On the usefulness of the Diebold-Mariano test in the selection of prediction models: Some Monte Carlo evidence

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

Presented at Österreichische Statistiktage,
Graz, September 2011

[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 revolutionized reporting forecast comparisons;
- Today most editors/referees demand DM or similar tests on top of every comparative forecast evaluation;
- Forecast procedures/models that beat the benchmark only 'insignificantly' are regarded as uninteresting: widespread preference for simple benchmark structures?

The Diebold-Mariano test

Diebold & Mariano (1995, JBES) 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.

Conceptual problems with the DM test

Two 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. 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.

Using the DM test for model selection

The main purpose of the DM test is not model selection. However, Diebold and Mariano 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."*

Later, doubts were raised on the validity of the null distribution for nested hypotheses.

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 information criterion;
- Inoue and Kilian (2006) consider evaluating over a fixed share of the sample, which yields an efficient information criterion.

Comparative forecast evaluation is a stronger tool than the apparently 'casual manner' referred to by DM.

Testing on top of a consistent IC: good news

Proposition

Suppose there exists a consistent information criterion τ_1 and an independent test-consistent significance test τ_2 at a given significance level α_2. Then, the joint decision from rejecting H_0 if both criteria prefer the alternative is a consistent model selection procedure.

Testing on top of a consistent IC: not so good news

Proposition

Suppose there exists a consistent information criterion τ_1 with implicit significance level $\alpha_1(T)$ at T, and an independent test-consistent significance test τ_2 at significance level α_2. Then, the joint test has critical level $\alpha_1(T)\alpha_2$.

In practice, of course, the two decisions will not be independent.

Testing on top of a consistent IC: bad news

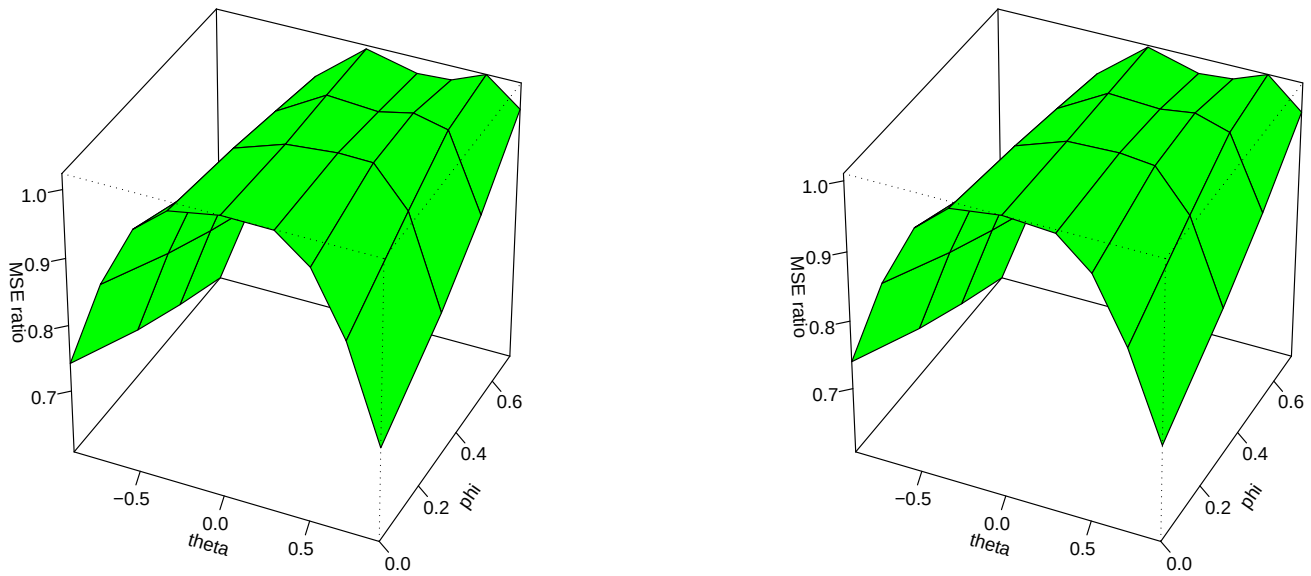
Proposition

Suppose there exists a consistent information criterion τ such that between two models M_1 and M_2 the event $\tau > 0$ indicates a preference for M_2, while $\tau \leq 0$ prefers M_1. Assume the user instead bases her decision on $\tau > \tau_0$ with $\tau_0 > 0$. This decision will be inconsistent in the sense that, as $T \to \infty$, the probability of preferring M_1 although M_2 is true, will not converge to 0.

