On the effects of the Diebold-Mariano test on the selection of prediction models

Mauro Costantini[1] and Robert M. Kunst[2]

Presented at the
Annual Conference of the Royal Statistical Society,
Newcastle, September 2013

[1]Brunel University London; mauro.costantini@brunel.ac.uk

[2]Institute for Advanced Studies, Vienna, Austria 1060, and University of Vienna; kunst@ihs.ac.at

Historical facts

- Diebold & Mariano (1995) revolutionized the reporting of forecast comparisons (potential priority by Meese and Rogoff, 1988);
- Today, comparative forecast evaluations are often seen as incomplete without DM or other significance tests;
- Since 1995, the literature has concentrated on size/power and distribution of test statistics (West, 1996; Clark and McCracken, 2001; Giacomini and White, 2006; survey Clark and McCracken, 2012) rather than on model selection aspects (Inoue and Kilian, 2006; Ing, 2007).

The Diebold-Mariano test

Diebold & Mariano considered the test statistic

$$S = \frac{\frac{1}{T}\sum_{t=1}^{T}\{g(e_{1t}) - g(e_{2t})\}}{\sqrt{\frac{2\pi \hat{f}_d(0)}{T}}},$$

with $g(e_{jt}), j = 1, 2$ denoting the loss from forecast error e_{jt} evolving from prediction model j.

The null hypothesis tested is $H_0 : \mathrm{E}g(e_{1t}) = \mathrm{E}g(e_{2t})$. Under H_0, S is asymptotically standard normal distributed.

Why test?

Clark and McCracken (2012) opine that

> *"Of the various rationales for forecast evaluation, the intention of evaluating the forecasts to assess the models for their actual value in forecasting should be the least controversial."*

This implies that all this inference aims at finding the best forecast model.

Or a substitute for a restriction test?

Diebold and Mariano (1995) conjecture that

> *"The ability to formally compare predictive accuracy afforded by our tests may prove useful as a model-specification diagnostic, as well as a means to test both nested and non-nested hypotheses under nonstandard conditions."*

The normal null distribution derived by DM is now known not to be valid for nested hypotheses.

Conceptual problems with the DM test

Main conceptual problems:

1. The null hypothesis, first criticized by Chatfield. In real-world applications, two simple prediction models rarely achieve the same loss moment, if the DGP is far more complex;
2. If forecasts are model-based and $H_0 : \mathrm{E}g(e_{1t}) = \mathrm{E}g(e_{2t})$ is seen as the loss achievable in a finite sample, the identity becomes even more implausible (conditional concept, Giacomini and White);
3. An out-of-sample comparison of $g(e_{jt})$ already is equivalent to an information-criterion (IC) evaluation. A test on top of an IC favors the simpler rival model, may correspond to a stronger complexity penalty.

Accuracy comparison: an information criterion

- Wei (1992) considers evaluating out-of-sample prediction expanding over (nearly) the whole sample, shows that this yields a *consistent* (BIC–type) information criterion;
- Inoue and Kilian (2006) consider evaluating over a fixed share of the sample, which yields an *efficient* (prediction-optimizing, AIC–type) information criterion;
- Ing (2007) shows that and how such equivalence properties depend on model assumptions.

Why test on top of an IC?

- If the IC is consistent, an additional test penalizes complexity in smaller samples and does not affect consistency in large samples;
- Such a 'simplicity booster' may assist in improving the (sometimes unsatisfactory) IC properties in small samples;
- Our simulations throw some light on which of the effects predominate in typical applications.

Experiment I: the concept

- Data are generated from ARMA(1,1) models $X_t = \phi X_{t-1} + \varepsilon_t - \theta \varepsilon_{t-1}$, with $\phi = j * 0.2, \theta = k * 0.2, \quad j, k \in \{-4, -3, \ldots, 4\}$.
- An AR(1) and an ARMA(1,1) model are fitted to the data, out-of-sample predictions are based on each of the two.
- The winner over the training sample (later 50% of the data) is evaluated and predicts the last observation.
- This winner prediction is compared to the forecast based on: ARMA(1,1) if 'significantly' (5%) better than the AR(1) 'benchmark', AR(1) otherwise.

