Tests of Conditional Predictive Ability: Existence, Size, and Power

Tests of Conditional Predictive Ability: Existence, Size, and Power*

Michael W. McCracken
Federal Reserve Bank of St. Louis

December 18, 2020

Abstract

We investigate a test of conditional predictive ability described in Giacomini and White (2006; *Econometrica*). Our main goal is simply to demonstrate existence of the null hypothesis and, in doing so, clarify just how unlikely it is for this hypothesis to hold. We do so using a simple example of point forecasting under quadratic loss. We then provide simulation evidence on the size and power of the test. While the test can be accurately sized we find that power is typically low.

JEL Nos.: C53, C12, C52
Keywords: prediction, out-of-sample, inference

*McCracken: Research Division; Federal Reserve Bank of St. Louis; P.O. Box 442; St. Louis, MO 63166; michael.w.mccracken@stls.frb.org. We are grateful to Ken West, Tatevik Sekhposyan, Minchul Shin, and Rafaella Giacomini for helpful comments as well as participants at the 2019 International Symposium on Forecasting, the 2019 International Association for Applied Econometrics, and the Federal Reserve System Committee on Econometrics Conference. The views expressed herein are solely those of the author and do not necessarily reflect the views of the Federal Reserve Bank of St. Louis or the Federal Reserve System.

1 Introduction

Tests of equal forecast accuracy are a now standard procedure in the forecasting literature. Early advocates include Diebold and Mariano (1995) who show that a t-statistic, based on sample averages of loss differentials, can be asymptotically normal. Asymptotically valid inference on the mean of the loss differentials is therefore straightforward. Subsequent work include adaptations to multiple testing (White 2000), comparisons between nested models (Clark and McCracken 2001, McCracken 2007), cointegrating relationships (Corradi, Olivetti, and Swanson, 2001), forecast breakdowns (Giacomini and Rossi, 2009), generated predictors (Goncalves, McCracken, and Perron 2017), generated predictands (Li and Patton, 2018), and many others.

Amidst many of these extensions is the issue of parameter estimation error. An early advocate of this issue is West (1996) who delineates conditions under which, when the forecasts include parameter estimates, not only is the t-statistic asymptotically normal, it is asymptotically standard normal. Specifically, he shows how to properly account for the effect parameter estimation error has on the asymptotic variance of the average loss differential.

Parameter estimation error is also a key feature of Giacomini and White (2006, henceforth GW) but their perspective on the issue is very different. They begin with a question: do we care about population-level forecast errors $u = u(\beta^*)$, which we can never observe, or do we care about finite-sample-level forecast errors $\hat{u} = u(\hat{\beta})$ which we do observe? As an example of this perspective, consider comparing the accuracy of h-step-ahead point forecasts from two parametric models ($i = 1, 2$) under quadratic loss. The bulk of the theoretical literature focuses on a null hypothesis of the form $E(u_{1,t+h}^2 - u_{2,t+h}^2) = 0$, while GW consider hypotheses of the form $E(\hat{u}_{1,t+h}^2 - \hat{u}_{2,t+h}^2) = 0$ and $E(\hat{u}_{1,t+h}^2 - \hat{u}_{2,t+h}^2 | \Im_t) = 0$. The first of these three hypotheses is a statement about the unconditional expectation of the population-level loss differential, while the second and third are statements about the unconditional and conditional (based on a time-t information set $\Im_t$) expectation of the finite-sample loss differential.[1]

In this paper we investigate the test of conditional predictive ability advocated by GW, that associated with the null hypothesis $E(\hat{u}_{1,t+h}^2 - \hat{u}_{2,t+h}^2 | \Im_t) = 0$.[2] Our first goal is to provide a proof-by-example that the null hypothesis can exist. We do so by delineating a data generating process (DGP) and parametric forecasting environment in which the null hypothesis holds. This may seem trivial

[1] In GW, the hypotheses are stated more broadly and apply to any well-behaved loss function and are not exclusive to point forecasts. Even so, for exposition purposes, we will focus on the most common applications in which quadratic loss is used to evaluate point forecasts.

[2] The unconditional test of equal predictive ability has been more thoroughly investigated. Examples include Coroneo and Iacone (2018), Giacomini and Rossi (2010), Rossi and Sekhposyan (2018), Inoue and Rossi (2012), Clark and McCracken (2013), McCracken (2020), and Zhu and Timmermann (2020).

but there exist no examples in which this null hypothesis can be shown to hold when parameters are estimated.[3] In large part this arises because the null hypothesis of equal conditional predictive ability is extremely narrow and imposes very strong restrictions on the DGP. As shown in GW, given two sequences of h-step-ahead point forecasts $(\hat{y}_{1,t+h}, \hat{y}_{2,t+h}) \in \Im_t$ and assuming a quadratic loss function, the null hypothesis requires that the DGP for y satisfies $E(y_{t+h}|\Im_t) = \frac{1}{2}(\hat{y}_{1,t+h} + \hat{y}_{2,t+h})$. Zhu and Timmermann (2020) go one step further and show that this relationship is both necessary and sufficient.

Despite these difficulties, we are able to delineate such a DGP but only for an application in which, under quadratic loss, a no-change forecast is being compared to one formed using a location model. We have not been able to establish an example that permits a broader collection of regression models and it remains unclear if such exist. In particular, Zhu and Timmermann (2020) prove that if one of the models is not finite order Markov, such as an MA or ARMA model, the null cannot hold.

Having provided an example in which the null can hold, our next goal is to investigate the finite sample size and power of the test based on two Wald statistics proposed in GW. We find that when the forecast horizon is 1, the nominal size of the test can be quite good. As the horizon grows, size distortions arise due to the need to estimate long-run variances. In accordance with the theory, these size distortions dissipate as the sample size increases.

We then provide an array of results on the finite sample power of the test using the same Wald statistics. To do so we emphasize the fact that, in the GW framework, the null hypothesis will hold only if the forecasting agent selects a forecasting method that aligns with the DGP in very specific ways we discuss later. We find that for most sample sizes and deviations from the null, the actual power of the test aligns closely with rejection frequencies observed under the null. Reasonable levels of finite sample power typically arise only when the out-of-sample period is quite large and even then it rarely exceeds 70%.

The remainder proceeds as follows. Section 2 delineates DGPs and forecasting environments which satisfies the null hypothesis $E(\hat{u}^2_{1,t+h} - \hat{u}^2_{2,t+h}|\Im_t) = 0$. Section 3 provides evidence on the actual size and power of of tests of the null hypothesis. Section 4 concludes.

