IHS Economics Series
Working Paper 341
July 2018

On Using Predictive-ability Tests in the Selection of Time-series Prediction Models: A Monte Carlo Evaluation

Mauro Costantini
Robert M. Kunst

Institut für Höhere Studien
Institute for Advanced Studies
Vienna

Impressum

Author(s):
Mauro Costantini, Robert M. Kunst

Title:
On Using Predictive-ability Tests in the Selection of Time-series Prediction Models: A Monte Carlo Evaluation

ISSN: 1605-7996

On using predictive-ability tests in the selection of time-series prediction models: A Monte Carlo evaluation

Mauro Costantini*
Department of Economics and Finance, Brunel University, London
and
Robert M. Kunst
Institute for Advanced Studies, Vienna, and University of Vienna

August 7, 2018

Abstract

Comparative ex-ante prediction experiments over expanding subsamples are a popular tool for the task of selecting the best forecasting model class in finite samples of practical relevance. Flanking such a horse race by predictive-accuracy tests, such as the test by Diebold and Mariano (1995), tends to increase support for the simpler structure. We are concerned with the question whether such simplicity boosting actually benefits predictive accuracy in finite samples. We consider two variants of the DM test, one with naive normal critical values and one with bootstrapped critical values, the predictive-ability test by Giacomini and White (2006), which continues to be valid in nested problems, the F test by Clark and McCracken (2001), and also model selection via the AIC as a benchmark strategy. Our Monte Carlo simulations focus on basic univariate time-series specifications, such as linear (ARMA) and nonlinear (SETAR) generating processes.

Keywords: forecasting, time series, predictive accuracy, model selection
JEL Code: C22, C52, C53

*The authors gratefully acknowledge helpful comments by Leopold Soegner.

1 Introduction

If two model-based forecasts for the same time-series variable are available over some time range where they can also be compared to actual realizations, it appears natural to use the forecast with the better track record in order to predict the physical future. It has become customary, however, to subject the outcome of horse races over training samples to various significance tests, following the seminal contributions by Diebold and Mariano (1995) and West (1996) or one of the numerous later developed procedures, for example Clark and McCracken (2001, 2005, 2012) and Giacomini and White (2006, GW).

Here, we are interested in the consequences of basing the preference for a forecasting model on the result of such a significance test, using the simpler model unless it is rejected at a 5% level. We are concerned by the possibility that such a strategy becomes too conservative, with an undue support for the simpler rival. Our argument is well grounded in the literature on statistical model selection (Wei, 1992; Inoue and Kilian, 2006; Ing, 2007), which has shown that the model choice determined by minimizing prediction errors over a test sample is, under conditions, asymptotically equivalent to traditional information criteria, such as AIC and BIC. The asymptotic implications of selecting models by information criteria on forecasting performance are a well-explored topic. Roughly, selecting models based on AIC optimizes prediction (Shibata, 1980), whereas BIC chooses the correct model, assuming it is in the choice set, at the cost of slightly larger prediction errors. This fact implies that subjecting the selection to any further criterion on top of the track record may involve the risk of becoming more 'conservative' than appears to be optimal.

Our contribution is a systematic Monte Carlo evaluation of these effects. In some designs, one of the rival forecasting models belongs to the same class as the generator, whereas in others, the generator is more complex than any of the rival models. We note that, while we build on the related literature, we also digress from it in several important aspects. First, the two recent decades have seen a strong emphasis on questions such as the asymptotic or finite-sample distributions of forecast accuracy test statistics and the power properties of the thus defined tests. By contrast, we see these aspects as a means to the end of selecting the model that optimizes forecast accuracy. Unlike statistical hypothesis testing proper, forecast model selection cannot choose a size level freely but has to determine the

size of a test tool in such a way that it benefits the decision based on the test.

Second, most of the literature targets the decision based on true or pseudo-true parameter values, in the sense that a nesting model will forecast at least as precisely as a nested model. By contrast, we are interested in forecast model selection at empirically relevant sample sizes. A simple structure can outperform a complex class that may even contain the generating structure if all parameters are unknown and are estimated. Similarly, GW argued that the null hypothesis of the DM and Clark-McCracken tests may not support the forecaster's aim. Consider the task of forecasting an economic variable, say aggregate investment, by its past and potentially by another variable, say an interest rate. In the world of the DM and Clark-McCracken tests, the complex alternative is conceptually preferred whenever the coefficients of the interest rate in a joint model are non-zero. Furthermore, a univariate model for investment can never forecast better than the joint model, even if the coefficients on interest are very small and their estimates have large sampling variation in finite samples. The concept of GW accounts for this sampling variation, so if the coefficients are non-zero but small, the forecaster is better off by ignoring the interest rate. Notwithstanding the important distinction of forecastability and predictability by Hendry (1997), the GW approach was revolutionary, and we feel that its impact has not fully been considered yet. For example, hitherto empirical investigations on Granger causality do not really build on predictability of a variable Y using another variable X but on conditional distributions, often regression t and F tests. In order to represent predictability in a finite sample, the possibility has to be taken into account that the forecast for Y may deteriorate if lags of X are used as additional predictors, with empirical non-causality representing a borderline between causality and anti-causality. By construction, this approach typically yields an even more conservative selection procedure than the DM test, thus aggravating our original concerns.

Third, most of the literature uses simulation designs that build on Granger-causal and bivariate ideas, with a target variable dynamically dependent on an explanatory source. Such designs may correspond to typical macroeconomic applications, and we also take them up in one design. Primarily, however, we start from a rigorous time-series design, with an emphasis on the most natural and elementary univariate models, such as AR(1), MA(1),

ARMA(1,1). We see this as the adequate starting point for all further analysis.

Within this paper, we restrict attention to binary comparisons between a comparatively simple time-series model and a more sophisticated rival. Main features should also be valid for the general case of comparing a larger set of rival models, with one of them chosen as the benchmark. Following some discussion on the background of the problem, we present results of several simulation experiments in order to explore the effects for sample sizes that are typical in econometrics.

The remainder of this paper is organized as follows. Section 2 reviews some of the fundamental theoretical properties of the problem of testing for relative predictive accuracy following a training-set comparison. Section 3 reports a basic Monte Carlo experiment with a purely univariate nested linear time-series design. To the best of our knowledge and somewhat surprisingly, our study is the first one that examines these competing prediction strategies systematically in a purely univariate ARMA(1,1) design, which we see as the natural starting point. Section 4 uses three more Monte Carlo designs: one with a non-nested linear design, one with a SETAR design that was suggested in the literature (Tiao and Tsay, 1994) to describe the dynamic behavior of a U.S. output series, and one with a design based on a three-variable vector autoregression that was fitted to macroeconomic U.K. data by Costantini and Kunst (2011). Section 5 concludes.

2 The theoretical background

Typically, the Diebold-Mariano (DM) test and comparable tests are performed on accuracy measures such as MSE (mean squared errors) following an out-of-sample forecasting experiment, in which a portion of size S from a sample of size N is predicted on the basis of expanding windows. In a notation close to DM, the null hypothesis of such tests is

$$\mathrm{E}g(e_1) = \mathrm{E}g(e_2),$$

where $e_j, j = 1, 2$ denote the prediction errors for the two rival forecasts, $g(.)$ is some function—for example, $g(x) = x^2$ for the MSE—and E denotes the expectation operator. In other words, both models use the true or pseudo-true (probability limit of estimates)

parameter θ. Alternatively, GW consider the null hypothesis

$$\mathrm{E}\{g(e_1)|\mathfrak{F}\} = \mathrm{E}\{g(e_2)|\mathfrak{F}\},$$

where $\mathfrak{F}$ denotes some information set, for example the history of the time series. In other words, both models use sample parameter estimates $\hat{\theta}_1, \hat{\theta}_2$. Thus, whereas DM consider a true-model setup, with the null rejected even in the presence of an arbitrarily small advantage for the alternative model, GW focus on the forecaster's situation who has to estimate all model parameters and has to take sampling variation into account.

A model-selection decision based on an out-of-sample prediction experiment (TS in the following for training-sample evaluation) without any further check on the significance of accuracy gains works like a decision based on an information criterion. The asymptotic properties of this TS criterion depend on regularity assumptions on the data-generating process, as usual, but critically on the large-sample assumptions on S/N.

If S/N converges to a constant in the open interval $(0, 1)$, Inoue and Kilian (2006) show that the implied TS criterion is comparable to traditional 'efficient' criteria such as AIC. The wording 'efficient' is due to Shibata (1980) and McQuarrie and Tsai (1998) and relates to the property of optimizing predictive performance at the cost of a slight large-sample inconsistency in the sense that profligate (though valid) models are selected too often as $N \to \infty$.

If $S/N \to 1$, Wei (1992) shows the consistency of the implied TS criterion in the sense that it selects the true model, assuming such a one exists, with probability one as $N \to \infty$. Wei (1992) essentially assumes that *all* available observations are predicted, excluding the sample start, where the estimation of a time-series model is not yet possible.

If a consistent model-selection procedure is flanked by a further hypothesis test that has the traditional test-consistency property, in the sense that it achieves its nominal significance level on its null and rejection with probability one on its alternative as $N \to \infty$, this clearly does not affect the asymptotic property of selection consistency if the criterion and the flanking test are independent or tend to decide similarly. Only if the two decisions counteract each other, one may construct cases where the application of the flanking test destroys selection consistency. In summary, the procedure that is of interest here, a model decision based on TS and an additional test jointly, is consistent in those cases where

TS alone is consistent, so nothing is gained in large samples. For this reason, it is the empirically relevant sample sizes that are of interest, and these are in the focus of our Monte Carlo.

Like other information criteria, TS entails an implicit significance level at which a traditional restriction test performs the same model selection as the criterion. For all consistent information criteria, this implicit significance level depends on N and approaches 0 as $N \to \infty$. On the other hand, efficient criteria approach a non-zero implicit significance level. For example, the asymptotic implicit significance level for AIC is surprisingly liberal at almost 16%. This value can be determined analytically as $2(1 - \Phi(\sqrt{2})$ with Φ the normal c.d.f., following the argument of Pötscher (1991).

Thus, the suggestion to base a decision on choosing a prediction model on a sequence of a TS comparison and a predictive-ability test makes little sense in large samples. In small samples, it acts as a *simplicity booster* that puts more emphasis on the simpler model than the simple TS evaluation. Our simulations are meant to shed some light on the benefits or drawbacks of such boosting of simplicity in typical situations.