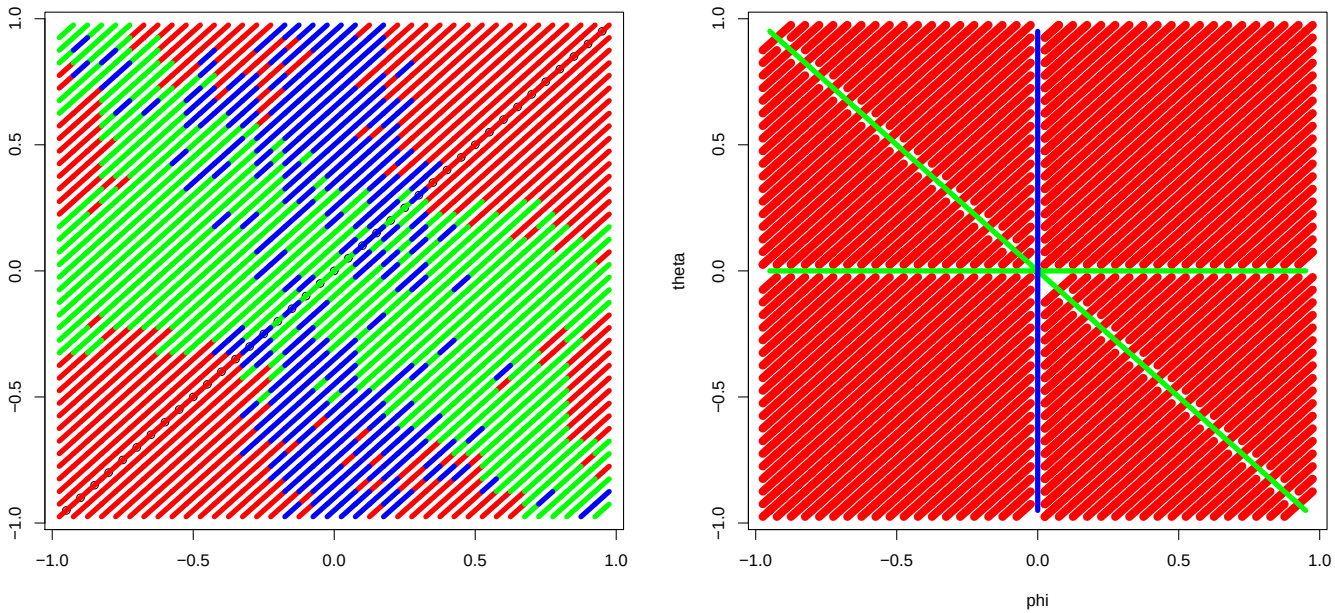
Bootstrap in bootstrap, b-in-b

Two possible assumptions for the experiment:

1. The hypothetical forecaster uses the (here, incorrect) asymptotic normal distribution for the significance decision;
2. The forecaster generates a bootstrap distribution for her (simulated) data.