Experiment I: the concept

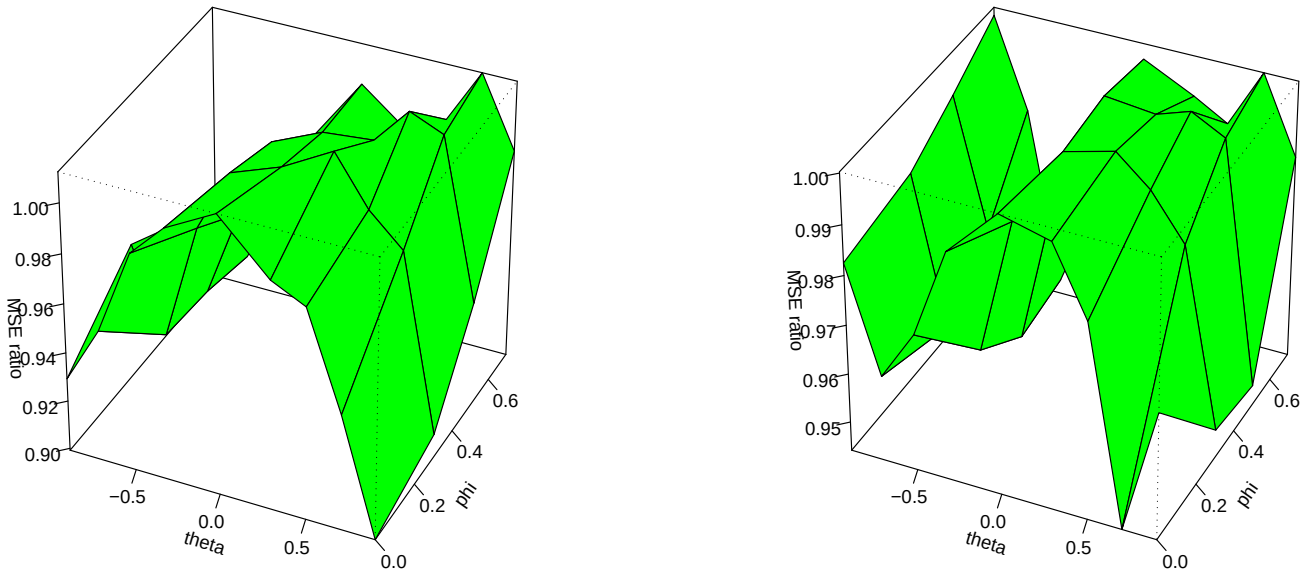
- Data are generated from ARMA(1,1) models $X_t = \phi X_{t-1} + \varepsilon_t - \theta \varepsilon_{t-1}$, with $\phi \in \{0, 0.3, 0.5, 0.7\}$ and $\theta \in \{-0.9, -0.7, -0.5, -0.3, 0, 0.3, 0.5, 0.7, 0.9\}$. 1000 replications.
- An AR(1) and an ARMA(1,1) model are fitted to the data, out-of-sample 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.

ARMA versus AR forecast over the training samples



MSE ratio ARMA forecast divided by AR forecast. $N = 100$ and $N = 200$.

