Working Paper Series 712
(ISSN 1211-3298)

Optimal Out-of-Sample Forecast Evaluation Under Stationarity

Filip Staněk

CERGE-EI
Prague, November 2021

ISBN 978-80-7343-519-6 (Univerzita Karlova, Centrum pro ekonomický výzkum a doktorské studium)
ISBN 978-80-7344-614-7 (Národohospodářský ústav AV ČR, v. v. i.)

Optimal Out-of-Sample Forecast Evaluation Under Stationarity

Filip Staněk[*]

CERGE-EI[†]

November 19, 2021

Abstract

It is common practice to split time-series into in-sample and pseudo out-of-sample segments and to estimate the out-of-sample loss of a given statistical model by evaluating forecasting performance over the pseudo out-of-sample segment. We propose an alternative estimator of the out-of-sample loss which, contrary to conventional wisdom, utilizes both measured in-sample and out-of-sample performance via a carefully constructed system of affine weights. We prove that, provided that the time-series is stationary, the proposed estimator is the best linear unbiased estimator of the out-of-sample loss and outperforms the conventional estimator in terms of sampling variance. Applying the optimal estimator to Diebold-Mariano type tests of predictive ability leads to a substantial power gain without worsening finite sample level distortions. An extensive evaluation on real world time-series from the M4 forecasting competition confirms the superiority of the proposed estimator and also demonstrates a substantial robustness to the violation of the underlying assumption of stationarity.

Keywords: Loss Estimation, Forecast Evaluation, Cross-Validation, Model Selection
JEL classification codes: C22, C52, C53

[*] E-mail: filip.stanek@cerge-ei.cz. Financial support from Grantová agentura UK under grant 264120 is gratefully acknowledged. Furthermore, I'm thankful to prof. Stanislav Anatolyev, the audience at IWEEE2020, ISF2020, and to participants in the Econometrics Reading Group of CERGE-EI for numerous suggestions.

[†] CERGE-EI, a joint workplace of Center for Economic Research and Graduate Education, Charles University and the Economics Institute of the Czech Academy of Sciences, Politických vězňů 7, P.O. Box 882, 111 21 Prague 1, Czech Republic

1 Introduction

In the field of time-series forecasting, researchers are typically concerned with the expected performance of a particular statistical model on yet unseen data, the so called out-of-sample loss. Researchers use the out-of-sample loss to assess whether a proposed model statistically significantly outperforms an already established benchmark model. Likewise, in practical forecasting tasks, the out-of-sample loss is frequently used to select a model that is likely to deliver the best forecasting performance from a set of competing models.

Out-of-sample loss is defined as the expected value of a contrast function that measures the discrepancy between the prediction and the observed value (e.g., the expected value of squared error). Thus, it is by definition unknown and needs to be estimated. This is typically achieved by excluding the most recent segment of the observed time-series from the estimation and performing a sequence of predictions for these observations instead, essentially mimicking the process of actual out-of-sample forecasting. The estimate of the out-of-sample loss is then obtained simply by averaging the precision of individual predictions as measured by the contrast function, i.e., the so called empirical contrasts (e.g. squared errors).[1] While there are many such pseudo out-of-sample evaluation schemes (for a survey, see Tashman, 2000), we restrict our attention to two prominent variants; the rolling scheme and the fixed scheme. When performing an evaluation under the rolling scheme, the model is repeatedly estimated on a rolling window of a fixed length and predictions are made for the subsequent observations. In the fixed scheme, the model is estimated only once on the first segment of the data and is then used to predict all remaining observations (see e.g. Clark and McCracken, 2013).

A common drawback of all such pseudo out-of-sample evaluation schemes and corresponding estimators is the relatively high sampling variance, as the estimate is computed based on only the most recent observations reserved for the pseudo out-of-sample evaluation (Bergmeir and Benítez, 2012; Bergmeir et al., 2014; Schnaubelt, 2019; Cerqueira et al., 2020). Moreover, this issue of scarcity of pseudo out-of-sample observations and consequently of high sampling variance is not limited to situations with few observations, but also afflicts longer time-series. This is because there is an inevitable trade-off between the size of the data-sets designated to be in-sample and pseudo out-of-sample. The former allows for a more faithful approximation of the loss when the whole data-set is used for estimation, whilst the latter allows for more precise estimation of the loss (see Arlot and Celisse, 2010).

To alleviate this issue, we propose an alternative estimator of the out-of-sample loss that utilizes in-sample performance to aid the estimation of the out-of-sample loss, a practice often considered taboo in the forecasting community. In particular, we use in-sample empirical contrasts to partially eliminate the idiosyncratic noise present in observations designated for the out-of-sample evaluation, via a carefully constructed system of optimal affine weights. We prove that, under stationarity, the

[1] There is another class of evaluation schemes that do not respect the temporal ordering of the data and perform out-of-sample evaluation not dissimilar to the canonical cross-validation for independent processes, see e.g., Burman et al. (1994), Racine (2000), and Bergmeir et al. (2018). However, these are not as widely used in practice and hence are not considered in this paper.

proposed estimator of the out-of-sample loss is optimal in terms of the sampling variance within the class of unbiased linear estimators, to which the conventional estimator also belongs. The proposed estimator hence offers a lower sampling variance relative to the conventional estimator, all without introducing any bias. In turn, this allows for a finer assessment of forecasting ability, more powerful predictive ability inference, and more precise model selection.

The proposed optimal estimator is obtained by finding weights that minimize the sampling variance, subject to constraints that guarantee unbiasedness. Importantly, both in- and out-of-sample contrasts can be included with non-zero weights, and weights are allowed to be negative, unlike the conventional estimator, which simply places equal positive weight only on out-of-sample contrasts. In practice, this translates to assigning negative weights to in-sample empirical contrasts that are positively correlated with out-of-sample empirical contrasts, and positive weights to in-sample empirical contrasts that are uncorrelated with out-of-sample empirical contrasts. At the same time, sums of weights of ex-ante identical in-sample contrasts are equal to zero, which ensures that the inclusion of in-sample contrasts does not alter the expected value of the estimator, and hence does not cause bias. From a more general standpoint, the possibility to reduce the sampling variance arises because time-series out-of-sample evaluation schemes are inherently unbalanced in the sense of Shao (1993). That is, these schemes generally do not treat observations equally in terms of in-/out-of-sample usage. The proposed optimal weighting partially rectifies this unbalanced design.

Aside from the optimal estimator itself, we also propose modifications of the canonical Diebold-Mariano test (Diebold and Mariano, 1995) and of the sub-sampling test of equal predictive ability (Zhu and Timmermann, 2020; Ibragimov and Müller, 2010). Both modified tests leverage the proposed optimal weighting for estimation of the loss differential. We show that these tests are asymptotically valid and demonstrate that they exhibit a substantially higher power in detecting deviations from the null hypothesis of equal predictive ability relative to their respective benchmarks.

Finally, to assess the real-life applicability and the robustness of the proposed estimator, we perform an extensive evaluation on 100,000 time-series from the M4 forecasting competition (Makridakis et al., 2020) ranging from yearly to hourly frequency. The proposed estimator delivers more than a 10% reduction in mean squared error relative to the conventional estimator when tasked with predicting the incurred loss on the test segments of time-series. Moreover, when selecting the model by comparing estimated losses, the proposed optimal estimator is more likely to select the best performing model and delivers a smaller overall incurred loss. Importantly, we take no special care to ensure that the time-series are stationary in this evaluation. In fact, most series in the M4 competition do exhibit either some trend, or seasonality, or both. Despite this adverse setting, the proposed estimator still substantially outperforms the conventional estimator, exhibiting a remarkable robustness to the violation of the underlying assumption of stationarity. This clearly demonstrates that the theoretical superiority of the proposed estimator does extend to actual forecasting applications, even with all the difficulties that forecasting real time-series entails.

The next section introduces the statistical framework and provides formal definitions of out-of-sample evaluation schemes and corresponding estimators. Section 3 introduces the proposed estimator of the out-of-sample loss, proves its optimality, and demonstrates its efficiency gains in a simulated environment. Section 4 introduces modified tests of equal out-of-sample predictive ability that utilize the optimal estimator, and demonstrates their power advantage relative to benchmarks. Section 5 compares the performance of the conventional estimator and the proposed optimal estimator on real world time-series from the M4 forecasting competition. Section 6 concludes. Appendices A and B contain proofs and auxiliary results. A ready-to-use implementation of the estimator and tests is provided as an R package ACV[2] and extends the widely used *forecast* package of Hyndman et al. (2020).

2 Conventional Estimator of the Loss

Consider a d-variate sequence $\{X_t\}_1^T \in \left(\mathbb{R}^d\right)^T$ from a stationary random process X_t for a given $T \in \mathbb{N}$. Following the notation of Arlot and Celisse (2010), a statistical model $\mathcal{M} = \{s,\, \widehat{\theta}\}$ is composed of two functions; the estimator $\widehat{\theta}\left(\{X_t\}_1^m\right)$ generating estimated parameters and the forecasting function $s\left(\{X_t\}_1^k;\, \theta\right)$, which predicts the upcoming value of the process based on past values and parameters θ.[3] In particular, $\widehat{\theta} : \cup_{m\in\mathbb{N}} \left(\mathbb{R}^d\right)^m \to \Theta$ where Θ is a parameter space and m is the number of observations used for the estimation. Predictions about the τ-th upcoming value of the process are made via function $s : \{\left(\mathbb{R}^d\right)^k;\, \Theta\} \to \Psi$, where τ is the forecasting horizon, Ψ represent the space of all such possible predictions, and k is the number of past values of the process used to predict the τ-th upcoming value. To assess the quality of a model $\mathcal{M}$, we use a contrast function $\gamma : \{\mathbb{R}^d,\, \Psi\} \to \mathbb{R}$ that measures the discrepancy between a prediction $\psi \in \Psi$ and the actual realization of the process.

This rather general framework allows us to simultaneously consider many typical applications encountered in time-series forecasting. For example, in the case of uni-variate step ahead mean forecasting, $\tau = 1$, and the space of possible predictions $\Psi = \mathbb{R}$. A model $\mathcal{M}$ could be AR(k) with the corresponding least square estimator, in which case the parameter space $\Theta = \mathbb{R}^k$ and prediction $\psi = s(\{X_t\}_1^k;\, \widehat{\theta}) = \sum_1^k X_k \widehat{\theta}_k$ where $\widehat{\theta}$ is an OLS estimator. A contrast function is typically a squared error, and hence $\gamma\left(X_{k+1},\, \psi\right) = \left(X_{k+1} - \psi\right)^2 = \left(X_{k+1} - \sum_1^k X_k \widehat{\theta}_k\right)^2$. In the case of univariate conditional density forecasting, a model $\mathcal{M}$ could be a class of densities and a corresponding estimator $\widehat{\theta}$ for its parameters, set Ψ is a space of density functions and $\psi(q) = s\left(\{X_t\}_1^k;\, \widehat{\theta}\right)(q) = \widehat{f}\left(q|\, \{X_t\}_1^k; \widehat{\theta}\right)$ is the predicted density at point q. One may take $\gamma\left(X_{k+\tau}, \psi\right) = -ln\left(\psi\left(X_{k+\tau}\right)\right) = -ln\left(\widehat{f}\left(X_{k+\tau}|\, \{X_t\}_1^k;\, \widehat{\theta}\right)\right)$ to obtain the Kullback-Leibler divergence (Kullback and Leibler, 1951) as a measure of precision of ψ.

Finally, let us denote the loss of model $\mathcal{M} = \{s,\, \widehat{\theta}\}$ when estimated on a sequence of length m

[2] Available at: `https://github.com/stanek-fi/ACV`.

[3] To facilitate the exposition, we take the liberty of representing the model as a prediction and estimation function pair $\mathcal{M} = \{s,\, \widehat{\theta}\}$ rather than a single function $\mathcal{A}$ representing a statistical algorithm as in Arlot and Celisse (2010), hence focusing on parametric models. All results can nonetheless be extended to non-parametric models by using the identity $\mathcal{A}\left(\{X_t\}_1^m\right)\left(\{X_t\}_{j-k-\tau+1}^{j-\tau}\right) = s\left(\{X_t\}_{j-k-\tau+1}^{j-\tau};\, \widehat{\theta}\left(\{X_t\}_1^m\right)\right)$.

and when faced with forecasting the period $j > m$ using observations $\{X_t\}_{j-k-\tau+1}^{j-\tau}$ as

$$\mathcal{L}_j^m(\mathcal{M}) = \mathbb{E}\left[\gamma\left(X_j, s\left(\{X_t\}_{j-k-\tau+1}^{j-\tau}; \widehat{\theta}\left(\{X_t\}_1^m\right)\right)\right)\right]. \tag{1}$$

Note that the expectation is taken over the whole segment $\{X_t\}_1^j$, i.e., both the forecasted observation X_j and its predecessors, including the estimation window $\{X_t\}_1^m$. We are therefore interested in the performance of model $\mathcal{M}$ rather than that of some particular forecasting function $s\left(\{X_t\}_{j-k-\tau+1}^{j-\tau}; \theta_0\right)$ with fixed $\theta_0 \in \Theta$ (i.e., Question 6 from Dietterich's (1998) taxonomy).

Further, for a "shifting" index $i : 0 \leq i \leq T - m$, we also denote the out-of-sample empirical contrast of model $\mathcal{M}$ when estimated on a sequence $\{X_t\}_{i+1}^{i+m}$ and evaluated at the $(i+j)$-th period with $j > m$ as

$$l_j^{m\,i}(\mathcal{M}) = \gamma\left(X_{i+j}, s\left(\{X_t\}_{i+j-k-\tau+1}^{i+j-\tau}; \widehat{\theta}\left(\{X_t\}_{i+1}^{i+m}\right)\right)\right). \tag{2}$$

The assumption of stationarity then immediately implies

$$\mathbb{E}\left[l_j^{m,\,i}(\mathcal{M})\right] = \mathcal{L}_j^m(\mathcal{M}). \tag{3}$$

In this text, we focus on the pseudo out-of-sample evaluation with step-size v (see e.g., Callen et al. (1996) and Swanson and White (1997)). The procedure is as follows. The model is estimated on a segment of data of length m and forecasts are iteratively made on v consecutive periods for which empirical contrasts are recorded. After that, the estimation window is moved forward by v, and the process is repeated until the end of the sample is reached. The estimate of the out-of-sample loss is then computed simply by averaging all pseudo out-of-sample empirical contrasts incurred. Figure 1a provides a diagram of such a procedure. More formally, the estimator is expressed as[4]

$$\widehat{\mathcal{L}}_{CV} = \frac{1}{n}\sum_{i=1}^{n/v}\sum_{j=1}^{v} l_{m+j}^{m,\,(i-1)v} \tag{4}$$

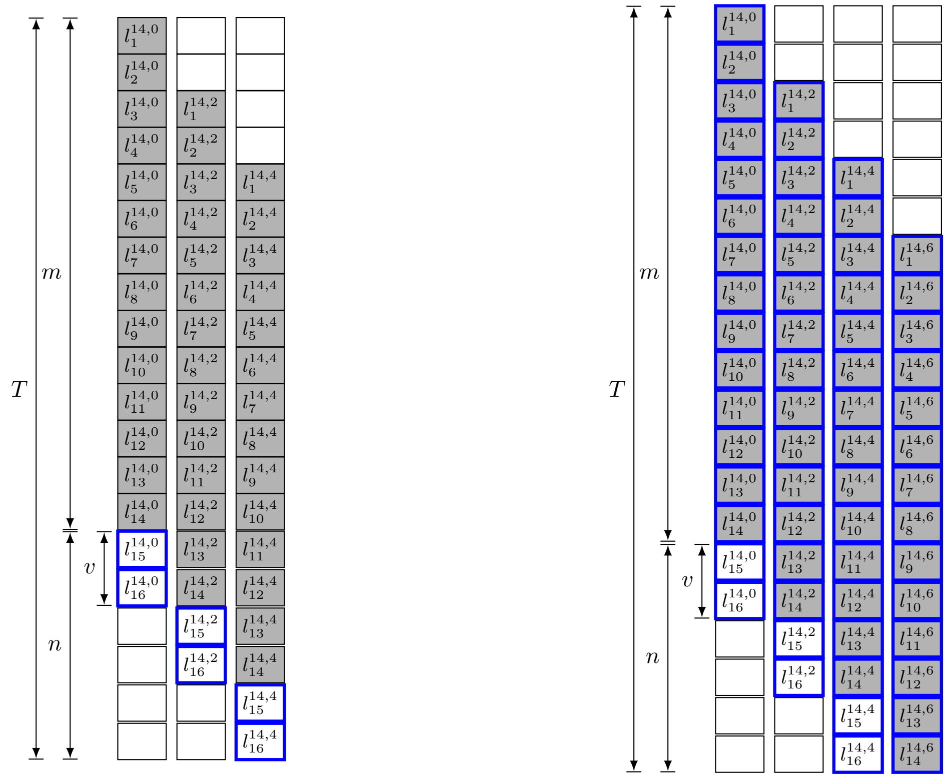
where $n \equiv T - m$ is the number of observations designated for the pseudo out-of-sample evaluation.[5] This specification nests the two most common variants of pseudo out-of-sample evaluation. By setting $v = n$, we obtain the fixed scheme evaluation, which is popular because of its low computational requirements and simplicity. On the other hand, by setting $v = 1$, we obtain the rolling scheme evaluation, which requires repeated re-estimations, but is presumably more theoretically appealing (Swanson and White, 1997).

From Eq. 3, it follows that

$$\mathbb{E}\left[\widehat{\mathcal{L}}_{CV}\right] = \frac{1}{v}\sum_{j=1}^{v} \mathcal{L}_{m+j}^m \equiv \mathcal{L}_{CV} \tag{5}$$

[4] Due to space considerations, we omit $\mathcal{M}$ from the argument of empirical contrasts, losses, and estimators when it causes no confusion.