3 Simulations with a nested ARMA(1,1) design

This section presents the results for our basic ARMA design. We first describe its background. The optimal decision between AR(1) and ARMA(1,1)—with 'optimal' always referring to the best out-of-sample prediction performance—for a given sample size can be determined exactly by simulation. Monte Carlo can deliver the boundary curve in the (ϕ, θ) space for generated ARMA trajectories with coefficients ϕ and θ, along which AR(1) and ARMA(1,1) forecasts with estimated coefficients yield the same forecast accuracy. Then, we compare the prediction strategies pairwise and close with a short summary of our general impression.

3.1 The background

The simplest and maybe most intuitive design for investigating model selection procedures in time-series analysis is the ARMA(1,1) model. In the parameterization

$$X_t = \phi X_{t-1} + \varepsilon_t - \theta \varepsilon_{t-1},$$

ARMA(1,1) models are known to be stable and uniquely defined on

$$\Omega_{ARMA} = \{(\phi, \theta) \in (-1, 1) \times (-1, 1) : \phi \neq \theta \vee (\phi, \theta) = (0, 0)\},$$

a sliced open square in the $\mathbb{R}^2$ plane. Assumptions on the process (ε_t) vary somewhat in the literature, in line with specific needs of applications or of theorems. Usually, *iid* ε_t and a standard symmetric distribution with finite second moments are assumed (e.g., Lütkepohl, 2005), although some authors relax assumptions considerably. We do not study generalizations in these directions here, and we use Gaussian *iid* ε_t throughout.

Among the simplest time-series models, ARMA(1,1) competes with the model classes AR(1) and MA(1) as prediction tools, both with only one coefficient parameter.

In more detail, the open square $(-1, 1) \times (-1, 1)$ consists of the following regions that play a role in our simulation experiments:

1. The punctured diagonal is not part of Ω_{ARMA}. Along this diagonal, processes are equivalent to white noise $(0, 0)$. We simulate along the diagonal in order to see whether reaction remains unaffected;

2. The origin $(\phi, \theta) = (0, 0)$ represents white noise. ARMA(1,1) has two redundant parameters, while AR(1) or MA(1) have one each. These simpler models are expected to perform better;

3. The punctured x–axis $\theta = 0, \phi \neq 0$ contains pure AR(1) models. ARMA(1,1) is over-parameterized and is expected to perform worse than AR(1);

4. The punctured y–axis $\phi = 0, \theta \neq 0$ contains pure MA(1) models. ARMA(1,1) is over-parameterized and is expected to perform worse than MA(1). AR(1) is misspecified here, so for large samples ARMA(1,1) should outperform AR(1) here;

5. On the remainder $\{(\phi, \theta) : \phi \neq 0, \theta \neq 0, \phi \neq \theta\}$, the ARMA(1,1) model is correctly specified, whereas the restricted AR(1) and MA(1) are incorrect. As $N \to \infty$, ARMA(1,1) should dominate its incorrect rivals. The ranking is uncertain for small N and in areas close to the other four regions.

With some simplification, we consider for later reference

$$\Theta_R = \{(\phi, \theta) | \theta = 0 \,\text{or}\, \theta = \phi\},$$

as the area of the open square where AR(1) is expected to outperform ARMA(1,1) in large samples, consisting of the diagonal # 1, white noise # 2, and the AR(1) models # 3. On Θ_R, either the AR(1) is correct or a simpler white-noise structure. On the remaining part of the open square # 4 and # 5, the AR(1) model is mis-specified, and the ARMA(1,1) model should dominate in very large samples.

Obviously, if the coefficient values $(\phi, \theta) \in \Theta \backslash \Theta_R$ were known, it would be optimal to use this ARMA(1,1) model for prediction. The situation is less obvious if the values of the coefficients are not known. Presumably, AR(1) models will still be preferable if θ is close but not identical to 0, such that the true model is ARMA(1,1), with insufficient information on θ in the sample that might permit reliable estimation. The same should be true for MA(1) models as prediction models and a small value for ϕ. We note that this distinction corresponds to the respective hypotheses considered by DM and by GW.

The so-called Greenland graphs, such as those in Figure 1, permit to make this statement more precise. On a sizable portion of the admissible parameter space, AR(1) defeats ARMA(1,1) as a forecasting model, even though θ is not exactly zero, i.e. parameter values are outside Θ_R. This area—we call it Greenland according to an original color version of the graph—shrinks as the sample size grows. The sizable area of AR dominance is not paralleled by a similar area of MA dominance. The MA(1) model is a comparatively poor forecast tool, and we will exclude it from further simulation experiments. Although some interesting facts about these preference areas can be determined by analytical means (see, e.g., the rule of thumb in Hendry, 1997), for example the poor performance of the MA forecasts is an issue of the algorithm and can be explored by simulation only.

The graph relies on 1000 replications with Gaussian errors. All trajectories were generated with burn-ins of 100 observations. From samples of size $N = 50$ and $N = 100$,

forecasts are generated from AR(1), MA(1), and ARMA(1,1) models with estimated parameters, and the squared prediction error for X_{51} and, respectively, X_{101} is evaluated. We generated similar graphs for smaller and slightly larger sample sizes. We certainly do not claim that we are the first to run such simulations, but the graphs serve as a valuable reference for the remainder of the paper and there does not seem to exist an easily accessible source for comparable graphs in the literature. It is obvious that some smoothing or higher numbers of replications will produce clear shapes of the preference areas. We feel, however, that its ragged appearance conveys a good impression of areas where preference for any of the three models is not very pronounced.

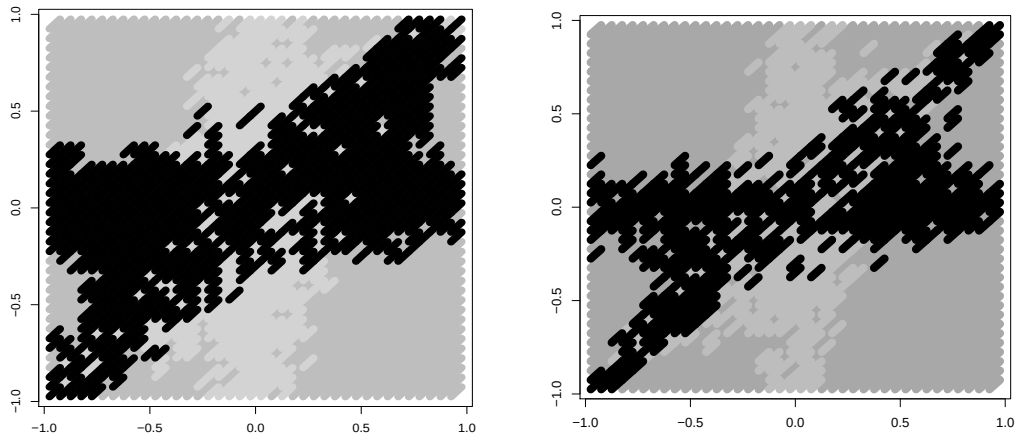

Figure 1: Forecasting data generated by ARMA models with autoregressive coefficient ϕ on the x–axis and moving-average coefficient θ on the y–axis. $N = 50$ (left) and $N = 100$ (right). Comparison of MSE according to AR, ARMA, and MA models with estimated coefficients. Black area has lowest MSE for AR models, gray area for ARMA, and light gray area for MA models.

For empirical work, the applicability of Figure 1 is limited, as it shows the optimal selection of prediction models for given and true parameter values. For a hypothetical researcher who observes ARMA(1,1) data, this decision is not available. If the true values were known, it would be optimal to use them in forecasting. On the other hand, it is not generally possible to draw a comparable figure that has on its axes values of estimates $(\hat{\phi}, \hat{\theta})$ that are available to the hypothetical observer. This would require convening a prior

distribution on the parameter space, thus adopting a Bayesian framework.

For our simulation study, however, these graphs are important benchmarks, as the almost ideal selection procedure between AR(1) and ARMA(1,1) would select AR(1) on the dark area and ARMA(1,1) on the remainder. An ideal procedure may be able to beat this benchmark by varying the classification of specific trajectories over the two regions, but it is unlikely that such improvements are practically relevant. We note explicitly that preferring AR(1) on Θ_R only does not lead to optimal forecasting decisions, in contrast to the underlying statistical hypothesis testing problem.

In all our experiments the outcome may depend critically on the estimation procedure used for AR as well as for ARMA models. Generally, we use the estimation routines implemented in R. We feel that the possible dependence of our results on the specific estimation routine need not be seen as a drawback, as we are interested in the typical situation of a forecaster who considers given samples and popular estimation options. In other words, even if other estimation routines perform much better, the R routines are more likely to be relevant as forecasters may tend to use them or similar algorithms.

In detail, we consider the following five strategies and report on pairwise comparisons between them:

1. Training-sample evaluation (TS) over 50% of the available time range. The model with the smaller MSE over the training sample is chosen as the forecasting model;

2. Training-sample evaluation (TS-DM-N) as in # 1, but followed by a Diebold-Mariano (DM) test. The more complex (here, the ARMA) model is only chosen if the DM test statistic is significant at a nominal N(0,1) 5% level;

3. Training-sample evaluation (TS-DM-B) as in # 2, but the significance of the DM statistic is evaluated against a carefully bootstrapped value;

4. Training-sample evaluation (TS-F-B) followed by an F–test evaluation according to Clark and McCracken (2001,2005). This statistic does not follow any standard distribution even in simple cases, so this strategy is evaluated with bootstrapping only;

5. AIC evaluation over the full available sample and choosing the model with the lower AIC value;

6. Training-sample evaluation followed by an evaluation of the concomitant GW statistic over moving windows (TS-GW). The more complex model is chosen only if the GW test statistic is significant at a nominal 5% significance level.

Other predictive accuracy tests can be considered, but choosing these as alternative selection strategies is hardly likely to affect our results. For example, Clark and McCracken (2005) show that in nested applications encompassing tests following Harvey et al. (1998) and DM tests are asymptotically equivalent, and the discriminatory power of their F tests is also close to the other tests, as all of them process the same information.

Basically, all our simulations follow the same pattern with expanding windows. For $N = 100$, observations $t = 52, \ldots, 99$ are used as a test sample in the sense that models are estimated from training samples $t = 1, \ldots, T$ and the mean squared error of one-step out-of-sample forecasts for observations X_{T+1} is evaluated by averaging over $T = 51, \ldots, 98$. In the pure TS strategy, the model with the smaller average MSE is selected as the one whose out-of-sample forecast for the observation at N based on the sample $t = 1, \ldots, N-1$ is considered. In the DM strategy, the more complex ARMA model is selected only if the DM test rejects. Otherwise, the DM strategy chooses the simpler AR(1) model to forecast the observation at N.