2 Existence

In this section we delineate a DGP and forecast environment that satisfies the null hypothesis $E(\hat{u}^2_{1,t+h} - \hat{u}^2_{2,t+h}|\Im_t) = 0$ for all $t = R, ..., T-h$. Before doing so, recall that GW require that a

[3] GW provide Monte Carlo evidence on size and power by simulating the loss differentials directly rather than delineating a DGP and forecasting environment that would imply those loss differentials.

finite number of observations (R) are used to estimate the parameters. This prevents the parameter estimates, and subsequent forecast errors, from converging to their population counterparts.[4] Specifically, either the estimates depend on a rolling window of observations x such that $\hat{\beta}_t = \beta(x_{t-R+1}, ..., x_t)$ or a fixed window of observations such that $\hat{\beta}_t = \beta(x_1, ..., x_R)$. Because of this restriction, the DGP that we describe is specific to whether the fixed or rolling scheme is used. This means that if the DGP is designed relative to the fixed (rolling) scheme, then the fixed (rolling) scheme must be used for forecasting.

Our DGPs are developed based on a simple application in which a no-change forecast is compared to that from a location model with a single estimated parameter. That is, $\hat{y}_{1,t+h} = 0$, while $\hat{y}_{2,t+h} = \bar{y}_t$, where $\bar{y}_t = R^{-1} \sum_{s=t-R+1}^{t} y_s$ under the rolling scheme and $\bar{y}_t = \bar{y}_R$ for all t under the fixed scheme. Holding the estimation scheme constant, the following two bullets describe the relevant DGPs.

- Fixed DGP. For $t = 1, ..., R$, set $y_t = 2\eta_t$, where $\eta_t = \varepsilon_t + \sum_{j=1}^{h-1} \theta_j \varepsilon_{t-j}$ with $\varepsilon_t \sim i.i.d.N(0, \sigma^2)$. For $t = R+1, ..., T$, set $y_t = \frac{1}{2}\bar{y}_R + \eta_t$.

- Rolling DGP. For $t = 1, ..., T$, y_t forms a stationary $ARMA(R+h-1, h-1)$ with autoregressive parameters α_j set so that the first $h-1$ values are zero (i.e., $\alpha_j = 0$ for $j = 1, ..., h-1$) and the remaining R values are $\frac{1}{2R}$ (i.e., $\alpha_j = \frac{1}{2R}$ for $j = h, ..., R+h-1$). The MA component takes the form $\varepsilon_t + \sum_{j=1}^{h-1} \theta_j \varepsilon_{t-j}$ with $\varepsilon_t \sim i.i.d.N(0, \sigma^2)$.

Proposition 2.1 *For each DGP and the corresponding parameter estimation scheme,* $E(\hat{u}^2_{1,t+h} - \hat{u}^2_{2,t+h} | \Im_t) = 0$.

Proof. In each case the proof is a straightforward application of the observation that for $t = R, ..., T-h$, the DGP for y satisfies $E(y_{t+h}|\Im_t) = \frac{1}{2}(\hat{y}_{1,t+h} + \hat{y}_{2,t+h})$. ■

The proposition provides examples in which the null hypothesis holds. And yet it is hard to not feel that the examples are strange and very unlikely to ever exist in economic data. As an extreme example, consider a case in which $\hat{y}_{1,t+h}$ and $\hat{y}_{2,t+h}$ denote two sequences of month-over-month U.S. PCE-based inflation forecasts made by the research teams at Morgan Stanley and J.P. Morgan respectively. The null hypothesis implies that in each period t, the conditional mean of month-over-month U.S. PCE-based inflation is the sample average of their two forecasts. This seems very unlikely and so much so that it is strange to even consider testing the null hypothesis. In addition, if $\hat{y}_{3,t+h}$ denotes the inflation forecast from Goldman-Sachs, it is logically inconsistent that both $E(\hat{u}^2_{1,t+h} - \hat{u}^2_{2,t+h} | \Im_t) = 0$ and

[4] As a technical matter, they permit R to vary across forecast origins, but the maximum is bounded from above. For brevity we assume it is constant.

$E(\hat{u}^2_{1,t+h} - \hat{u}^2_{3,t+h}|\Im_t) = 0$ hold simultaneously unless the J.P. Morgan and Goldman-Sachs forecasts are always identical.[5]

One reaction to this extreme example is that it is perfectly reasonable that the null hypothesis does not hold. This is true, but then we would want tests to have substantial power to detect deviations from the null hypothesis. For that reason, in the following section we provide Monte Carlo experiments designed to characterize the finite sample size and power of two test statistics proposed in GW.

3 Actual Size and Power

In this section we operationalize the examples in the previous section in order to investigate the finite sample size and power properties of two test statistics proposed by GW. Given a sequence of loss differentials $\hat{d}_{t+h} = \hat{u}^2_{1,t+h} - \hat{u}^2_{2,t+h}$, $t = R, ..., T-h = R+P-1$, the first is a t-statistic of the form $P^{-1/2}\sum_{t=R}^{T-h}\hat{d}_{t+h}/\hat{\omega}$ where $\hat{\omega}^2$ is a consistent estimate of the long-run variance of $\hat{d}_{t+h}$, ω^2. Under standard moment and memory conditions on the loss differentials, this t-statistic is asymptotically standard normal under the null. For the second, GW note that since the null hypothesis implies $\hat{d}_{t+h}$ is uncorrelated with any observable $z_t \in \Im_t$, $P^{-1/2}\sum_{t=R}^{T-h}\hat{d}_{t+h}z_t$ will also be asymptotically normal under the null. Since z_t need not be scalar, they recommend a Wald statistic of the form $(P^{-1/2}\sum_{t=R}^{T-h}\hat{d}_{t+h}z_t)'\hat{\Omega}^{-1}(P^{-1/2}\sum_{t=R}^{T-h}\hat{d}_{t+h}z_t)$. This statistic is asymptotically chi-square with $\dim(z_t)$ degrees of freedom under the null so long as $\hat{\Omega}$ is consistent for the full rank long-run variance of $\hat{d}_{t+h}z_t$, Ω. Following the suggestion in GW, we investigate the usefulness of this statistic using $z_t = (1, \hat{d}_t)'$ as the test function.

As there are no empirical applications that use the fixed scheme and test this null hypothesis, we only consider the case in which model parameters are estimated using a rolling window of observations.[6] That said, we do report results based on the fixed window DGP. We do so since, in that case, rolling window estimates of model parameters constitutes a deviation from the null and thus inform us on one type of deviation from the null hypothesis. We also consider two other deviations from the null. For the first, the window size used to estimate the parameters ($\tilde{R}$) does not align with that maintained in the DGP (R).[7] For the second, the chosen forecast horizon ($\tilde{h}$) does not align with that maintained by the DGP (h).