[5] Throughout this text, we assume that n is divisible by v, i.e. $n \bmod v = 0$.



(a) Conventional estimator $\widehat{\mathcal{L}}_{CV}$.

(b) Optimal estimator $\widehat{\mathcal{L}}_{ACV}$.

Figure 1: A diagram illustrating estimators of the out-of-sample loss.
The example is for $T = 20$ observations, length of the estimation window $m = 14$, and step size $v = 2$.
The gray background indicates whether the observation X_t is used in the estimation of parameters θ.
The blue outline indicates whether the empirical contrast $l_j^{m,i}$ is used when computing the estimate of the out-of-sample loss.

where $\mathcal{L}_{CV}$ is the quantity of interest. Note that $\mathcal{L}_{CV}$ depends not only on model $\mathcal{M}$ but also τ, v, and m. Indeed, different losses $\mathcal{L}_{CV}$ might be relevant to different applications, depending on the desired horizon, the ability to update the model, and the length of the available data. However, irrespective of the particular $\mathcal{L}_{CV}$ to be estimated, we show that the conventional estimator $\widehat{\mathcal{L}}_{CV}$ is sub-optimal for that task. In the next section, we derive the optimal estimator of $\mathcal{L}_{CV}$ which, under the assumption of stationarity, outperforms the conventional estimator in terms of the sampling variance while retaining its unbiasedness.

3 Optimal Estimator of the Loss

Analogically to out-of-sample empirical contrasts, in-sample empirical contrasts can be expressed as

$$l_j^{m\,i}\left(\mathcal{M}\right) = \gamma\left(X_{i+j}, s\left(\{X_t\}_{i+j-k-\tau+1}^{i+j-\tau}; \widehat{\theta}\left(\{X_t\}_{i+1}^{i+m}\right)\right)\right) \tag{6}$$

with the only difference being that $j \leq m$.[6] To construct the optimal estimator, we leverage two facts. First, the correlation between out-of-sample contrast $l_j^{m,\,i}$ and in-sample contrast $l_{j'}^{m,\,i'}$ varies, generally being the strongest when $j + i = j' + i'$, i.e. when the in-sample empirical contrast is computed from the same observation X_{i+j} as the out-of-sample contrast, and hence is influenced by the same idiosyncratic noise. Second, for any pair i and i' it holds that $\mathbb{E}[l_j^{m,\,i}] = \mathbb{E}[l_j^{m,\,i'}]$. Consequently, we can construct affine combinations of in-sample contrasts $l_j^{m,\,i}$, which are of zero mean, but are still negatively correlated with $\widehat{\mathcal{L}}_{CV}$, and whose inclusion reduces the sampling variance without introducing any bias.

To provide a precise description of how such affine combinations should be obtained, we denote the vector of in-sample and out-of-sample contrasts of a model estimated on $\{X_t\}_{i+1}^{i+m}$ by $\boldsymbol{l}_{in}^{m,\,i}$ and $\boldsymbol{l}_{out}^{m,\,i}$ respectively, i.e.

$$\boldsymbol{l}_{in}^{m,\,i} = \left(l_1^{m,\,i},\, l_2^{m,\,i}, \ldots,\, l_m^{m,\,i}\right)^\top \tag{7}$$

$$\boldsymbol{l}_{out}^{m,\,i} = \left(l_{m+1}^{m,\,i},\, l_{m+2}^{m,\,i}, \ldots,\, l_{m+v}^{m,\,i}\right)^\top. \tag{8}$$

We can then collect all measured in-sample and out-of-sample contrasts across different window locations i to a single column vector ϕ, i.e.

$$\phi = \left(\begin{pmatrix}\boldsymbol{l}_{in}^{m,\,0v}\\ \boldsymbol{l}_{out}^{m,\,0v}\end{pmatrix}^\top, \begin{pmatrix}\boldsymbol{l}_{in}^{m,\,1v}\\ \boldsymbol{l}_{out}^{m,\,1v}\end{pmatrix}^\top, \ldots, \begin{pmatrix}\boldsymbol{l}_{in}^{m,\,(\frac{n}{v}-1)v}\\ \boldsymbol{l}_{out}^{m,\,(\frac{n}{v}-1)v}\end{pmatrix}^\top, \left(\boldsymbol{l}_{in}^{m,\,n}\right)^\top\right)^\top. \tag{9}$$

Throughout this paper, we consider estimators linear in measured empirical contrasts, i.e.

$$\lambda^\top \phi \quad \text{with} \quad \lambda \in \mathbb{R}^{\mathrm{card}(\phi)} \tag{10}$$

where, following the work of Lavancier and Rochet (2016) on optimal weighting of estimators, λ is a vector of weights for individual elements of ϕ. Note that the conventional estimator $\widehat{\mathcal{L}}_{CV}$ can likewise be expressed as in Eq. 10; by defining[7]

$$\lambda_{CV\,q} = \begin{cases} \dfrac{1}{n} & \text{for } q \text{ coresponding to elements } l_j^{m,\,iv} \text{ with } 0 \leq i \leq \frac{n}{v} \text{ and } j > m \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

it follows that

$$\widehat{\mathcal{L}}_{CV} = (\lambda_{CV})^\top \phi. \tag{12}$$

This automatically poses the question of whether the vector of weights λ_{CV} is optimal in terms

[6] The derivations remain valid even if the definition in Eq. 6 is replaced with a measurable model-specific function $\kappa_j\left(\{X_t\}_{i+1}^{i+m}\right)$ proxying the in-sample contrasts as defined in Eq. 6. This allows us to also consider applications in which the forecasting function s uses all available observations up to $X_{j-\tau}$ in order to predict X_j, i.e., when $k = m$.

[7] We follow convention and denote q-th element of vector a by a_q and the row (resp. column) subset of matrix A by $A_{Q,:}$ (resp. $A_{:,Q}$) where Q is the set of indices to be kept. Furthermore, we denote the identity matrix by I and column vectors of ones (resp. zeroes) of length k by $\mathbf{1}_k$ (resp. $\mathbf{0}_k$).

of mean squared error

$$\mathbb{E}\left[\left(\lambda^\top \phi - \mathcal{L}_{CV}\right)^2\right] = \lambda^\top \Sigma_\phi \lambda \tag{13}$$

where

$$\Sigma_\phi = \mathbb{E}\left[(\phi - \mathcal{L}_{CV}\mathbf{1}_{\mathrm{card}(\phi)})(\phi - \mathcal{L}_{CV}\mathbf{1}_{\mathrm{card}(\phi)})^\top\right]. \tag{14}$$

In the following proposition, we derive the optimal linear unbiased estimator of $\mathcal{L}_{CV}$ (denoted by $\widehat{\mathcal{L}}_{ACV^*}$ where the "A" stands for affine) and show that the conventional estimator $\widehat{\mathcal{L}}_{CV}$ is generally not optimal.

Proposition 1 *Let $\{X_t\}$ be a stationary process and let V_ϕ be a positive definite covariance matrix of vector ϕ. It then holds that the set of all linear estimators of $\mathcal{L}_{CV}$ that are guaranteed to be unbiased is given as*

$$\mathbb{E}[\lambda^\top \phi] = \mathcal{L}_{CV} \qquad \iff \qquad \lambda \in \Lambda_{ACV} \equiv \left\{ x \in \mathbb{R}^{\mathrm{card}(\phi)} \middle| Bx = b \right\} \tag{15}$$

with

$$B = \left(\mathbf{1}_{n/v}^\top \otimes I,\, I_{:,M}\right) \qquad b = \begin{pmatrix} \mathbf{0}_m \\ \frac{1}{v}\mathbf{1}_v \end{pmatrix} \tag{16}$$

where $M = (1,\, 2,\, \ldots,\, m)$. Furthermore, for estimator

$$\widehat{\mathcal{L}}_{ACV^*} = (\lambda_{ACV})^\top \phi \qquad \textit{with} \qquad \lambda_{ACV} = V_\phi^{-1} B^\top \left(B V_\phi^{-1} B^\top\right)^{-1} b \tag{17}$$

it holds that

$$\mathbb{E}\left[\widehat{\mathcal{L}}_{ACV^*}\right] = \mathcal{L}_{CV}, \tag{18}$$

$$Var\left(\widehat{\mathcal{L}}_{ACV^*}\right) < Var\left(\lambda^\top \phi\right) \qquad \textit{with} \qquad \lambda \in \Lambda_{ACV},\, \lambda \neq \lambda_{ACV}, \tag{19}$$

and also

$$Var\left(\widehat{\mathcal{L}}_{ACV^*}\right) \leq Var\left(\widehat{\mathcal{L}}_{CV}\right). \tag{20}$$

In Proposition 1, we first show that, for all linear unbiased estimators, it holds that $\lambda \in \Lambda_{ACV}$. We then derive the variance minimizing weights λ_{ACV} within Λ_{ACV}. The corresponding optimal estimator $\widehat{\mathcal{L}}_{ACV^*} = (\lambda_{ACV})^\top \phi$ is preferred to the conventional estimator $\widehat{\mathcal{L}}_{CV}$ as it is also unbiased and $Var\left(\widehat{\mathcal{L}}_{ACV^*}\right) \leq Var\left(\widehat{\mathcal{L}}_{CV}\right)$.

It is worth noting that the efficiency gains do not necessarily stem from the stationarity per se, but rather from the existence of some partition (in addition to the partition of singletons) of vector ϕ where contrasts within components of that partition share a common mean. Consequently, analogous estimators can also be constructed for non-stationary series, provided that there is such a partition, i.e., as long as there is at least some degree of regularity. For example, by partitioning ϕ so $l_j^{m,\,iv}$ and $l_{j'}^{m,\,i'v}$ share a common component of the partition if and only if $j = j'$ and both contrasts are from the same day of the week, we can construct the optimal estimator for time-series with a day-of-the-week seasonality.

3.1 Feasible Approximate Optimal Estimator of the Loss

Obviously, the estimator $\widehat{\mathcal{L}}_{ACV^*}$ as presented in Eq. 17 is not feasible, as V_ϕ is not known and needs to be estimated. Given the large size of matrix V_ϕ relative to the amount of data available, some restrictions on its structure are necessary. Furthermore, computational resources needed for the storage of V_ϕ, and even more so for its inversion, grow very quickly, making the computation of optimal weights λ_{ACV} directly via Eq. 17 infeasible for even moderately sized applications.[8]

Consequently, to make the proposed estimator practical, it is essential to develop the estimator $\widehat{V}_\phi$ jointly with an algorithm for computation of weights $\widehat{\lambda}_{ACV}$, so it is not prohibitively computationally expensive. To achieve this, we assume the following covariance structure:

$$Cov(l_j^{m,\,iv}, l_{j'}^{m,\,i'v}) = \begin{cases} 0 & \text{for } i + jv \neq i' + j'v \\ \sigma^2 \rho^{|i-i'|} & \text{for } i + jv = i' + j'v \end{cases}, \tag{21}$$

i.e., only contrasts computed from the same period are mutually correlated, and the strength of that correlation increases in the overlap between respective estimation windows. We can then express $\widehat{V}_\phi$ as

$$\widehat{V}_\phi = \hat{\sigma}^2 \begin{pmatrix} I & A_L^1 & A_L^2 & \dots & A_L^{\frac{n}{v}-2} & A_L^{\frac{n}{v}-1} & (A_L^{\frac{n}{v}})_{:,M} \\ A_U^1 & I & A_L^1 & \ddots & & A_L^{\frac{n}{v}-2} & (A_L^{\frac{n}{v}-1})_{:,M} \\ A_U^2 & A_U^1 & I & \ddots & & & (A_L^{\frac{n}{v}-2})_{:,M} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ A_U^{\frac{n}{v}-2} & & & \ddots & I & A_L^1 & (A_L^2)_{:,M} \\ A_U^{\frac{n}{v}-1} & A_U^{\frac{n}{v}-2} & & \ddots & A_U^1 & I & (A_L^1)_{:,M} \\ (A_U^{\frac{n}{v}})_{M,:} & (A_U^{\frac{n}{v}-1})_{M,:} & (A_U^{\frac{n}{v}-2})_{M,:} & \dots & (A_U^2)_{M,:} & (A_U^1)_{M,:} & (I)_{M,M} \end{pmatrix} \tag{22}$$

where

- $A_U^i = (\hat{\rho} U^v)^i$
- $A_L^i = (\hat{\rho} L^v)^i$

and $M = (1,\, 2,\, \dots,\, m)$. Matrices $U, L \in \mathbb{R}^{(m+v)^2}$ are upper and lower shift matrices, i.e., matrices

[8] For applications as small as $T = 600$, $m = 400$, and $v = 1$, approximately 109 GB of RAM would be needed merely for the storage of V_ϕ (assuming double precision). Inversion of such a matrix is practically impossible via regularly available CPUs, as it requires $O\left(\left((m+v)\frac{n}{v} + m\right)^3\right)$ floating-point operations.

with ones on the superdiagonal and subdiagonal, respectively:

$$U_{i,j} = \begin{cases} 0 & \text{for } i-j \neq -1 \\ 1 & \text{for } i-j = -1 \end{cases} \qquad L_{i,j} = \begin{cases} 0 & \text{for } i-j \neq 1 \\ 1 & \text{for } i-j = 1 \end{cases}. \tag{23}$$

The convenient structure of $\widehat{V}_\phi$ from Eq. 22 admits a closed-form inverse as shown in Lemma 2. Consequently, we can estimate parameters ρ and σ^2 via GMM and compute a feasible and approximately optimal analog of $\widehat{\mathcal{L}}_{ACV^*}$; estimator $\widehat{\mathcal{L}}_{ACV}$ with weights

$$\widehat{\lambda}_{ACV} = \widehat{V}_\phi^{-1} B^\top \left(B \widehat{V}_\phi^{-1} B^\top \right)^{-1} b, \tag{24}$$

without the need to store or numerically invert $\widehat{V}_\phi$.[9]

Admittedly, the parametrization via ρ and σ^2 is rather restrictive and might not fully account for all complexities of the true V_ϕ. However, since the covariances of contrasts from the same period are generally larger than other entries of V_ϕ by an order of magnitude, and since they tend to decay approximately exponentially, $\widehat{V}_\phi$ as defined in Eq. 22 successfully captures the key properties relevant for optimal weighting. Consequently, it is able to reap a major share of the available variance reduction as demonstrated in Sub-section 3.2. This is in line with the observation of Lavancier and Rochet (2016) that the weighting of estimators is often beneficial, even when based on an imperfect variance estimator. Furthermore, the estimator $\widehat{\mathcal{L}}_{CV}$ retains unbiasedness irrespective of how well $\widehat{V}_\phi$ approximates the true V_ϕ, as by definition $\widehat{\lambda}_{ACV} \in \Lambda_{ACV}$. Therefore, only the magnitude of the sampling variance reduction is at risk when V_ϕ is imprecisely estimated.

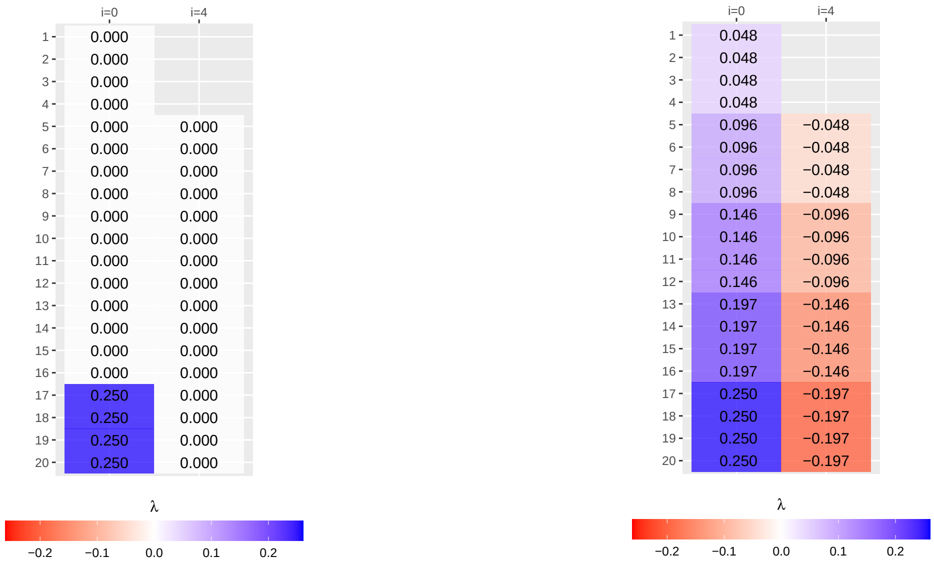
3.2 Simulations

We first illustrate the core mechanism that leads to the variance reduction. Figures 2 and 3 display weights λ_{CV} and $\widehat{\lambda}_{ACV}$ for an illustrative simulated scenario with $T = 20$, $m = 16$, $n = 4$, and simple AR(1) process/model for the fixed and the rolling schemes, respectively. As is apparent from the figures, $\widehat{\lambda}_{ACV}$ includes in-sample empirical contrasts from periods $17 - 20$ with negative weights to eliminate a part of the idiosyncratic noise present in out-of-sample empirical contrasts. In turn, it is necessary to include other in-sample contrasts with positive weights to retain unbiasedness, creating a chain of positive and negative weights that gradually approach zero as we move towards the beginning of the sample. Obviously, such a small sample application is rarely encountered in practice, but it serves well for illustrative purposes, as the basic mechanics are the same regardless of the sample size.