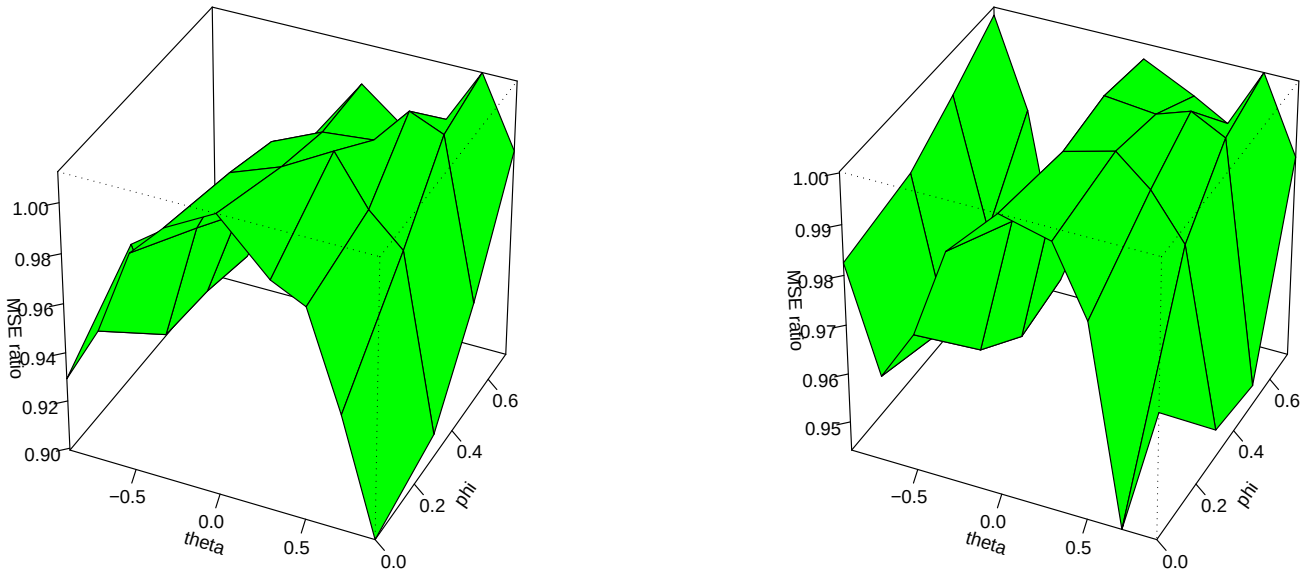
The latter option has to confront the known small-sample bias of time-series estimates: first bootstrap to correct for the bias, then another bootstrap to get the distribution of the DM statistic. The method follows Busetti *et al.* (2009) and Kilian (1998).

Forecasting ARMA models: known patterns (Greenland)



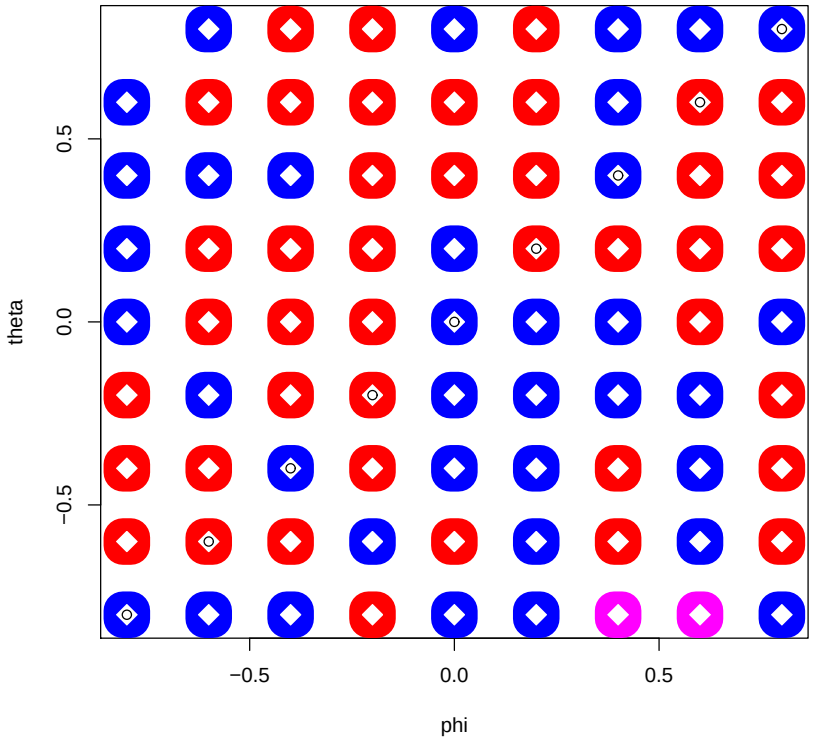
Generating model is ARMA(1,1) $x_t = \phi x_{t-1} + \varepsilon_t - \theta \varepsilon_{t-1}$. Green area prefers AR(1) forecast, red area prefers ARMA(1,1) forecast, blue area prefers MA(1) forecast. Left figure $N = 50$, right figure true structures.

Graphical summary for experiment I, naive distribution



MSE ratio AR or ARMA model selected by training sample divided by selected model after DM testing. $N = 100$ (left) and $N = 200$ (right).