3.2 The bootstrapped DM test and training

Inoue and Kilian (2006) established the result that, roughly, TS works like an information criterion asymptotically. Depending on whether the share of the training sample in the available sample converges to unity or not, the information criterion can be a consistent one like the BIC by Schwarz or a prediction-optimizing efficient one like the AIC by Akaike.

In small samples, it is now widely recognized that the AIC tends to be too 'liberal' in the sense that it leans toward over-parameterization (see McQuarrie and Tsai, 1998), which issue will be evaluated in the next subsection. Thus, the strategy not to accept the more general ARMA(1,1) model as a prediction model unless the DM test additionally rejects its null may benefit prediction.

The TS strategy has been considered, for example, by Inoue and Kilian (2006), who were interested in the question whether it be outperformed by BIC selection. There are

arguments in favor of both TS and BIC. BIC uses the whole sample, while TS restricts attention to the portion that is used for the training evaluation. On the other hand, the property of BIC consistency is asymptotic and need not imply optimality in a small sample. Inoue and Kilian find that BIC dominates TS over large portions of the admissible parameter space. We note that their simulations differ from ours by the non-nested nature of decisions on single realizations: some trajectories may be classified as ARMA by one strategy but as AR by the other, while other trajectories experience the reverse discrepancy. We do not consider BIC selection in our experiments.

It is well known that the normal significance points of the DM test are invalid if nested models are compared (see Clark and McCracken, 2001, 2012). For our experiment, we obtained correct critical values using the bootstrap-in-bootstrap procedure suggested by Kilian (1998). The known estimation bias for AR models is bootstrapped out in a first run, and another bootstrap with reduced bias is then conducted to deliver significance points. Unfortunately, the bootstrap is time-consuming, thus only 100 bootstrap iterations can be performed for this experiment. Nonetheless, correspondence to the targeted size of 5% is satisfactory.

The left graph in Figure 2 shows the result of our Monte Carlo at $N = 50$. It corresponds to expectations at least with regard to the behavior around the horizontal axis, where the AR model is true. Here, DM testing changes the implicit significance level of the TS procedure of around 20% to 5%. Less AR trajectories are classified incorrectly as ARMA, and forecasting precision improves. At some distance from the axis, TS dominates due to its larger implicit 'power' that attains 100% at $\theta = 0.8$. In the graph, the size of the filled circles is used to indicate the intensity of the discrepancy in mean-squared errors. The maximum tilt in favor of DM testing is achieved at $(\phi, \theta) = (-0.4, -0.2)$, the maximum in favor of pure TS occurs along the margins of the square, for extreme values of ϕ and of θ.

An unexpected feature of Figure 2 is the asymmetry of the preference areas: the northwest area with negative ϕ and negative correlation among residuals after a preliminary AR fit appears to be more promising for the DM test than the southeast area with positive ϕ and negative residual correlation after AR fitting. This effect is not easily explained. The Greenland graph at $N = 50$ is approximately symmetric. Support for the pure AR model

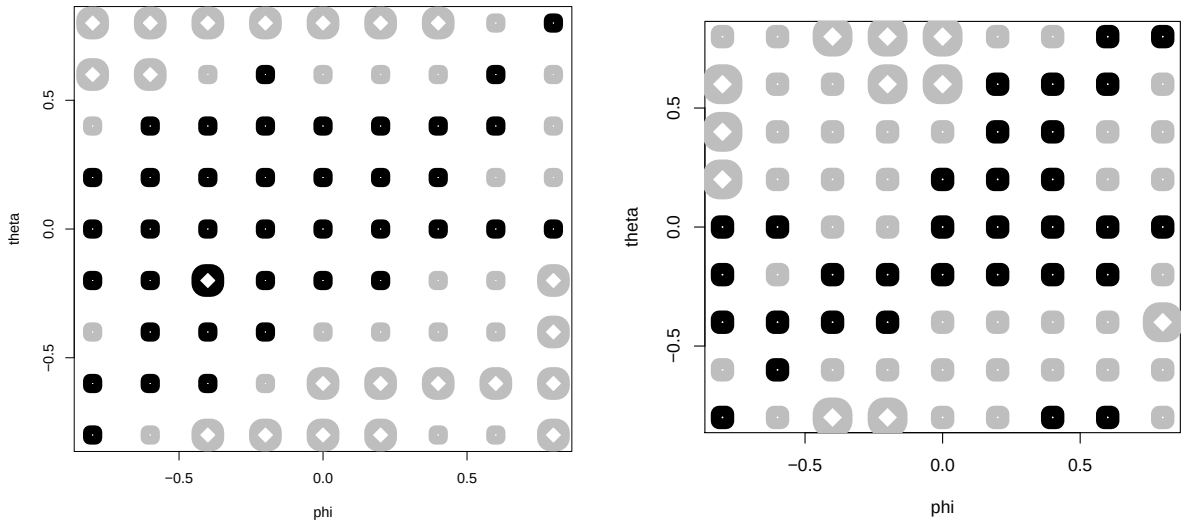

Figure 2: AR/ARMA model selected by MSE miminization over a training sample (TS) versus selection based on a Diebold-Mariano test (TS-DM-B) with bootstrapped significance points. Left graph for $N = 50$, right graph for $N = 100$. Gray dots express a preference for TS, black dots one for DM. 1000 replications.

is slightly stronger in the northwest than in the southeast, according to TS and to DM, with a difference of around 10 percentage points.

The right graph of Figure 2 shows the results for $N = 100$. For most parts of the scheme, the version without DM appears to be the preferred strategy. DM dominance has receded to an area approximately matching Θ_R, with even two perverse spots along the x–axis. Like the isolated area in the southeast at $(0.4, -0.8)$ and $(0.6, -0.8)$, we interpret them as artifacts. In these areas, the difference in performance between TS and DM is so small that a larger number of replications would be required to bring them out clearly.

We also ran some unreported exploratory experiments with larger N. The preference area for DM versus TS shrinks faster than Greenland as the sample size increases. The DM test yields a reliable decision procedure for AR within ARMA for those who are interested in theoretical data-generating processes, but it does not help in selecting prediction models. In summary, a tendency is palpable that for even larger N any support for the DM version tends to disappear. The change from an implicit 15-20% test to a 5% test does not benefit forecasting properties.

Generally, we note some typical features of our graphical visualization of the simulations. A conservative procedure, i.e. one that tends to stay with the simpler AR model, will dominate on the dark area in the Greenland plot, as there it is beneficial to use the AR model as a forecasting tool. A rather liberal procedure, i.e. one that tends to prefer the ARMA model and has 'high power' in the traditional sense of hypothesis testing, will dominate on the outer light area of the Greenland plot. In this interpretation, a strategy that dominates on the outer area and on a portion of the inner area can be seen as liberal and relatively promising, while a strategy that inhabits a narrow outer margin is too liberal, and one that lives on a narrow band around Θ_R is too conservative to be efficient.

3.3 The bootstrapped F test and training

If a parallel experiment to the previous one is run on the basis of the F test due to Clark and McCracken (2001,2005) that replaces the DM statistic by

$$\frac{N}{2} \times \frac{\sum_{t=N/2}^{N-1}(e_{1,t}^2 - e_{2,t}^2)}{\sum_{t=N/2}^{N-1} e_{2,t}^2},$$

with prediction errors from the two models $e_{j,t}, j = 1, 2$, this yields the outcome shown in Figure 3. For $N = 50$, the additional testing step boosts forecasting accuracy on Θ_R and in some areas that may be artifacts. For $N = 100$, the procedures with and without the testing step become so close that the picture fades. In summary, the additional testing step helps in small samples if the generating model is AR(1) or at least very close to AR(1), which is not surprising, whereas it does not help at all in larger samples.

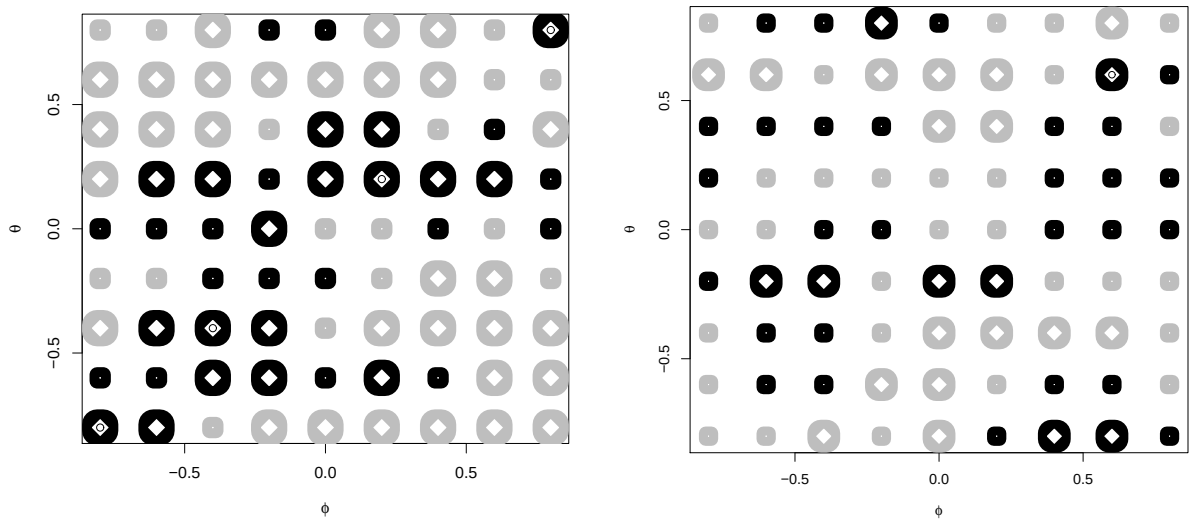

Figure 3: AR/ARMA model selected by MSE minimization over a training sample (TS) versus selection based on an F test (TS-F-B) with bootstrapped significance points. Left graph for $N = 50$, right graph for $N = 100$. Gray dots express a preference for TS, black dots one for the F test. 1000 replications.

3.4 Nested models and naive normal distribution

As we mentioned above, the (normal) DM test is known to suffer from severe distortions in nested model situations, see Clark and McCracken (2001, 2012). Nevertheless, it has been used repeatedly in the empirical forecasting literature, and the typical handling of stochastic properties may be somewhere in between the correct bootstrap used in the previous subsection and the naive N(0,1) distribution suggested in the original DM paper.