To assess the magnitude of the variance reduction, we perform a series of simulations with the AR(1) data generating process ($\varphi_1 = 0.9$) and an AR(1) model estimated via OLS. For varying m and n, we repeatedly[10] estimate the loss of the model by $\widehat{\mathcal{L}}_{CV}$ and $\widehat{\mathcal{L}}_{ACV}$ under a fixed scheme, and

[9] A description of the GMM estimation procedure and the (partially analytic) computational implementation, which does not require exorbitant computational resources, is available in the documentation of the *ACV* software package.

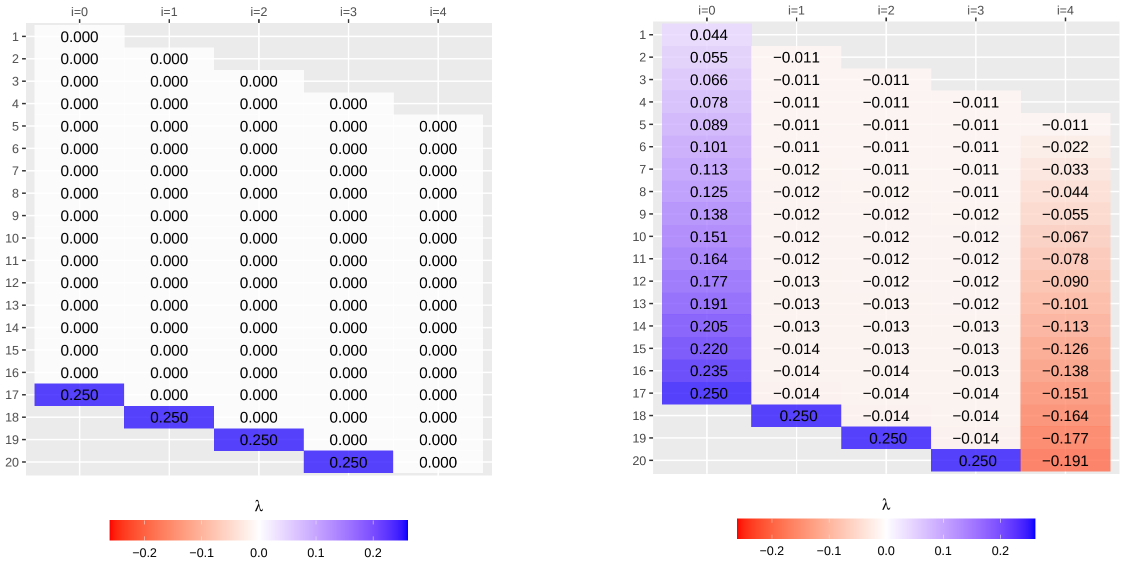
[10] 1,000 repetitions for each combination of m and n.



(a) Weights λ_{CV}

(b) Weights $\widehat{\lambda}_{ACV}$

Figure 2: A side by side comparison of weights λ_{CV} and $\widehat{\lambda}_{ACV}$ for the fixed scheme.



(a) Weights λ_{CV}

(b) Weights $\widehat{\lambda}_{ACV}$

Figure 3: A side by side comparison of weights λ_{CV} and $\widehat{\lambda}_{ACV}$ for the rolling scheme.

measure the variance of each estimator. Furthermore, to assess how well the feasible approximate estimator $\widehat{V}_\phi$ matches the true V_ϕ, we also compute the true V_ϕ by means of simulations, which then allows us to compute the unfeasible $\widehat{\mathcal{L}}_{ACV^*}$ and its variance as a reference point.[11]

Figure 4 displays ratios $\frac{Var(\widehat{\mathcal{L}}_{ACV})}{Var(\widehat{\mathcal{L}}_{CV})}$ for different combinations of m and n. Clearly, the improvement brought by $\widehat{\mathcal{L}}_{ACV}$ relative to $\widehat{\mathcal{L}}_{CV}$ decreases in n and increases in m. This is because the larger the n, the more precise the $\widehat{\mathcal{L}}_{CV}$ and the lesser the potential of reducing the variance by optimal weighting. On the other hand, the larger the m, the stronger the correlation ρ, which in turn allows for better utilization of in-sample contrasts and larger variance reduction. Consequently, for commonly used in-/out-of-sample splitting rules that maintain a fixed ratio of n and m, $\widehat{\mathcal{L}}_{ACV}$ delivers a variance reduction that is approximately constant in the sample size T. Variance ratios range from ~ 0.4, when $1/3$ of the sample is reserved for the out-of-sample evaluation, to ~ 0.1, when $1/10$ of the sample is reserved for the out-of-sample evaluation. This clearly demonstrates that the variance reduction is sizable and not limited to small sample applications.

Furthermore, the estimator $\widehat{V}_\phi$, despite its parsimonious parametrization, approximates the true matrix V_ϕ relatively well, as measured by the performance of $\widehat{\mathcal{L}}_{ACV}$ relative to $\widehat{\mathcal{L}}_{ACV^*}$. Indeed, the feasible estimator $\widehat{\mathcal{L}}_{ACV}$ is able to reap more than 90% of the available variance reduction relative to the optimal unfeasible estimator $\widehat{\mathcal{L}}_{ACV^*}$, as is apparent from the ratios $\frac{Var(\widehat{\mathcal{L}}_{CV})-Var(\widehat{\mathcal{L}}_{ACV})}{Var(\widehat{\mathcal{L}}_{CV})-Var(\widehat{\mathcal{L}}_{ACV^*})}$.

[11] We choose a fixed scheme as the computation of the true matrix V_ϕ is prohibitively computationally expensive in the case of a rolling scheme. With respect to the variance reduction of $\widehat{\mathcal{L}}_{ACV}$ relative to $\widehat{\mathcal{L}}_{CV}$, the results are comparable (available upon request).

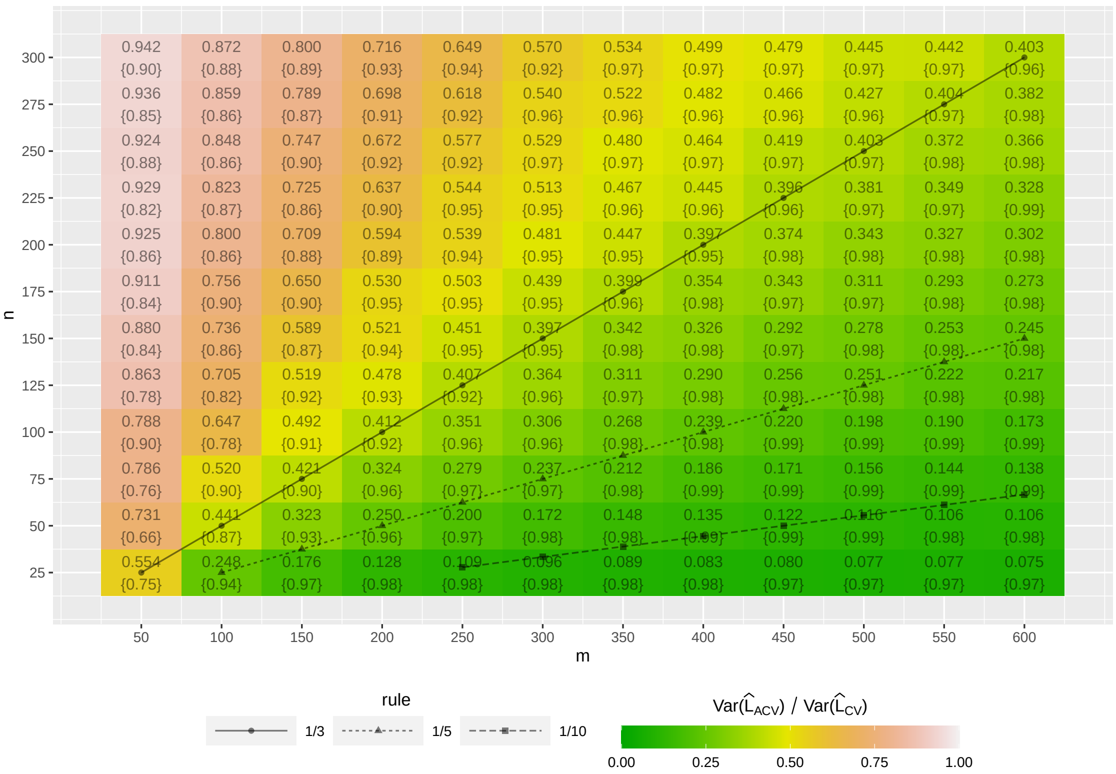


Figure 4: Ratios $\frac{Var(\widehat{\mathcal{L}}_{ACV})}{Var(\widehat{\mathcal{L}}_{CV})}$ for different combinations of m and n.
Numbers in brackets measure the optimality of the feasible estimator relative to the true unfeasible optimal estimator, that is $\frac{Var(\widehat{\mathcal{L}}_{CV})-Var(\widehat{\mathcal{L}}_{ACV})}{Var(\widehat{\mathcal{L}}_{CV})-Var(\widehat{\mathcal{L}}_{ACV*})}$. Common in-/out-of-sample splitting rules $\{1/3,\ 1/5,\ 1/10\}$ are highlighted.

4 Predictive Ability Inference

The lower variance of the proposed estimator might also translate to a substantial power advantage when performing loss inference. Since Diebold and Mariano's (1995) pioneering work, many studies have been devoted to the field of predictive ability inference (see West (2006) or Clark and McCracken (2013) for a comprehensive survey). Following the taxonomy of Clark and McCracken (2013), these tests can be broadly divided into two families. First, there are the tests of population-level predictive ability (e.g. West, 1996; Clark and McCracken, 2001), which are concerned with the null hypothesis about prediction errors of models evaluated at the true, unknown parameters. Second, there are the tests of finite-sample predictive ability (e.g. Giacomini and White, 2006; Clark and McCracken, 2015), which are concerned with the null hypothesis about prediction errors of models with parameters that are themselves a function of a finitely sized window of observed data.

In this section, we apply the optimal estimator to an inference about finite-sample predictive ability, i.e., asymptotics $n \to \infty$ with m considered fixed. The reasons for adoption of this asymptotic framework are threefold. First, the null hypothesis addressed by the test of finite-sample predictive ability appeals to practitioners, as it takes into consideration the bias/variance trade-off inherent to comparing models of different complexity at a given sample size (Clark and McCracken, 2013). Second, unlike for tests of population-level predictive ability, the null hypothesis cannot be addressed with full-sample methods, which tend to dominate pseudo out-of-sample methods in terms of power if applicable (Diebold, 2015). Lastly, the finite-sample predictive ability inference is very general and can be used for both parametric/non-parametric and nested/non-nested models, which is in sharp contrast to tests of population-level predictive ability, where special care has to be taken to address individual cases (West, 2006).

We restrict our attention to the rolling window (i.e. $v = 1$) one-step ahead unconditional test of equal predictive ability, i.e. the test of null hypothesis $H_0 : \mathcal{L}^m_{m+1}(\mathcal{M}_1) = \mathcal{L}^m_{m+1}(\mathcal{M}_2)$ for models $\mathcal{M}_1$ and $\mathcal{M}_2$. This narrower scope is motivated by recent findings showing that the null hypothesis of equal conditional predictive ability can occur only under very specific data generating processes (Zhu and Timmermann, 2020) and findings that the inference under the fixed scheme (i.e. $v = n$) fails to address the desired null hypothesis about models $\mathcal{M}_1$ and $\mathcal{M}_2$ (McCracken, 2020).

Let $\Delta\widehat{\mathcal{L}}_{CV} \equiv \widehat{\mathcal{L}}_{CV}(\mathcal{M}_2) - \widehat{\mathcal{L}}_{CV}(\mathcal{M}_1)$ and let $\widehat{\sigma}^2_{CV}$ be a HAC estimator of its asymptotic variance; $\sigma^2_{CV} \equiv Var\left(\sqrt{n}\Delta\widehat{\mathcal{L}}_{CV}\right)$. As shown in Giacomini and White (2006), the following proposition applies.

Proposition 2 *Provided that:*

(i) $\{X_t\}$ *is mixing with* ϕ *of size* $-r/(2r-2)$, $r \geq 2$ *or* α *of size* $-r/(r-2)$, $r > 2$.

(ii) $\mathbb{E}\left[|\Delta l^{m,v}_{m+1}|^{2r}\right] < \infty$ *for all* v.

(iii) $\sigma^2_{CV} \equiv Var\left(\sqrt{n}\Delta\widehat{\mathcal{L}}_{CV}\right) > 0$ *for all* n *sufficiently large.*

Then under H_0

$$t_{DM} \equiv \frac{\Delta\widehat{\mathcal{L}}_{CV}}{\widehat{\sigma}_{CV}/\sqrt{n}} = \frac{(\lambda_{CV})^\top \Delta\phi}{\widehat{\sigma}_{CV}/\sqrt{n}} \xrightarrow{\text{d}} N(0,1) \tag{25}$$

where $\Delta\phi = \phi(\mathcal{M}_1) - \phi(\mathcal{M}_1)$ *and under* $H_A : |\mathbb{E}\left[\Delta\widehat{\mathcal{L}}_{CV}\right]| \geq \delta > 0$ *for all* n *sufficiently large*

$$P\left(|t_{DM}| > c\right) \longrightarrow 1. \tag{26}$$

We denote the test statistic by a subscript DM as it coincides exactly with the canonical Diebold and Mariano (1995) test (henceforth DM test).

Provided that $\{X_t\}$ is stationary, the third expression in Equation 25 motivates an alternative test statistic that utilizes the optimal weights $\widehat{\lambda}_{ACV}$ to gain more power. Note that, unlike in Section 3, here the weights are optimal for minimizing the variance of the loss differential rather than that of individual estimators of $\mathcal{L}^m_{m+1}(\mathcal{M}_1)$ and $\mathcal{L}^m_{m+1}(\mathcal{M}_2)$, which is generally not the same task. We propose the following modification of the DM test, which uses the optimal affine weighting (ADM test henceforth).

Proposition 3 *Provided that* $\{X_t\}$ *is stationary,* $plim(\hat{\rho}) \neq 1$*, and (i)-(iii) holds, then*

$$t_{ADM} \equiv \frac{\Delta\widehat{\mathcal{L}}_{ACV}}{\widehat{\sigma}_{ACV}/\sqrt{n}} = \frac{(\widehat{\lambda}_{ACV})^\top \Delta\phi}{\widehat{\sigma}_{ACV}/\sqrt{n}} \xrightarrow{\text{d}} N(0,1) \tag{27}$$

where $\widehat{\sigma}_{ACV} = \widehat{\sigma}_{CV}\frac{\widehat{\lambda}_{ACV}^\top \widehat{V}_{\Delta\phi}\widehat{\lambda}_{ACV}}{\widehat{\lambda}_{CV}^\top \widehat{V}_{\Delta\phi}\widehat{\lambda}_{CV}}$ *and under* $H_A : |\mathbb{E}\left[\Delta\widehat{\mathcal{L}}_{CV}\right]| \geq \delta > 0$ *for all* n *sufficiently large*

$$P\left(|t_{ADM}| > c\right) \longrightarrow 1. \tag{28}$$

While widely adopted, the DM test is known to suffer from level distortions in small samples, stemming from the estimation of the long-run variance (see Clark and McCracken, 2013). To mitigate this issue, Zhu and Timmermann (2020) propose to use Ibragimov and Müller's (2010) subsampling t-test (IM test henceforth), which does not require a variance estimation. In particular, Zhu and Timmermann (2020) prove the following proposition.

Proposition 4 *Suppose that* $\{X_t\}$ *is stationary and* $E[\Delta l^{m,i}_{m+1}] = 0$*. Assume that* $E|\Delta l^{m,i}_{m+1}|^r = 0$ *is bounded for some* $r > 2$ *and* $\Delta l^{m,i}_{m+1}$ *is strong mixing of size* $-r/(r-2)$*. Then, for fixed* $K > 1$

$$t_{IM} = \frac{\overline{\Delta\widehat{\mathcal{L}}_{CV}}}{\sqrt{(K-1)\sum_{k=1}^{K}\left(\widehat{\mathcal{L}}^{(k)}_{CV} - \overline{\Delta\widehat{\mathcal{L}}_{CV}}\right)^2/\sqrt{K}}} \xrightarrow{\text{d}} t_{K-1} \tag{29}$$

where $\widehat{\mathcal{L}}^{(k)}_{CV}$ *is the loss estimate computed from the k-th block of data of size* $\tilde{n} = n/K$*, that is* $\widehat{\mathcal{L}}^{(k)}_{CV} = \tilde{n}^{-1}\sum_{i=0}^{\tilde{n}-1}\Delta l^{m,i+\tilde{n}(k-1)}_{m+1} = \lambda^{(k)}_{CV}\Delta\phi^{(k)}$ *where* $\Delta\phi^{(k)} = \Delta\phi_M$ *with* $M = \{i\}_{i=1+\tilde{n}*(m+1)(k-1)}^{(\tilde{n}+1)*(m+1)-1+\tilde{n}*(m+1)(k-1)}$*, and where* $\overline{\Delta\widehat{\mathcal{L}}_{CV}} = K^{-1}\sum_{k=1}^{K}\widehat{\mathcal{L}}^{(k)}_{CV}$*.*

Similarly to the DM test, the IM test also immediately lends itself to a modified version that exploits the optimal weighting $\widehat{\lambda}_{ACV}$ (AIM test henceforth).