Graphical summary for experiment I, b-in-b



Preferences (lower MSE) for direct training-sample comparison (red) and for test-based comparison (blue). Ties in magenta. $N = 100$.

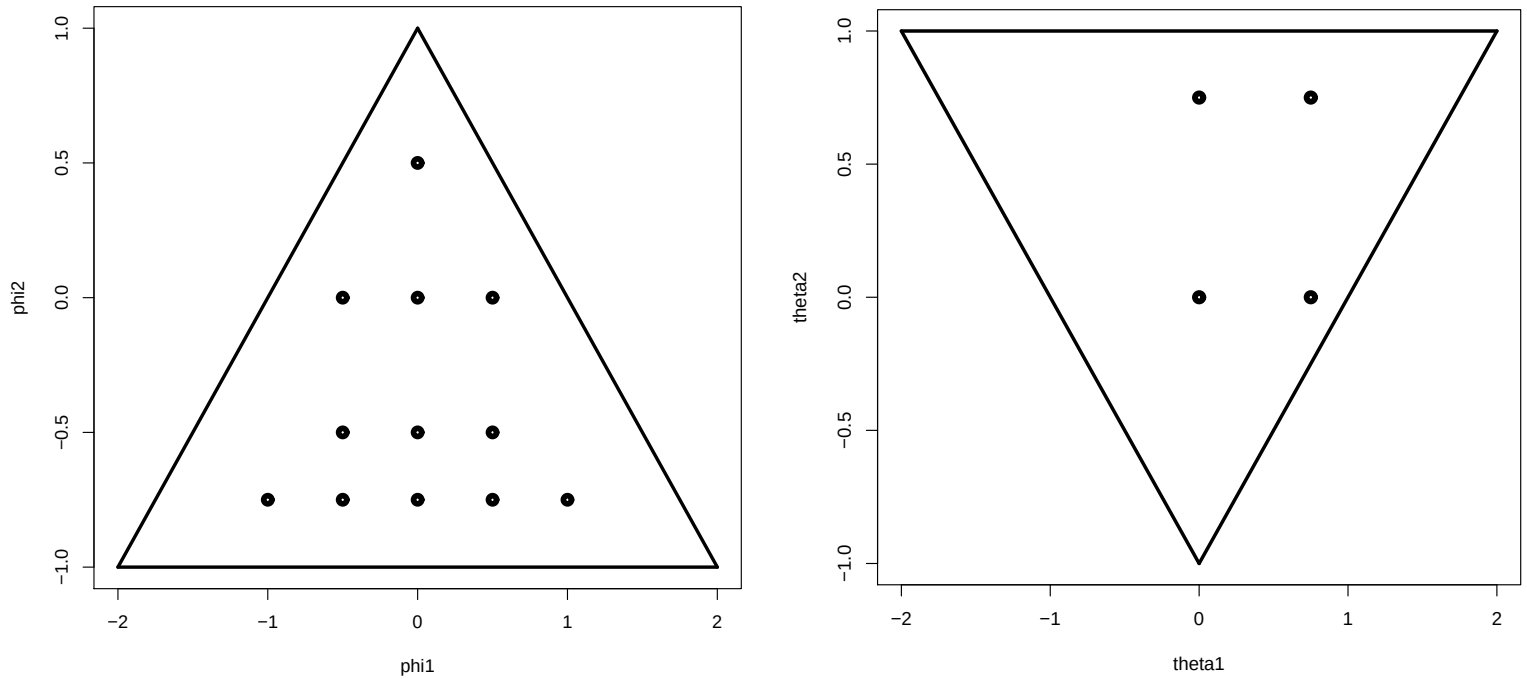
Summary of results from experiment I

- Basing the decision on the incorrect normal distribution makes little sense. The infrequent rejection of the AR null does not benefit the forecasts. This option corresponds to testing at a 1% level;
- Basing the decision on the bootstrap benefits the forecasts occasionally but not systematically. The DM simplicity boost corresponds to an increased penalty term in an information criterion. This option corresponds to testing at a 5% level;
- The training-sample comparison corresponds to testing at a 25-30% level;
- Inoue and Kilian (2006) demonstrate that BIC–guided selection often outperforms training-sample comparisons. The DM simplicity boost may assist in tuning the implicit information criterion closer to BIC.

