}\right), \end{aligned} \tag{23}$$

where the last equality follows from the fact that $H_{0,B}$: $\beta\frac{1}{P}\sum_{t=R}^{T}\overline{\mathcal{L}}_t = 0$ and Assumption 4(b) imply $\beta = 0$, and $\Phi_P \equiv (\frac{1}{P}\sum_{t=R}^{T}\widehat{\mathcal{L}}_t)(\frac{1}{P}\sum_{t=R}^{T}\widehat{\mathcal{L}}_t^2)^{-1}$. Note that $U_P = \frac{1}{P}\sum_{t=R}^{T}\widehat{L}_{t+h} - B_P$, thus

[15] Let $\mathcal{L} = [\mathcal{L}_R, ..., \mathcal{L}_T]'$, $L = [L_{R+h}, ..., L_{T+h}]'$, $\mathcal{P}_P = \mathcal{L}(\mathcal{L}'\mathcal{L})^{-1}\mathcal{L}'$, $\mathcal{M}_P \equiv I - \mathcal{P}_P$, $\mathbf{B} = \mathcal{P}_P L$, $\boldsymbol{U} = \mathcal{M}_P L$. Then, $B_P = P^{-1}\iota'\mathbf{B}$ and $U_P = P^{-1}\iota'\boldsymbol{U}$, where ι is a $(P \times 1)$ vector of ones. Note that, under either $H_{0,B}$ or $H_{0,U}$, either B_P or U_P have zero mean. Also note that $Cov(B_P, U_P) = E(B_P'U_P) = P^{-2}E\left(\boldsymbol{B}'\iota\ \iota'\boldsymbol{U}\right) = P^{-1}E\left(\boldsymbol{B}'\boldsymbol{U}\right) = P^{-1}E\left(L_{t+h}'\mathcal{P}_P\mathcal{M}_P L_{t+h}\right) = 0$. Therefore, B_P and U_P are asymptotically uncorrelated.

$U_P - \overline{U}_P = \frac{1}{P}\sum_{t=R}^{T}\left(\widehat{L}_{t+h} - \overline{L}_{t+h}\right) - \left(B_P - \overline{B}_P\right)$. Then, from Assumptions 1,5, we have

$$P^{1/2}\begin{pmatrix} A_{\tau,P} - \overline{A}_{\tau,P} \\ B_P - \overline{B}_P \\ U_P - \overline{U}_P \end{pmatrix} = \begin{pmatrix} \frac{P}{m}\frac{1}{\sqrt{P}}\sum_{t=R+\tau-m}^{R+\tau-1}\left(\widehat{L}_{t+h} - \overline{L}_{t+h}\right) - \frac{1}{\sqrt{P}}\sum_{t=R}^{T}\left(\widehat{L}_{t+h} - \overline{L}_{t+h}\right) \\ \begin{bmatrix} 0 & \Phi_P \\ 1 & -\Phi_P \end{bmatrix}\begin{bmatrix} \frac{1}{\sqrt{P}}\sum_{t=R}^{T}\left(\widehat{L}_{t+h} - \overline{L}_{t+h}\right) \\ \frac{1}{\sqrt{P}}\sum_{t=R}^{T}\widehat{\mathcal{L}}_t\widehat{L}_{t+h} \end{bmatrix} \end{pmatrix}$$

$$\Rightarrow \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & \Phi \\ 0 & 1 & -\Phi \end{pmatrix}\begin{pmatrix} \frac{1}{\mu}\left[\mathcal{B}_1(\lambda) - \mathcal{B}_1(\lambda-\mu)\right] - \mathcal{B}_1(1) \\ \mathcal{B}_1(1) \\ \mathcal{B}_2(1) \end{pmatrix}, \tag{24}$$

where $\frac{1}{P}\sum_{t=R}^{T}\widehat{\mathcal{L}}_t \underset{p}{\to} \lim_{T\to\infty}\frac{1}{P}\sum_{t=R}^{T}\overline{\mathcal{L}}_t \neq 0$ by Assumption 4(b), and $\Phi_P \underset{p}{\to} \Phi$ by Assumption 1, where $\Phi \equiv \left(\lim_{T\to\infty}\frac{1}{P}\sum_{t=R}^{T}\overline{\mathcal{L}}_t\right)\left(\lim_{T\to\infty}\frac{1}{P}\sum_{t=R}^{T}E\widehat{\mathcal{L}}_t^2\right)^{-1}$. Note that

$$\begin{aligned} & Cov\left(\frac{1}{\mu}\left[\mathcal{B}_1(\lambda) - \mathcal{B}_1(\lambda-\mu)\right] - \mathcal{B}_1(1), \Phi\mathcal{B}_2(1)\right) \\ = \; & Cov\left(\frac{1}{\mu}\mathcal{B}_1(\lambda), \Phi\mathcal{B}_2(1)\right) - Cov\left(\frac{1}{\mu}\mathcal{B}_1(\lambda-\mu), \Phi\mathcal{B}_2(1)\right) - \Omega_{(1,2),roll}\Phi \\ = \; & \Omega_{(1,2),roll}\Phi\left(\frac{\lambda}{\mu} - \frac{\lambda-\mu}{\mu} - 1\right) = 0. \end{aligned}$$