Results from experiment I: graphical summary



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

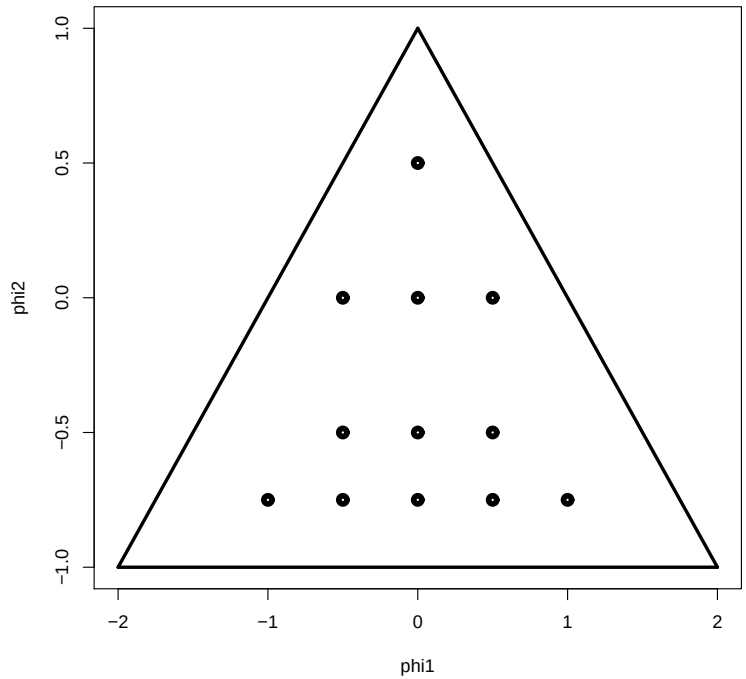
Summary of results from experiment I

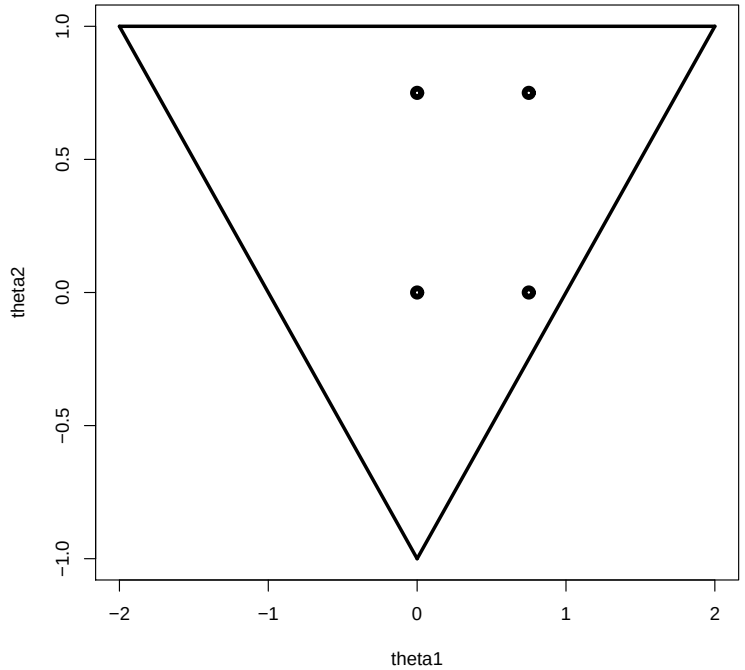
- For most designs, forecasting performance deteriorates if the DM test is applied on top of the training-sample evaluation;
- Even for cases where the AR model is correctly specified, such as $\theta = 0$ and canceling roots, there are no benefits from using the DM test in selecting the prediction model;
- The deleterious influence of the DM test shrinks as the sample size increases.