Proposition 5 *Suppose that $\{X_t\}$ is stationary, $plim(\hat{\rho}) \neq 1$, and $E[\Delta l_{m+1}^{m,i}] = 0$. Assume that $E|\Delta l_{m+1}^{m,i}|^r = 0$ is bounded for some $r > 2$ and $\Delta l_{m+1}^{m,i}$ is strong mixing of size $-r/(r-2)$. Then, for fixed $K > 1$*

$$t_{AIM} = \frac{\overline{\Delta\widehat{\mathcal{L}}_{ACV}}}{\sqrt{(K-1)\sum_{k=1}^{K}\left(\widehat{\mathcal{L}}_{ACV}^{(k)} - \overline{\Delta\widehat{\mathcal{L}}_{ACV}}\right)^2/\sqrt{K}}} \xrightarrow{\text{d}} t_{K-1} \tag{30}$$

where $\widehat{\mathcal{L}}_{ACV}^{(k)}$ is the loss estimate computed from the k-th block of data of size $\tilde{n} = n/K$, that is $\widehat{\mathcal{L}}_{ACV}^{(k)} = \widehat{\lambda}_{ACV}^{(k)}\Delta\phi^{(k)}$ where $\Delta\phi^{(k)} = \Delta\phi_M$ with $M = \{i\}_{i=1+\tilde{n}*(m+1)(k-1)}^{(\tilde{n}+1)*(m+1)-1+\tilde{n}*(m+1)(k-1)}$, and where $\overline{\Delta\widehat{\mathcal{L}}_{ACV}} = K^{-1}\sum_{k=1}^{K}\widehat{\mathcal{L}}_{ACV}^{(k)}$.

4.1 Power and Level Properties

To evaluate the power and level properties of the proposed tests, we adapt the simulation environment of Giacomini and White (2006). We consider a process $\{X_t\} = \{Y_t, Z_t\}$ following the law of motion

$$Y_{t+1} = c + Z_t + \varepsilon_{t+1} \qquad \varepsilon_{t+1} \sim N(0, \sigma^2) \tag{31}$$

and two models $\mathcal{M}_1 = \{s_1, \widehat{\theta}_1\}$ and $\mathcal{M}_2 = \{s_2, \widehat{\theta}_2\}$ producing point predictions of Y_{t+1}:

$$s_1\left(X_t;\, \widehat{\beta}_t^1\right) = \widehat{\beta}_{1,t}^1 Z_t \quad \text{with} \quad \widehat{\beta}_t^1 = \left\{\widehat{\beta}_{1,t}^1\right\} = \widehat{\theta}_1\left(\{X_t\}_{t-m+1}^t\right) \tag{32}$$

and

$$s_2\left(X_t;\, \widehat{\beta}_t^2\right) = \widehat{\beta}_{0,t}^2 + \widehat{\beta}_{1,t}^2 Z_t \quad \text{with} \quad \widehat{\beta}_t^2 = \left\{\widehat{\beta}_{0,t}^2,\, \widehat{\beta}_{1,t}^2\right\} = \widehat{\theta}_2\left(\{X_t\}_{t-m+1}^t\right) \tag{33}$$

where $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are OLS estimators of the slope, and of the intercept and the slope, respectively. Model $\mathcal{M}_1$ is hence misspecified in that it omits the intercept. Under the mean squared error contrast function (henceforth MSE), we put forward the following proposition, which is a slight generalization of Proposition 5 from Giacomini and White (2006). It allows us to explore not only level distortions but also the power of the proposed tests.

Proposition 6 *Let $\overline{Z}_t = \frac{1}{m-1}\sum_{j=t-m+1}^{t-1} Z_j, \overline{Z^2}_t = \frac{1}{m-1}\sum_{j=t-m+1}^{t-1} Z_j^2$, and $S_t = \sum_{j=t-m+1}^{t-1} Z_j^2 - (m-1)\overline{Z}_t^2$. Then for given $\varsigma \geq 1$ and*

$$c = \sigma\left(\frac{\sum_{t=m}^{T-1}\left(1 + \frac{\overline{Z^2}_t}{S_t} + \frac{Z_t^2}{S_t} - 2\frac{\overline{Z}_t}{S_t}Z_t\right)\varsigma - \sum_{t=m}^{T-1}\left(1 + \frac{Z_t^2}{(m-1)\overline{Z^2}_t}\right)}{\sum_{t=m}^{T-1}\left(1 - \frac{\overline{Z}_t}{\overline{Z^2}_t}Z_t\right)^2}\right)^{0.5} \tag{34}$$

it holds that

$$\varsigma = \frac{\mathcal{L}_{m+1}^m(\mathcal{M}_1)}{\mathcal{L}_{m+1}^m(\mathcal{M}_2)}. \tag{35}$$

By setting $\varsigma = 1$, Proposition 6 allows us to simulate data under H_0, that is with c such that the omission of the intercept will cause an increase of the loss that is exactly offset by the reduction of the loss stemming from a more precisely estimated slope coefficient. Furthermore, we also consider values $\varsigma > 1$, in which the omission of the intercept will result in worse predictions. In the exercise below, we follow the setup of Giacomini and White (2006) and take $\{Z_t\}$ to be a second log difference of the US CPI index for the 1959-01 – 1998-12 period (U.S. Bureau of Labor Statistics, 2020).[12] The truncation lag of Newey and West's (1987) HAC estimator in DM and ADM tests is chosen according to the commonly used rule $\left\lfloor \frac{3}{4} n^{\frac{1}{3}} \right\rfloor$ (see e.g. Lazarus et al., 2018). The number of groups K in IM and AIM tests is 2 as in Zhu and Timmermann (2020).

We simulate the process from Eq. 31 with constant c corresponding to values of $\varsigma \in \{$ 1, 1.03125, 1.0625, 1.125, 1.25, 1.375, 1.5, 1.75, 2 $\}$ for $n \in \{10, 20, 50, 100, 200, 300\}$ and $m = 100$. As can be seen in Figure 5, the proposed ADM and AIM tests exhibit substantially higher power relative to their conventional counterparts. In accordance with results from Section 3.2, the power gain is especially sizable in scenarios with small n relative to m. The power gain also appears to be more pronounced in the case of the IM type tests, which tend to sacrifice power in exchange for lesser finite sample level distortions, creating a greater opportunity for improvement.

To better explore level properties, we repeat the exercise with $\varsigma = 1$, levels $p \in \{0.01, 0.05, 0.1\}$ for values $n \in \{10, 20, 50, 100, 200, 300\}$ and $m = 100$. Inspecting Table 1, it is apparent that while we do observe the same small sample level distortions for DM type tests as documented in the literature, their magnitude is, in fact, smaller for the proposed ADM test. This shows that the power gain is achieved despite better level properties, not because of them. As expected, rejection probabilities for IM type tests are closer to the desired levels. The AIM test exhibits larger level distortions in small samples relative to the conventional IM test. These distortions stem from stronger finite sample dependencies between individual estimators $\widehat{\mathcal{L}}_{ACV}^{(k)}$ introduced by the affine weighting. However, given the substantially higher power of AIM relative to IM, these finite sample distortions seem acceptable.

		$p = 0.01$				$p = 0.05$				$p = 0.10$		
n	DM	ADM	IM	AIM	DM	ADM	IM	AIM	DM	ADM	IM	AIM
10	0.048	0.034	0.011	0.017	0.122	0.085	0.050	0.086	0.190	0.135	0.095	0.164
20	0.029	0.025	0.011	0.016	0.098	0.072	0.046	0.077	0.167	0.117	0.092	0.162
50	0.017	0.015	0.010	0.014	0.070	0.055	0.050	0.079	0.129	0.099	0.099	0.152
100	0.013	0.016	0.010	0.014	0.059	0.056	0.044	0.068	0.111	0.100	0.096	0.142
200	0.010	0.012	0.011	0.015	0.048	0.046	0.054	0.073	0.091	0.083	0.107	0.140
300	0.010	0.009	0.012	0.013	0.045	0.040	0.053	0.070	0.082	0.075	0.106	0.136

Table 1: A table of rejection probabilities for DM, ADM, IM, and AIM tests for different levels p.

[12] As a robustness check, we also repeat the exercise with the first log difference of CPI and the log of CPI. These results are qualitatively similar and are available in Appendix B.

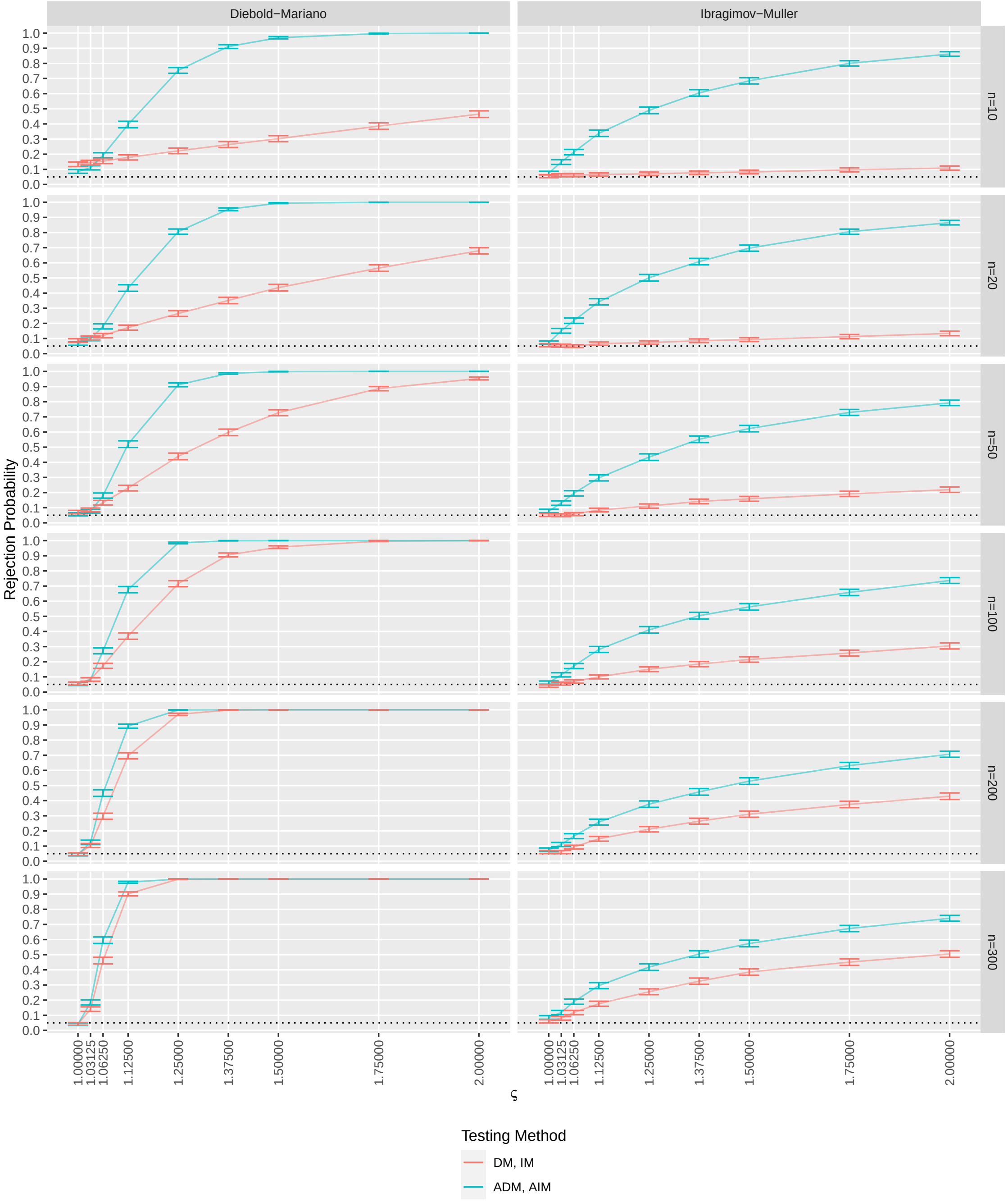


Figure 5: Plots of rejection probabilities for DM, IM, ADM, and AIM tests at level 0.05.

5 Empirical Evaluation

To demonstrate that the theoretical superiority of the proposed estimator also translates to real-life forecasting tasks, we perform an extensive evaluation on the M4 competition (Makridakis et al., 2020), which is currently the largest time-series forecasting competition, with 100,000 time-series ranging from yearly to hourly frequency. Participants in the M4 competition were asked to produce forecasts for each of the series for the upcoming 6/8/18/13/14/48 periods for yearly/quarterly/monthly/weekly/daily/hourly frequency, respectively. The organizers withheld the most recent segment of each series of corresponding length (test segments, henceforth). Submitted forecasts were then compared with test segments to evaluate their precision.

To assess the performance of $\widehat{\mathcal{L}}_{ACV}$ we consider two canonical models that were used as standards for comparison in the M4 competition; the ETS (Hyndman et al., 2002), which automatically selects the optimal form of exponential smoothing via the information criterion, and the autoARIMA (Hyndman and Khandakar, 2008), which selects the most appropriate ARIMA specification via the information criterion. Both these models are frequently used in practice and performed comparably well in the M4 competition, making them ideal candidates. Similarly to the competition, the performance of each model is assessed on the test segment of series using the sMAPE contrast function:[13]

$$\gamma\left(X_t,\, \widehat{X}_t\right) = \frac{|X_t - \widehat{X}_t|}{\frac{1}{2}|X_t| + \frac{1}{2}|\widehat{X}_t|}100. \tag{36}$$

Unlike in the M4 competition however, our interest is not in the performance of individual models per se, but rather in our ability to predict the out-of-sample performance $\widetilde{\mathcal{L}}_{CV,s}(\mathcal{M})$[14] on the test segment of a series s with the use of in-sample data only. To do so, we perform pseudo out-of-sample evaluations under the rolling scheme (i.e. $v = 1$) with the same number of pseudo out-of-sample observations as in the test segment (i.e. $n \in \{6,\, 8,\, 18,\, 13,\, 14,\, 48\}$). For each series s, we compute the estimates $\widehat{\mathcal{L}}_{CV,s}(\mathcal{M})$ and $\widehat{\mathcal{L}}_{ACV,s}(\mathcal{M})$ and compare them with the actual out-of-sample loss $\widetilde{\mathcal{L}}_{CV,s}(\mathcal{M})$ incurred on the test segment. The overall precision of the estimator is computed as

$$MSE_{CV}(\mathcal{M}) = \frac{1}{|S|}\sum_{s\in S}\left(\widetilde{\mathcal{L}}_{CV,s}(\mathcal{M}) - \widehat{\mathcal{L}}_{CV,s}(\mathcal{M})\right)^2 \tag{37}$$

and

$$MSE_{ACV}(\mathcal{M}) = \frac{1}{|S|}\sum_{s\in S}\left(\widetilde{\mathcal{L}}_{CV,s}(\mathcal{M}) - \widehat{\mathcal{L}}_{ACV,s}(\mathcal{M})\right)^2 \tag{38}$$

with $S \subset \{1,\, 2,\, \ldots,\, 100000\}$ being a subset of time-series under consideration. To better assess the performance on different types of series, we also subject each series to a non-parametric CS test for

[13] This contrast function was chosen by organizers so that losses of series on different scales are approximately comparable. As a robustness check, we also repeat the exercise with MAE and MSE contrast functions with prior normalization and obtain comparable results (available upon request).

[14] We use the notation $\widetilde{\mathcal{L}}_{CV}$ rather than $\widehat{\mathcal{L}}_{CV}$ to highlight that this is the loss incurred on the test segment (i.e., the true out-of-sample evaluation). However, as the test segment is of finite length, this is still only an estimate of the true theoretical loss $\mathcal{L}_{CV}$. The subscript CV indicates that the conventional estimator is used to compute the loss incurred on the test segment.

the presence of a trend (Cox and Stuart, 1955) and a QS test for the presence of seasonality (Ljung and Box, 1978).

Table 2 depicts MSE_{CV} and MSE_{ACV} for both models across all frequencies, further broken down by the results of the CS and QS tests (for both, the threshold $p = 0.05$ is considered). For each model, percentage improvements of $\widehat{\mathcal{L}}_{ACV}$ over $\widehat{\mathcal{L}}_{CV}$ in terms of MSE are shown alongside their statistical significance. As is apparent, the use of $\widehat{\mathcal{L}}_{ACV}$ leads to a substantially more precise estimation of the incurred out-of-sample loss $\widetilde{\mathcal{L}}_{CV}$, in particular to a reduction of MSE by 13.0% and 10.6% on average for ETS and autoARIMA, respectively. It is worth highlighting that this MSE reduction likely underestimates the true magnitude of the sampling variance reduction, as the comparison is made with respect to the estimate of loss $\widetilde{\mathcal{L}}_{CV}$ rather than the true theoretical loss $\mathcal{L}_{CV}$; hence, the corresponding part of the MSE in principle cannot be reduced. A back of the envelope calculation suggests that the theoretical MSE reduction, if computed against the true loss rather than its estimate, is actually twice the size.

Furthermore, the majority of series in the M4 competition are not-stationary, exhibiting either a trend (90%), seasonality (39%), or both (36%). Despite this adverse setting, $\widehat{\mathcal{L}}_{ACV}$ still offers a substantial advantage over $\widehat{\mathcal{L}}_{CV}$, although it should be noted the MSE reduction is not as sizable for series that exhibit seasonality. The fact that $\widehat{\mathcal{L}}_{ACV}$ exhibits superior performance relative to $\widehat{\mathcal{L}}_{CV}$, even when applied indiscriminately to a wide range of time-series without any regards for stationarity, clearly demonstrates its robustness and practical applicability.