[5] In contrast, it is easy to imagine that all three research teams are equally accurate unconditionally and hence $E(\hat{u}^2_{1,t+h} - \hat{u}^2_{j,t+h}) = 0$, $j = 2, 3$.

[6] Unreported simulations that estimate model parameters using a fixed window of observations yield comparable rejection frequencies and are omitted for brevity.

[7] In the context of tests of equal unconditional finite-sample predictive ability, these deviations from the null are noted, but not investigated, by Rossi and Sekhposyan (2019). Inoue and Rossi (2012) consider the issue but from the standpoint of data snooping. They develop a test that is robust to the choice of $\tilde{R}$. The choice of $\tilde{R}$ is not treated as a potential deviation from the null.

The following bullets delineate the remaining elements of our simulation design.

- We use sample sizes in which R, $\tilde{R}$, and P range across 25, 75, 25, and 175 – similar to those used in GW. We also consider $P = 1000$. In each simulation, the total sample size T is $\tilde{R} + P + h - 1$.

- Throughout we set the variance σ^2 of the primitive shocks ε_t to 1. Unreported results that set σ^2 to 0.1 and 10 are quite similar and excluded for brevity. The MA coefficients are set to $\theta_j = (0.5)^j$ in all simulations.

- For most results we consider horizons in which $h = \tilde{h}$ and h ranges across 1, 3, and 12. For power experiments in which $h \neq \tilde{h}$ we also consider a longer horizon of 24.[8]

- Under the null, when $h = 1$ there is no serial correlation in either $\hat{d}_{t+1}$ or $\hat{d}_{t+1} z_t$ and hence their variance is the long-run variance. We therefore construct $\hat{\Omega}$ as $P^{-1} \sum_{t=\tilde{R}}^{T-1} \hat{d}_{t+1}^2 z_t z_t'$ and $\hat{\omega}^2$ as $P^{-1} \sum_{t=\tilde{R}}^{T-1} \hat{d}_{t+1}^2$ when $\tilde{h} = 1$. For the longer horizons, we use the Bartlett kernel to estimate ω^2 and Ω (i.e., Newey and West, 1987). The choice of bandwidth has an affect: hence, we experimented with a few, including fixed values of $\tilde{h}$ and a sample-dependent version $\lfloor 4(P/100)^{2/9} \rfloor + 1$ (Newey and West, 1994; Andrews and Monihan, 1992).[9] Since the actual sizes of the tests were weakly better using the data dependent rule, we use that bandwidth when $\tilde{h} > 1$.

- In all tables, we report results associated with a nominal size of 5%. Unreported results associated with 1% and 10% levels provide comparable results and are excluded for brevity.

- For the rolling DGP, the initial conditions are set to zero and a burn-in of $10,000$ is used.

- In all experiments, the number of replications is $5,000$.

3.1 t-statistic

In Table 1 we provide rejection frequencies associated with the t-statistic version of the test when the rolling scheme is used in the DGP and in the modeling. Values in bold denote the actual size. All other values are the actual power since for these, $\tilde{R} \neq R$. When $h = 1$, actual size is quite good for all sample sizes. As we increase the forecast horizon, the need to estimate long-run variances induces size distortions especially for the smallest sample sizes P. Even so, it is comforting to see that for $h = 3$, the distortions diminish quickly as the sample size increases. At the longest horizon $h = 12$, the actual size improves as the sample increases but remains slightly elevated even when $P = 1000$.

[8] In unreported results we consider actual size of the test when $h = 24$. The results were comparable to those for $h = 12$ and hence were excluded for brevity.

[9] Note that when using $\hat{d}_t$ as a test function, $\tilde{h}$ observations are lost when constructing the second test statistic. Hence when estimating Ω, we use the bandwidth $\lfloor 4((P - \tilde{h})/100)^{2/9} \rfloor + 1$.

Under the alternative in which $\tilde{R} \neq R$, (Table 1, not bold), the actual power is poor, with rejection frequencies generally well below 25% for all but the largest sample size P. For the largest sample size, the actual power can be as high as 80% but remains as low as 10% for some permutations of $\tilde{R} \neq R$. For a fixed value of R, actual power tends to be larger when $\tilde{R}$ is smaller, rather than larger, than R.

In Table 2 we report rejection frequencies when rolling windows are used to estimate model parameters but the DGP is associated with the fixed scheme. It is therefore the case that all values represent actual power. For this alternative, actual power is comparable to that in Table 1 except for the largest sample size where it can be modestly higher. Across all horizons, the actual power is rarely above 30% unless $P = 1000$. At the largest sample size, the rejection frequencies approach 75% but do so only when $\tilde{R}$ is, as we saw in Table 1, much smaller than R. Somewhat oddly, there are times when the rejection frequencies are u-shaped as P increases for a fixed $\tilde{R}$. For example, when $h = 12$, $R = 125$, and $\tilde{R} = 175$, the rejection frequencies vary from 18%, 13%, 12%, 14%, to 25% as P increases from 25, 75, 125, 175, to 1000 respectively.

Tables 3 and 4 report rejection frequencies when we vary $\tilde{h}$ across 1, 3, 12, and 24 for a fixed value of $h = 1$, 3, 12. Values in bold denote the actual size while those not in bold denote actual power. In both Tables, $R = \tilde{R}$ but while Table 3 reports results when the rolling DGP is used, Table 4 provides results when the fixed DGP is used and hence there are types of deviation from the null. Subject to Monte Carlo variation and rounding, in Table 3 the actual size results align with those in Table 1. In terms of actual power, both Tables are broadly similar. For all sample sizes less than 1000, rejection frequencies are rarely higher than 15% with the exception that when $R = 25$, actual power leaps to values greater than 90% when $\tilde{h} = 24$ especially when $h = 1$. As h increases, actual power decreases monotonically holding P constant when $\tilde{h} = 24$ and $R = 25$. In Table 3, when $P = 1000$, actual power is typically quite poor with rejection frequencies that often align with the actual size. In Table 4, for which their are two types of deviations from the null, actual power is comparable to that in Table 3 except for the largest sample size where it can be modestly higher.

3.2 χ^2-statistic

In this section, the simulations parallel those in the previous subsection but now use the χ^2-statistic-version of the test with $z_t = (1, \hat{d}_t)'$ as the test function. In Table 5, bolded values indicate that the test is modestly-to-severely oversized. At the longer horizons, the actual size of the test is typically poor for the smaller sample sizes, with rejection frequencies as high as 40%. Even so, consistent with the theory, at all horizons, the actual size generally improves as the sample size increases. In fact, when $P = 1000$ and $h = 1$, the actual size perfectly aligns with the nominal size.