It follows that $[A_{\tau,P} - \overline{A}_{\tau,P}]$ and $[B_P - \overline{B}_P]$ are asymptotically uncorrelated. A similar proof shows that $[A_{\tau,P} - \overline{A}_{\tau,P}]$ and $[U_P - \overline{U}_P]$ are asymptotically uncorrelated:

$$\begin{aligned} & Cov\left(\frac{1}{\mu}\left[\mathcal{B}_1(\lambda) - \mathcal{B}_1(\lambda-\mu)\right] - \mathcal{B}_1(1), \mathcal{B}_1(1) - \Phi\mathcal{B}_2(1)\right) \\ = \; & Cov\left(\frac{1}{\mu}\mathcal{B}_1(\lambda), \mathcal{B}_1(1)\right) - Cov\left(\frac{1}{\mu}\mathcal{B}_1(\lambda-\mu), \mathcal{B}_1(1)\right) - \Omega_{(1,1),roll} \\ & -Cov\left(\frac{1}{\mu}\mathcal{B}_1(\lambda), \Phi\mathcal{B}_2(1)\right) + Cov\left(\frac{1}{\mu}\mathcal{B}_1(\lambda-\mu), \Phi\mathcal{B}_2(1)\right) + \Omega_{(1,2),roll}\Phi \\ = \; & \Omega_{(1,1),roll}\left(\frac{\lambda}{\mu} - \frac{\lambda-\mu}{\mu} - 1\right) - \Omega_{(1,2),roll}\Phi\left(\frac{\lambda}{\mu} - \frac{\lambda-\mu}{\mu} - 1\right) = 0. \end{aligned}$$

Thus, $\sup_{\tau=m,..,P}[A_{\tau,P} - \overline{A}_{\tau,P}]$ is asymptotically uncorrelated with $B_P - \overline{B}_P$ and $U_P - \overline{U}_P$.
(b) For the recursive window estimation, the results follows directly from the proof of Proposition (6) by noting that, when $D = 0$, $\Omega(\lambda)_{rec}$ is independent of λ. Thus (20) is exactly the

same as (10). ■

Proof of Proposition 4. (a) For the fixed rolling window estimation case, we first show that $\widehat{\sigma}_A^2$, $\widehat{\sigma}_B^2$, $\widehat{\sigma}_U^2$ are consistent estimators for σ_A^2, σ_B^2, σ_U^2. From (23), it follows that $\sigma_B^2 \equiv \lim_{T\to\infty} Var\left(P^{1/2}B_P\right) = \lim_{T\to\infty} \Phi_P^2 Var(\frac{1}{\sqrt{P}}\sum_{t=R}^{T}\widehat{\mathcal{L}}_t\widehat{L}_{t+h}) = \Phi^2 \lim_{T\to\infty} Var(\frac{1}{\sqrt{P}}\sum_{t=R}^{T}\widehat{\mathcal{L}}_t\widehat{L}_{t+h})$, which can be consistently estimated by $\widehat{\sigma}_B^2 = \Phi_P^2\widehat{\Omega}_{(2,2),roll}$. Also, the proof of Proposition 3 ensures that, under either $H_{0,B}$ or $H_{0,U}$, B_P and U_P are uncorrelated. Then, $\widehat{\sigma}_U^2 = \widehat{\sigma}_A^2 - \widehat{\sigma}_B^2$ is a consistent estimate of $Var\left(P^{1/2}U_P\right)$. Then, consistency of $\widehat{\sigma}_A^2$, $\widehat{\sigma}_B^2$, $\widehat{\sigma}_U^2$ follows from standard arguments (see Newey and West, 1987). Part (i) follows directly from the proof of Proposition 3, eq. (24). Part (ii) follows from $\sqrt{P}B_P \Rightarrow \sigma_B\mathcal{W}_2(1)$ and thus $\sqrt{P}B_P/\widehat{\sigma}_B \Rightarrow \mathcal{W}_2(1) \equiv N(0,1)$. Similar arguments apply for U_P.

(b) For the recursive window estimation case, the result follows similarly. ■

Proof of Proposition 6. Consistent with the definition in the proof of Propositions 2 and 5, let $m_{t+h}(k) = \Omega(k)_{rec}^{-1/2}[\widehat{l}_{t+h} - E(l_{t+h})]$ and $M_{t+h}(k) = \Omega(k)_{(1,1),rec}^{-1/2}[\widehat{L}_{t+h} - \overline{L}_{t+h}]$.

Proposition 5 and Assumptions 3,5, where $\overline{A}_{\tau,P} \equiv E(\widetilde{A}_{\tau,P})$, imply

$$\begin{bmatrix} \frac{P}{m}\left(\frac{1}{\sqrt{P}}\sum_{t=R}^{R+\tau-1} m_{t+h}(\lambda) - \frac{1}{\sqrt{P}}\sum_{t=R}^{R+\tau-m-1} m_{t+h}(\lambda-\mu)\right) - \frac{1}{\sqrt{P}}\sum_{t=R}^{T} m_{t+h}(1) \\ \frac{1}{\sqrt{P}}\sum_{t=R}^{T} m_{t+h}(1) \end{bmatrix}$$
$$\Rightarrow \begin{bmatrix} \frac{1}{\mu}(\mathcal{W}(\lambda) - \mathcal{W}(\lambda-\mu)) - \mathcal{W}(1) \\ \mathcal{W}(1) \end{bmatrix}.$$