Again, we simulate ARMA(1,1) series of length $N = 50, 100$, with Gaussian N(0,1) noise (ε_t) and the identical design as above. Out-of-sample forecasts for the latter half of the sample are generated on the basis of AR(1) and ARMA(1,1) models with estimated parameters, and the model with the lower MSE is used to predict the observation at position $N + 1$. Significance of the DM test statistic, however, is now checked against the theoretically incorrect $N(0, 1)$ distribution instead of the carefully bootstrapped correct null distribution.

The relative performance of the pure TS strategy and of the DM–test strategy is evaluated graphically in Figure 4. The area of preference for the DM strategy appears to be a subset of the inner Greenland area, which implies that the selection strategy based on the DM test with normal quantiles is suboptimal.

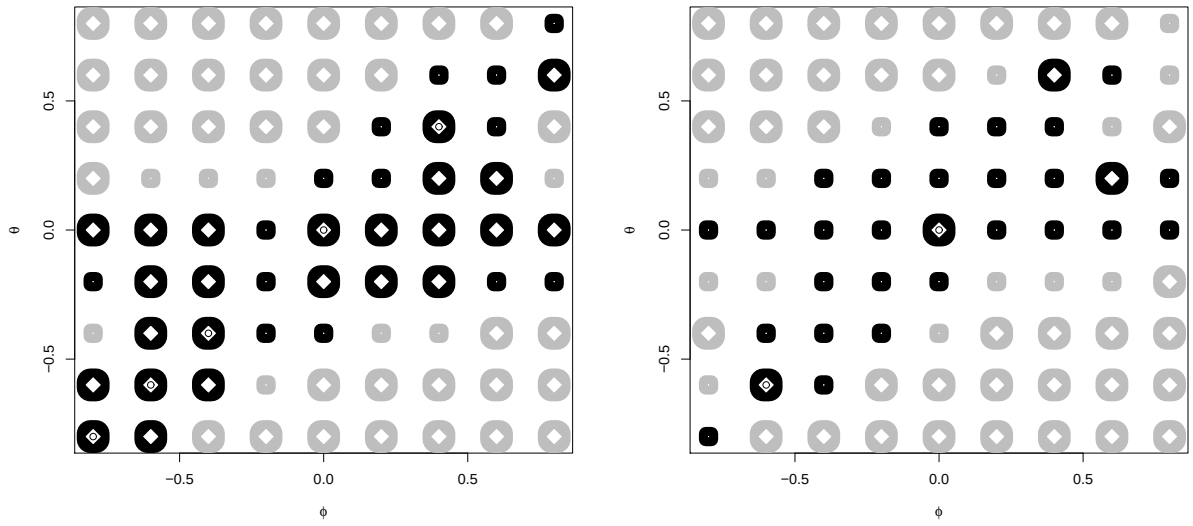

Figure 4: AR/ARMA model selected by MSE minimization over a training sample (TS) versus selection based on a Diebold-Mariano test (TS-DM-N) with 'naive' 5% $N(0, 1)$ significance points. Left graph for $N = 50$, right graph for $N = 100$. Gray dots express a preference for TS, black dots one for DM. 1000 replications.

3.5 AIC and training

According to Inoue and Kilian (2006), TS and AIC will be equivalent in large samples if the training sample grows linearly with the sample size. In our experiments, we set the training sample at 50% of the complete sample, i.e. slightly more than economic forecasters tend to use although less than would be suggested by an asymptotic approximation to BIC.

The arguments considered by Inoue and Kilian (2006) again apply here. Information criteria tend to exploit the information in the entire sample, while TS concentrates on the part that is used as a training sample. As GW argue, this latter property constitutes an advantage if the generating mechanism changes slowly over time, but our generating models are exclusively time-homogeneous. Rather, TS focuses on the specific aim of the forecasting exercise, while AIC has been derived on grounds of asymptotic properties and is known to perform poorly in small samples (see McQuarrie and Tsai, 1998).

Figure 5 shows the regions where TS and AIC dominate. Among the strategies, AIC is the exception, as it is the only procedure that tends to be more liberal than TS. For this reason, the colors are turned on their heads, and TS dominates around Θ_R, whereas AIC dominates in the corners, where it classifies considerably more trajectories into the ARMA(1,1) class. For the larger samples $N = 100$, AIC gains ground, maybe due to its more efficient processing of the sample information. For the smaller samples $N = 50$, TS dominance around Θ_R is quite pronounced. On the whole, AIC–based forecast model selection evolves as the most serious rival strategy to pure TS.

3.6 Giacomini-White and training

GW considered forecasts based on *moving* windows of fixed size m. In a sample of size N, $N-m-2$ such one-step forecasts are available, if the last observation X_N is to be reserved for a final evaluation. GW call their test statistic 'Wald-type' and define it formally as

$$T_{m,n} = n \left[n^{-1} \sum_{t=m}^{N-2} h_t \left\{ g(e_{1,t+1}) - g(e_{2,t+1}) \right\} \right]' \hat{\Omega}_n^{-1} \left[n^{-1} \sum_{t=m}^{N-2} h_t \left\{ (g(e_{1,t+1}) - g(e_{2,t+1}) \right\} \right],$$

with $n = N-m-2$ and $g(x) = x^2$ in our case and h_t a 2–vector of 'test functions', typically specified as a constant 1 in the first position and the lagged discrepancy $g(e_{1,t}) - g(e_{2,t})$

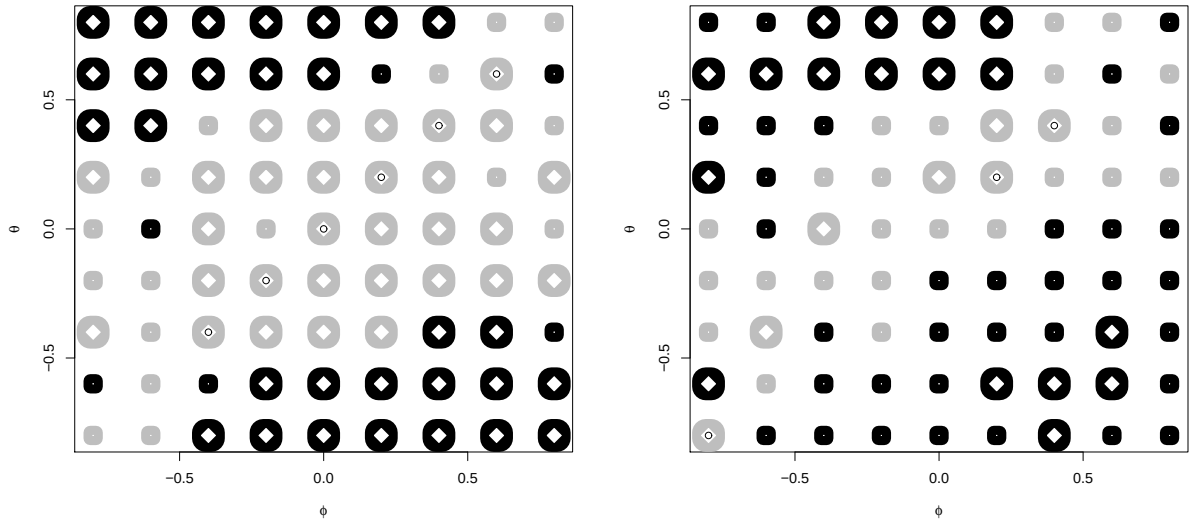

Figure 5: Smaller MSE if either TS or AIC are used to select between ARMA(1,1) and AR(1) as prediction models. Left graph for $N = 50$, right graph for $N = 100$. Light spots prefer TS, dark spots prefer AIC.

in its second position. Denoting the summand terms by $Z_{m,t+1}$, the 2×2–matrix $\hat{\Omega}_n$ is defined as

$$\hat{\Omega}_n = n^{-1} \sum_{t=m}^{N-2} Z_{m,t+1} Z'_{m,t+1}.$$

GW show that under their null this statistic $T_{m,n}$ will be distributed as χ^2_2. We note that the construction of this test statistic is not symmetric, and we typically are interested in the alternative of method 2 outperforming method 1, such that the discrepancies tend to be positive.

We consider the efficiency of this test as a simplicity booster in the following sense. The AR(1) and the ARMA(1,1) forecast are evaluated comparatively on *expanding* subsamples as before. If the AR(1) forecast wins, it is selected. If the ARMA(1,1) forecast wins, it is selected only in those cases where the GW test rejects at 5%. In line with the simulations presented by GW, we specify $m = N/3$ for the window width.

Figure 6 shows that advantages for the GW step are restricted to the Greenland area at $N = 50$ and weaken for $N = 100$. It may be argued that running the GW test at risk

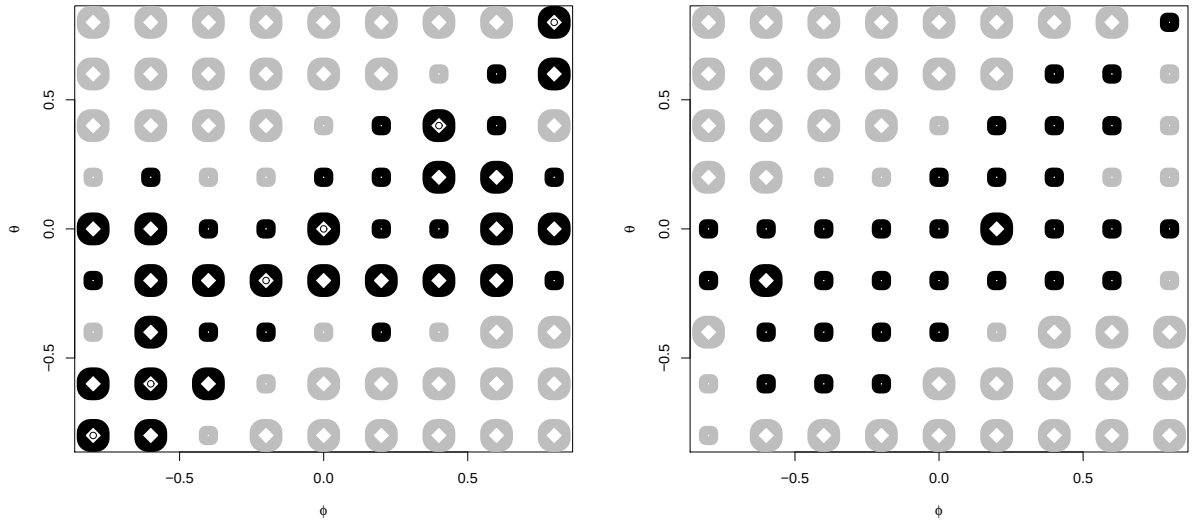

Figure 6: Smaller MSE if either pure TS or TS jointly with the Giacomini-White test at 5% are used to select between ARMA(1,1) and AR(1) as prediction models. Left graph for $N = 50$, right graph for $N = 100$. Light spots prefer pure TS, dark spots prefer TS with GW.

levels around 50% may benefit forecasting. We would like, however, to keep to the way that the procedures are used in current practice, and a conventional risk level is part of the TS-GW procedure in our design.