Experiment II: the concept

- Data are generated from ARMA(2,2) models.
- An AR(2) and an ARMA(1,1) model are fitted to the data, out-of-sample predictions are based on each of the two.
- The winner over the training sample (later 50% of the data) is evaluated and predicts the last observation.
- This winner prediction is compared to the forecast based on: ARMA(1,1) if 'significantly' better than the AR(2) 'benchmark', the AR(2) otherwise.

ARMA(2,2) parameters



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.

Results of the simulation for $N = 100$

ϕ_1	ϕ_2	θ_1	θ_2	MSE(ARMA)	MSE(AR)	MSE(tr)	MSE(DM)
0	0.5	0	0	[illegible]	[illegible]	[illegible]	[illegible]
-0.5	0	0	0	[illegible]	[illegible]	0.995	0.995
0	0	0	0	[illegible]	[illegible]	[illegible]	[illegible]
0.5	0	0	0	1.047	1.047	[illegible]	[illegible]
-0.5	-0.5	0	0	[illegible]	[illegible]	[illegible]	[illegible]
0	-0.5	0	0	[illegible]	[illegible]	[illegible]	[illegible]
0.5	-0.5	0	0	[illegible]	[illegible]	[illegible]	[illegible]
-1	-0.75	0	0	[illegible]	[illegible]	[illegible]	[illegible]
-0.5	-0.75	0	0	[illegible]	[illegible]	[illegible]	[illegible]
0	-0.75	0	0	[illegible]	[illegible]	[illegible]	[illegible]
0.5	-0.75	0	0	[illegible]	[illegible]	[illegible]	[illegible]
1	-0.75	0	0	[illegible]	[illegible]	[illegible]	[illegible]
0	0.5	0	0.75	[illegible]	[illegible]	1.318	1.318
-0.5	0	0	0.75	[illegible]	[illegible]	[illegible]	[illegible]
0	0	0	0.75	[illegible]	[illegible]	1.166	1.166
0.5	0	0	0.75	[illegible]	[illegible]	[illegible]	[illegible]
-0.5	-0.5	0	0.75	[illegible]	[illegible]	[illegible]	[illegible]
0	-0.5	0	0.75	[illegible]	[illegible]	[illegible]	[illegible]
0.5	-0.5	0	0.75	[illegible]	[illegible]	[illegible]	[illegible]
-1	-0.75	0	0.75	[illegible]	[illegible]	1.366	1.366
-0.5	-0.75	0	0.75	[illegible]	[illegible]	[illegible]	[illegible]
0	-0.75	0	0.75	[illegible]	[illegible]	[illegible]	[illegible]
0.5	-0.75	0	0.75	[illegible]	[illegible]	[illegible]	[illegible]
1	-0.75	0	0.75	[illegible]	[illegible]	[illegible]	[illegible]
0	0.5	0.75	0	[illegible]	[illegible]	[illegible]	[illegible]
-0.5	0	0.75	0	[illegible]	[illegible]	[illegible]	[illegible]
0	0	0.75	0	[illegible]	[illegible]	[illegible]	[illegible]
0.5	0	0.75	0	[illegible]	[illegible]	[illegible]	[illegible]
-0.5	-0.5	0.75	0	[illegible]	[illegible]	[illegible]	[illegible]
0	-0.5	0.75	0	[illegible]	[illegible]	[illegible]	[illegible]
0.5	-0.5	0.75	0	[illegible]	[illegible]	[illegible]	[illegible]
-1	-0.75	0.75	0	[illegible]	[illegible]	[illegible]	[illegible]
-0.5	-0.75	0.75	0	[illegible]	[illegible]	[illegible]	[illegible]
0	-0.75	0.75	0	[illegible]	[illegible]	[illegible]	[illegible]
0.5	-0.75	0.75	0	[illegible]	[illegible]	[illegible]	[illegible]
1	-0.75	0.75	0	[illegible]	[illegible]	[illegible]	[illegible]
0	0.5	0.75	0.75	[illegible]	[illegible]	[illegible]	[illegible]
-0.5	0	0.75	0.75	[illegible]	[illegible]	[illegible]	[illegible]
0	0	0.75	0.75	[illegible]	[illegible]	[illegible]	[illegible]
0.5	0	0.75	0.75	[illegible]	[illegible]	[illegible]	[illegible]
-0.5	-0.5	0.75	0.75	[illegible]	[illegible]	[illegible]	[illegible]
0	-0.5	0.75	0.75	[illegible]	[illegible]	[illegible]	[illegible]
0.5	-0.5	0.75	0.75	[illegible]	[illegible]	[illegible]	[illegible]
-1	-0.75	0.75	0.75	[illegible]	[illegible]	[illegible]	[illegible]
-0.5	-0.75	0.75	0.75	[illegible]	[illegible]	[illegible]	[illegible]
0	-0.75	0.75	0.75	[illegible]	[illegible]	[illegible]	[illegible]
0.5	-0.75	0.75	0.75	[illegible]	[illegible]	[illegible]	[illegible]
1	-0.75	0.75	0.75	[illegible]	[illegible]	[illegible]	[illegible]