Assumptions 4(b),5, where $\overline{A}_{\tau,P} \equiv E(\widetilde{A}_{\tau,P})$, and eqs. (20), (23), (24) imply:

$$P^{1/2}\begin{bmatrix} \widetilde{A}_{\tau,P} - E(\widetilde{A}_{\tau,P}) \\ \widetilde{B}_P - E(\widetilde{B}_P) \\ \widetilde{U}_P - E(\widetilde{U}_P) \end{bmatrix} = \begin{bmatrix} \frac{P}{m\sqrt{P}}\left(\sum_{t=R}^{R+\tau-1} M_{t+h}(\lambda) - \sum_{t=R}^{R+\tau-m-1} M_{t+h}(\lambda-\mu)\right) - \frac{1}{\sqrt{P}}\sum_{t=R}^{T} M_{t+h}(1) \\ \Omega(1)_{(1,1),rec}^{-1/2}\begin{pmatrix} 0 & \Phi_P \\ 1 & -\Phi_P \end{pmatrix}\begin{pmatrix} \frac{1}{\sqrt{P}}\sum_{t=R}^{T}(\widehat{L}_{t+h} - \overline{L}_{t+h}) \\ \frac{1}{\sqrt{P}}\sum_{t=R}^{T}\widehat{\mathcal{L}}_t\widehat{L}_{t+h} \end{pmatrix} \end{bmatrix}$$
$$\Rightarrow \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & \Phi \\ 0 & 1 & -\Phi \end{pmatrix}\begin{pmatrix} \frac{1}{\mu}\left[\Omega(\lambda)_{(1,1),rec}^{-1/2}\mathcal{B}_1(\lambda) - \Omega(\lambda-\mu)_{(1,1),rec}^{-1/2}\mathcal{B}_1(\lambda-\mu)\right] - \Omega(1)_{(1,1),rec}^{-1/2}\mathcal{B}_1(1) \\ \Omega(1)_{(1,1),rec}^{-1/2}\mathcal{B}_1(1) \\ \Omega(1)_{(1,1),rec}^{-1/2}\mathcal{B}_2(1) \end{pmatrix}.$$

since $\Phi_P \equiv (\frac{1}{P}\sum_{t=R}^{T}\widehat{\mathcal{L}}_t)(\frac{1}{P}\sum_{t=R}^{T}\widehat{\mathcal{L}}_t^2)^{-1} \underset{p}{\to} \Phi \equiv \left(\lim_{T\to\infty}\frac{1}{P}\sum_{t=R}^{T}\overline{\mathcal{L}}_t\right)\left(\lim_{T\to\infty}\frac{1}{P}\sum_{t=R}^{T}E\widehat{\mathcal{L}}_t^2\right)^{-1}$.

Also, note that

$$Cov\{\frac{1}{\mu}\left[\Omega(\lambda)^{-1/2}_{(1,1),rec}\mathcal{B}_1(\lambda)-\Omega(\lambda-\mu)^{-1/2}_{(1,1),rec}\mathcal{B}_1(\lambda-\mu)\right]-\Omega(1)^{-1/2}_{(1,1),rec}\mathcal{B}_1(1),\Phi\Omega(1)^{-1/2}_{(1,1),rec}\mathcal{B}_2(1)\}=$$

$$Cov\{\frac{1}{\mu}\mathcal{W}_1(\lambda),\Phi\Omega(1)^{-1/2}_{(1,1),rec}\Omega(1)^{-1/2}_{(2,2),rec}\mathcal{W}_2(1)\}-Cov\{\frac{1}{\mu}\mathcal{W}_1(\lambda-\mu)\Phi\Omega(1)^{-1/2}_{(1,1),rec}\Omega(1)^{-1/2}_{(2,2),rec}\mathcal{W}_2(1)\}$$

$$-Cov\{\mathcal{W}_1(1),\Phi\Omega(1)^{-1/2}_{(1,1),rec}\Omega(1)^{-1/2}_{(2,2),rec}\mathcal{W}_2(1)\}=\Phi(\frac{\lambda}{\mu}-\frac{\lambda-\mu}{\mu}-1)\Omega(\lambda)_{(1,2),rec}\Omega(\lambda)^{-1/2}_{(1,1),rec}\Omega(\lambda)^{-1/2}_{(2,2),rec}.$$

It follows that $\widetilde{A}_{\tau,P}-E(\widetilde{A}_{\tau,P})$ is asymptotically uncorrelated with $\widetilde{B}_P-E(\widetilde{B}_P)$. Similar arguments to those in the proof of Proposition 3 establish that $\widetilde{A}_{\tau,P}-E(\widetilde{A}_{\tau,P})$ is asymptotically uncorrelated with $\widetilde{U}_P-E(\widetilde{U}_P)$. ■

Proof of Proposition 7. The proof follows from Proposition 6 and arguments similar to those in the proof of Proposition 4. ■

Appendix C. Data Description

1. Exchange rates. We use the bilateral end-of-period exchange rates for the Swiss franc (CHF), Canadian dollar (CAD), and Japanese yen (JPY). We use the bilateral end-of-period exchange rates for the German mark (DEM - EUR) using the fixed conversion factor adjusted euro rates after 1999. The conversion factor is 1.95583 marks per euro. For the British pound (GBP), we use the U.S. dollar per British pound rate to construct the British pound to U.S. dollar rate. The series are taken from IFS and correspond to lines "146..AE.ZF..." for CHF, "112..AG.ZF..." for BP, "156..AE.ZF..." for CAD, "158..AE.ZF..." for JPY and "134..AE.ZF..." for DEM, and "163..AE.ZF..." for the EUR.