3.7 Summary of the nested experiment

Overall, it appears that the pure TS strategy that decides on the model to be eventually used for prediction on the basis of a straightforward training-sample evaluation is hard to beat. Any significance test decision on top of it in the sense of simplicity boosting tends to worsen the results over a sizable portion of the parameter space. The least attractive ideas appear to be DM testing based on the incorrect normal distribution and GW. The most competitive idea appears to be the direct usage of information criteria.

In detail, if we compare the MSE values for TS with the bootstrapped DM and for AIC in a graph that is comparable to the hitherto shown figures, no recognizable pattern emerges. The difference among the two strategies appears to be dominated by the sampling variation

in the Monte Carlo, and the two procedures have comparable quality. TS without any further test does slightly worse but still comes pretty close to the two graphically analyzed strategies. The other suggestions, TS-DM-N and GW, perform considerably worse.

In search of the reasons for the differences in performance among the strategies, one may surmise that these are rooted in the frequency at which either of the two models is selected. Figure 7 shows that this cannot be the complete explanation. For a reasonable visualization, we restrict attention to a vertical slice through our maps, i.e. we show how the strategies behave in the presence of pure MA generating models. For low θ, an AR(1) may be a reasonable approximation for MA(1) behavior, and the implied forecasts may be rather accurate. Thus, TS–based strategies may find it difficult to discriminate among the two models. The primary impression, however, is dominated by a strong three-way classification among strategies: TS and AIC are 'liberal' procedures that are locally equivalent to hypothesis tests at significance levels of around 20%; the two test-based strategies with bootstrap approximately match the targeted 5% rate; the naive DM and the GW test are extremely conservative and opt for AR models even in the presence of sizeable MA coefficients. Within the three classes, differences are only slight: the F test performs 'better' than DM for negative θ and 'worse' for $\theta > 0$; AIC appears to dominate TS; and GW is even more conservative than naive DM.

Naive normal DM and Giacomini-White suffer from similar problems. The attempt of the GW test to attain 5% significance at the Greenland boundary instead of the population null hypothesis Θ_R implies that the GW–based strategy has a much too strong preference for simplicity. On the other hand, TS and AIC have a comparable tendency toward the more complex model. AIC tends to dominate TS, however, as it selects the better trajectories while the selection frequency is similar. This may be rooted in a more efficient processing of sample information by taking the entire sample into account instead of concentrating on the latter half. Quite successful strategies are TS-DM-B and TS-F. In particular for small parameter values, these strategies boost simplicity efficiently and thus succeed in overcoming the tendency of pure TS to classify trajectories with only weak evidence against AR(1) as ARMA(1,1). Even if the generating models definitely are not AR(1), it remains efficient to see them as AR(1), to estimate just an autoregressive coefficient, and to evaluate

the concomitant forecast. Selection based on the bootstrapped tests and on AIC dominates, and, as AIC is much faster and easier to calculate, the bottom line may be some preference for traditional AIC.

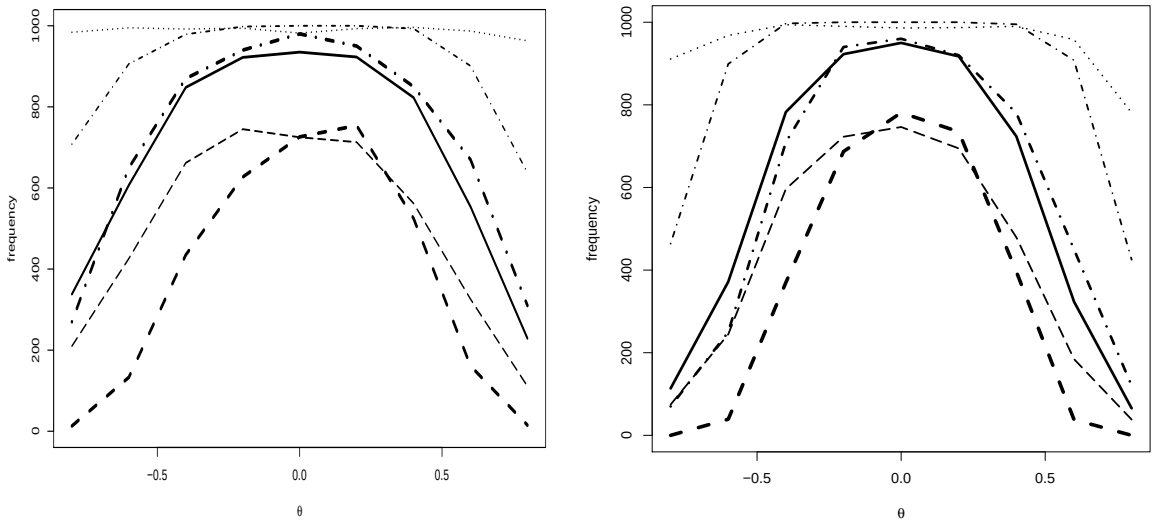

Figure 7: Frequency of choosing the AR(1) model rather than ARMA(1,1) if generating model is MA with given coefficient θ. Curves stand for bootstrapped Diebold-Mariano (bold solid), bootstrapped F (bold dash-dotted), AIC (bold dashed), Giacomini-White (dotted), unconstrained TS (dashed), naive normal Diebold-Mariano (dash-dotted). Left graph for $N = 50$, right graph for $N = 100$.

4 Other designs

4.1 A non-nested ARMA design

In this experiment, data are generated from ARMA(2,2) processes. There are twelve pairs of AR coefficients. The left graph in Figure 8 shows their distribution across the stability region. Eight pairs yield complex conjugates in the roots of the characteristic AR polynomial and hence cyclical behavior in the generated processes. Three pairs imply real roots, and one case is the origin in order to cover pure MA structures. We feel that this design

exhausts the interesting cases in the stability region, avoiding near-nonstationary cases that may impair the estimation step.

These autoregressive designs are combined with the moving-average specifications given in the right graph of Figure 8: a benchmark case without MA component, a first-order MA model, and an MA(2) model with $\theta_1 = 0$. Like in our other experiments, errors are generated as Gaussian white noise.

This design is plausible. Second-order models are often considered for economics variables, as they are the simplest linear models that generate cycles. Thus, AR(2) models are not unlikely empirical candidates for data generated from ARMA(2,2): the dependence structure rejects white noise, autoregressive models can be fitted by simple least squares. Similarly, ARMA(1,1) may be good candidates if a reliable ARMA estimator is available: often, ARMA models are found to provide a more parsimonious fit than pure autoregressions.

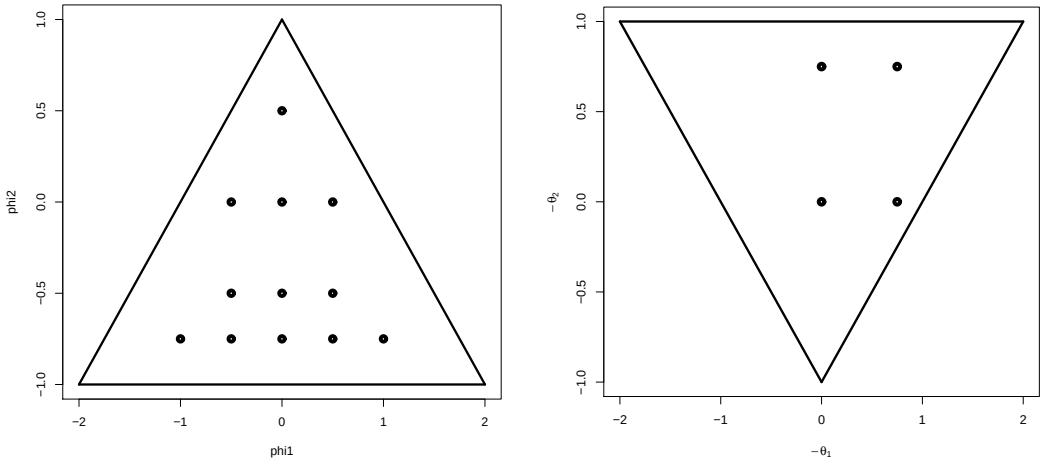

Figure 8: Parameter values for the autoregressive part of the generated ARMA models within the triangular region of stable AR models and values for the MA part within the invertibility region for MA(2) models.

The columns headed ARMA and AR in Tables 1 and 2 show the MSE for predictions using the ARMA(1,1) and the AR(2) models, respectively, if the data-generating process is ARMA(2,2). We note that the prediction models are misspecified for most though not

all parameter values. The first twelve lines correspond to the design $(\theta_1, \theta_2) = (0, 0)$, when the AR(2) model is correctly specified.

The prevailing impression is that the AR(2) model dominates at most parameter values. This dominance is partly caused by the comparatively simpler MA part of the generating processes, but it may also indicate greater robustness in the estimation of autoregressive models as compared to mixed models. The relative performance of the two rival models, measured by the ratio of MSE(AR) and MSE(ARMA), remains almost constant as N increases from 100 to 200, which indicates that the large-sample ratios may already have been attained. The absolute performance, however, improves perceptibly as the sample size increases.

The columns headed TS and DM-N report the MSE based on the direct evaluation of a training sample and on the additional DM step on the basis of Gaussian significance points in line with the non-nested design. In pure AR(2) designs, there are mostly gains for imposing the DM step. The null model of the test is the true model, and the extra step helps in supporting it. For strong MA effects, the DM step tends to incur some deterioration.

Another column (DM-B) refers to the DM-based selection using bootstrapped significance points. This bootstrapped version classifies substantially more trajectories as ARMA(1,1) than DM-N or TS, which incurs a deterioration in performance. Note that bootstrapping has been conducted for the test null model, i.e. the AR(2) model, which is not the data-generating mechanism that is presumed unknown to the forecaster. This situation may be representative for empirical situations where the data-generating mechanism is also unknown and the null distribution used for the bootstrap is unreliable. It is of some interest that DM-B does not even perform satisfactorily when AR(2) is the generating model. For most parameter constellations, DM-B performs worst among all competing procedures.