Results of the simulation for $N = 200$

ϕ_1	ϕ_2	θ_1	θ_2	MSE(ARMA)	MSE(AR)	MSE(tr)	MSE(DM)
0	0.5	0	0	[illegible]	[illegible]	[illegible]	[illegible]
-0.5	0	0	0	[illegible]	[illegible]	0.986	0.986
0	0	0	0	[illegible]	[illegible]	[illegible]	[illegible]
0.5	0	0	0	1.024	1.024	[illegible]	[illegible]
-0.5	-0.5	0	0	[illegible]	[illegible]	[illegible]	[illegible]
0	-0.5	0	0	[illegible]	[illegible]	0.995	0.995
0.5	-0.5	0	0	[illegible]	[illegible]	[illegible]	[illegible]
-1	-0.75	0	0	[illegible]	[illegible]	[illegible]	[illegible]
-0.5	-0.75	0	0	[illegible]	[illegible]	0.993	0.993
0	-0.75	0	0	[illegible]	[illegible]	0.994	0.994
0.5	-0.75	0	0	[illegible]	[illegible]	0.993	0.993
1	-0.75	0	0	[illegible]	[illegible]	[illegible]	[illegible]
0	0.5	0	0.75	[illegible]	[illegible]	1.344	1.344
-0.5	0	0	0.75	[illegible]	[illegible]	[illegible]	[illegible]
0	0	0	0.75	[illegible]	[illegible]	[illegible]	[illegible]
0.5	0	0	0.75	[illegible]	[illegible]	[illegible]	[illegible]
-0.5	-0.5	0	0.75	[illegible]	[illegible]	[illegible]	[illegible]
0	-0.5	0	0.75	[illegible]	[illegible]	[illegible]	[illegible]
0.5	-0.5	0	0.75	[illegible]	[illegible]	[illegible]	[illegible]
-1	-0.75	0	0.75	[illegible]	[illegible]	1.408	1.408
-0.5	-0.75	0	0.75	[illegible]	[illegible]	[illegible]	[illegible]
0	-0.75	0	0.75	[illegible]	[illegible]	[illegible]	[illegible]
0.5	-0.75	0	0.75	[illegible]	[illegible]	[illegible]	[illegible]
1	-0.75	0	0.75	[illegible]	[illegible]	[illegible]	[illegible]
0	0.5	0.75	0	1.027	1.027	0.987	0.987
-0.5	0	0.75	0	[illegible]	[illegible]	[illegible]	[illegible]
0	0	0.75	0	[illegible]	[illegible]	[illegible]	[illegible]
0.5	0	0.75	0	[illegible]	[illegible]	[illegible]	[illegible]
-0.5	-0.5	0.75	0	[illegible]	[illegible]	[illegible]	[illegible]
0	-0.5	0.75	0	[illegible]	[illegible]	[illegible]	[illegible]
0.5	-0.5	0.75	0	[illegible]	[illegible]	[illegible]	[illegible]
-1	-0.75	0.75	0	[illegible]	[illegible]	1.309	1.309
-0.5	-0.75	0.75	0	[illegible]	[illegible]	1.362	1.362
0	-0.75	0.75	0	[illegible]	[illegible]	[illegible]	[illegible]
0.5	-0.75	0.75	0	[illegible]	[illegible]	[illegible]	[illegible]
1	-0.75	0.75	0	[illegible]	[illegible]	[illegible]	[illegible]
0	0.5	0.75	0.75	[illegible]	[illegible]	[illegible]	[illegible]
-0.5	0	0.75	0.75	[illegible]	[illegible]	[illegible]	[illegible]
0	0	0.75	0.75	[illegible]	[illegible]	[illegible]	[illegible]
0.5	0	0.75	0.75	[illegible]	[illegible]	[illegible]	[illegible]
-0.5	-0.5	0.75	0.75	[illegible]	[illegible]	[illegible]	[illegible]
0	-0.5	0.75	0.75	[illegible]	[illegible]	[illegible]	[illegible]
0.5	-0.5	0.75	0.75	[illegible]	[illegible]	[illegible]	[illegible]
-1	-0.75	0.75	0.75	[illegible]	[illegible]	[illegible]	[illegible]
-0.5	-0.75	0.75	0.75	[illegible]	[illegible]	[illegible]	[illegible]
0	-0.75	0.75	0.75	[illegible]	[illegible]	[illegible]	[illegible]
0.5	-0.75	0.75	0.75	[illegible]	[illegible]	[illegible]	[illegible]
1	-0.75	0.75	0.75	[illegible]	[illegible]	1.695	1.695