2. Money supply. The money supply data for U.S., Japan, and Germany are measured in seasonally adjusted values of M1 (IFS line items 11159MACZF..., 15859MACZF..., and 13459MACZF... accordingly). The seasonally adjusted value for the M1 money supply for the Euro Area is taken from the Eurostat and used as a value for Germany after 1999. The money supply for Canada is the seasonally adjusted value of the Narrow Money (M1) Index from the OECD Main Economic Indicators (MEI). Money supply for UK is measured in the seasonally adjusted series of the Average Total Sterling notes taken from the Bank of England. We use the IFS line item "14634...ZF..." as a money supply value for Switzerland. The latter is not seasonally adjusted and we seasonally adjust the data using monthly dummies.

3. Industrial production. We use the seasonally adjusted value of the industrial production index taken from IFS and it corresponds to the line items "11166..CZF...", "14666..BZF...", "11266..CZF...", "15666..CZF...", "15866..CZF...", and "13466..CZF..." for the U.S., Switzerland, United Kingdom, Canada, Japan, and Germany correspondingly.

4. Unemployment rate. The unemployment rate corresponds to the seasonally adjusted value of the "Harmonised Unemployment Rate" taken from the OECD Main Economic Indicators for all countries except Germany. For Germany we use the value from Datastream (mnemonic WGUN%TOTQ) that covers the unemployment rate of West Germany only over time.

5. Interest rates. The interest rates are taken from IFS and correspond to line items "11160B..ZF...", "14660B..ZF...", "11260B..ZF...", "15660B..ZF...", "15860B..ZF...", and "13460B..ZF..." for the U.S., Switzerland, United Kingdom, Canada, Japan, and Germany correspondingly.

6. Commodity prices. The average crude oil price is taken from IFS line item "00176AAZZF...". Country-specific "Total, all commodities" index for Canada is from CANSIM database.

References

[1] Andrews, D.W.K., 1991, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation. Econometrica 59(3), 817-858.

[2] Andrews, D.W.K., 1993, Tests for Parameter Instability and Structural Change with Unknown Change Point. Econometrica 61(4), 821-856.

[3] Bacchetta, P., E. van Wincoop and T. Beutler, 2010, Can Parameter Instability Explain the Meese-Rogoff Puzzle? in: L. Reichlin, K.D. West, (Eds.). NBER International Seminar on Macroeconomics 2009. University of Chicago Press, Chicago, pp. 125-173.

[4] Clark, T.E. and M.W. McCracken, 2001, Tests of Equal Forecast Accuracy and Encompassing for Nested Models. Journal of Econometrics 105(1), 85-110.

[5] Clark, T.E. and M.W. McCracken, 2005, The Power of Tests of Predictive Ability in the Presence of Structural Breaks. Journal of Econometrics 124(1), 1-31.

[6] Clark, T.E. and M.W. McCracken, 2006, The Predictive Content of the Output Gap for Inflation: Resolving In-Sample and Out-of-Sample Evidence. Journal of Money, Credit & Banking 38(5), 1127-1148.

[7] Clark, T.E. and K.D. West, 2006, Using Out-of-sample Mean Squared Prediction Errors to Test the Martingale Difference Hypothesis. Journal of Econometrics 135(1-2), 155-186.

[8] Clark, T.E. and K.D. West, 2007, Approximately Normal Tests for Equal Predictive Accuracy in Nested Models, Journal of Econometrics 138(1), 291-311.

[9] Diebold, F.X. and R. S. Mariano, 1995, Comparing Predictive Accuracy. Journal of Business and Economic Statistics 13(3), 253-263.

[10] Elliott, G. and A. Timmermann, 2008, Economic Forecasting. Journal of Economic Literature 46(1), 3-56.

[11] Engel, C., N.C. Mark and K.D. West, 2008, Exchange Rate Models Are Not as Bad as You Think. NBER Macroeconomics Annual 2007(22), 381-441.

[12] Giacomini, R. and B. Rossi, 2009, Detecting and Predicting Forecast Breakdowns. Review of Economic Studies 76(2), 669-705.

[13] Giacomini, R. and B. Rossi, 2010, Forecast Comparisons in Unstable Environments. Journal of Applied Econometrics 25(4), 595-620.

[14] Giacomini, R. and H. White, 2006, Tests of Conditional Predictive Ability. Econometrica 74(6), 1545-1578.

[15] Hansen, P.R., 2008, In-Sample Fit and Out-of-Sample Fit: Their Joint Distribution and Its Implications for Model Selection. Mimeo.

[16] McCracken, M., 2000, Robust Out-of-Sample Inference. Journal of Econometrics 99(2), 195-223.

[17] Meese, R.A. and K. Rogoff, 1983a, Empirical Exchange Rate Models of the Seventies: Do They Fit Out Of Sample? Journal of International Economics 14, 3-24.

[18] Meese, R.A. and K. Rogoff, 1983b, The Out-Of-Sample Failure of Empirical Exchange Rate Models: Sampling Error or Misspecification? in: J.A. Frenkel, (Eds.), Exchange Rates and International Macroeconomics. University of Chicago Press, Chicago, pp. 67-112.