It is worth reiterating that, in Table 5, heteroskedasticity and autocorrelation consistent methods are being used to estimate the long-run variance of $\hat{d}_{t+h}z_t$. While this obviously captures the role of serial correlation in the long-run variance, it also captures the presence of any conditional heteroskedasticity. This is particularly important given that $\hat{d}_t$ is an element of z_t. To understand why, suppose $h = 1$, let $\tilde{R} = R$, and consider the second element of $P^{-1/2}\sum_{t=R}^{T-1}\hat{d}_{t+1}z_t$. Straightforward algebra reveals that $P^{-1/2}\sum_{t=R}^{T-1}\hat{d}_{t+1}\hat{d}_t$ is asymptotically normal with zero mean and variance $\Omega_{22} = E\hat{d}_{t+1}^2\hat{d}_t^2$. Since $\hat{d}_{t+1} = 2\varepsilon_{t+1}\bar{y}_t$ and $\varepsilon_{t+1} \sim i.i.d.N(0,1)$, we obtain $E\hat{d}_{t+1}^2\hat{d}_t^2 = 16E(\varepsilon_t^2\bar{y}_t^2\bar{y}_{t-1}^2) \neq 16E(\bar{y}_t^2)E(\bar{y}_{t-1}^2) = E\hat{d}_{t+1}^2E\hat{d}_t^2$ and thus $\hat{d}_{t+1}$ exhibits conditional heteroskedasticity. In unreported results, akin to those for $h = 1$ in Table 5, we obtained rejection frequencies ranging from 16% to 19% for $P = 1000$ when we did not account for conditional heteroskedasticity. These are well above the corresponding values in the first panel of Table 5 where the actual size aligns perfectly with the nominal size of the test.

Under the alternative in which $\tilde{R} \neq R$, (Table 5, not bold), the actual power remains poor, with no obvious improvements relative to those observed in Table 1 when using the t-statistic. For the largest sample size, the actual power can be as large as 70% but, again, remains as low as 10% for some permutations of $\tilde{R} \neq R$. For a fixed value of R, actual power continues to be larger when $\tilde{R}$ is smaller, rather than larger, than R.

Table 6 reports rejection frequencies when the fixed DGP does not align with the rolling window model. As we saw before, power remains relatively weak with rejection frequencies typically below 25% for all but the largest sample size. For the largest sample size $P = 1000$, the actual power improves to as much as 70%. There is a tendency for actual power to increase as h increases holding $(\tilde{R}, P)$ constant. For all horizons, actual power again improves as $\tilde{R}$ declines. In addition, we continue to observe instances in which the rejection frequencies are u-shaped for a fixed value of $\tilde{R}$, as P increases.

In Table 7, actual power is driven by the fact that $h \neq \tilde{h}$ while in Table 8 it is also the case that the DGP is designed for the fixed scheme. One feature of Tables 7 and 8 that differs from Tables 3 and 4 are the missing values when $\tilde{h} = 24$ and $P = 25$. These are missing because, when the test function includes $\hat{d}_t$, $\tilde{h}$ observations are lost in construction of the χ^2-statistic. In the extreme case, there exist only one remaining observation and that is insufficient to construct the test statistic.

Subject to Monte Carlo variation and rounding, the bolded actual size results in Table 7 align with those in Table 5. In terms of actual power, both Tables 7 and 8 report many instances in which rejection frequencies are comparable to those under the null. Even so, we again find that when $R = 25$ and $\tilde{h} = 24$, there is substantially higher actual power that diminishes as h increases. One major difference from the results in Tables 3 and 4 arises when $\tilde{h}$ is larger than h. Here we find our first

evidence that using the χ^2-statistic (rather than the t-statistic) provides a substantial improvement in actual power. In the second and third panels of both tables we find actual power as high as 90% for all values of R and all but the smallest value of P. In contrast, when using the t-statistic the rejection frequencies are closer to 20%.

4 Conclusion

In this paper we have modest goals. The first is simply to provide examples of DGPs and forecasting environments in which the null of equal conditional predictive ability holds. We are able to do so but only in very limited and *unusual* environments. In one of the two examples the dependent variable forms an $ARMA(R+h-1, h-1)$ with a very unrealistic pattern of coefficients. Nevertheless, the example is sufficient to prove existence of the null hypothesis which was our goal.

Given these DGPs, we then provide simulation-based evidence on the finite sample size and power of the proposed statistics used to test the null. We find that the tests can be accurately sized given large enough samples, especially at shorter horizons. When lagged values of the loss differential are used as test functions we find that obtaining accurately sized tests also requires properly accounting for conditional heteroskedasticity when estimating the long-run variance. In all experiments power is low, and even poor, unless the sample sizes are quite large and even then rejection frequencies rarely rise above 70%.

References

Andrews, Donald W. K., and J. Christopher Monahan (1992), "An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator," *Econometrica* 60, 953-966.

Clark, Todd E., and Michael W. McCracken (2001), "Tests of Equal Forecast Accuracy and Encompassing for Nested Models," *Journal of Econometrics* 105, 85-110.

Clark, Todd E., and Michael W. McCracken (2013), "Advances in Forecast Evaluation," Elsevier Handbook of Economic Forecasting, Vol. 2B, 1107-1202.

Coroneo, Laura, and Fabrizio Iacone (2018), "Comparing Predictive Accuracy in Small Samples Using Fixed-Smoothing Asymptotics," manuscript.

Corradi, Valentina, Olivetti, Claudia, and Norman R. Swanson (2001), "Predictive Ability with Cointegrated Variables," *Journal of Econometrics* 104, 315-358.

Diebold, Francis X., and Roberto S. Mariano (1995), "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics* 13, 253-263.

Giacomini, Rafaella, and Barbara Rossi (2009), "Detecting and Predicting Forecast Breakdowns," *Review of Economic Studies* 76, 669-705.

Giacomini, Rafaella, and Barbara Rossi (2010), "Forecast Comparisons in Unstable Environments," *Journal of Applied Econometrics* 25, 595-620.

Giacomini, Rafaella, and Halbert White (2006), "Tests of Conditional Predictive Ability," *Econometrica* 74, 1545-1578.

Goncalves, Silvia, Perron, Benoit, and Michael W. McCracken (2017), "Tests of Equal Accuracy for Nested Models with Estimated Factors," *Journal of Econometrics* 198, 231-252.

Inoue, Atsushi, and Barbara Rossi (2012), "Out-of-Sample Forecast Tests Robust to the Choice of Window Size," *Journal of Business & Economic Statistics* 30, 432-453.

Li, Jia, and Andrew Patton (2018), "Asymptotic Inference about Predictive Accuracy Using High Frequency Data," *Journal of Econometrics* 203, 223-240.