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	1.2941	1.0381	1.0381	1.0381
-0.5	0	0	0	1.0041	1.0041	0.995	0.995
0	0	0	0	1.0381	1.0381	1.0381	1.0381
0.5	0	0	0	1.047	1.047	1.0381	1.0381
-0.5	-0.5	0	0	1.2181	1.0381	1.0141	1.0381
0	-0.5	0	0	1.3522	1.0381	1.0141	1.0381
0.5	-0.5	0	0	1.1831	1.0381	1.0201	1.0381
-1	-0.75	0	0	1.6041	1.0381	1.0051	1.0381
-0.5	-0.75	0	0	1.7677	1.0381	1.0101	1.0381
0	-0.75	0	0	1.4151	1.0381	1.0191	1.0381
0.5	-0.75	0	0	1.7431	1.0381	1.0241	1.0381
1	-0.75	0	0	1.5724	1.0381	1.0361	1.0381
0	0.5	0	0.75	1.6531	1.4621	1.318	1.318
-0.5	0	0	0.75	1.2751	1.0621	1.0284	1.0284
0	0	0	0.75	1.5281	1.3081	1.166	1.166
0.5	0	0	0.75	1.4741	1.0381	1.0284	1.0284
-0.5	-0.5	0	0.75	1.4581	1.2481	1.1741	1.1761
0	-0.5	0	0.75	1.5171	1.0381	1.0241	1.0241
0.5	-0.5	0	0.75	1.4581	1.2481	1.1681	1.1674
-1	-0.75	0	0.75	2.0221	1.4681	1.366	1.366
-0.5	-0.75	0	0.75	1.8451	1.2681	1.1841	1.1841
0	-0.75	0	0.75	1.8531	1.2681	1.0981	1.0981
0.5	-0.75	0	0.75	1.8211	1.2681	1.1941	1.1941
1	-0.75	0	0.75	2.0011	1.4681	1.3921	1.3921
0	0.5	0.75	0	1.0581	1.0381	1.0381	1.0381
-0.5	0	0.75	0	1.0581	1.0701	1.0121	1.0381
0	0	0.75	0	1.0581	1.1401	1.0381	1.0381
0.5	0	0.75	0	1.0581	1.2181	1.0141	1.0381
-0.5	-0.5	0.75	0	1.3501	1.0381	1.2201	1.0381
0	-0.5	0.75	0	1.3581	1.3381	1.2381	1.2041
0.5	-0.5	0.75	0	1.3341	1.3701	1.2381	1.2381
-1	-0.75	0.75	0	1.8221	1.4381	1.2401	1.2401
-0.5	-0.75	0.75	0	1.5141	1.4381	1.3141	1.3381
0	-0.75	0.75	0	1.8341	1.4381	1.4741	1.4381
0.5	-0.75	0.75	0	1.5141	1.4381	1.4161	1.4381
1	-0.75	0.75	0	1.8141	1.4381	1.4401	1.4381
0	0.5	0.75	0.75	1.0381	1.0381	1.0381	1.0381
-0.5	0	0.75	0.75	1.0381	1.0381	1.0381	1.0381
0	0	0.75	0.75	1.0381	1.1381	1.0381	1.0381
0.5	0	0.75	0.75	1.0381	1.2381	1.0141	1.0381
-0.5	-0.5	0.75	0.75	1.0381	1.0381	1.0381	1.0381
0	-0.5	0.75	0.75	1.3161	1.0381	1.1381	1.0381
0.5	-0.5	0.75	0.75	1.6381	1.0381	1.3381	1.0381
-1	-0.75	0.75	0.75	1.0381	1.0381	1.0381	1.0381
-0.5	-0.75	0.75	0.75	1.1381	1.0381	1.0741	1.0381
0	-0.75	0.75	0.75	1.0381	1.0381	1.0381	1.0381
0.5	-0.75	0.75	0.75	1.4141	1.0381	1.2381	1.0381
1	-0.75	0.75	0.75	1.0381	1.0381	1.0381	1.0381