The column AIC selects the forecasting model based on the likelihood, as both rival models have two free parameters. In most cases, AR(2) incurs the better likelihood than ARMA(1,1), the forecasts remain close to the AR forecasts, and AIC wins in approximately one third of all cases, on a par with DM-N and with GW.

Table 1: Results of the simulation for $N = 100$.

design parameter values				mean squared errors							
ϕ_1	ϕ_2	θ_1	θ_2	ARMA	AR	TS	DM-N	DM-B	AIC	GW	F-B
0	0.5	0	0	1.145	0.981	0.983	0.981	1.055	0.981	0.982	1.000
-0.5	0	0	0	0.989	0.995	0.995	0.995	1.048	0.993	0.995	0.995
0	0	0	0	0.995	0.991	0.998	0.994	1.052	0.998	0.992	0.990
0.5	0	0	0	0.983	0.984	0.981	0.984	1.047	0.983	0.984	0.985
-0.5	-0.5	0	0	1.158	1.002	1.014	1.004	1.066	1.010	1.002	0.991
0	-0.5	0	0	1.345	1.011	1.014	1.011	1.185	1.011	1.011	1.029
0.5	-0.5	0	0	1.161	1.006	1.020	1.006	1.103	1.010	1.006	1.025
-1	-0.75	0	0	1.544	0.997	1.003	0.997	1.047	1.001	0.994	0.996
-0.5	-0.75	0	0	1.752	1.006	1.009	1.006	1.491	1.006	1.006	0.984
0	-0.75	0	0	2.242	1.018	1.019	1.018	1.487	1.018	1.018	1.057
0.5	-0.75	0	0	1.738	1.019	1.026	1.019	1.155	1.019	1.019	1.037
1	-0.75	0	0	1.483	0.991	0.996	0.992	1.064	0.991	0.991	1.031
0	0.5	0	0.75	2.651	1.318	1.318	1.318	1.527	1.318	1.318	1.318
-0.5	0	0	0.75	1.336	1.279	1.284	1.279	1.386	1.279	1.279	1.280
0	0	0	0.75	1.380	1.167	1.166	1.166	1.258	1.167	1.167	1.181
0.5	0	0	0.75	1.370	1.286	1.293	1.286	1.395	1.286	1.289	1.293
-0.5	-0.5	0	0.75	1.169	1.178	1.171	1.176	1.278	1.178	1.176	1.181
0	-0.5	0	0.75	1.169	1.018	1.024	1.017	1.102	1.016	1.019	1.020
0.5	-0.5	0	0.75	1.171	1.167	1.165	1.167	1.266	1.167	1.167	1.163
-1	-0.75	0	0.75	1.845	1.365	1.366	1.366	2.171	1.365	1.363	1.364
-0.5	-0.75	0	0.75	1.248	1.177	1.184	1.182	1.262	1.179	1.177	1.173
0	-0.75	0	0.75	0.995	0.991	0.998	0.994	1.046	0.998	0.993	0.991
0.5	-0.75	0	0.75	1.264	1.185	1.193	1.184	1.306	1.185	1.185	1.209
1	-0.75	0	0.75	1.867	1.323	1.325	1.323	1.477	1.323	1.323	1.330

design parameter values				mean squared errors							
ϕ_1	ϕ_2	θ_1	θ_2	ARMA	AR	TS	DM-N	DM-B	AIC	GW	F-B
0	0.5	0.75	0	0.980	0.984	0.982	0.985	1.053	0.982	0.984	0.983
-0.5	0	0.75	0	1.010	1.009	1.012	1.006	1.080	1.012	1.016	1.008
0	0	0.75	0	1.006	1.086	1.020	1.063	1.047	1.012	1.083	1.040
0.5	0	0.75	0	0.995	1.147	1.017	1.048	1.092	1.003	1.120	1.039
-0.5	-0.5	0.75	0	1.565	1.212	1.220	1.211	1.410	1.218	1.212	1.234
0	-0.5	0.75	0	1.316	1.304	1.278	1.294	1.337	1.275	1.302	1.316
0.5	-0.5	0.75	0	1.260	1.317	1.281	1.264	1.331	1.297	1.303	1.294
-1	-0.75	0.75	0	1.829	1.270	1.280	1.270	1.389	1.274	1.270	1.271
-0.5	-0.75	0.75	0	2.553	1.360	1.371	1.360	2.189	1.360	1.360	1.370
0	-0.75	0.75	0	2.284	1.436	1.475	1.438	2.019	1.444	1.434	1.704
0.5	-0.75	0.75	0	2.047	1.435	1.496	1.443	2.028	1.450	1.437	1.626
1	-0.75	0.75	0	2.000	1.413	1.441	1.410	1.961	1.425	1.402	1.639
0	0.5	0.75	0.75	1.856	1.635	1.657	1.639	1.758	1.635	1.636	1.653
-0.5	0	0.75	0.75	1.443	1.277	1.283	1.280	1.353	1.277	1.277	1.296
0	0	0.75	0.75	1.300	1.293	1.290	1.294	1.383	1.293	1.292	1.303
0.5	0	0.75	0.75	1.386	1.272	1.281	1.275	1.361	1.272	1.272	1.271
-0.5	-0.5	0.75	0.75	1.067	1.069	1.066	1.068	1.120	1.072	1.069	1.070
0	-0.5	0.75	0.75	1.288	1.179	1.189	1.183	1.264	1.179	1.178	1.214
0.5	-0.5	0.75	0.75	1.618	1.305	1.329	1.306	1.377	1.305	1.310	1.337
-1	-0.75	0.75	0.75	1.051	1.054	1.051	1.053	1.141	1.054	1.054	1.054
-0.5	-0.75	0.75	0.75	1.117	1.067	1.073	1.067	1.123	1.068	1.069	1.067
0	-0.75	0.75	0.75	1.680	1.326	1.353	1.324	1.413	1.327	1.327	1.360
0.5	-0.75	0.75	0.75	2.366	1.515	1.535	1.520	1.826	1.515	1.526	1.836
1	-0.75	0.75	0.75	3.058	1.630	1.636	1.629	2.631	1.630	1.632	1.727

Table 2: Results of the simulation for $N = 200$.

design parameter values				mean squared errors							
ϕ_1	ϕ_2	θ_1	θ_2	ARMA	AR	TS	DM-N	DM-B	AIC	GW	F-B
0	0.5	0	0	1.180	0.978	0.981	0.978	0.984	0.978	0.978	0.977
-0.5	0	0	0	0.986	0.986	0.986	0.986	0.986	0.986	0.986	0.986
0	0	0	0	0.988	0.983	0.986	0.983	1.136	0.986	0.983	1.142
0.5	0	0	0	0.984	0.982	0.984	0.983	0.981	0.984	0.982	0.982
-0.5	-0.5	0	0	1.143	0.990	0.985	0.991	0.997	0.989	0.990	1.006
0	-0.5	0	0	1.278	0.995	0.995	0.995	1.194	0.995	0.995	1.002
0.5	-0.5	0	0	1.135	0.991	0.992	0.991	1.100	0.991	0.991	1.006
-1	-0.75	0	0	1.517	0.985	0.990	0.985	0.985	0.985	0.985	1.019
-0.5	-0.75	0	0	1.700	0.993	0.993	0.993	1.684	0.993	0.993	1.013
0	-0.75	0	0	2.053	0.994	0.994	0.994	1.813	0.994	0.994	1.047
0.5	-0.75	0	0	1.704	0.993	0.993	0.993	1.214	0.993	0.993	1.014
1	-0.75	0	0	1.466	0.993	0.991	0.993	1.040	0.993	0.993	1.022
0	0.5	0	0.75	2.480	1.344	1.344	1.344	1.711	1.344	1.344	1.342
-0.5	0	0	0.75	1.384	1.278	1.281	1.278	1.258	1.278	1.278	1.262
0	0	0	0.75	1.426	1.163	1.164	1.163	1.170	1.163	1.163	1.167
0.5	0	0	0.75	1.409	1.332	1.331	1.332	1.330	1.332	1.332	1.324
-0.5	-0.5	0	0.75	1.179	1.181	1.181	1.180	1.177	1.181	1.181	1.182
0	-0.5	0	0.75	1.039	1.009	1.011	1.009	1.013	1.011	1.010	1.013
0.5	-0.5	0	0.75	1.183	1.180	1.180	1.179	1.175	1.180	1.181	1.188
-1	-0.75	0	0.75	1.931	1.409	1.408	1.408	1.412	1.409	1.409	1.412
-0.5	-0.75	0	0.75	1.233	1.171	1.173	1.169	1.179	1.171	1.171	1.177
0	-0.75	0	0.75	0.988	0.983	0.987	0.982	1.023	0.986	0.983	1.019
0.5	-0.75	0	0.75	1.277	1.193	1.197	1.192	1.262	1.193	1.193	1.186
1	-0.75	0	0.75	1.915	1.370	1.376	1.370	1.391	1.370	1.370	1.386

design parameter values				mean squared errors							
ϕ_1	ϕ_2	θ_1	θ_2	ARMA	AR	TS	DM-N	DM-B	AIC	GW	F-B
0	0.5	0.75	0	0.985	0.987	0.987	0.987	0.984	0.987	0.987	0.987
-0.5	0	0.75	0	0.989	1.008	0.993	0.999	0.989	0.989	1.007	1.009
0	0	0.75	0	0.990	1.097	0.993	1.029	0.991	0.988	1.076	1.035
0.5	0	0.75	0	0.982	1.164	0.986	1.040	0.983	0.982	1.089	0.985
-0.5	-0.5	0.75	0	1.497	1.220	1.233	1.219	1.477	1.214	1.220	1.204
0	-0.5	0.75	0	1.271	1.307	1.266	1.290	1.277	1.254	1.308	1.288
0.5	-0.5	0.75	0	1.269	1.353	1.276	1.332	1.253	1.241	1.347	1.349
-1	-0.75	0.75	0	1.736	1.309	1.309	1.309	1.309	1.309	1.309	1.309
-0.5	-0.75	0.75	0	2.445	1.362	1.362	1.362	2.304	1.362	1.362	1.362
0	-0.75	0.75	0	2.080	1.447	1.446	1.447	2.083	1.446	1.446	1.749
0.5	-0.75	0.75	0	2.094	1.472	1.496	1.469	2.095	1.475	1.473	1.799
1	-0.75	0.75	0	2.038	1.481	1.512	1.484	2.035	1.493	1.477	1.765
0	0.5	0.75	0.75	1.931	1.730	1.728	1.730	1.746	1.730	1.723	1.723
-0.5	0	0.75	0.75	1.491	1.294	1.299	1.294	1.335	1.295	1.294	1.307
0	0	0.75	0.75	1.323	1.317	1.320	1.318	1.324	1.317	1.318	1.285
0.5	0	0.75	0.75	1.428	1.293	1.290	1.293	1.296	1.293	1.293	1.294
-0.5	-0.5	0.75	0.75	1.060	1.055	1.055	1.056	1.057	1.056	1.055	1.056
0	-0.5	0.75	0.75	1.289	1.177	1.179	1.177	1.242	1.177	1.177	1.175
0.5	-0.5	0.75	0.75	1.663	1.344	1.351	1.344	1.352	1.344	1.346	1.356
-1	-0.75	0.75	0.75	1.088	1.079	1.085	1.079	1.092	1.079	1.078	1.071
-0.5	-0.75	0.75	0.75	1.082	1.054	1.059	1.055	1.065	1.057	1.054	1.060
0	-0.75	0.75	0.75	1.656	1.349	1.352	1.349	1.443	1.349	1.350	1.409
0.5	-0.75	0.75	0.75	2.440	1.590	1.599	1.592	2.213	1.590	1.591	1.746
1	-0.75	0.75	0.75	3.214	1.697	1.695	1.695	3.198	1.697	1.697	1.827