McCracken, Michael W. (2007), "Asymptotics for Out-of-Sample Tests of Granger Causality," *Journal of Econometrics* 140, 719-752.

Newey, Whitney K., and Kenneth D. West (1987). "A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica* 55, 703-708.

Newey, Whitney K., and Kenneth D. West (1994), "Automatic Lag Selection in Covariance Matrix Estimation," *Review of Economic Studies* 61, 631-653.

Rossi, Barbara, and Tatevik Sekhposyan (2019). "Alternative tests for correct specification of conditional predictive densities," *Journal of Econometrics* 208, 638-657.

West, Kenneth D. (1996), "Asymptotic Inference About Predictive Ability," *Econometrica* 64, 1067-1084.

White, Halbert (2000), "A Reality Check for Data Snooping," *Econometrica* 68, 1097-1127.

White, Halbert and Ian Domowitz (1984), "Nonlinear Regression with Dependent Observations," *Econometrica* 52, 142-162.

Zhu, Yinchu, and Allan Timmermann (2020), "Can Two Tests Have the Same Conditional Expected Accuracy?" manuscript.

Tables

Table 1: Rolling DGP and Rolling Model, t-statistic

		$h=1$					$h=3$					$h=12$				
		P					P					P				
R	$\tilde{R}$	25	75	125	175	1000	25	75	125	175	1000	25	75	125	175	1000
25	**25**	**0.05**	**0.06**	**0.05**	**0.05**	**0.05**	**0.12**	**0.09**	**0.08**	**0.07**	**0.07**	**0.17**	**0.10**	**0.09**	**0.08**	**0.07**
	75	0.11	0.10	0.08	0.08	0.07	0.18	0.14	0.11	0.10	0.07	0.20	0.15	0.12	0.11	0.09
	125	0.13	0.14	0.13	0.12	0.10	0.22	0.20	0.16	0.15	0.10	0.23	0.20	0.17	0.16	0.12
	175	0.14	0.18	0.16	0.14	0.12	0.24	0.23	0.21	0.18	0.12	0.25	0.24	0.20	0.19	0.15
75	25	0.07	0.09	0.10	0.15	0.49	0.12	0.14	0.16	0.20	0.53	0.18	0.15	0.18	0.20	0.50
	75	**0.06**	**0.05**	**0.05**	**0.04**	**0.05**	**0.13**	**0.10**	**0.08**	**0.08**	**0.06**	**0.17**	**0.11**	**0.09**	**0.08**	**0.08**
	125	0.07	0.07	0.06	0.07	0.05	0.15	0.11	0.09	0.10	0.06	0.18	0.14	0.11	0.09	0.08
	175	0.08	0.08	0.09	0.08	0.07	0.17	0.14	0.13	0.12	0.08	0.20	0.16	0.15	0.13	0.08
125	25	0.07	0.10	0.13	0.17	0.68	0.13	0.15	0.20	0.24	0.71	0.19	0.18	0.20	0.23	0.66
	75	0.06	0.06	0.07	0.08	0.14	0.13	0.10	0.10	0.10	0.16	0.17	0.12	0.11	0.12	0.17
	125	**0.06**	**0.05**	**0.05**	**0.05**	**0.05**	**0.14**	**0.10**	**0.09**	**0.08**	**0.06**	**0.19**	**0.11**	**0.09**	**0.10**	**0.07**
	175	0.07	0.07	0.07	0.05	0.05	0.15	0.11	0.10	0.09	0.07	0.19	0.13	0.12	0.11	0.08
175	25	0.07	0.10	0.14	0.20	0.77	0.13	0.15	0.19	0.25	0.80	0.19	0.18	0.21	0.26	0.74
	75	0.06	0.06	0.07	0.08	0.19	0.14	0.10	0.10	0.11	0.23	0.18	0.12	0.12	0.13	0.25
	125	0.07	0.06	0.05	0.06	0.09	0.14	0.10	0.09	0.10	0.11	0.19	0.13	0.11	0.10	0.12
	175	**0.07**	**0.06**	**0.05**	**0.06**	**0.05**	**0.14**	**0.10**	**0.08**	**0.07**	**0.06**	**0.19**	**0.13**	**0.10**	**0.08**	**0.07**

Notes: R is the window size used to generate the time series, $\tilde{R}$ is the window size used to estimate the model parameters before generating a forecast, h is the forecast horizon, and P is the number of forecast periods. Each entry in the table represents the fraction of replications where the null hypothesis was rejected at the 5% level using the standard normal critical values of a two-sided test for a given R, $\tilde{R}$, h, and P. Bolded rows indicate when the null hypothesis holds (i.e., when $R = \tilde{R}$). The test statistic takes the form of the t-statistic described in the first paragraph of section 3.

Table 2: Fixed DGP and Rolling Model, t-statistic

		$h=1$					$h=3$					$h=12$				
		P					P					P				
R	$\tilde{R}$	25	75	125	175	1000	25	75	125	175	1000	25	75	125	175	1000
25	25	0.06	0.08	0.11	0.14	0.54	0.13	0.13	0.16	0.20	0.57	0.20	0.17	0.19	0.23	0.56
	75	0.07	0.08	0.11	0.14	0.42	0.15	0.13	0.17	0.19	0.44	0.21	0.18	0.18	0.22	0.46
	125	0.07	0.11	0.13	0.16	0.42	0.16	0.16	0.18	0.20	0.44	0.22	0.21	0.21	0.23	0.46
	175	0.07	0.11	0.15	0.18	0.44	0.17	0.18	0.19	0.22	0.47	0.22	0.21	0.24	0.25	0.48
75	25	0.07	0.12	0.15	0.19	0.63	0.13	0.18	0.20	0.25	0.67	0.20	0.21	0.22	0.27	0.66
	75	0.06	0.05	0.06	0.06	0.25	0.13	0.10	0.10	0.10	0.30	0.19	0.13	0.12	0.12	0.32
	125	0.05	0.06	0.06	0.07	0.22	0.13	0.10	0.10	0.10	0.25	0.19	0.14	0.12	0.12	0.27
	175	0.06	0.07	0.06	0.07	0.22	0.14	0.10	0.10	0.11	0.25	0.19	0.15	0.13	0.13	0.26
125	25	0.07	0.12	0.17	0.23	0.69	0.14	0.17	0.24	0.28	0.71	0.20	0.20	0.26	0.30	0.70
	75	0.06	0.09	0.09	0.10	0.29	0.13	0.11	0.13	0.14	0.32	0.18	0.15	0.16	0.16	0.35
	125	0.06	0.06	0.05	0.06	0.17	0.13	0.09	0.09	0.09	0.20	0.17	0.12	0.11	0.11	0.21
	175	0.06	0.05	0.06	0.05	0.14	0.13	0.09	0.09	0.09	0.17	0.18	0.13	0.11	0.11	0.20
175	25	0.07	0.12	0.18	0.25	0.72	0.14	0.18	0.25	0.32	0.75	0.20	0.20	0.25	0.33	0.74
	75	0.06	0.07	0.09	0.11	0.30	0.13	0.11	0.13	0.14	0.35	0.18	0.14	0.15	0.18	0.36
	125	0.06	0.07	0.08	0.08	0.20	0.13	0.11	0.10	0.11	0.24	0.18	0.13	0.12	0.14	0.25
	175	0.06	0.05	0.05	0.05	0.13	0.13	0.08	0.09	0.08	0.15	0.18	0.12	0.12	0.11	0.18

Notes: See notes to Table 1. No values are bolded, because the null hypothesis does not hold for any entry.