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	1.0249	1.0249	1.0249	1.0249
-0.5	0	0	0	1.0026	1.0026	0.986	0.986
0	0	0	0	1.0027	1.0027	0.9861	0.9861
0.5	0	0	0	1.024	1.024	1.0064	1.0064
-0.5	-0.5	0	0	1.1741	1.0027	1.0027	1.0027
0	-0.5	0	0	1.3033	1.0026	0.995	0.995
0.5	-0.5	0	0	1.1641	1.0027	1.0027	1.0027
-1	-0.75	0	0	1.5351	1.0026	0.9901	0.9901
-0.5	-0.75	0	0	1.7101	1.0027	0.993	0.993
0	-0.75	0	0	1.6501	1.0026	0.994	0.994
0.5	-0.75	0	0	1.6331	1.0027	0.993	0.993
1	-0.75	0	0	1.5301	1.0026	0.9901	0.9901
0	0.5	0	0.75	1.6601	1.0026	1.344	1.344
-0.5	0	0	0.75	1.4501	1.0027	1.2831	1.2831
0	0	0	0.75	1.5089	1.0027	1.1641	1.1641
0.5	0	0	0.75	1.4531	1.0026	1.0341	1.0341
-0.5	-0.5	0	0.75	1.7831	1.2341	1.1831	1.1831
0	-0.5	0	0.75	1.6841	1.0027	1.0131	1.0131
0.5	-0.5	0	0.75	1.7831	1.2341	1.1801	1.1801
-1	-0.75	0	0.75	1.9571	1.0026	1.408	1.408
-0.5	-0.75	0	0.75	1.8101	1.0026	1.1741	1.1741
0	-0.75	0	0.75	1.8027	1.0027	0.9871	0.9871
0.5	-0.75	0	0.75	1.8061	1.0026	1.1497	1.1497
1	-0.75	0	0.75	1.9035	1.0026	1.3761	1.3761
0	0.5	0.75	0	1.027	1.027	0.987	0.987
-0.5	0	0.75	0	1.0026	1.0249	1.0026	1.0026
0	0	0.75	0	1.0027	1.1224	1.0027	1.0749
0.5	0	0.75	0	1.0249	1.1824	1.0249	1.0401
-0.5	-0.5	0.75	0	1.1741	1.0249	1.2431	1.2431
0	-0.5	0.75	0	1.3033	1.1641	1.2401	1.2401
0.5	-0.5	0.75	0	1.1641	1.0249	1.0341	1.0341
-1	-0.75	0.75	0	1.7431	1.0249	1.309	1.309
-0.5	-0.75	0.75	0	1.6444	1.0249	1.362	1.362
0	-0.75	0.75	0	1.7131	1.0249	1.4401	1.4474
0.5	-0.75	0.75	0	1.6641	1.0249	1.4401	1.4401
1	-0.75	0.75	0	1.6041	1.0249	1.5124	1.6601
0	0.5	0.75	0.75	1.9013	1.0249	1.4401	1.7401
-0.5	0	0.75	0.75	1.5261	1.0249	1.2491	1.2491
0	0	0.75	0.75	1.5007	1.1241	1.3401	1.3401
0.5	0	0.75	0.75	1.4241	1.0249	1.2401	1.2431
-0.5	-0.5	0.75	0.75	1.6831	1.0026	1.0026	1.0026
0	-0.5	0.75	0.75	1.2491	1.0249	1.1791	1.1791
0.5	-0.5	0.75	0.75	1.6241	1.0249	1.3631	1.3631
-1	-0.75	0.75	0.75	1.8141	1.0249	1.0401	1.0401
-0.5	-0.75	0.75	0.75	1.8147	1.0249	1.0591	1.0591
0	-0.75	0.75	0.75	1.6241	1.0249	1.3591	1.3591
0.5	-0.75	0.75	0.75	1.3941	1.0249	1.3591	1.3591
1	-0.75	0.75	0.75	1.6401	1.0249	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 $t = N$.

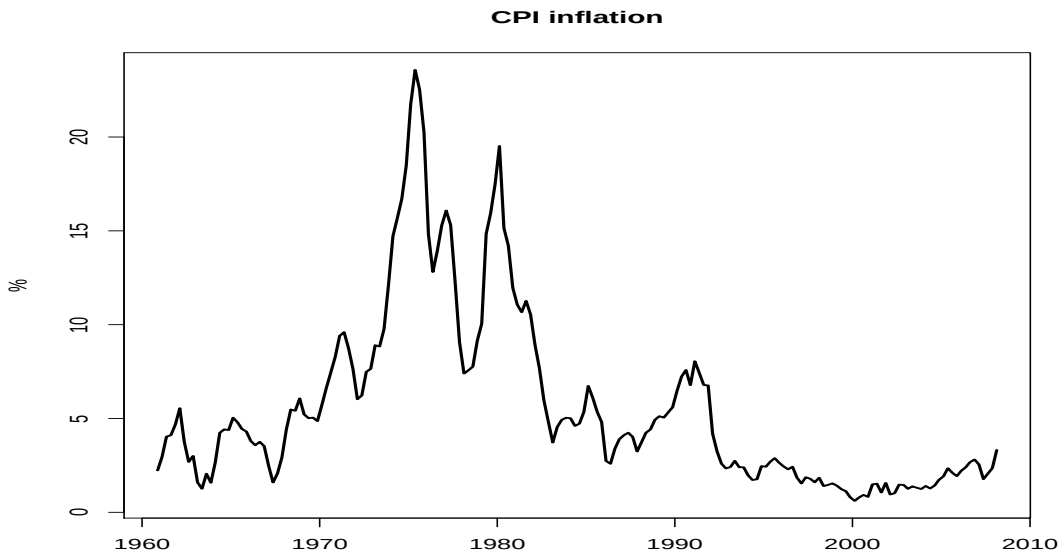
Summary of results from experiment III

- 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	0.191	NYA	0.47	NYA
ARMA	0.180	NYA	0.53	NYA
50% training				
lower MSE	0.179	NYA	0.28	NYA
DM–based	0.194	NYA	0.25	NYA

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 will be studied.

References

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

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

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.

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

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

Thank you for your attention