Lastly, we assess the performance of $\widehat{\mathcal{L}}_{ACV}$ in terms of model selection. In this exercise, the task is to use the loss estimate to select the model $\mathcal{M}$ that will perform best on the test segment of a given series, i.e., to identify the model with the smallest $\widetilde{\mathcal{L}}_{CV,s}(\mathcal{M})$. Table 3 shows the average incurred loss $\widetilde{\mathcal{L}}_{CV}$ and the probability of selecting the best model, whether this is done according to AIC (Akaike, 1998) or the sign of the loss differential estimated via $\widehat{\mathcal{L}}_{CV}$ and $\widehat{\mathcal{L}}_{ACV}$. The table also includes the average loss that would be incurred if we knew which model was the best-performing on the test segment.[15] Obviously, such a selection is not feasible in practice but it provides a useful benchmark, as it represents the best possible outcome that can be achieved via model selection alone. Compared to AIC, $\widehat{\mathcal{L}}_{ACV}$ achieves a 23.7% incurred loss reduction relative to what is achievable and is more likely to select the best model by 4.9% points.[16] Compared to $\widehat{\mathcal{L}}_{CV}$, the relative loss reduction is more modest, only 1.4%, but still statistically significant. The estimator $\widehat{\mathcal{L}}_{ACV}$ is 0.3% points more likely to select the best model than $\widehat{\mathcal{L}}_{CV}$.

While the gains from more accurate model selection via $\widehat{\mathcal{L}}_{ACV}$ rather than $\widehat{\mathcal{L}}_{CV}$ are not as sizable, it should be noted that the variance minimizing weights of $\widehat{\mathcal{L}}_{ACV}$ are not necessarily optimal in terms of selecting a model so that its incurred loss is the lowest in expectation. By computing

[15] We denoted these incurred losses and probabilities of selecting the best model by $\widetilde{\mathcal{L}}_{CV}(\mathcal{M}_x)$ and $P(best)_x$, respectively, where $x \in \{AIC, CV, ACV, ex-post\,opt.\}$.

[16] The dominance of CV and ACV over AIC likely stems from violations of stationarity, which more heavily penalize the AIC than the ACV, and/or the fact that the sMAPE contrast function in Eq. 36 is not aligned with the MSE contrast function, for which the AIC is designed. A thorough theoretical comparison of the AIC and pseudo out-of-sample estimators such as $\widehat{\mathcal{L}}_{CV}$ or $\widehat{\mathcal{L}}_{ACV}$ is beyond the scope of this article. A detailed analysis can, however, be found in Inoue and Kilian (2006).

	Time-series				ETS			autoARIMA	
Period	Trending	Seasonal	N	MSE_{CV}	MSE_{ACV}	$\Delta MSE\,[\%]$	MSE_{CV}	MSE_{ACV}	$\Delta MSE\,[\%]$
Yearly			23000	48.68	41.47	-14.8***	57.05	51.37	-10.0***
				(1.65)	(1.49)		(2.42)	(2.38)	
	F	F	2214	139.93	126.41	-9.7***	194.26	187.93	-3.3
				(10.60)	(10.27)		(16.17)	(16.36)	
	F	T	267	20.22	19.86	-1.8	24.08	23.84	-1.0
				(5.49)	(5.43)		(6.22)	(5.76)	
	T	F	15076	49.62	41.06	-17.3***	54.29	46.57	-14.2***
				(1.92)	(1.65)		(2.74)	(2.63)	
	T	T	5443	10.35	9.12	-11.9**	10.51	10.44	-0.7
				(0.90)	(0.81)		(1.38)	(1.37)	
Quarterly			24000	28.70	24.18	-15.8***	33.87	29.30	-13.5***
				(1.16)	(0.97)		(1.41)	(1.22)	
	F	F	1561	92.17	78.02	-15.4**	101.50	81.68	-19.5***
				(8.96)	(7.75)		(9.51)	(7.78)	
	F	T	681	65.29	49.90	-23.6	90.95	81.41	-10.5
				(14.91)	(8.83)		(16.93)	(14.93)	
	T	F	14115	26.34	21.68	-17.7***	29.50	25.23	-14.5***
				(1.32)	(1.09)		(1.49)	(1.18)	
	T	T	7643	16.82	15.50	-7.9	23.02	21.50	-6.6
				(1.48)	(1.38)		(2.43)	(2.33)	
Monthly			48000	19.32	17.65	-8.6***	21.68	19.69	-9.2***
				(0.47)	(0.45)		(0.57)	(0.55)	
	F	F	2574	78.64	63.56	-19.2***	87.64	73.09	-16.6***
				(5.02)	(4.42)		(5.89)	(5.28)	
	F	T	1964	21.89	19.70	-10.0*	24.63	21.33	-13.4**
				(1.99)	(1.88)		(2.67)	(2.38)	
	T	F	21613	23.60	22.27	-5.6**	26.57	24.78	-6.8***
				(0.72)	(0.75)		(0.91)	(0.94)	
	T	T	21849	7.85	7.49	-4.7*	8.81	8.23	-6.6**
				(0.36)	(0.36)		(0.41)	(0.40)	
Weekly			359	8.81	5.55	-37.0***	6.47	5.95	-8.0
				(1.40)	(0.99)		(0.86)	(1.13)	
	F	F	54	13.18	10.50	-20.4	9.50	7.51	-21.0
				(3.05)	(4.09)		(2.27)	(1.67)	
	F	T	3	2.72	1.85	-32.1	0.81	0.67	-18.0
				(2.09)	(1.47)		(0.80)	(0.64)	
	T	F	257	8.81	5.15	-41.5***	6.39	6.35	-0.7
				(1.81)	(1.06)		(1.06)	(1.53)	
	T	T	45	4.01	2.14	-46.8	3.61	2.12	-41.2
				(1.99)	(0.91)		(1.65)	(0.84)	
Daily			4227	1.62	1.56	-3.5	2.11	2.15	1.7
				(0.33)	(0.37)		(0.51)	(0.54)	
	F	F	226	2.71	3.98	47.1	4.05	4.36	7.6
				(2.33)	(3.73)		(3.53)	(4.01)	
	F	T	19	0.39	0.43	10.8	0.33	0.42	26.5
				(0.19)	(0.24)		(0.18)	(0.21)	
	T	F	3535	0.89	0.71	-19.3*	0.89	0.77	-13.4*
				(0.22)	(0.19)		(0.23)	(0.22)	
	T	T	447	6.94	7.11	2.5	10.92	12.05	10.3**
				(2.29)	(2.45)		(4.07)	(4.31)	
Hourly			414	12.71	8.55	-32.7***	53.11	45.62	-14.1
				(2.27)	(1.47)		(11.36)	(12.16)	
	F	F	1	0.24	0.09	-60.4	0.06	0.02	-56.3
				(NA)	(NA)		(NA)	(NA)	
	F	T	125	29.67	17.97	-39.4***	90.33	59.04	-34.6***
				(6.65)	(3.89)		(21.18)	(15.35)	
	T	F	5	2.12	2.70	27.1	1.56	1.86	18.6
				(1.93)	(2.54)		(1.30)	(1.49)	
	T	T	283	5.45	4.53	-16.8	37.77	40.63	7.6
				(1.36)	(1.22)		(13.64)	(16.44)	
All			**100000**	**27.51**	**23.94**	**-13.0*****	**31.99**	**28.60**	**-10.6*****
				(0.52)	**(0.47)**		**(0.71)**	**(0.68)**	

Table 2: Comparison of $\widehat{\mathcal{L}}_{CV}$ and $\widehat{\mathcal{L}}_{ACV}$ in terms of the loss estimation. $\Delta MSE\,[\%] = \frac{MSE_{ACV} - MSE_{CV}}{MSE_{CV}} 100$. Standard errors in brackets, $***p < 0.001$, $**\ p < 0.01$, $*p < 0.05$.

Time-series		ex-post opt.	AIC		CV		ACV		AIC vs ACV	CV vs ACV
Period	N	$\widetilde{\mathcal{L}}$	$P(best)$	$\widetilde{\mathcal{L}}$	$P(best)$	$\widetilde{\mathcal{L}}$	$P(best)$	$\widetilde{\mathcal{L}}$	$\Delta\widetilde{\mathcal{L}}\,[\%]$	$\Delta\widetilde{\mathcal{L}}\,[\%]$
Yearly	23000	6.489	0.513	7.186	0.528	7.096	0.526	7.089	-13.9***	-1.2
		(0.056)	(0.003)	(0.065)	(0.003)	(0.063)	(0.003)	(0.062)		
Quarterly	24000	5.602	0.484	6.198	0.548	6.007	0.551	6.002	-32.8***	-1.0
		(0.055)	(0.003)	(0.061)	(0.003)	(0.059)	(0.003)	(0.059)		
Monthly	48000	6.513	0.525	6.944	0.578	6.858	0.585	6.852	-21.3***	-1.7
		(0.043)	(0.002)	(0.046)	(0.002)	(0.045)	(0.002)	(0.045)		
Weekly	359	5.033	0.616	5.162	0.526	5.245	0.577	5.229	52.1	-7.3
		(0.298)	(0.026)	(0.303)	(0.026)	(0.316)	(0.026)	(0.316)		
Daily	4227	1.013	0.516	1.052	0.522	1.030	0.509	1.031	-53.6***	4.4*
		(0.027)	(0.008)	(0.031)	(0.008)	(0.028)	(0.008)	(0.028)		
Hourly	414	6.765	0.551	9.261	0.804	6.911	0.819	6.869	-95.9***	-28.9
		(0.443)	(0.024)	(0.655)	(0.020)	(0.452)	(0.019)	(0.450)		
All	**100000**	**6.052**	**0.512**	**6.575**	**0.558**	**6.456**	**0.561**	**6.451**	**-23.7*** **	**-1.4***
		(0.028)	**(0.002)**	**(0.031)**	**(0.002)**	**(0.030)**	**(0.002)**	**(0.030)**		

Table 3: Comparison of AIC, $\widehat{\mathcal{L}}_{CV}$ and $\widehat{\mathcal{L}}_{ACV}$ in terms of model selection.
For $x \in \{AIC, CV\}$, $\Delta\widetilde{\mathcal{L}}\,[\%] = \frac{\widetilde{\mathcal{L}}_{CV}(\mathcal{M}_{ACV})-\widetilde{\mathcal{L}}_{CV}(\mathcal{M}_x)}{\widetilde{\mathcal{L}}_{CV}(\mathcal{M}_x)-\widetilde{\mathcal{L}}_{CV}(\mathcal{M}_{ex-post\,opt.})}100$. Standard errors in brackets,
$***p < 0.001$, $**\ p < 0.01$, $*p < 0.05$.

multiple sets of weights jointly, so that they are optimal in terms of model selection, we could presumably attain even better results. This promising research direction is, however, beyond the scope of this paper.

6 Conclusion

We challenge the notion that a model's in-sample performance cannot be utilized when assessing its out-of-sample performance. We propose an alternative estimator of the out-of-sample loss that optimally utilizes both in-sample and out-of-sample empirical contrasts via a system of affine weights. We prove that under stationarity, the proposed (unfeasible) estimator is the best unbiased linear estimator of the out-of-sample loss and that it dominates the conventional estimator in terms of the sampling variance. We also propose an approximate feasible variant of the estimator, which closely matches the performance of the unfeasible optimal estimator, and which exhibits a substantially smaller sampling variance relative to the conventional estimator, by a factor of ~ 0.4 to ~ 0.1 in our simulations. The variance reduction is most sizable in situations where few observations are designated for the out-of-sample evaluation relative to the number of in-sample observations.

The proposed optimal estimator can also be applied to the inference about predictive ability. We put forward modifications of Diebold and Mariano's (1995) test and of Ibragimov and Müller's (2010) test and show that utilization of the optimal estimator leads to a substantial power gain (often by a factor > 2) in detecting deviations from the null hypothesis of equal predictive ability. In addition, the finite sample level distortions of Diebold and Mariano's (1995) test frequently documented in the literature seem to be attenuated, rather than exacerbated, by the system of optimal affine weights.

Finally, to assess the real-life applicability of the estimator and its robustness, we perform an extensive evaluation on time-series from the M4 forecasting competition (Makridakis et al., 2020). In line with the theoretical derivations and the simulation evidence, the proposed estimator more precisely estimates the losses incurred on the test segments of series ($> 10\%$ MSE reduction relative to the conventional estimator). Furthermore, selecting a model based on the proposed estimator leads to a higher probability of selecting the ex-post optimal model and also to an overall lower loss relative to that which would be incurred if the model were selected according to the conventional estimator. Importantly, these improvements are achieved despite the majority of time-series in the M4 competition exhibiting some form of non-stationarity, and hence not strictly satisfying requirements for the application of the proposed estimator. This demonstrates a remarkable robustness of the proposed estimator and even potential for forecasting time-series where the assumption of stationarity might be in question.

References

Akaike, H., 1998. Information Theory and an Extension of the Maximum Likelihood Principle, in: Parzen, E., Tanabe, K., Kitagawa, G. (Eds.), Selected Papers of Hirotugu Akaike. Springer, New York, NY. Springer Series in Statistics, pp. 199–213. doi:10.1007/978-1-4612-1694-0_15.

Arlot, S., Celisse, A., 2010. A survey of cross-validation procedures for model selection. Statistics surveys 4, 40–79.

Bergmeir, C., Benítez, J.M., 2012. On the use of cross-validation for time series predictor evaluation. Information Sciences 191, 192–213. doi:10.1016/j.ins.2011.12.028.

Bergmeir, C., Costantini, M., Benítez, J.M., 2014. On the usefulness of cross-validation for directional forecast evaluation. Computational Statistics & Data Analysis 76, 132–143. doi:10.1016/j.csda.2014.02.001.

Bergmeir, C., Hyndman, R.J., Koo, B., 2018. A note on the validity of cross-validation for evaluating autoregressive time series prediction. Computational Statistics & Data Analysis 120, 70–83. doi:10.1016/j.csda.2017.11.003.

Burman, P., Chow, E., Nolan, D., 1994. A cross-validatory method for dependent data. Biometrika 81, 351–358. doi:10.1093/biomet/81.2.351.

Callen, J.L., Kwan, C.C.Y., Yip, P.C.Y., Yuan, Y., 1996. Neural network forecasting of quarterly accounting earnings. International Journal of Forecasting 12, 475–482. doi:10.1016/S0169-2070(96)00706-6.

Cerqueira, V., Torgo, L., Mozetič, I., 2020. Evaluating time series forecasting models: An empirical study on performance estimation methods. Machine Learning 109, 1997–2028. doi:10.1007/s10994-020-05910-7.

Clark, T., McCracken, M., 2013. Chapter 20 - Advances in Forecast Evaluation, in: Elliott, G., Timmermann, A. (Eds.), Handbook of Economic Forecasting. Elsevier. volume 2 of *Handbook of Economic Forecasting*, pp. 1107–1201. doi:10.1016/B978-0-444-62731-5.00020-8.

Clark, T.E., McCracken, M.W., 2001. Tests of equal forecast accuracy and encompassing for nested models. Journal of Econometrics 105, 85–110. doi:10.1016/S0304-4076(01)00071-9.

Clark, T.E., McCracken, M.W., 2015. Nested forecast model comparisons: A new approach to testing equal accuracy. Journal of Econometrics 186, 160–177. doi:10.1016/j.jeconom.2014.06.016.

Cox, D.R., Stuart, A., 1955. Some Quick Sign Tests for Trend in Location and Dispersion. Biometrika 42, 80–95. doi:10.2307/2333424.

Diebold, F.X., 2015. Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold–Mariano Tests. Journal of Business & Economic Statistics 33, 1–1. doi:10.1080/07350015.2014.983236.

Diebold, F.X., Mariano, R.S., 1995. Comparing Predictive Accuracy. Journal of Business & Economic Statistics 13.

Dietterich, T.G., 1998. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Computation 10, 1895–1923. doi:10.1162/089976698300017197.

Giacomini, R., White, H., 2006. Tests of Conditional Predictive Ability. Econometrica 74, 1545–1578. doi:10.1111/j.1468-0262.2006.00718.x.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., Yasmeen, F., 2020. Forecast: Forecasting Functions for Time Series and Linear Models.

Hyndman, R.J., Khandakar, Y., 2008. Automatic Time Series Forecasting: The forecast Package for R. Journal of Statistical Software 27, 1–22. doi:10.18637/jss.v027.i03.

Hyndman, R.J., Koehler, A.B., Snyder, R.D., Grose, S., 2002. A state space framework for automatic forecasting using exponential smoothing methods. International Journal of Forecasting 18, 439–454. doi:10.1016/S0169-2070(01)00110-8.

Ibragimov, R., Müller, U.K., 2010. T-Statistic based correlation and heterogeneity robust inference. Journal of Business & Economic Statistics 28, 453–468.

Inoue, A., Kilian, L., 2006. On the selection of forecasting models. Journal of Econometrics 130, 273–306. doi:10.1016/j.jeconom.2005.03.003.

Johnson, D.H., 2020. Statistical signal processing .

Kullback, S., Leibler, R.A., 1951. On Information and Sufficiency. The Annals of Mathematical Statistics 22, 79–86. doi:10.1214/aoms/1177729694.

Lavancier, F., Rochet, P., 2016. A general procedure to combine estimators. Computational Statistics & Data Analysis 94, 175–192. doi:10.1016/j.csda.2015.08.001.

Lazarus, E., Lewis, D.J., Stock, J.H., Watson, M.W., 2018. HAR Inference: Recommendations for Practice. Journal of Business & Economic Statistics 36, 541–559. doi:10.1080/07350015.2018.1506926.

Ljung, G.M., Box, G.E.P., 1978. On a measure of lack of fit in time series models. Biometrika 65, 297–303. doi:10.1093/biomet/65.2.297.

Makridakis, S., Spiliotis, E., Assimakopoulos, V., 2020. The M4 Competition: 100,000 time series and 61 forecasting methods. International Journal of Forecasting 36, 54–74. doi:10.1016/j.ijforecast.2019.04.014.

McCracken, M.W., 2020. Diverging Tests of Equal Predictive Ability. Econometrica 88, 1753–1754. doi:10.3982/ECTA17523.

Newey, W.K., West, K.D., 1987. A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. Econometrica 55, 703–708. doi:10.2307/1913610.