[19] Mincer, J.A. and V. Zarnowitz, 1969, The Evaluation of Economic Forecasts. in: J.A. Mincer, (Eds.), Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance. NBER, pp. 1-46.

[20] Newey, W.K. and K.D. West, 1987, A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. Econometrica 55(3), 703-708.

[21] West, K. D., 1994, Asymptotic Inference about Predictive Ability, An Additional Appendix. Mimeo.

[22] West, K. D., 1996, Asymptotic Inference about Predictive Ability. Econometrica 64(5), 1067-1084.

[23] West, K.D. and M.W. McCracken, 1998, Regression-Based Tests of Predictive Ability. International Economic Review 39(4), 817-840.

[24] Wooldridge, J.M. and H. White, 1988, Some Invariance Principles and Central Limit Theorems for Dependent Heterogeneous Processes. Econometric Theory 4(2), 210-230.

Tables and Figures

Table 1. Critical Values for the $\Gamma_P^{(A)}$ Test

μ	α :	0.10	0.05	0.025	0.01	μ	α :	0.10	0.05	0.025	0.01
0.10		9.844	10.496	11.094	11.858	0.55		2.449	2.700	2.913	3.175
0.15		7.510	8.087	8.612	9.197	0.60		2.196	2.412	2.615	2.842
0.20		6.122	6.609	7.026	7.542	0.65		1.958	2.153	2.343	2.550
0.25		5.141	5.594	5.992	6.528	0.70		1.720	1.900	2.060	2.259
0.30		4.449	4.842	5.182	5.612	0.75		1.503	1.655	1.804	1.961
0.35		3.855	4.212	4.554	4.941	0.80		1.305	1.446	1.575	1.719
0.40		3.405	3.738	4.035	4.376	0.85		1.075	1.192	1.290	1.408
0.45		3.034	3.333	3.602	3.935	0.90		0.853	0.952	1.027	1.131
0.50		2.729	2.984	3.226	3.514						

Note. The table reports critical values k_α for the test statistic $\Gamma_P^{(A)}$ at significance levels $\alpha = 0.10, 0.05, 0.025, 0.01$.

Table 2. DGP 1: Size Results

		Panel A. IID			*Panel B. Serial Corr.*		
R	P	$\Gamma_P^{(A)}$	$\Gamma_P^{(B)}$	$\Gamma_P^{(U)}$	$\Gamma_P^{(A)}$	$\Gamma_P^{(B)}$	$\Gamma_P^{(U)}$
20	150	0.02	0.07	0.03	0.02	0.08	0.04
	200	0.01	0.06	0.03	0.01	0.08	0.04
	300	0.01	0.05	0.03	0.01	0.07	0.03
50	150	0.03	0.07	0.03	0.02	0.08	0.04
	200	0.02	0.06	0.04	0.01	0.08	0.04
	300	0.03	0.06	0.03	0.02	0.08	0.03
100	150	0.04	0.06	0.05	0.03	0.09	0.04
	200	0.03	0.07	0.04	0.02	0.08	0.04
	300	0.03	0.06	0.04	0.03	0.07	0.04
200	150	0.04	0.06	0.06	0.03	0.08	0.05
	200	0.03	0.06	0.05	0.03	0.08	0.05
	300	0.04	0.06	0.04	0.04	0.07	0.04

Note. The table reports empirical rejection frequencies of the test statistics $\Gamma_P^{(A)}$, $\Gamma_P^{(B)}$, $\Gamma_P^{(U)}$ for various window and sample sizes (see DGP 1 in Section 5 for details). $m = 100$. Nominal size is 0.05.

Table 3. DGP 2: Time Variation Case

	Panel A. IID			*Panel B. Serial Corr.*		
b	$\Gamma_P^{(A)}$	$\Gamma_P^{(B)}$	$\Gamma_P^{(U)}$	$\Gamma_P^{(A)}$	$\Gamma_P^{(B)}$	$\Gamma_P^{(U)}$
0	0.07	0.06	0.04	0.06	0.07	0.04
0.1	0.06	0.06	0.05	0.06	0.07	0.04
0.2	0.07	0.07	0.05	0.06	0.07	0.05
0.3	0.08	0.07	0.06	0.06	0.07	0.05
0.4	0.10	0.07	0.06	0.06	0.07	0.06
0.5	0.14	0.07	0.07	0.07	0.08	0.06
0.6	0.19	0.07	0.07	0.09	0.08	0.06
0.7	0.23	0.07	0.08	0.10	0.08	0.06
0.8	0.28	0.07	0.08	0.13	0.08	0.07
0.9	0.35	0.07	0.09	0.15	0.08	0.07
0.1	0.41	0.07	0.09	0.18	0.08	0.07

Note. The table reports empirical rejection frequencies of the test statistics $\Gamma_P^{(A)}$, $\Gamma_P^{(B)}$, $\Gamma_P^{(U)}$ in the presence of time variation in the relative performance (see DGP 2 in Section 5). $R = 100$, $P = 300$, $m = 60$. The nominal size is 0.05.