The information conveyed in Tables 1 and 2 can be summarized as follows. For $N = 100$, the procedures DM-N, AIC, and GW are approximately equivalent, and they dominate the simple TS comparison. The bootstrapped DM-B is not competitive, and this includes those designs where actually AR(2) is the correctly specified model. For $N = 200$, the simple TS comparison comes closer to the dominant procedures DM-N, AIC, GW, whereas DM-B still fails to convince. In summary, TS is not optimal as a decision guideline, and it pays to boost simplicity by making the selection procedure more conservative. Among the three conservative tests, GW is the most conservative one, and it approximates the pure AR(2) strategy at least for $N = 200$. DM-B does not fail because it is too liberal, but rather because it chooses the wrong trajectories. In other words, ARMA is selected in those cases where AR would have generated the better forecast even though it is not the generating model. We also note that by optimizing the selection among trajectories, the pure strategies can be improved upon, and that similarly a bad selection strategy can perform worse than the worse pure strategy. For example, white noise for $N = 200$ is predicted better by AR(2) than by ARMA(1,1), and GW comes close to AR(2) by almost never rejecting its null, whereas DM-B performs much worse than the pure ARMA(1,1) strategy.

4.2 A nonlinear generation mechanism

In this experiment, the data are generated by a nonlinear time-series process that has been suggested by Tiao and Tsay (1994) for the growth rate of U.S. gross national output. Their self-exciting threshold autoregressive (SETAR) model defines four regimes that correspond to whether an economy is in a recession or in an expansion and on whether the recessive or expansive tendencies are accelerating or decelerating.

Define X_t as the growth rate of U.S. output. With parameter values directly taken from the model fitted by Tiao and Tsay (1994), the model reads

$$X_t = \begin{cases} -0.015 - 1.076X_{t-1} + \varepsilon_{1,t}, & X_{t-1} \leq X_{t-2} \leq 0, \\ -0.006 + 0.630X_{t-1} - 0.756X_{t-2} + \varepsilon_{2,t}, & X_{t-1} > X_{t-2},\ X_{t-2} \leq 0, \\ 0.006 + 0.438X_{t-1} + \varepsilon_{3,t}, & X_{t-1} \leq X_{t-2},\ X_{t-2} > 0, \\ 0.004 + 0.443X_{t-1} + \varepsilon_{4,t}, & X_{t-1} > X_{t-2} > 0. \end{cases}$$

Errors $\varepsilon_{j,t}$ are Gaussian white noise. Their standard deviations $\sigma_j = \sqrt{\mathrm{E}\varepsilon_{j,t}^2}$, $\sigma_1 = 0.0062$, $\sigma_2 = 0.0132$, $\sigma_3 = 0.0094$, and $\sigma_4 = 0.0082$, are an important part of the parametric structure. In contrast to linear models, threshold models may behave quite differently if the relative scales of the error processes change.

For a more recent summary of known results on the statistical properties of this model class, see Fan and Yao (2005). Some further characteristics are revealed easily by some simulation and inspection. Within regime 1, which corresponds to a deepening economic recession, the model is 'locally unstable', as the coefficient is less than -1. Nevertheless, the model is 'globally stable'. In fact, it is the large negative coefficient in regime 1, where lagged growth rates are by definition negative, which pushes the economy quickly out of a recession. X_t tends to remain in regimes 3 and 4 for much longer time spans than in regime 2, and it spends the shortest episodes in the deepening recession of regime 1. Thus, the exercise of fitting linear time-series models to simulated trajectories often leads to coefficient estimates that are close to those for regimes 3 and 4.

For our prediction experiment, we use samples drawn from the SETAR process with $N = 100, 200$. Burn-in samples of 1000 observations are generated and discarded, as the distribution of the nonlinear generating process may be affected by starting conditions. 1000 replications are performed. The hypothetical forecaster is supposed to be unaware of the nonlinear nature of the DGP, and she fits AR(p) and ARMA(p, p) models to the time series. In analogy to the other experimental designs, the models deliver out-of-sample forecasts for the latter half of the observation range, excepting the very last time point. Either the model with better performance in the training sample or the one that is 'significantly' better according to a test is used to forecast this last time point. We also compare the accuracy of the strategies to the forecasts that always use the autoregressive or the ARMA model.

A main difference to the former experiments is that, rather than imposing a fixed lag order p on the time-series models, we determine an optimal $\hat{p}$ by minimizing AIC over the range $1, \ldots, p^*$. The ARMA model uses twice as many parameters as the AR model, so maximum lag orders are set at the popular rules of thumb $2\sqrt{N}/3$ for the AR and at $\sqrt{N}/3$ for the ARMA model. This choice is not very influential, as AIC minimization typically

implies low lag orders in most replications.

Although it is of little relevance for our focus, we conjecture that the null hypothesis of the DM test may hold. The model is stationary and thus admits a Wold representation that in turn may be approximated to an arbitrary precision by ARMA models and, provided the Wold representation does not come 'close' to moving-average unit roots, also by AR models. Thus, in population both models entail the same predictive accuracy, assuming the mentioned condition is fulfilled, which is difficult to check but is insinuated by the construction of the data-generating process. By contrast, the null of the GW test does not hold, as we demonstrate in our simulations that show a small but quite persistent advantage for the AR model. If we view the GW test as one-sided, however, ARMA can definitely not outperform AR significantly in the sense of the GW test. Of course, these features are of little interest if we focus on the prediction properties of the strategies.

Table 3 gives the resulting values for the mean squared errors. For $N = 100$, the pure AR appears to approximate better than the ARMA model. Choosing the better model on the basis of a pure comparison of performance over the training sample (TS) yields an MSE that is slightly better than forecasting by always using the AR model. This average hides some specific features in single replications. For example, the AR model is preferred on the basis of the training sample in 697 out of 1000 replications, while in the remaining 303 cases the ARMA model can be substantially better. Applying the DM test in order to revise the comparison reduces the cases of selecting ARMA from 303 to 58. This TS-DM-N strategy incurs a further slight improvement in accuracy.

Other strategies deserve being mentioned. For example, always staying with the AR model dominates always choosing the ARMA model in a mere 52% of the cases, in line with the only moderate improvement by the TS–based choice. Using the Giacomini-White test on top of the TS choice turns out to be nearly equivalent to the pure AR strategy, as the ARMA forecast is significantly better than the AR forecast in only 3 out of 1000 replications. In these 3 cases, ARMA forecasting is a lot better than AR forecasting, thus the GW–based choice improves upon the pure AR forecast. With $N = 200$ observations, GW rejection becomes more frequent, but curiously enough GW–based forecasting improves upon pure AR for single-step predictions only.

Table 3: Results of the SETAR experiment: one-step forecasts.

	MSE$\times 10^{-4}$		frequency $\succ$	
	$N = 100$	$N = 200$	$N = 100$	$N = 200$
AR	1.115	1.037	0.518	0.479
ARMA	1.133	1.044	0.482	0.521
TS	1.113	1.041	0.123	0.118
TS-DM-N	1.112	1.038	0.122	0.106
TS-DM-B	1.123	1.040		
TS-F-B	1.124	1.038		
AIC	1.131	1.041		
GW	1.114	1.040		

Notes: 'frequency $\succ$' gives the empirical frequency of the model yielding the better prediction for the observation at $t = N$.

Determining the DM significance points by time-consuming bootstrap yields a classification comparable to pure TS, with some 35% of the replications choosing the ARMA model. Performance is also close to TS. Also AIC generates a comparable probability of preference among rival models, but it tends to select trajectories less efficiently, such that the MSE increases relative to TS or TS-DM. AIC is no panacea, and it may be dominated by smarter selection methods, if none of the rivals is based on correct specification. Modifications of AIC in the presence of misspecification were considered, e.g., by Reschenhofer (1999).

Note that Table 3 provides 'percentage better' for pairwise comparisons only. AR and ARMA forecasts can be compared by frequencies, and so can TS and TS-DM-N, where results are identical for around 75% of all trajectories. The remaining procedures TS-DM-B, AIC, and GW are also defeated in such pairwise comparisons with TS.

When the sample size increases to $N = 200$, the effect in favor of DM testing weakens. Both test-based approaches are beaten by the pure AR model. There is still a slight advantage for the DM–based search. The frequency of significant rejections decreases slightly

to 3.5%. Even in these cases do the ARMA models offer no systematic improvement of forecasting accuracy. This result is in keeping with the previous experiment, where the beneficial effect of a flanking test weakens in larger samples.

For distributions with high variance, MSE may not be the most reliable evaluation criterion. When the cases of improvement among the replications are counted, even the slight advantage for test-based selection is turned on its head. At $N = 200$, in 118 cases is the pure training-sample comparison better, while there are only 106 cases with the opposite ranking. By construction, the forecasts are identical for the remaining 776 cases. At $N = 100$, wins and losses are fairly identical: the test-based procedure wins 123 times, and the comparison without flanking test 122 times. Application of the DM test helps as much as tossing a coin.

It is interesting that a similar remark holds, however, with respect to the ranking among the AR and ARMA forecasts. For $N = 200$, the ARMA model forecasts better in 521 out of 1000 cases, even though it yields the larger MSE. Note that the strong preference for the AR model by the training samples is based on an MSE comparison. Counting cases would yield a different selection. With the smaller sample of $N = 100$, support for the AR model is more unanimous. It yields the smaller MSE as well as the better head count, though with a comparatively small preponderance of 518 cases.