Table 3: Rolling DGP and Rolling Model, t-statistic (alternating $\tilde{h}$)

		$h=1$					$h=3$					$h=12$				
		P					P					P				
R	$\tilde{h}$	25	75	125	175	1000	25	75	125	175	1000	25	75	125	175	1000
25	1	**0.06**	**0.05**	**0.04**	**0.05**	**0.05**	0.16	0.13	0.15	0.16	0.48	0.20	0.16	0.18	0.19	0.55
	3	0.07	0.06	0.05	0.06	0.05	**0.11**	**0.09**	**0.09**	**0.08**	**0.07**	0.13	0.09	0.07	0.08	0.08
	12	0.09	0.10	0.11	0.12	0.31	0.15	0.12	0.11	0.12	0.22	**0.17**	**0.10**	**0.08**	**0.08**	**0.07**
	24	0.34	0.72	0.91	0.97	1.00	0.24	0.40	0.54	0.67	1.00	0.19	0.15	0.15	0.17	0.43
75	1	**0.06**	**0.06**	**0.05**	**0.05**	**0.04**	0.19	0.16	0.13	0.14	0.23	0.23	0.19	0.17	0.16	0.29
	3	0.08	0.05	0.05	0.05	0.06	**0.14**	**0.09**	**0.07**	**0.08**	**0.07**	0.16	0.09	0.08	0.08	0.06
	12	0.09	0.06	0.06	0.06	0.07	0.15	0.10	0.08	0.07	0.08	**0.17**	**0.11**	**0.09**	**0.08**	**0.07**
	24	0.09	0.07	0.07	0.07	0.10	0.14	0.11	0.09	0.09	0.12	0.19	0.13	0.10	0.09	0.10
125	1	**0.07**	**0.06**	**0.05**	**0.05**	**0.05**	0.20	0.16	0.15	0.13	0.16	0.25	0.20	0.18	0.18	0.23
	3	0.08	0.06	0.05	0.06	0.05	**0.13**	**0.09**	**0.08**	**0.08**	**0.06**	0.17	0.10	0.09	0.09	0.06
	12	0.08	0.06	0.06	0.06	0.05	0.14	0.10	0.08	0.07	0.06	**0.18**	**0.11**	**0.10**	**0.09**	**0.07**
	24	0.09	0.07	0.06	0.06	0.06	0.15	0.11	0.09	0.09	0.08	0.19	0.13	0.11	0.10	0.08
175	1	**0.06**	**0.06**	**0.05**	**0.05**	**0.05**	0.19	0.18	0.16	0.15	0.15	0.25	0.22	0.22	0.19	0.19
	3	0.08	0.06	0.06	0.05	0.05	**0.15**	**0.10**	**0.08**	**0.08**	**0.07**	0.18	0.11	0.09	0.08	0.07
	12	0.09	0.06	0.05	0.06	0.05	0.14	0.10	0.08	0.08	0.07	**0.18**	**0.12**	**0.10**	**0.09**	**0.07**
	24	0.09	0.07	0.06	0.06	0.06	0.14	0.11	0.09	0.08	0.07	0.19	0.13	0.10	0.09	0.09

Notes: R is the window size, h is the forecast horizon used to generate the time series, $\tilde{h}$ is the forecast horizon used in the forecasts, and P is the number of forecast periods. Each entry in the table represents the fraction of replications where the null hypothesis was rejected at the 5% level using the standard normal critical values of a two-sided test for a given R, h, $\tilde{h}$ and P. Bolded values indicate when the null hypothesis holds (i.e., when $h = \tilde{h}$). The test statistic takes the form of the t-statistic described in the first paragraph of section 3.

Table 4: Fixed DGP and Rolling Model, t-statistic (alternating $\tilde{h}$)

		$h=1$					$h=3$					$h=12$				
		P					P					P				
R	$\tilde{h}$	25	75	125	175	1000	25	75	125	175	1000	25	75	125	175	1000
25	1	0.06	0.08	0.11	0.14	0.55	0.17	0.16	0.20	0.23	0.48	0.24	0.21	0.27	0.29	0.50
	3	0.07	0.09	0.12	0.16	0.57	0.13	0.12	0.16	0.20	0.57	0.16	0.14	0.15	0.19	0.48
	12	0.13	0.15	0.20	0.24	0.68	0.18	0.18	0.21	0.25	0.66	0.19	0.17	0.19	0.23	0.58
	24	0.75	0.90	0.96	0.98	1.00	0.46	0.57	0.69	0.77	0.94	0.23	0.23	0.27	0.33	0.70
75	1	0.05	0.06	0.06	0.07	0.25	0.20	0.17	0.15	0.15	0.30	0.26	0.22	0.20	0.20	0.35
	3	0.07	0.06	0.06	0.08	0.27	0.12	0.10	0.09	0.10	0.30	0.16	0.11	0.10	0.09	0.26
	12	0.08	0.07	0.07	0.08	0.29	0.14	0.10	0.09	0.10	0.31	0.17	0.13	0.12	0.12	0.32
	24	0.08	0.09	0.10	0.11	0.35	0.14	0.12	0.12	0.13	0.36	0.18	0.15	0.13	0.14	0.36
125	1	0.06	0.06	0.06	0.06	0.17	0.20	0.16	0.17	0.16	0.24	0.25	0.23	0.22	0.20	0.29
	3	0.07	0.06	0.07	0.07	0.16	0.14	0.09	0.08	0.09	0.20	0.16	0.11	0.10	0.10	0.18
	12	0.08	0.06	0.06	0.07	0.19	0.13	0.10	0.09	0.09	0.20	0.18	0.12	0.11	0.11	0.21
	24	0.08	0.07	0.08	0.08	0.19	0.14	0.10	0.10	0.10	0.21	0.19	0.14	0.12	0.12	0.24
175	1	0.05	0.05	0.05	0.06	0.21	0.18	0.17	0.16	0.21	0.08	0.26	0.22	0.22	0.22	0.25
	3	0.07	0.06	0.06	0.06	0.12	0.13	0.09	0.08	0.08	0.15	0.16	0.11	0.10	0.09	0.14
	12	0.08	0.06	0.06	0.06	0.12	0.13	0.10	0.08	0.08	0.16	0.17	0.12	0.11	0.11	0.18
	24	0.08	0.06	0.05	0.07	0.14	0.14	0.10	0.09	0.09	0.18	0.19	0.13	0.12	0.11	0.18

Notes: See notes to Table 3. No values are bolded, because the null hypothesis does not hold for any entry.