Table 4. DGP 3: Stronger Predictive Content Case

	Panel A. IID			*Panel B. Serial Corr.*		
b	$\Gamma_P^{(A)}$	$\Gamma_P^{(B)}$	$\Gamma_P^{(U)}$	$\Gamma_P^{(A)}$	$\Gamma_P^{(B)}$	$\Gamma_P^{(U)}$
0	0.03	0.06	0.04	0.03	0.07	0.04
0.1	0.05	0.06	0.11	0.04	0.07	0.05
0.2	0.05	0.07	0.58	0.04	0.07	0.20
0.3	0.04	0.10	0.94	0.04	0.07	0.50
0.4	0.04	0.16	1	0.04	0.08	0.79
0.5	0.04	0.29	1	0.04	0.11	0.95
0.6	0.04	0.47	1	0.04	0.14	0.99
0.7	0.04	0.68	1	0.04	0.18	1
0.8	0.04	0.86	1	0.04	0.24	1
0.9	0.04	0.96	1	0.04	0.32	1
1	0.04	0.99	1	0.04	0.40	1

Note. The table reports empirical rejection frequencies of the test statistics $\Gamma_P^{(A)}$, $\Gamma_P^{(B)}$, $\Gamma_P^{(U)}$ in the case of an increasingly stronger predictive content of the explanatory variable (see DGP 3 in Section 5). $R = 100$, $P = 300$, $m = 100$. Nominal size is 0.05.

Table 5. DGP 4: Over-fitting Case

	Panel A. IID			*Panel B. Serial Corr.*		
p	$\Gamma_P^{(A)}$	$\Gamma_P^{(B)}$	$\Gamma_P^{(U)}$	$\Gamma_P^{(A)}$	$\Gamma_P^{(B)}$	$\Gamma_P^{(U)}$
0	0.03	0.06	0.04	0.03	0.07	0.04
1	0.03	0.06	0.08	0.02	0.06	0.08
2	0.02	0.06	0.15	0.02	0.07	0.14
5	0.02	0.07	0.41	0.01	0.07	0.42
10	0.01	0.08	0.80	0.01	0.09	0.80
15	0.01	0.12	0.95	0.01	0.13	0.95
20	0.02	0.17	1	0.01	0.19	0.99
25	0.01	0.24	1	0.01	0.27	1
30	0.02	0.34	1	0.02	0.37	1
35	0.02	0.46	1	0.02	0.49	1
40	0.02	0.60	1	0.02	0.63	1

Note. The table reports empirical rejection frequencies of the test statistics $\Gamma_P^{(A)}$, $\Gamma_P^{(B)}$, $\Gamma_P^{(U)}$ in the presence of over-fitting (see DGP 4 in Section 5), where p is the number of redundant regressors included in the largest model. $R = 100$, $P = 200$, $m = 100$. Nominal size is 0.05.

Table 6. Empirical Results

Countries:		One-month Ahead	One-year Ahead Multistep
Switzerland	DMW	1.110	1.420
	$\Gamma_P^{(A)}$	2.375	3.800
	$\Gamma_P^{(B)}$	2.704*	-0.658
	$\Gamma_P^{(U)}$	1.077	1.788
United Kingdom	DMW	2.095*	2.593*
	$\Gamma_P^{(A)}$	3.166	4.595*
	$\Gamma_P^{(B)}$	0.600	1.362
	$\Gamma_P^{(U)}$	2.074*	2.283*
Canada	DMW	0.335	-0.831
	$\Gamma_P^{(A)}$	3.923	4.908*
	$\Gamma_P^{(B)}$	2.594*	-2.167*
	$\Gamma_P^{(U)}$	0.279	1.029
Japan	DMW	1.409	0.677
	$\Gamma_P^{(A)}$	2.541	3.222
	$\Gamma_P^{(B)}$	2.034*	-1.861
	$\Gamma_P^{(U)}$	1.251	1.677
Germany	DMW	1.909	1.290
	$\Gamma_P^{(A)}$	2.188	1.969
	$\Gamma_P^{(B)}$	-2.247*	0.109
	$\Gamma_P^{(U)}$	1.945	1.285
Commodity Prices	DMW	-1.116	1.420
	$\Gamma_P^{(A)}$	2.842	2.300
	$\Gamma_P^{(B)}$	-2.241*	-1.459
	$\Gamma_P^{(U)}$	-1.352	1.656

Note. The table reports the estimated values of the statistics $\Gamma_P^{(A)}$, $\Gamma_P^{(B)}$, $\Gamma_P^{(U)}$. DMW denotes the Diebold and Mariano (1995) and West (1996) test statistic. "$*$" denotes significance at the 5% level. Significance of the DMW test follows from Giacomini and White's (2006) critical values. The results are based on window sizes $R = m = 100$.

Figure 1: Out-of-Sample Fit in Data: $MSFE^{Model}/MSE^{RW}$, P = 200 - non-overlapping