Racine, J., 2000. Consistent cross-validatory model-selection for dependent data: Hv-block cross-validation. Journal of Econometrics 99, 39–61. doi:10.1016/S0304-4076(00)00030-0.

Schnaubelt, M., 2019. A Comparison of Machine Learning Model Validation Schemes for Non-Stationary Time Series Data. Working Paper 11/2019. FAU Discussion Papers in Economics.

Shao, J., 1993. Linear Model Selection by Cross-validation. Journal of the American Statistical Association 88, 486–494. doi:10.1080/01621459.1993.10476299.

Swanson, N.R., White, H., 1997. Forecasting economic time series using flexible versus fixed specification and linear versus nonlinear econometric models. International Journal of Forecasting 13, 439–461. doi:10.1016/S0169-2070(97)00030-7.

Tashman, L.J., 2000. Out-of-sample tests of forecasting accuracy: An analysis and review. International Journal of Forecasting 16, 437–450. doi:10.1016/S0169-2070(00)00065-0.

U.S. Bureau of Labor Statistics, 2020. Consumer Price Index for All Urban Consumers: All Items in U.S. City Average. https://fred.stlouisfed.org/series/CPIAUCSL.

Usmani, R.A., 1994. Inversion of a tridiagonal Jacobi matrix. Linear Algebra and its Applications 212, 413–414.

West, K.D., 1996. Asymptotic Inference about Predictive Ability. Econometrica 64, 1067–1084. doi:10.2307/2171956.

West, K.D., 2006. Chapter 3 Forecast Evaluation, in: Elliott, G., Granger, C.W.J., Timmermann, A. (Eds.), Handbook of Economic Forecasting. Elsevier. volume 1, pp. 99–134. doi:10.1016/S1574-0706(05)01003-7.

Zhu, Y., Timmermann, A., 2020. Can Two Forecasts Have the Same Conditional Expected Accuracy? arXiv:2006.03238 [stat] arXiv:2006.03238.

Appendices

A Proofs

Lemma 1 *Let* $P = \left\{P_1, P_2, \ldots, P_{\mathrm{card}(P)}\right\}$ *be a partition of* $\{1, 2, \ldots, \mathrm{card}(\phi)\}$ *such that* $\forall j \in \{1, 2, \ldots, card(P)\}$ $\forall i, i' \in P_j : \mathbb{E}[\phi_i] = \mathbb{E}[\phi_{i'}]$*. Then for* $\lambda \in \Lambda_{ACV}$ *where*

$$\Lambda_{ACV} = \left\{ \lambda_{CV} + x \middle| x \in \mathbb{R}^{\mathrm{card}(\phi)} \wedge \forall j \in \{1, 2, \ldots, card(P)\} : \sum_{i \in P_j} x_i = 0 \right\}, \tag{39}$$

it holds that

$$\mathbb{E}[\lambda^\top \phi] = \mathcal{L}_{CV} \tag{40}$$

and

$$\lambda^\top \Sigma_\phi \lambda = \lambda^\top V_\phi \lambda \tag{41}$$

where $\Sigma_\phi = \mathbb{E}\left[(\phi - \mathcal{L}_{CV}\mathbf{1})(\phi - \mathcal{L}_{CV}\mathbf{1})^\top\right]$ *and* $V_\phi = Var(\phi)$*.*

Proof of Lemma 1 *To prove this lemma, consider*

$$\begin{aligned} \mathbb{E}[\lambda^\top \phi] &= \mathbb{E}[(\lambda_{CV} + x)^\top \phi] \\ &= \mathbb{E}[(\lambda_{CV})^\top \phi] + \mathbb{E}[x^\top \phi] \\ &= \mathcal{L}_{CV} + \sum_{j=1}^{\mathrm{card}(P)} \underbrace{\sum_{i \in P_j} x_i \mathbb{E}[\phi_i]}_{=0} \\ &= \mathcal{L}_{CV}. \end{aligned} \tag{42}$$

Furthermore

$$\begin{aligned} \Sigma_\phi &= \mathbb{E}[(\phi - \mathcal{L}_{CV}\mathbf{1})\,(\phi - \mathcal{L}_{CV}\mathbf{1})^\top] \\ &= \mathbb{E}[((\phi - \mathbb{E}[\phi]) + (\mathbb{E}[\phi] - \mathcal{L}_{CV}\mathbf{1}))\,((\phi - \mathbb{E}[\phi]) + (\mathbb{E}[\phi] - \mathcal{L}_{CV}\mathbf{1}))^\top] \\ &= Var(\phi) + (\mathbb{E}[\phi] - \mathcal{L}_{CV}\mathbf{1})\,(\mathbb{E}[\phi] - \mathcal{L}_{CV}\mathbf{1})^\top \end{aligned} \tag{43}$$

and

$$\begin{aligned} \lambda^\top \Sigma_\phi \lambda &= \lambda^\top \left(Var(\phi) + (\mathbb{E}[\phi] - \mathcal{L}_{CV}\mathbf{1})\,(\mathbb{E}[\phi] - \mathcal{L}_{CV}\mathbf{1})^\top \right) \lambda \\ &= \lambda^\top Var(\phi) \lambda + \lambda^\top\,(\mathbb{E}[\phi] - \mathcal{L}_{CV}\mathbf{1})\,(\mathbb{E}[\phi] - \mathcal{L}_{CV}\mathbf{1})^\top\,\lambda \\ &= \lambda^\top Var(\phi) \lambda \end{aligned} \tag{44}$$

as

$$\begin{aligned}\lambda^\top \left(\mathbb{E}[\phi] - \mathcal{L}_{CV}\mathbf{1}\right) &= \left(\lambda_{CV} + x\right)^\top \left(\mathbb{E}[\phi] - \mathcal{L}_{CV}\mathbf{1}\right) \\ &= \left(\lambda_{CV}\right)^\top \mathbb{E}[\phi] - \left(\lambda_{CV}\right)^\top \mathcal{L}_{CV}\mathbf{1} + x^\top \mathbb{E}[\phi] - x^\top \mathcal{L}_{CV}\mathbf{1} \\ &= \mathcal{L}_{CV} - \mathcal{L}_{CV} + \sum_{j=1}^{\operatorname{card}(P)} \underbrace{\sum_{i\in P_j} x_i \mathbb{E}[\phi_i]}_{=0} - \mathcal{L}_{CV} \sum_{j=1}^{\operatorname{card}(P)} \underbrace{\sum_{i\in P_j} x_i \mathbf{1}_i}_{=0} \qquad (45) \\ &= 0,\end{aligned}$$

which completes the proof.

Proof of Proposition 1 *Let $P = \{P_1, P_2, \ldots, P_{m+v}\}$ be a partition of $\{1, 2, \ldots, \operatorname{card}(\phi)\}$ such that $\forall j \in \{1, 2, \ldots, m+v\} \, \forall i \in \left\{0, 1, \ldots, \frac{n}{v}\right\} : l_j^{m,\, iv} \in P_j$. Due to stationarity, it holds that $\forall j \in \{1, 2, \ldots, card(P)\} \; \forall i, i' \in P_j : \mathbb{E}[\phi_i] = \mathbb{E}[\phi_{i'}]$ and hence Lemma 1 can be applied. Also note that the set Λ_{ACV} from Lemma 1 can be equivalently expressed as*

$$\lambda \in \Lambda_{ACV} \qquad \Longleftrightarrow \qquad B\lambda = b \tag{46}$$

with

$$B = \left(\mathbf{1}_{n/v}^\top \otimes I,\, I_{:,M}\right) \qquad b = \begin{pmatrix} \mathbf{0}_m \\ \frac{1}{v}\mathbf{1}_v \end{pmatrix} \tag{47}$$

where $M = (1, 2, \ldots, m)$.

By virtue of Proposition 1, for any $\lambda \in \Lambda_{ACV}$, it holds that

$$\mathbb{E}[\lambda^\top \phi] = \mathcal{L}_{CV} \tag{48}$$

and

$$\lambda^\top \Sigma_\phi \lambda = \lambda^\top V_\phi \lambda, \tag{49}$$

i.e., all estimators with weights in Λ_{ACV} are unbiased estimators of $\mathcal{L}_{CV}$ and their mean squared error is equal to their variance. We are interested in the best possible estimator (in terms of mean squared error/variance) in the set Λ_{ACV}. Formally:

$$\operatorname*{argmin}_{\lambda} \lambda^\top V_\phi \lambda \;\text{ s.t} : B\lambda = b. \tag{50}$$

This is an elementary problem of quadratic programming and its solution can be found in many texts related to that field (see e.g. Johnson, 2020, p: 53). Below we present a short outline of the proof.

The Lagrangian associated with the problem is given by

$$L(\lambda, \alpha) = \lambda^\top V_\phi \lambda - \alpha^\top (B\lambda - b). \tag{51}$$

Necessary conditions for pair $\{\lambda, \alpha\}$ to be solution to Eq. 50 are

$$\frac{\partial L(\lambda, \alpha)}{\partial \lambda} = 2V_\phi \lambda - B^\top \alpha = 0, \tag{52}$$

$$\frac{\partial L(\lambda, \alpha)}{\partial \alpha} = B\lambda - b = 0. \tag{53}$$

From Eq. 52, it follows

$$\lambda = \frac{1}{2} V_\phi^{-1} B^\top \alpha, \tag{54}$$

combining that with Eq. 53 leads to

$$\alpha = 2 \left(B V_\phi^{-1} B^\top \right)^{-1} b \tag{55}$$

and consequently

$$\lambda = V_\phi^{-1} B^\top \left(B V_\phi^{-1} B^\top \right)^{-1} b. \tag{56}$$

The invertibility of matrix V_ϕ and $\left(B V_\phi^{-1} B^\top \right)$ follows from positive-definiteness of V_ϕ and full rank of B. The sufficient conditions then follows from the fact that $\lambda^\top V_\phi \lambda$ is strictly convex function as V_ϕ is positive definite. We denote the optimum weights as λ_{ACV} and the corresponding estimator by $\widehat{\mathcal{L}}_{ACV^*}$, i.e.

$$\widehat{\mathcal{L}}_{ACV^*} = (\lambda_{ACV})^\top \phi \qquad \text{with} \qquad \lambda_{ACV} = V_\phi^{-1} B^\top \left(B V_\phi^{-1} B^\top \right)^{-1} b. \tag{57}$$

The statement

$$\mathbb{E}[\widehat{\mathcal{L}}_{ACV^*}] = \mathcal{L}_{CV} \tag{58}$$

stems directly from $\lambda_{ACV} \in \Lambda_{ACV}$ and Proposition 1. Statements

$$Var(\widehat{\mathcal{L}}_{ACV^*}) < Var(\lambda^\top \phi) \qquad \text{with} \qquad \lambda \in \Lambda_{ACV},\, \lambda \neq \lambda_{ACV} \tag{59}$$

and

$$Var(\widehat{\mathcal{L}}_{ACV^*}) \leq Var(\widehat{\mathcal{L}}_{CV}) \tag{60}$$

follows from strict convexity of function $\lambda^\top V_\phi \lambda$ and $\lambda_{CV} \in \Lambda_{ACV}$, respectively.

It remains to show that there is no $\lambda' \notin \Lambda_{ACV}$ such that it is guaranteed that $\mathbb{E}[(\lambda')^\top \phi] = \mathcal{L}_{CV}$. Suppose that there is such λ' and let $x = \lambda' - \lambda_{CV}$. From $\lambda' \notin \Lambda_{ACV}$ it follows that $\exists j' : \sum_{i \in P_{j'}} x_i = c \neq 0$. Suppose that $\forall j \in \{1, 2, ..., m+v\}, j \neq j' : \mathcal{L}_j^m = 0$ and $\mathcal{L}_{j'}^m \neq 0$. Then

$$\mathbb{E}[(\lambda')^\top \phi] = \mathbb{E}[\lambda_{CV}^\top \phi] + \mathbb{E}[x^\top \phi] = \mathcal{L}_{CV} + c\mathcal{L}_{j'}^m \neq \mathcal{L}_{CV}, \tag{61}$$

which is a contradiction.

Lemma 2 *Provided that $\hat{\rho} \neq 1$, matrix $\widehat{V}_\phi$ defined as:*

$$\widehat{V}_\phi = \hat{\sigma}^2 \begin{pmatrix} I & A_L^1 & A_L^2 & \dots & A_L^{\frac{n}{v}-2} & A_L^{\frac{n}{v}-1} & (A_L^{\frac{n}{v}})_{:,M} \\ A_U^1 & I & A_L^1 & \ddots & & A_L^{\frac{n}{v}-2} & (A_L^{\frac{n}{v}-1})_{:,M} \\ A_U^2 & A_U^1 & I & \ddots & & & (A_L^{\frac{n}{v}-2})_{:,M} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ A_U^{\frac{n}{v}-2} & & & \ddots & I & A_L^1 & (A_L^2)_{:,M} \\ A_U^{\frac{n}{v}-1} & A_U^{\frac{n}{v}-2} & & \ddots & A_U^1 & I & (A_L^1)_{:,M} \\ (A_U^{\frac{n}{v}})_{M,:} & (A_U^{\frac{n}{v}-1})_{M,:} & (A_U^{\frac{n}{v}-2})_{M,:} & \dots & (A_U^2)_{M,:} & (A_U^1)_{M,:} & (I)_{M,M} \end{pmatrix} \tag{62}$$

with

- $A_U^i = (\hat{\rho} U^v)^i$
- $A_L^i = (\hat{\rho} L^v)^i$
- $M = (1,\, 2,\, \dots,\, m)$

is invertible and its inverse is given by:

$$\widehat{V}_\phi^{-1} = \frac{1}{\hat{\sigma}^2} \begin{pmatrix} Z_1 & Z_L & 0 & \dots & 0 & 0 & (0)_{:,M} \\ Z_U & Z_2 & Z_L & \ddots & & 0 & (0)_{:,M} \\ 0 & Z_U & Z_2 & \ddots & & & (0)_{:,M} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & & & \ddots & Z_2 & Z_L & (0)_{:,M} \\ 0 & 0 & & \ddots & Z_U & Z_2 & (Z_L)_{:,M} \\ (0)_{M,:} & (0)_{M,:} & (0)_{M,:} & \dots & (0)_{M,:} & (Z_U)_{M,:} & (Z_3)_{M,M} \end{pmatrix} \tag{63}$$

with

- $Z_1 = I + \frac{\hat{\rho}^2}{1-\hat{\rho}^2} L^v U^v$
- $Z_2 = I + \frac{\hat{\rho}^2}{1-\hat{\rho}^2} \left(L^v U^v + U^v L^v\right)$
- $Z_3 = \frac{1}{1-\hat{\rho}^2} I$
- $Z_U = \frac{-\hat{\rho}}{1-\hat{\rho}^2} U^v$
- $Z_L = \frac{-\hat{\rho}}{1-\hat{\rho}^2} L^v.$

Proof of Lemma 2 *To prove this lemma, we check individual sub-matrices of $\widehat{V}_\phi \widehat{V}_\phi^{-1}$ to verify that, together, they indeed constitute an identity matrix:*

- $[i,i] : i = 1$

$$\begin{aligned} IZ_1 + A_L^1 Z_U &= I\left(I + \frac{\hat{\rho}^2}{1-\hat{\rho}^2} L^v U^v\right) + \hat{\rho} L^v \frac{-\hat{\rho}}{1-\hat{\rho}^2} U^v \\ &= I \end{aligned} \tag{64}$$

- $[i,i] : 1 < i \leq \frac{n}{v}$

$$\begin{aligned} A_U^1 Z_L + IZ_2 + A_L^1 Z_U &= \hat{\rho} U^v \frac{-\hat{\rho}}{1-\hat{\rho}^2} L^v + I\left(I + \frac{\hat{\rho}^2}{1-\hat{\rho}^2}\left(L^v U^v + U^v L^v\right)\right) + \hat{\rho} L^v \frac{-\hat{\rho}}{1-\hat{\rho}^2} U^v \\ &= I \end{aligned} \tag{65}$$

- $[i,i] : i = \frac{n}{v} + 1$

$$\begin{aligned} (A_U^1)_{M,:}(Z_L)_{:,M} + (I)_{M,M}(Z_3)_{M,M} &= \hat{\rho}(U^v)_{M,:}\frac{-\hat{\rho}}{1-\hat{\rho}^2}(L^v)_{:,M} + (I)_{M,M}\frac{1}{1-\hat{\rho}^2}(I)_{M,M} \\ &= \frac{-\hat{\rho}^2}{1-\hat{\rho}^2}(I)_{M,M} + \frac{1}{1-\hat{\rho}^2}(I)_{M,M} \\ &= (I)_{M,M} \end{aligned} \tag{66}$$

- $[i,j] : 1 < i \leq \frac{n}{v}, j = 1$

$$\begin{aligned} A_U^{i-1} Z_1 + A_U^{i-2} Z_U &= \left(\hat{\rho} U^v\right)^{i-2}\left(\hat{\rho} U^v\left(I + \frac{\hat{\rho}^2}{1-\hat{\rho}^2} L^v U^v\right) + \frac{-\hat{\rho}}{1-\hat{\rho}^2} U^v\right) \\ &= \left(\hat{\rho} U^v\right)^{i-2}\frac{1}{1-\hat{\rho}^2}\left(\left(\hat{\rho} - \hat{\rho}^3\right) U^v + \hat{\rho}^3 U^v - \hat{\rho} U^v\right) \\ &= 0 \end{aligned} \tag{67}$$

- $[i,j] : i = \frac{n}{v} + 1, j = 1$

$$\begin{aligned} (A_U^{i-1})_{M,:} Z_1 + (A_U^{i-2})_{M,:} Z_U &= (A_U^{i-1} Z_1 + A_U^{i-2} Z_U)_{M,:} \\ &= (0)_{M,:} \end{aligned} \tag{68}$$

- $[i,j] : j < i < \frac{n}{v}, 1 < j \leq \frac{n}{v}$

$$\begin{aligned} &A_U^{i-j+1} Z_L + A_U^{i-j} Z_2 + A_U^{i-j-1} Z_U = \\ &= \left(\hat{\rho} U^v\right)^{i-j-1}\left(\left(\hat{\rho} U^v\right)^2 \frac{-\hat{\rho}}{1-\hat{\rho}^2} L^v + \hat{\rho} U^v\left(I + \frac{\hat{\rho}^2}{1-\hat{\rho}^2}\left(L^v U^v + U^v L^v\right)\right) + \frac{-\hat{\rho}}{1-\hat{\rho}^2} U^v\right) \\ &= \left(\hat{\rho} U^v\right)^{i-2}\frac{1}{1-\hat{\rho}^2}\left(-\hat{\rho}^3 U^{2v} L^v + \left(\hat{\rho} - \hat{\rho}^3\right) U^v + \hat{\rho}^3 U^v L^v U^v + \hat{\rho}^3 U^{2v} L^v - \hat{\rho} U^v\right) \\ &= 0 \end{aligned} \tag{69}$$

- $[i, j] : i = \frac{n}{v} + 1, 1 < j \leq \frac{n}{v}$

$$\begin{aligned}(A_U^{i-j+1})_{M,:} Z_L + (A_U^{i-j})_{M,:} Z_2 + (A_U^{i-j-1})_{M,:} Z_U &= (A_U^{i-j+1} Z_L + A_U^{i-j} Z_2 + A_U^{i-j-1} Z_U)_{M,:} \\ &= (0)_{M,:} \,. \end{aligned} \tag{70}$$

The fact that remaining submatrices above the diagonal equal 0 follows from the symmetry of $\widehat{V}_\phi$.