Table 5: Rolling DGP and Rolling Model, Wald-statistic

		$h=1$					$h=3$					$h=12$				
		P					P					P				
R	$\tilde{R}$	25	75	125	175	1000	25	75	125	175	1000	25	75	125	175	1000
25	**25**	**0.11**	**0.08**	**0.08**	**0.07**	**0.05**	**0.24**	**0.15**	**0.15**	**0.13**	**0.09**	**0.37**	**0.16**	**0.12**	**0.11**	**0.09**
	75	0.15	0.11	0.09	0.09	0.08	0.30	0.20	0.17	0.15	0.11	0.41	0.21	0.17	0.15	0.11
	125	0.17	0.14	0.13	0.12	0.12	0.32	0.24	0.21	0.18	0.16	0.43	0.25	0.20	0.18	0.18
	175	0.17	0.17	0.16	0.15	0.14	0.33	0.26	0.24	0.22	0.19	0.42	0.27	0.25	0.22	0.21
75	25	0.12	0.10	0.11	0.13	0.40	0.23	0.20	0.20	0.21	0.46	0.38	0.17	0.17	0.19	0.41
	75	**0.11**	**0.08**	**0.07**	**0.06**	**0.05**	**0.25**	**0.16**	**0.13**	**0.13**	**0.08**	**0.40**	**0.16**	**0.13**	**0.12**	**0.08**
	125	0.11	0.09	0.08	0.07	0.06	0.26	0.17	0.15	0.13	0.09	0.41	0.18	0.15	0.13	0.09
	175	0.12	0.11	0.09	0.09	0.06	0.28	0.19	0.17	0.15	0.09	0.41	0.20	0.17	0.15	0.10
125	25	0.12	0.10	0.12	0.14	0.58	0.24	0.19	0.21	0.23	0.63	0.38	0.19	0.18	0.21	0.57
	75	0.11	0.08	0.08	0.08	0.11	0.26	0.17	0.15	0.14	0.16	0.40	0.18	0.15	0.14	0.16
	125	**0.11**	**0.07**	**0.07**	**0.06**	**0.05**	**0.26**	**0.16**	**0.14**	**0.12**	**0.08**	**0.40**	**0.17**	**0.14**	**0.12**	**0.09**
	175	0.12	0.08	0.08	0.07	0.05	0.27	0.16	0.14	0.13	0.09	0.40	0.19	0.15	0.13	0.09
175	25	0.12	0.11	0.13	0.16	0.67	0.23	0.20	0.21	0.23	0.72	0.38	0.19	0.20	0.22	0.66
	75	0.11	0.08	0.08	0.08	0.14	0.26	0.15	0.15	0.14	0.20	0.40	0.18	0.15	0.14	0.21
	125	0.11	0.08	0.08	0.07	0.08	0.25	0.17	0.14	0.14	0.12	0.41	0.18	0.14	0.13	0.12
	175	**0.11**	**0.08**	**0.07**	**0.07**	**0.05**	**0.25**	**0.17**	**0.13**	**0.12**	**0.09**	**0.41**	**0.18**	**0.14**	**0.13**	**0.08**

Notes: See notes to Table 1. The test statistic takes the form of the Wald-statistic described in the first paragraph of section 3. Consequently, $\chi^2(2)$ critical values are used for inference (as opposed to the standard normal critical values used in Table 1). Note that h observations were lost because the test function includes an h-lagged loss differential.

Table 6: Fixed DGP and Rolling Model, Wald-statistic

		$h=1$					$h=3$					$h=12$				
		P					P					P				
R	$\tilde{R}$	25	75	125	175	1000	25	75	125	175	1000	25	75	125	175	1000
25	25	0.12	0.10	0.11	0.13	0.48	0.25	0.18	0.19	0.22	0.55	0.40	0.21	0.20	0.23	0.53
	75	0.12	0.10	0.11	0.13	0.36	0.28	0.21	0.20	0.20	0.40	0.42	0.23	0.22	0.21	0.43
	125	0.13	0.11	0.13	0.15	0.38	0.27	0.20	0.20	0.22	0.43	0.41	0.24	0.24	0.23	0.43
	175	0.12	0.12	0.14	0.16	0.39	0.28	0.21	0.22	0.22	0.43	0.44	0.25	0.25	0.25	0.45
75	25	0.11	0.13	0.15	0.16	0.57	0.23	0.21	0.23	0.24	0.62	0.38	0.22	0.21	0.22	0.59
	75	0.11	0.08	0.08	0.07	0.20	0.25	0.16	0.15	0.15	0.26	0.40	0.18	0.15	0.15	0.27
	125	0.12	0.09	0.08	0.07	0.18	0.27	0.15	0.15	0.13	0.22	0.40	0.18	0.16	0.15	0.24
	175	0.11	0.08	0.08	0.08	0.18	0.26	0.16	0.15	0.14	0.23	0.41	0.18	0.17	0.16	0.25
125	25	0.12	0.12	0.15	0.19	0.59	0.23	0.20	0.25	0.27	0.67	0.37	0.20	0.23	0.25	0.64
	75	0.11	0.09	0.11	0.10	0.24	0.24	0.17	0.18	0.17	0.29	0.39	0.18	0.17	0.15	0.30
	125	0.11	0.07	0.07	0.07	0.13	0.25	0.15	0.15	0.13	0.19	0.39	0.17	0.14	0.14	0.19
	175	0.11	0.07	0.07	0.07	0.11	0.26	0.15	0.14	0.12	0.15	0.39	0.18	0.15	0.14	0.19
175	25	0.11	0.12	0.15	0.20	0.62	0.23	0.20	0.26	0.29	0.69	0.38	0.20	0.22	0.27	0.66
	75	0.12	0.08	0.09	0.09	0.23	0.24	0.17	0.16	0.17	0.31	0.39	0.18	0.16	0.17	0.29
	125	0.12	0.09	0.08	0.09	0.17	0.25	0.16	0.16	0.16	0.21	0.40	0.17	0.17	0.15	0.21
	175	0.11	0.07	0.07	0.06	0.11	0.25	0.15	0.14	0.13	0.14	0.40	0.19	0.14	0.13	0.16

Notes: See notes to Table 5. No values are bolded, because the null hypothesis does not hold for any entry.