Summary of results from experiment II

- For most designs, forecasting performance improves if the DM test is applied on top of the training-sample evaluation;
- The benefits from using the DM test decrease as the sample size increases;
- Both models have identical parameter dimension. ARMA(1,1) could be chosen as the benchmark null. Then, the DM test incurs a deterioration of performance;
- The AR(2) model forecasts better due to (a) better approximation to the DGP and (b) numerically better estimation. It makes sense to view AR(2) as the benchmark.

Experiment III: the concept

- Data are generated from a SETAR model.
- AR(p) and ARMA(q, q) models are fitted to the data, with p and q determined by AIC. Out-of-sample predictions are based on each of the two.
- The winner over the training sample (25% and 50% of the data) is evaluated and predicts the last observation.
- This winner prediction is compared to the forecast based on: ARMA(q, q) if 'significantly' better than the AR(p) 'benchmark', the AR(p) otherwise.

The SETAR model used as the DGP

A SETAR model has been suggested by Tiao and Tsay (1994) for the growth rate of U.S. GNP:

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

Standard deviations of errors are $\sigma_1 = 0.0062$, $\sigma_2 = 0.0132$, $\sigma_3 = 0.0094$, and $\sigma_4 = 0.0082$.

Results of the simulations for experiment III

	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
50% training				
lower MSE	1.113	1.041	0.123	0.118
DM–based	1.112	1.038	0.122	0.106
25% training				
lower MSE	1.106	1.042	0.144	0.144
DM–based	1.114	1.035	0.127	0.137