Particularly in this experiment, we also considered different specifications for the relative length of training and test sets. The empirical literature often uses shorter test sets, and we accordingly reduced them from 50% to 25% of the data. For $N = 100$, this indeed induces a slight improvement in predictive accuracy, with a stronger effect on the method without additional DM test. For $N = 200$, this variant entails no change in MSE. Again, selection without DM testing wins with regard to the count of cases. These rather ambiguous effects of shortening the test sample are a bit surprising, as the simulation design involves switches among regimes with locally linear behavior, such that a shorter test set increases the chance that the whole set remains within a regime, which may benefit prediction. Our general impression is that there is little motivation for working with short test sets. This impression is confirmed by some unreported simulation variants for the other experimental designs.

Table 4: Results of the SETAR experiment: two-step forecasts.

	MSE$\times 10^{-4}$		frequency $\succ$	
	$N = 100$	$N = 200$	$N = 100$	$N = 200$
AR	1.248	1.191	0.513	0.478
ARMA	1.285	1.198	0.487	0.522
TS	1.258	1.189	0.131	0.115
TS-DM-N	1.247	1.192	0.122	0.108
TS-DM-B	1.259	1.193		
TS-F-B	1.263	1.197		
AIC	1.291	1.195		
GW	1.247	1.191		

Notes: 'frequency $\succ$' gives the empirical frequency of the model yielding the better prediction for the observation at $t = N$.

Similarly, we also considered changing the significance level for the DM test to 10%. This implies that more cases of improved MSE become significant and that the procedure approaches pure selection. Indeed, this helps in improving average MSE for $N = 100$, while there is no change for $N = 200$ relative to the 5% procedure.

In particular for nonlinear models, larger forecast horizons may also be of interest. Table 4 displays the results for two-step forecasts. The DM-based procedure shows some merits for the smaller sample $N = 100$, while it fails to improve the results for $N = 200$. For $N = 200$, the pure comparison yields a slightly lower MSE than the AR forecast, which indicates that it successfully singles out trajectories that benefit from using the ARMA model. The count of cases favors skipping the DM–testing step in all variants.

Again, we re-ran this experiment with significance levels other than 5%, though we do not report detailed results. Tuning the DM decision to a looser significance level implies a slight deterioration for $N = 200$.

4.3 A variable in a macroeconomic core VAR

Our first two experiments target an exhaustive exploration of the admissible parameter space. There is no indication which designs are close to empirical economic situations. By contrast, the SETAR experiment and this last experiment are inspired by dynamic patterns that appear in actual economic data. In this last design, data are generated from a small vector autoregression. VAR models typically imply univariate ARMA models for their components (e.g., see Lütkepohl, 2005). To these ARMA data, AR and ARMA models are fitted using information criteria. This results in the situation of incomplete nesting or overlapping in the terminology of Vuong (1989). In order to attain a good representativeness of economic data, we adopt a design from an empirical forecasting project by Costantini and Kunst (2011).

Costantini and Kunst (2011) fit vector autoregressions (VAR) to three-variable macroeconomic core sets for the French and U.K. economies. From their sets, we select the British VAR as a generating mechanism and focus on the rate of price inflation among its components. Our choice has been guided by the dynamic dependence structures of the components, which turned out to be strongest and thus most interesting for the inflation series.

Table 5 shows that the AR forecasts are better than the ARMA forecasts at both $N = 100$ and $N = 200$. We note that the ARMA forecasts are not necessarily based on the true model, as the AIC lag selection tends to find lower orders than the theoretically correct ARMA model class. If the pure training-sample comparison is used, approximately two out of three replications favor the simpler AR model at $N = 200$, and on average the MSE is in between the smaller AR and the larger ARMA numbers. Subjecting this decision to a DM test on the basis of normal 95% quantiles leads to a very conservative procedure that chooses AR in 93% ($N = 100$) to 99% ($N = 200$) of all cases. This implies a small gain in precision for $N = 100$, where the procedure is too conservative, while for $N = 200$ it implies a value close to the AR minimum. Using the Giacomini-White test instead yields an even more conservative decision (98% for $N = 100$ and 95% for $N = 200$) that turns out to be optimal here, as always using the AR model would be the dominant strategy. Basing the choice between AR and ARMA on AIC instead implies good performance for $N = 100$

Table 5: Results of the core VAR experiment.

	MSE		frequency $\succ$	
	$N = 100$	$N = 200$	$N = 100$	$N = 200$
AR	0.174	0.183	0.55	0.55
ARMA	0.197	0.206	0.45	0.45
TS	0.183	0.191		
TS-F	0.182			
DM-N	0.181	0.183	0.25:0.22	0.18:0.15
GW	0.175	0.183	0.34:0.34	0.27:0.24
AIC	0.177	0.190	0.27:0.25	0.18:0.17

Notes: 'frequency $\succ$' gives the empirical frequency of the model yielding a better prediction versus TS for the observation at $t = N$.

but a not quite so strong showing for $N = 200$. AIC selects the simpler AR model in 83% ($N = 100$) to 88% ($N = 200$) of all cases.

This experiment was also conducted for larger prediction horizons. The results are generally in line with the reported case of one-step forecasts.

5 Summary and conclusion

Our work was inspired by concerns that the widespread usage of predictive-ability tests may entail an unhealthy preference toward simple prediction models that are dominated by better models though not significantly according to test results. Our simulations have confirmed that such concerns may be well-founded if forecasters actually base their selection on test outcomes.

We view our first design as the most important one, although literally correct specifications may not be common in forecasting applications. If the generating model is ARMA(1,1), it is indeed profitable to use AR(1) for prediction if the coefficient parameters are unknown, at least for a sizable portion of the parameter space. Nonetheless, flanking a training-sample comparison by a Diebold-Mariano test results in an excessively

conservative selection, as long as the incorrect N(0,1) significance points are used. If the distribution is bootstrapped, selection improves, but a simple AIC evaluation serves the same purpose and is much less time-consuming. The Giacomini-White test, although we concede that it addresses the main issue of interest, is too conservative in this and in all other experiments.

In our second design, data are generated from an ARMA(2,2) model, and AR(2) and ARMA(1,1) are considered as forecasting devices. This situation of underspecification by all rivals may be relevant in applications. Here, it evolves that indeed does the DM test often help in boosting the more resilient AR(2) model in many specifications, but bootstrapping the DM distribution becomes counter-productive. We note that in practice it is never known whether models are correctly specified, thus the observation that bootstrapping an incorrect model implies bad model selection may also be relevant.

The third design uses a mildly nonlinear generation process that may be quite realistic, and considers AIC–fitted AR and ARMA models as prediction tools. There is actually not much to choose between the two classes, although there is some preference for AR prediction in smaller samples. Model selection based on a training sample hardly improves by DM testing with normal significance points, and deteriorates by bootstrapped DM testing, which again points to the danger of bootstrapping misspecified models. AIC performs poorly, suggesting that it is no panacea if none of the rivals is optimally specified.

In a fourth design, we use a VAR with coefficients fitted to macroeconomic data for the United Kingdom and we focus on predicting the component with the strongest time dependence structure, the rate of inflation. In a VAR(2), components follow 'marginal' univariate ARMA models, so the design resembles the second experiment. However, in this experiment we entertained AR and ARMA prediction models guided by an AIC search. Then, the DM step implies a deterioration of prediction accuracy in all considered variants.

Our general impression from the prediction experiments is that adding a significance test to a selection of prediction models guided by a training sample fails to systematically improve predictive accuracy. The evaluation of prediction accuracy of rival models over a substantial part of the available sample is a strong selection tool in itself that hardly needs another significance test to additionally support the simpler model.

References

Chatfield, C. (2002). *Time-series forecasting.* Chapman & Hall.

Clark, T.E. and McCracken, M.W. (2001). Tests of equal forecast accuracy and encompassing for nested models . *Journal of Econometrics 105*, 85–110.

Clark, T.E. and McCracken, M.W. (2005). Evaluating direct multistep forecasts , *Econometric Reviews 24*, 369–404.

Clark, T.E. and McCracken, M.W. (2012). *Advances in Forecast Evaluation.* Working Paper 2011-025B, Federal Research Bank of St. Louis.

Costantini, M. and Kunst, R.M. (2011). Combining forecasts based on multiple encompassing tests in a macroeconomic core system. *Journal of Forecasting 30*, 579–596.

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*, 1-9.

Diebold, F.X. and Mariano, R.S. (1995). Comparing Predictive Accuracy. *Journal of Business and Economic Statistics 13*, 253–263.

Fan, J. and Yao, Q. (2005). *Nonlinear Time Series.* Springer.

Giacomini, R. and White, H. (2006). Tests of Conditional Predictive Ability. *Econometrica 74*, 1545–1578.

Hendry, D.H.F. (1997). The Econometrics of Macroeconomic Forecasting. *The Economic Journal 107*, 1330–1357.

Ing, C.K. (2007). Accumulated prediction errors, information criteria and optimal forecasting for autoregressive time series. *Annals of Statistics 35*, 1238–1277.

Inoue, A. and Kilian, L. (2006). On the selection of forecasting models. *Journal of Econometrics 130*, 273–306.

Kilian, L. (1998) Small-sample confidence intervals for impulse response functions. *Review of Economics and Statistics 80*, 218–230.

Linhart, H. (1988). A test whether two AIC?s differ significantly. *South African Statistical Journal 22*, 153–161.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*, Springer-Verlag.

McQuarrie, A.D.R. and Tsai, C.-L. (1998). *Regression and Time Series Model Selection*. World Scientific.

Pötscher, B.M. (1991). Effects of Model Selection on Inference. *Econometric Theory 7*, 163–185.

Reschenhofer, E. (1999). Improved estimation of the expected Kullback-Leibler discrepancy in case of misspecification. *Econometric Theory 15*, 377–387.

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear model. *Annals of Statistics 8*, 147–164.

Tiao, G.C. and Tsay, R.S. (1994). Some Advances in Non Linear and Adaptive Modelling in Time Series. *Journal of Forecasting 13*, 109–131.

Vuong, Q.H. (1989). Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica 57*, 307–333.

Wei, C.Z. (1992). On predictive least squares principles. *Annals of Statistics 20*, 1–42.

West, K.D. (1996). Asymptotic inference about predictive ability. *Econometrica 64*, 1067–1084