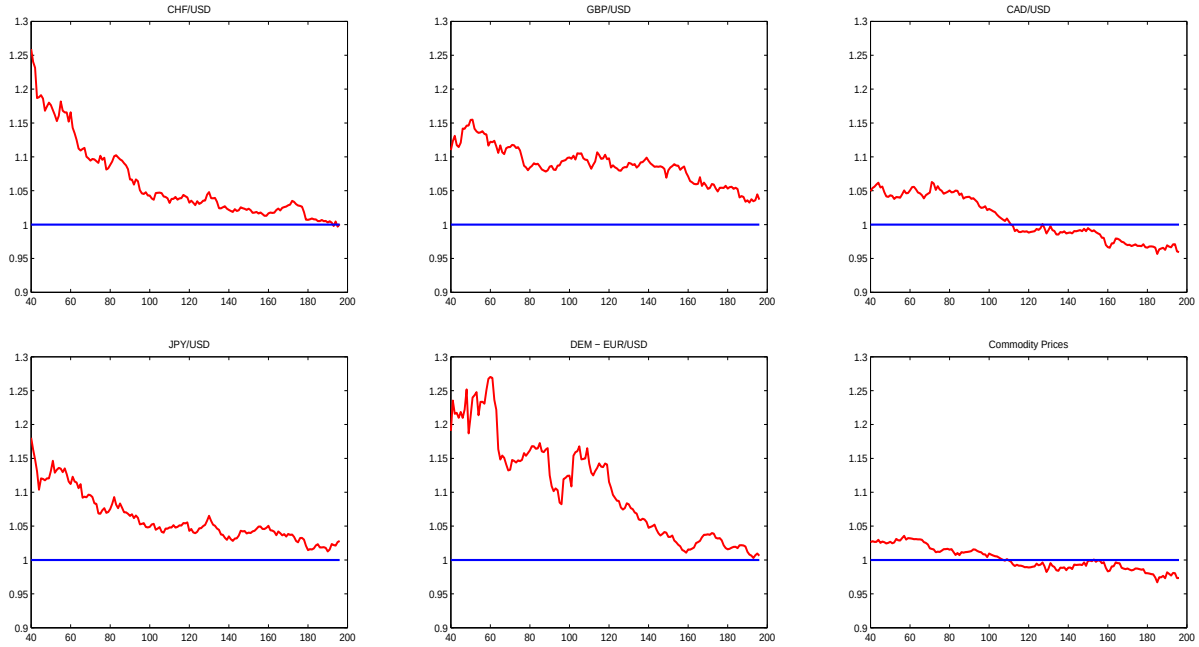
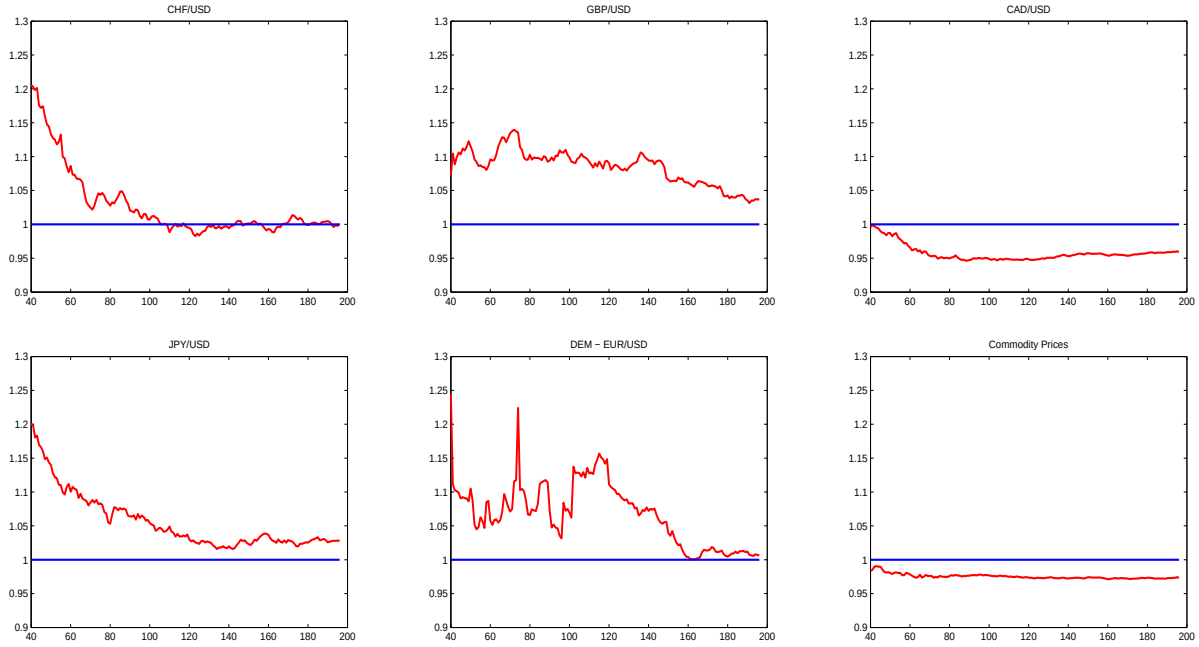


Figure 2: Out-of-Sample Fit in Data: $MSFE^{Model}/MSE^{RW}$, P = 200 - overlapping



Notes: The figures report the MSFE of one month ahead exchange rate forecasts for each country from the economic model relative to a random walk benchmark as a function of the estimation window (R). Figure 1 considers the first $P = 200$ out-of-sample periods following the estimation period, while figure 2 considers the last $P = 200$ out-of-sample periods which are overlapping for the estimation windows of different size.

Figure 3: Decomposition for One-step Ahead Forecast.