Table 7: Rolling DGP and Rolling Model, Wald-statistic (alternating $\tilde{h}$)

		$h=1$					$h=3$					$h=12$				
		P					P					P				
R	$\tilde{h}$	25	75	125	175	1000	25	75	125	175	1000	25	75	125	175	1000
25	1	**0.11**	**0.08**	**0.07**	**0.07**	**0.06**	0.48	0.78	0.91	0.96	1.00	0.54	0.83	0.94	0.98	1.00
	3	0.15	0.11	0.10	0.08	0.07	**0.25**	**0.16**	**0.14**	**0.13**	**0.09**	0.22	0.15	0.15	0.13	0.29
	12	0.24	0.10	0.11	0.10	0.26	0.34	0.15	0.11	0.12	0.18	**0.37**	**0.15**	**0.12**	**0.11**	**0.08**
	24		0.50	0.78	0.91	1.00		0.31	0.39	0.53	1.00		0.18	0.15	0.16	0.34
75	1	**0.11**	**0.08**	**0.07**	**0.06**	**0.05**	0.54	0.87	0.96	0.99	1.00	0.60	0.89	0.97	0.99	1.00
	3	0.16	0.10	0.10	0.08	0.06	**0.26**	**0.17**	**0.14**	**0.12**	**0.09**	0.24	0.17	0.18	0.18	0.41
	12	0.25	0.09	0.08	0.08	0.07	0.36	0.16	0.12	0.11	0.08	**0.41**	**0.17**	**0.14**	**0.12**	**0.09**
	24		0.11	0.08	0.08	0.09		0.18	0.13	0.11	0.10		0.20	0.15	0.12	0.10
125	1	**0.11**	**0.08**	**0.07**	**0.07**	**0.06**	0.56	0.89	0.97	0.99	1.00	0.61	0.91	0.98	0.99	1.00
	3	0.17	0.09	0.09	0.08	0.06	**0.26**	**0.16**	**0.13**	**0.13**	**0.09**	0.26	0.18	0.20	0.18	0.45
	12	0.25	0.10	0.09	0.08	0.06	0.36	0.15	0.12	0.11	0.09	**0.39**	**0.17**	**0.14**	**0.13**	**0.09**
	24		0.11	0.09	0.08	0.06		0.16	0.13	0.12	0.09		0.20	0.15	0.12	0.09
175	1	**0.12**	**0.08**	**0.07**	**0.06**	**0.05**	0.58	0.90	0.97	0.99	1.00	0.62	0.92	0.98	0.99	1.00
	3	0.15	0.10	0.09	0.07	0.07	**0.26**	**0.16**	**0.13**	**0.12**	**0.08**	0.25	0.19	0.19	0.21	0.47
	12	0.24	0.10	0.09	0.07	0.06	0.37	0.14	0.13	0.11	0.08	**0.40**	**0.17**	**0.14**	**0.13**	**0.08**
	24		0.10	0.10	0.08	0.05		0.17	0.13	0.10	0.07		0.20	0.15	0.13	0.09

Notes: See notes to Table 3. The test statistic takes the form of the Wald-statistic described in the first paragraph of section 3. Consequently, $\chi^2(2)$ critical values are used for inference (as opposed to the standard normal critical values used in Table 1). Note that $\tilde{h}$ observations were lost because the test function includes an $\tilde{h}$-lagged loss differential. The loss of these observations implies there are insufficient observations to construct the Wald-statistic when $P = 25$ and $\tilde{h} = 24$.

Table 8: Fixed DGP and Rolling Model, Wald-statistic (alternating $\tilde{h}$)

		$h=1$					$h=3$					$h=12$				
		P					P					P				
R	$\tilde{h}$	25	75	125	175	1000	25	75	125	175	1000	25	75	125	175	1000
25	1	0.12	0.09	0.11	0.14	0.48	0.51	0.82	0.94	0.98	1.00	0.56	0.87	0.96	0.99	1.00
	3	0.17	0.13	0.14	0.16	0.54	0.24	0.18	0.20	0.21	0.54	0.25	0.20	0.23	0.27	0.70
	12	0.25	0.13	0.17	0.19	0.61	0.36	0.18	0.19	0.23	0.59	0.39	0.20	0.19	0.22	0.53
	24		0.51	0.78	0.90	1.00		0.34	0.44	0.57	0.92		0.24	0.22	0.26	0.67
75	1	0.11	0.09	0.08	0.07	0.21	0.56	0.87	0.96	0.98	1.00	0.60	0.90	0.97	0.99	1.00
	3	0.16	0.10	0.09	0.09	0.22	0.24	0.16	0.15	0.14	0.26	0.25	0.19	0.20	0.22	0.61
	12	0.25	0.10	0.10	0.09	0.23	0.37	0.15	0.13	0.12	0.26	0.40	0.19	0.15	0.14	0.27
	24		0.12	0.10	0.10	0.27		0.18	0.14	0.13	0.30		0.20	0.16	0.16	0.30
125	1	0.12	0.08	0.07	0.07	0.13	0.55	0.90	0.97	0.99	1.00	0.64	0.91	0.97	0.99	1.00
	3	0.15	0.10	0.09	0.09	0.14	0.26	0.15	0.14	0.13	0.18	0.26	0.20	0.22	0.22	0.57
	12	0.25	0.10	0.09	0.08	0.15	0.36	0.15	0.12	0.12	0.17	0.41	0.17	0.14	0.13	0.20
	24		0.11	0.09	0.09	0.15		0.18	0.13	0.12	0.18		0.21	0.16	0.13	0.19
175	1	0.11	0.08	0.07	0.07	0.10	0.59	0.90	0.97	0.99	1.00	0.62	0.91	0.98	0.99	1.00
	3	0.16	0.11	0.08	0.09	0.11	0.25	0.15	0.14	0.12	0.14	0.26	0.20	0.21	0.21	0.58
	12	0.25	0.10	0.08	0.08	0.10	0.38	0.15	0.12	0.11	0.14	0.42	0.18	0.14	0.13	0.16
	24		0.11	0.08	0.08	0.12		0.16	0.13	0.11	0.14		0.20	0.14	0.14	0.17

Notes: See notes to Table 7. No values are bolded, because the null hypothesis does not hold for any entry.