Proof of Proposition 2 *The proof is provided in Giacomini and White (2006, p. 1575).*

Lemma 3 *Provided that $\{X_t\}$ is stationary, plim$(\hat{\rho}) \neq 1$, and $v = 1$, it holds that:*

$$\sqrt{n}(\widehat{\lambda}_{ACV} - \lambda_{CV})^\top \phi \xrightarrow{\text{p}} 0 \tag{71}$$

and

$$\frac{\widehat{\lambda}_{ACV}^\top \widehat{V}_\phi \widehat{\lambda}_{ACV}}{\lambda_{CV}^\top \widehat{V}_\phi \lambda_{CV}} \xrightarrow{\text{p}} 1. \tag{72}$$

Proof of Lemma 3 *To prove this lemma, we first express $\widehat{\lambda}_{ACV}$ as function of m, n and ρ. First let us recapitulate that*

$$\widehat{\lambda}_{ACV} = \widehat{V}_\phi^{-1} B^\top \left(B \widehat{V}_\phi^{-1} B^\top \right)^{-1} b \tag{73}$$

and note that for $v = 1$, the system of restriction B and b representing partition implied by stationarity is the following:

$$B = \begin{pmatrix} \mathbf{1}_n^\top \otimes I, \, I_{:,M} \end{pmatrix} \qquad b = \begin{pmatrix} \mathbf{0}_m \\ 1 \end{pmatrix} \tag{74}$$

where $M = (1, 2, \ldots, m)$.

Consider any $\hat{\rho} \neq 1$, using the Lemma 2, we can express

$$\widehat{V}_\phi^{-1} B^\top = \frac{1}{\hat{\sigma}^2} \begin{pmatrix} Z_1 + Z_L \\ \mathbf{1}_{n-1} \otimes (Z_U + Z_2 + Z_L) \\ (Z_U)_{M,:} + (Z_3)_{M,M} \, I_{M,:} \end{pmatrix} \tag{75}$$

and furthermore

$$
\begin{aligned}
B\widehat{V}_{\phi}^{-1}B^{\top} &= \frac{1}{\hat{\sigma}^2}\left(Z_1 + Z_L + (n-1)\left(Z_U + Z_2 + Z_L\right) + \underbrace{I_{:,M}\left(Z_U\right)_{M,:}}_{=Z_U} + \underbrace{I_{:,M}\left(Z_3\right)_{M,M} I_{M,:}}_{=\frac{1}{1-\hat{\rho}^2}U^v L^v}\right) \\
&= \frac{1}{\hat{\sigma}^2}\left(n\left(Z_U + Z_2 + Z_L\right) + Z_1 - Z_2 + \frac{1}{1-\hat{\rho}^2}U^v L^v\right) \\
&= \frac{1}{\hat{\sigma}^2}\left(n\left(Z_U + Z_2 + Z_L\right) + U^v L^v\right) \\
&= \frac{1}{\hat{\sigma}^2}\frac{1}{1-\hat{\rho}^2}\left(n\left((1-\hat{\rho}^2)I + \hat{\rho}^2(L^v U^v + U^v L^v) - \hat{\rho}(U^v + L^v)\right) + (1-\hat{\rho}^2)U^v L^v\right).
\end{aligned}
\tag{76}
$$

Under $v = 1$, the resulting matrix is tridiagonal, in particular:

$$
B\widehat{V}_{\phi}^{-1}B^{\top} = \frac{1}{\hat{\sigma}^2}\frac{1}{1-\hat{\rho}^2}\underbrace{\begin{pmatrix}
a_1 & c & 0 & \dots & 0 & 0 & 0 \\
c & a_2 & c & \ddots & & 0 & 0 \\
0 & c & a_3 & \ddots & & & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\
0 & & & \ddots & a_{m-1} & c & (0 \\
0 & 0 & & \ddots & c & a_m & c \\
0 & 0 & 0 & \dots & 0 & c & a_{m+1}
\end{pmatrix}}_{\equiv Y}
\tag{77}
$$

with

- $a_1 = n + 1 - \hat{\rho}^2$
- $a_j = (1+\hat{\rho}^2)n + 1 - \hat{\rho}^2$, $1 < j < m+1$
- $a_{m+1} = n$
- $c = -n\hat{\rho}$.

Using the results of Usmani (1994) on the inverse of tridiagonal matrices, we know that the left-most column of Y^{-1} can be expressed as

$$
\begin{aligned}
\left(Y^{-1}\right)_{j,m+1} &= (-1)^{j+(m+1)}\, c^{(m+1)-j}\, \frac{\theta_{j-1}}{\theta_{m+1}} * 1 \\
&= (n\hat{\rho})^{m+1-j}\, \frac{\theta_{j-1}}{\theta_{m+1}}
\end{aligned}
\tag{78}
$$

with $\theta_0 = 1$, $\theta_1 = a_1$, and $\theta_j = a_j\theta_{j-1} + c^2\theta_{j-2}$ with $2 \leq j \leq m+1$. In our particular case it then *follows that*

$$\theta_j = \begin{cases} n^j + O(n^{j-1}) & 0 \leq j \leq m \\ (1-\hat{\rho}^2)n^j + O(n^{j-1}) & j = m+1, \end{cases} \tag{79}$$

which can be proven by induction as $\theta_0 = 1$ *and* $\theta_1 = n + 1 - \hat{\rho}^2$ *and for* $2 \leq j \leq m$ *it holds that*

$$\begin{aligned} \theta_j &= a_j \theta_{j-1} + c^2 \theta_{j-2} \\ &= \left((1+\hat{\rho}^2)n + 1 - \hat{\rho}^2\right)\left(n^{j-1} + O(n^{j-2})\right) - (-n\hat{\rho})^2\left(n^{j-2} + O(n^{j-3})\right) \\ &= n^j + O(n^{j-1}) \end{aligned} \tag{80}$$

and consequently for $j = m+1$

$$\begin{aligned} \theta_j &= a_j \theta_{j-1} + c^2 \theta_{j-2} \\ &= (n)\left(n^{j-1} + O(n^{j-2})\right) - (-n\hat{\rho})^2\left(n^{j-2} + O(n^{j-3})\right) \\ &= (1-\hat{\rho}^2)n^j + O(n^{j-1}). \end{aligned} \tag{81}$$

Therefore

$$\begin{aligned} \left(Y^{-1}\right)_{j,m+1} &= (n\hat{\rho})^{m+1-j}\,\frac{n^{j-1} + O(n^{j-2})}{(1-\hat{\rho}^2)n^{m+1} + O(n^m)} \\ &= \frac{\hat{\rho}^{m+1-j}n^m + O(n^{m-1})}{(1-\hat{\rho}^2)n^{m+1} + O(n^m)} \\ &= \frac{\hat{\rho}^{m+1-j}}{1-\hat{\rho}^2}\frac{1}{n} + O\left(\frac{1}{n^2}\right) \end{aligned} \tag{82}$$

and finally

$$\begin{aligned} \left(\left(B\widehat{V}_\phi^{-1}B^\top\right)^{-1} b\right)_j &= \hat{\sigma}^2\left(1-\hat{\rho}^2\right)\left(Y^{-1}\right)_{j,m+1} \\ &= \hat{\sigma}^2\hat{\rho}^{m+1-j}\frac{1}{n} + O\left(\frac{1}{n^2}\right) \end{aligned} \tag{83}$$

and furthemore using the definitions of Z_j*,* Z_U*, and* Z_L

$$\widehat{\lambda}_{ACV} = \begin{pmatrix} Z_1 + Z_L \\ \mathbf{1}_{n-1} \otimes (Z_U + Z_2 + Z_L) \\ (Z_U)_{M,:} + (Z_3)_{M,M}\, I_{M,:} \end{pmatrix} \begin{pmatrix} \frac{1}{n}\hat{\rho}^m + O(\frac{1}{n^2}) \\ \frac{1}{n}\hat{\rho}^{m-1} + O(\frac{1}{n^2}) \\ \vdots \\ \frac{1}{n}\hat{\rho}^1 + O(\frac{1}{n^2}) \\ \frac{1}{n}\hat{\rho}^0 + O(\frac{1}{n^2}) \end{pmatrix} = \begin{pmatrix} \frac{1}{n}P + \epsilon_1(n) \\ \mathbf{1}_{n-1} \otimes \left(\frac{1}{n}\begin{pmatrix} \mathbf{0}_m \\ 1 \end{pmatrix} + \epsilon_2(n)\right) \\ \frac{1}{n}\mathbf{0}_m + \epsilon_3(n) \end{pmatrix} \tag{84}$$

where $P = \left(\hat{\rho}^m,\, \hat{\rho}^{m-1},\, ...,\, \hat{\rho}^1,\, \hat{\rho}^0\right)^\top$ *and* ϵ_k *for* $k \in \{1,\, 2,\, 3\}$ *is a vector function that is element-wise* $O(\frac{1}{n^2})$.

With the explicit, albeit approximate (up to $O(\frac{1}{n^2})$*), expression for* $\widehat{\lambda}_{ACV}$*, we proceed with proving* *the individual claims. Let us denote*

$$
\begin{aligned}
\lambda_\Delta \equiv \widehat{\lambda}_{ACV} - \lambda_{CV} &= \begin{pmatrix} \frac{1}{n}P + \epsilon_1(n) \\ \mathbf{1}_{n-1} \otimes \left(\frac{1}{n} \begin{pmatrix} \mathbf{0}_m \\ 1 \end{pmatrix} + \epsilon_2(n) \right) \\ \frac{1}{n}\mathbf{0}_m + \epsilon_3(n) \end{pmatrix} - \begin{pmatrix} \mathbf{1}_n \otimes \left(\frac{1}{n} \begin{pmatrix} \mathbf{0}_m \\ 1 \end{pmatrix} \right) \\ \frac{1}{n}\mathbf{0}_m \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{n}\left(P - \begin{pmatrix} \mathbf{0}_m \\ 1 \end{pmatrix} \right) + \epsilon_1(n) \\ \mathbf{1}_{n-1} \otimes \epsilon_2(n) \\ \epsilon_3(n) \end{pmatrix}
\end{aligned} \tag{85}
$$

and furthermore

$$
\lambda_\Delta^\top \phi = \sum_{j=1}^{m+1} \underbrace{\left(\left(\frac{1}{n}\hat{\rho}^{m+1-j} + \epsilon_1(n)_j \right) l_j^{m,0} + \sum_{i=1}^{n-1} \epsilon_2(n)_j l_j^{m,i} + \epsilon_3(n)_j l_j^{m,n} \mathbf{1}(j \le m) \right)}_{\equiv Q_j}. \tag{86}
$$

Consider any $j \in \{1, 2, ..., m+1\}$*. From the definition of* $\epsilon_k(n)$*,* $k \in \{1, 2, 3\}$ *it follows that* $\exists C, n_0 : \forall n \ge n_0$*:*

$$
\begin{aligned}
0 \le |\sqrt{n}Q_j| &\le \sqrt{n}\left(\left(|\frac{1}{n}\hat{\rho}^{m+1-j}| + |\epsilon_1(n)_j| \right) |l_j^{m,0}| + \sum_{i=1}^{n-1} |\epsilon_2(n)_j||l_j^{m,i}| + |\epsilon_3(n)_j||l_j^{m,n}|\mathbf{1}(j \le m) \right) \\
&\le \sqrt{n}\frac{1}{n}\hat{\rho}^{m+1-j}|l_j^{m,0}| + \sqrt{n}\sum_{i=1}^{n-1} C\frac{1}{n^2}|l_j^{m,i}| + \sqrt{n}C\frac{1}{n^2}|l_j^{m,n}|\mathbf{1}(j \le m) \\
&= \underbrace{\frac{1}{\sqrt{n}}\hat{\rho}^{m+1-j}|l_j^{m,0}|}_{\xrightarrow{\text{p}}0} + \underbrace{\frac{1}{\sqrt{n}}C\frac{1}{n}\sum_{i=1}^{n-1}|l_j^{m,i}|}_{\xrightarrow{\text{p}}0} + \underbrace{\frac{1}{\sqrt{n}}C\frac{1}{n}|l_j^{m,n}|\mathbf{1}(j \le m)}_{\xrightarrow{\text{p}}0} \xrightarrow{\text{p}} 0.
\end{aligned} \tag{87}
$$

Considering that

$$
-\sum_{j=1}^{m+1} |\sqrt{n}Q_j| \le -|\sqrt{n}\sum_{j=1}^{m+1} Q_j| \le \sqrt{n}\lambda_\Delta^\top \phi \le \sum_{j=1}^{m+1} |\sqrt{n}Q_j| \le |\sqrt{n}\sum_{j=1}^{m+1} Q_j| \tag{88}
$$

it follows that

$$
\sqrt{n}(\lambda_{ACV} - \widehat{\lambda}_{CV})^\top \phi \xrightarrow{\text{p}} 0 \tag{89}
$$

via Squeeze theorem. To prove the second claim, note that

$$\begin{aligned}\frac{\widehat{\lambda}_{ACV}^\top \widehat{V}_\phi \widehat{\lambda}_{ACV}}{\lambda_{CV}^\top \widehat{V}_\phi \lambda_{CV}} &= \frac{(\lambda_{CV} + \lambda_\Delta)^\top \widehat{V}_\phi (\lambda_{CV} + \lambda_\Delta)}{\lambda_{CV}^\top \widehat{V}_\phi \lambda_{CV}} \\ &= \frac{\lambda_{CV}^\top \widehat{V}_\phi \lambda_{CV} + \lambda_\Delta^\top \widehat{V}_\phi \lambda_{CV} + \lambda_{CV}^\top \widehat{V}_\phi \lambda_\Delta + \lambda_\Delta^\top \widehat{V}_\phi \lambda_\Delta}{\lambda_{CV}^\top \widehat{V}_\phi \lambda_{CV}}.\end{aligned} \tag{90}$$

Let us denote

$$\tilde{\epsilon}_1(n) = |\frac{1}{n}\left(P - \begin{pmatrix}\mathbf{0}_m \\ 1\end{pmatrix}\right) + \epsilon_1(n)| \tag{91}$$

$$e(n) = \frac{1}{n}\begin{pmatrix}\mathbf{0}_m \\ 1\end{pmatrix}. \tag{92}$$

From the definition of $\epsilon_k(n)$, $k \in \{1, 2, 3\}$ it follows that $\exists C, n_0 : \forall n \geq n_0$:

$$\begin{aligned} n\lambda_\Delta^\top \widehat{V}_\phi \lambda_{CV} &\leq n|\lambda_\Delta^\top||\widehat{V}_\phi||\lambda_{CV}| \\ &\leq n\hat{\sigma}^2 \left((2m+1)|\tilde{\epsilon}_1(n)|^\top J e(n) + n(2m+1)|\epsilon_2(n)|^\top J e(n) + (2m+1)|\epsilon_3(n)|^\top J e(n)\right) \\ &\leq n\hat{\sigma}^2(m+1)\left((2m+1)C\frac{1}{n}\frac{1}{n} + n(2m+1)C\frac{1}{n^2}\frac{1}{n} + (2m+1)C\frac{1}{n^2}\frac{1}{n}\right) \xrightarrow{\text{p}} 0. \end{aligned} \tag{93}$$

Where we utilized the fact that $\frac{1}{\hat{\sigma}^2}\widehat{V}_\phi$ can be bounded from above by a block-Toeplitz matrix with a matrix of ones (denoted by J) on the diagonal and first m sub/super-diagonals. Similarly for

$$\begin{aligned} n\lambda_\Delta^\top \widehat{V}_\phi \lambda_\Delta &\leq n|\lambda_\Delta^\top||\widehat{V}_\phi||\lambda_\Delta| \\ &\leq n\hat{\sigma}^2(|\tilde{\epsilon}_1(n)|^\top J|\tilde{\epsilon}_1(n)| + 2(n-1)|\tilde{\epsilon}_1(n)|^\top J|\epsilon_2(n)| + (n-1)^2|\epsilon_2(n)|^\top J|\epsilon_2(n)| + \\ &+ 2|\tilde{\epsilon}_1(n)|^\top J|\epsilon_3(n)| + 2(n-1)|\epsilon_2(n)|^\top J|\epsilon_3(n)| + 2|\epsilon_3(n)|^\top J|\epsilon_3(n)|) \\ &\leq n\hat{\sigma}^2(m+1)^2(C\frac{1}{n}\frac{1}{n} + 2(n-1)C\frac{1}{n}\frac{1}{n^2} + (n-1)^2 C\frac{1}{n^2}\frac{1}{n^2} + \\ &+ 2C\frac{1}{n}\frac{1}{n^2} + 2(n-1)C\frac{1}{n^2}\frac{1}{n^2} + C\frac{1}{n^2}\frac{1}{n^2}) \xrightarrow{\text{p}} 0. \end{aligned} \tag{94}$$