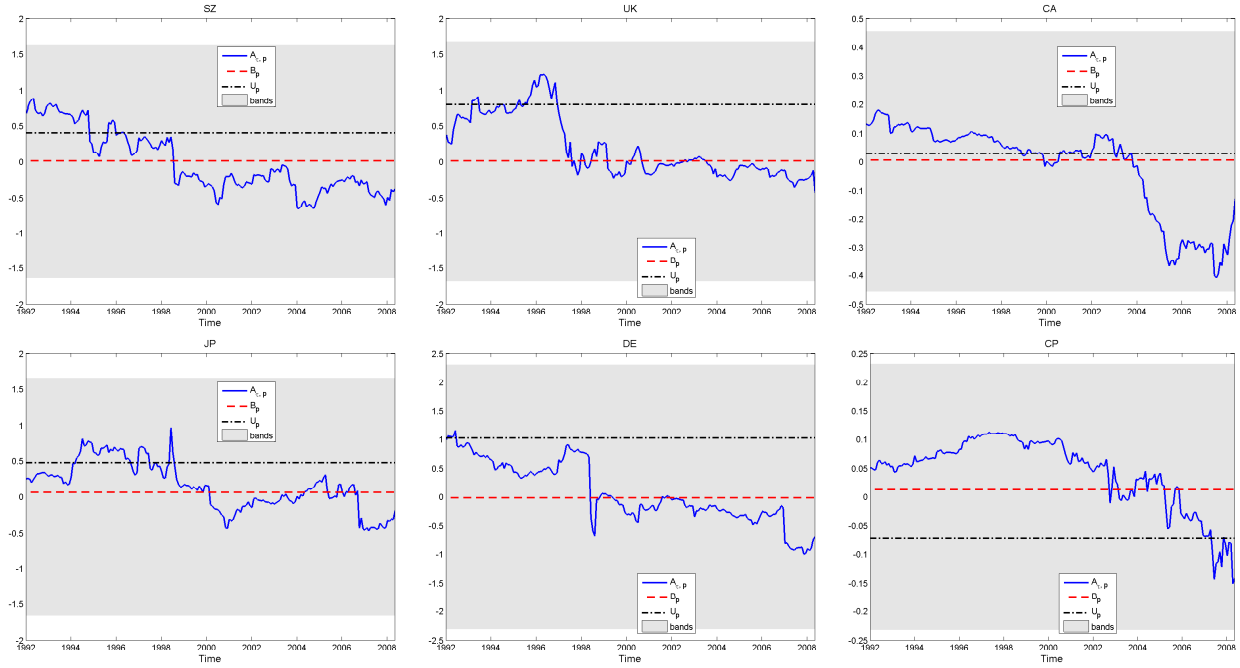
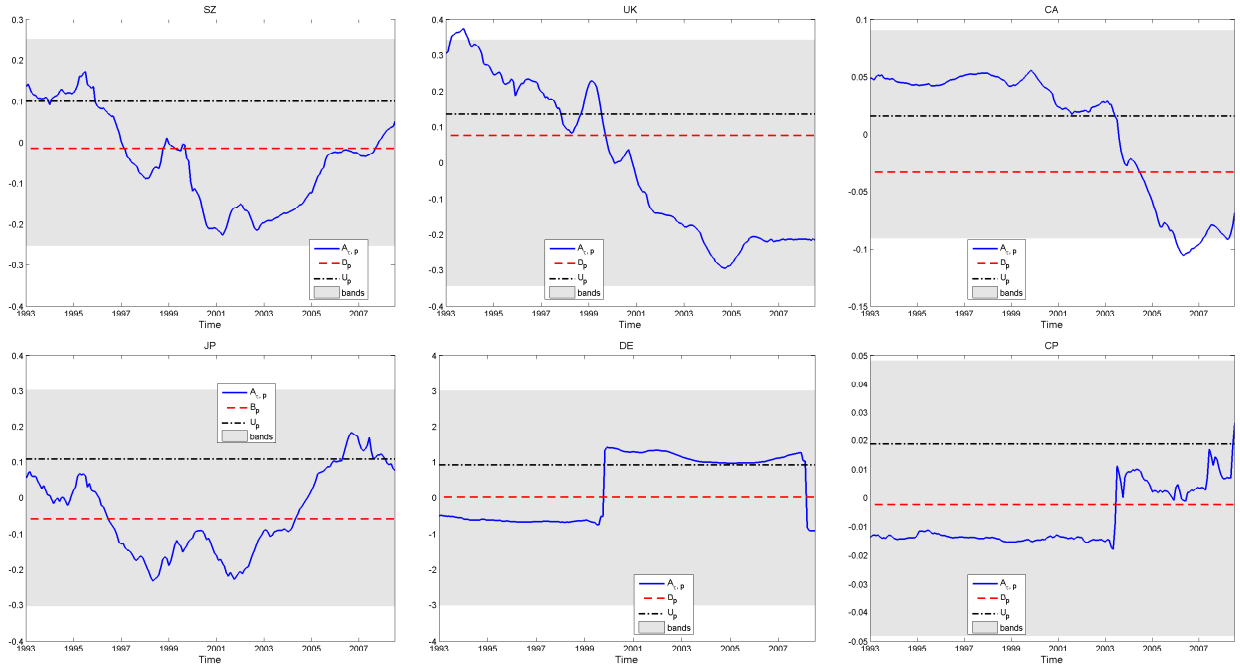


Figure 4: Decomposition for One-year Ahead Direct Multistep Forecast.



Notes: The figures report $A_{\tau,P}$, B_p, and U_p and 5% significance bands for $A_{\tau,P}$. The results are based on window sizes $R = m = 100$.