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

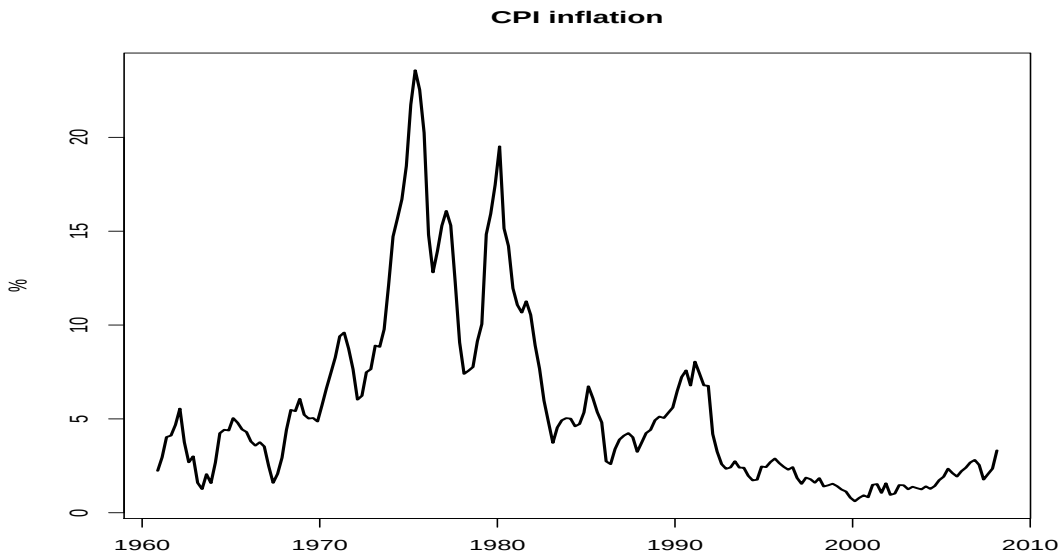
Summary of results from experiment III

- Performance across replications is quite heterogeneous due to the highly nonlinear DGP;
- It appears that DM testing benefits the MSE ranking that may not be a good criterion here;
- By contrast, pure training-sample evaluation appears to be preferable with regard to the probability of achieving the better prediction.

Experiment IV: the concept

- Data are generated as a component from a trivariate VAR model. The VAR is tuned to U.K. macroeconomic data;
- AR(p) and ARMA(q, q) models are fitted to the data, with p and q determined by AIC. Out-of-sample predictions are based on each of the two;
- The winner over the training sample (50% of the data) is evaluated and predicts the last observation;
- This winner prediction is compared to the forecast based on: ARMA(q, q) if 'significantly' better than the AR(p) 'benchmark', the AR(p) otherwise.

Experiment IV: some details



A VAR(2) is fitted to a system that comprises U.K. GDP growth, an interest rate, and CPI inflation, and the fitted VAR is generated with Gaussian errors. CPI inflation is forecasted. Its implied generating model is ARMA(2,2), thus the generating model is contained in the prediction toolbox.

Results of the simulations for experiment IV

	MSE		frequency $\succ$	
	$N = 100$	$N = 200$	$N = 100$	$N = 200$
AR	0.189	0.183	0.48	0.50
ARMA	0.179	0.180	0.52	0.50
50% training				
lower MSE	0.178	0.180	0.30	0.27
DM–based	0.187	0.182	0.26	0.27

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

Summary of results from experiment IV

- Forecasting performance deteriorates if the DM test is applied on top of the training-sample evaluation;
- The pure training-sample selection shows good performance;
- Usage of the DM test implies failure to reject the AR benchmark in 3/4 of the cases for $N = 100$.

General summary

- There are no systematic benefits from 'double testing';
- Double testing using the DM test may be beneficial if the benchmark has better prediction properties but this is trivial;
- Double testing may give undue support to a simple benchmark model and lead to ignoring the benefits from using more sophistication;
- Extensions to larger forecasting horizons have also been studied. They are generally in line with the single-step patterns.

References I

Busetti, F., J. Marcucci, and G. Veronese (2009) 'Comparing forecast accuracy: a Monte Carlo investigation,' Banca d'Italia working paper.

Chatfield, C. (2001) *Time-Series Forecasting,* Chapman & Hall.

Clark, T.E., and M.W. McCracken (2001) 'Tests of equal forecast accuracy and encompassing for nested models,' *Journal of Econometrics* **105**, 85–110.

Clark, T.E., and M.W. McCracken (2012) 'Advances in forecast evaluation,' St. Louis working paper.

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

Giacomini, R., and H. White (2006) 'Tests of conditional predictive ability,' *Econometrics* **74**, 1545–1578.

References II

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 L. Kilian (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.

Meese, R., and K.S. Rogoff (1988) 'Was it real? The exchange-rate interest differential relation over the modern floating-rate period,' *Journal of Finance* **43**, 933–948.

Tiao, G.C., and R.S. Tsay (1994) 'Some advances in non-linear and adaptive modelling in time series,' *Journal of Forecasting* **13**, 109–131.

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.

Thank you for your attention