Utilizing the Squeeze theorem, we obtain $n\lambda_\Delta^\top \widehat{V}_\phi \lambda_{CV} \xrightarrow{\text{p}} 0$ and $n\lambda_\Delta^\top \widehat{V}_\phi \lambda_\Delta \xrightarrow{\text{p}} 0$. By noting that $plim(n\lambda_{CV}^\top \widehat{V}_\phi \lambda_{CV}) = const$ we can invoke Slutsky's theorem to obtain

$$\frac{\widehat{\lambda}_{ACV}^\top \widehat{V}_\phi \widehat{\lambda}_{ACV}}{\lambda_{CV}^\top \widehat{V}_\phi \lambda_{CV}} = \frac{n\widehat{\lambda}_{ACV}^\top \widehat{V}_\phi \widehat{\lambda}_{ACV}}{n\lambda_{CV}^\top \widehat{V}_\phi \lambda_{CV}} \xrightarrow{\text{p}} 1. \tag{95}$$

Proof of Proposition 3 *Applying lemma 3 to the contrasts differential $\Delta\phi$, it follows that*

$$\sqrt{n}(\widehat{\lambda}_{ACV} - \lambda_{CV})^\top \Delta\phi \xrightarrow{\text{p}} 0 \tag{96}$$

$$\frac{\widehat{\lambda}_{ACV}^{\top}\widehat{V}_{\Delta\phi}\widehat{\lambda}_{ACV}}{\lambda_{CV}^{\top}\widehat{V}_{\Delta\phi}\lambda_{CV}} \xrightarrow{\text{p}} 1 \tag{97}$$

noting that

$$t_{ADM} \equiv \frac{(\widehat{\lambda}_{ACV})^{\top}\Delta\phi}{\widehat{\sigma}_{ACV}/\sqrt{n}} = \frac{\sqrt{n}(\lambda_{CV})^{\top}\Delta\phi + \sqrt{n}(\widehat{\lambda}_{ACV} - \lambda_{CV})^{\top}\Delta\phi}{\widehat{\sigma}_{CV}\,\frac{\widehat{\lambda}_{ACV}^{\top}\widehat{V}_{\Delta\phi}\widehat{\lambda}_{ACV}}{\lambda_{CV}^{\top}\widehat{V}_{\Delta\phi}\lambda_{CV}}} \tag{98}$$

and hence via Slutsky's theorem

$$plim(t_{ADM}) = plim(t_{DM}). \tag{99}$$

Combing this with already established results from Proposition 2, both

$$t_{ADM} \xrightarrow{\text{d}} N(0,1) \tag{100}$$

and

$$P\left(|t_{ADM}| > c\right) \longrightarrow 1 \tag{101}$$

immediately follow.

Proof of Proposition 4 *The proof is provided in Zhu and Timmermann (2020). Just note that stationarity of $\left\{\Delta l_{m+1}^{m,i}\right\}$ follows from the stationarity of $\{X_t\}$.*

Proof of Proposition 5 *From Lemma 3 it follows that $\forall k \in \{1, ..., K\}$:*

$$plim\left(\sqrt{\tilde{n}}\widehat{\mathcal{L}}_{CV}^{(k)}\right) = plim\left(\sqrt{\tilde{n}}\widehat{\mathcal{L}}_{ACV}^{(k)}\right). \tag{102}$$

As

$$\sqrt{\tilde{n}}\left(\widehat{\mathcal{L}}_{CV}^{(1)}, ..., \widehat{\mathcal{L}}_{CV}^{(K)}\right) \xrightarrow{\text{d}} N(0, c^2 I) \tag{103}$$

where $c^2 = E[\Delta l_{m+1}^{m,i}] + 2\sum_{s=1}^{\infty} E[\Delta l_{m+1}^{m,i}\Delta l_{m+1}^{m,i+s}]$ (see Zhu and Timmermann (2020)), it then immediately follows that also

$$\sqrt{\tilde{n}}\left(\widehat{\mathcal{L}}_{ACV}^{(1)}, ..., \widehat{\mathcal{L}}_{ACV}^{(K)}\right) \xrightarrow{\text{d}} N(0, c^2 I). \tag{104}$$

The rest of the proof coincides with Zhu and Timmermann (2020).

Proof of Proposition 6 *For for both models $i \in \{1, 2\}$ the losses (conditioned on sequence $\{Z_t\}$) can be written as:*

$$\mathcal{L}_{m+1}^{m}(\mathcal{M}_i) = \sum_{t=m}^{T-1} \mathbb{E}\left[Y_{t+1} - \widehat{f}_t^i\right]^2 + Var\left(Y_{t+1} - \widehat{f}_t^i\right). \tag{105}$$

For the bias term of model $\mathcal{M}_1$ we have

$$\mathbb{E}\left[Y_{t+1}-\widehat{f}_t^1\right]^2=\left(c+Z_t-\mathbb{E}\left[\widehat{\beta}_{1,t}^1\right]Z_t\right)^2=c^2\left(1-\frac{\overline{Z}_t}{\overline{Z^2}_t}Z_t\right)^2 \tag{106}$$

and for the variance term

$$Var\left(Y_{t+1}-\widehat{f}_t^1\right)=Var\left(Y_{t+1}-\widehat{\beta}_{1,t}^1Z_t\right)=\sigma^2\left(1+\frac{Z_t^2}{(m-1)\,\overline{Z^2}_t}\right). \tag{107}$$

For model $\mathcal{M}_2$ the bias term is equal to 0 *and the variance term is*

$$\begin{aligned} Var\left(Y_{t+1}-\widehat{f}_t^2\right)&=Var\left(Y_{t+1}-\widehat{\beta}_{0,t}^2-\widehat{\beta}_{1,t}^2Z_t\right)\\ &=\sigma^2+Var\left(\widehat{\beta}_{0,t}^2\right)+Z_t^2Var\left(\widehat{\beta}_{1,t}^2\right)+2Z_tCov\left(\widehat{\beta}_{0,t}^2,\widehat{\beta}_{1,t}^2\right)\\ &=\sigma^2\left(1+\frac{\overline{Z^2}_t}{S_t}+\frac{Z_t^2}{S_t}-2\frac{\overline{Z}_t}{S_t}Z_t\right). \end{aligned} \tag{108}$$

By setting

$$\varsigma=\frac{\mathcal{L}_{m+1}^m(\mathcal{M}_1)}{\mathcal{L}_{m+1}^m(\mathcal{M}_2)} \tag{109}$$

and solving for c, we obtain

$$c=\sigma\left(\frac{\sum_{t=m}^{T-1}\left(1+\frac{\overline{Z^2}_t}{S_t}+\frac{Z_t^2}{S_t}-2\frac{\overline{Z}_t}{S_t}Z_t\right)\varsigma-\sum_{t=m}^{T-1}\left(1+\frac{Z_t^2}{(m-1)\overline{Z^2}_t}\right)}{\sum_{t=m}^{T-1}\left(1-\frac{\overline{Z}_t}{\overline{Z^2}_t}Z_t\right)^2}\right)^{0.5}. \tag{110}$$

B Additional Results

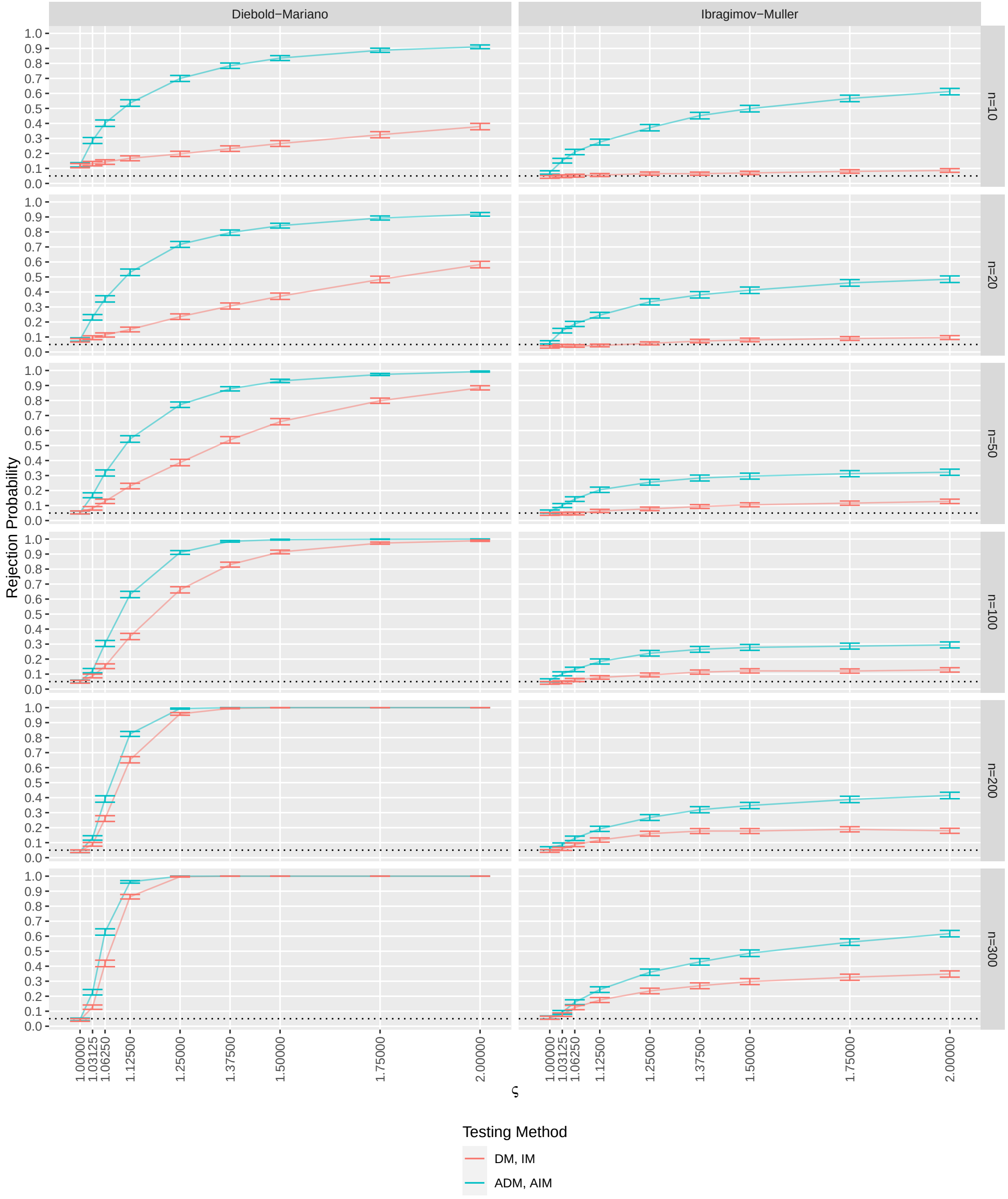


Figure 6: A plot of rejection probabilities for DM, IM, ADM, and AIM tests at level 0.05 for the first log difference of US CPI index.

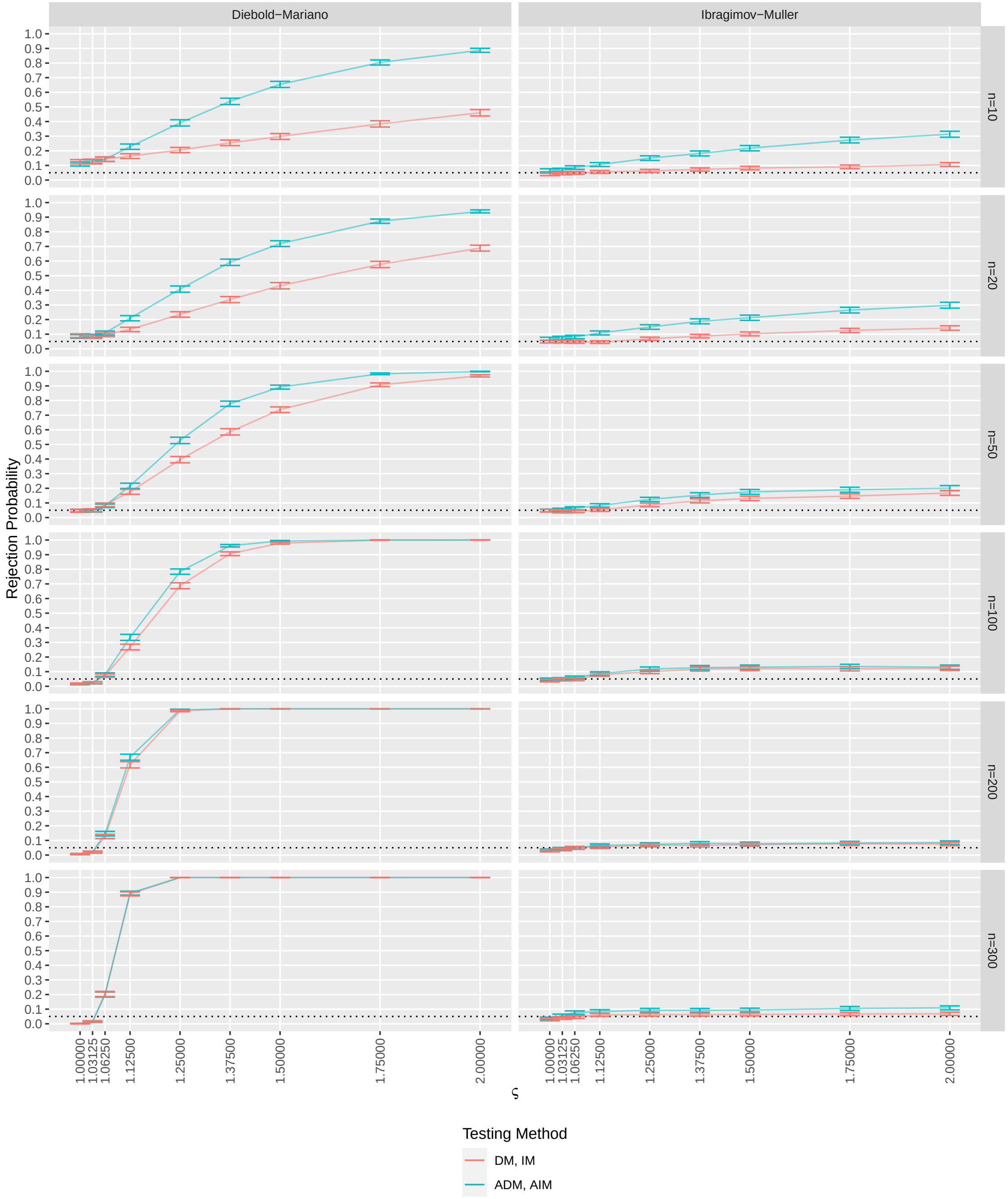


Figure 7: A plot of rejection probabilities for DM, IM, ADM, and AIM tests at level 0.05 for the log of US CPI index.

Abstrakt

Časová řady jsou často rozdělovány na estimační a evaluační část za účelem odhadnutí kvality předpovědí produkovaných daným statistickým modelem. Navrhujeme alternativní estimátor kvality předpovědí který, navzdory intuici, využívá pro estimaci kvalitu předpovědí jak z estimační, tak z evaluační části dat, a to pomocí specifického systému afinních vah. Dokážeme, že navrhovaný estimátor je optimální ve třídě nestranných lineárních estimátorů, a tudíž nabízí vyšší přesnost než konvenční estimátor. Aplikace navrhovaného estimátoru v Diebold-Mariano testech prediktivní schopnosti vede k vyšší síle testů při zachování stejné míry zkreslení v malých vzorcích. Evaluace navrhovaného estimátoru na časových řadách ze soutěže M4 potvrzuje superioritu vůči konvenčnímu estimátoru, a navíc ukazuje, že navrhovaný estimátor je poměrně robustní vůči porušení klíčového předpokladu stacionarity.

Working Paper Series
ISSN 1211-3298
Registration No. (Ministry of Culture): E 19443

Individual researchers, as well as the on-line and printed versions of the CERGE-EI Working Papers (including their dissemination) were supported from institutional support RVO 67985998 from Economics Institute of the CAS, v. v. i.

Specific research support and/or other grants the researchers/publications benefited from are acknowledged at the beginning of the Paper.

Published by
Charles University, Center for Economic Research and Graduate Education (CERGE)
and
Economics Institute of the CAS, v. v. i. (EI)
CERGE-EI, Politických vězňů 7, 111 21 Prague 1, tel.: +420 224 005 153, Czech Republic.
Printed by CERGE-EI, Prague
Subscription: CERGE-EI homepage: http://www.cerge-ei.cz

Phone: + 420 224 005 153
Email: office@cerge-ei.cz
Web: http://www.cerge-ei.cz

Editor: Byeongju Jeong

The paper is available online at http://www.cerge-ei.cz/publications/working_papers/.

ISBN 978-80-7343-519-6 (Univerzita Karlova, Centrum pro ekonomický výzkum a doktorské studium)
ISBN 978-80-7344-614-7 (Národohospodářský ústav AV ČR, v. v. i.)