Discussion of the paper "*Forecast evaluation of small nested model sets*" by Kirstin Hubrich and Kenneth West

Fabio Busetti

Banca d'Italia

19th January 2010

The tests ...

- ... consider m regression models $(i = 1, ..., m)$ each of them nesting a benchmark one $(i = 0)$. The null hypothesis is equal predictive accuracy (EPA) across all models, while the alternative postulates that at least one model has a lower mean square prediction error than the benchmark

$$
\begin{aligned}
H_0 &: \quad \sigma_0^2 = \sigma_1^2 = ... = \sigma_m^2, \\
H_A &: \quad \max\left(\sigma_0^2 - \sigma_1^2, ..., \sigma_0^2 - \sigma_m^2\right) > 0.
\end{aligned}
$$

The tests ...

- ... consider m regression models ($i = 1, ..., m$) each of them nesting a benchmark one ($i = 0$). The null hypothesis is equal predictive accuracy (EPA) across all models, while the alternative postulates that at least one model has a lower mean square prediction error than the benchmark

$$
\begin{aligned}
H_0 &: \sigma_0^2 = \sigma_1^2 = ... = \sigma_m^2, \\
H_A &: \max\left(\sigma_0^2 - \sigma_1^2, ..., \sigma_0^2 - \sigma_m^2\right) > 0.
\end{aligned}
$$

- *Why "small set" of nested models?* White (2000)'s "reality check" was mainly intended for guarding against extensive data mining. How does test performance deteriorate as m gets larger?

The tests ...

- ... consider m regression models $(i = 1, ..., m)$ each of them nesting a benchmark one $(i = 0)$. The null hypothesis is equal predictive accuracy (EPA) across all models, while the alternative postulates that at least one model has a lower mean square prediction error than the benchmark

$$
\begin{aligned}
H_0 &: \sigma_0^2 = \sigma_1^2 = ... = \sigma_m^2, \\
H_A &: \max\left(\sigma_0^2 - \sigma_1^2, ..., \sigma_0^2 - \sigma_m^2\right) > 0.
\end{aligned}
$$

- *Why "small set" of nested models?* White (2000)'s "reality check" was mainly intended for guarding against extensive data mining. How does test performance deteriorate as m gets larger?
- *Can use the test to select the "best" model?*

The statistics...

- ... are based on comparing the average squared prediction error plus some adjustment (that -for nested models- helps to re-center the limiting distribution),

$$\widehat{f}_{i,t+1} = \widehat{e}^2_{0,t+1} - \widehat{e}^2_{i,t+1} + \left(\widehat{y}_{0,t+1} - \widehat{y}_{i,t+1}\right)^2 .$$

The statistics...

- ... are based on comparing the average squared prediction error plus some adjustment (that -for nested models- helps to re-center the limiting distribution),

$$\widehat{f}_{i,t+1} = \widehat{e}^2_{0,t+1} - \widehat{e}^2_{i,t+1} + \left(\widehat{y}_{0,t+1} - \widehat{y}_{i,t+1}\right)^2 .$$

- As the limiting distribution of $\overline{f}_i \equiv P^{-1} \sum_{s=1}^{P} \widehat{f}_{i,t+s}$ is not too badly approximated by a Gaussian, two Wald-type statistics are proposed in the paper: a quadratic form in the vector of the $\overline{f}_i$'s (called $\chi^2\,(adj)$ statistic) and the $\widehat{z}$ statistic

$$\widehat{z} = \max\left(P^{1/2}\overline{f}_1/\sqrt{\widehat{\upsilon}_1}, ..., P^{1/2}\overline{f}_m/\sqrt{\widehat{\upsilon}_m}\right),$$

where $\widehat{\upsilon}_i$ is an estimate of the long-run variance of $\widehat{f}_{i,t+1}$. The approximate null distribution of the $\widehat{z}$ statistic can be easily simulated.

Size-power tradeoff of testing against many alternative models

- The simultaneous comparison allows to **control size** against the possibility of cherry-picking the best performing model: with pairwise comparisons you may end up rejecting the null hypothesis too often.

Size-power tradeoff of testing against many alternative models

- The simultaneous comparison allows to **control size** against the possibility of cherry-picking the best performing model: with pairwise comparisons you may end up rejecting the null hypothesis too often.
- However, as *m* gets larger the simultaneous comparison is "diluted" by adding a lot of randomness –> inevitable **loss of power**

Size-power tradeoff of testing against many alternative models

- The simultaneous comparison allows to **control size** against the possibility of cherry-picking the best performing model: with pairwise comparisons you may end up rejecting the null hypothesis too often.
- However, as *m* gets larger the simultaneous comparison is "diluted" by adding a lot of randomness –> inevitable **loss of power**
- I do not necessarily share the opinion (in the empirical section) that unemployment does not really help predicting euro-area inflation because the forecasts from that model are not significantly better in a 5-model comparison (while they appear better in pairwise tests).

Testing equal predictive ability and testing forecast encompassing ...

- ... is equivalent for the case of nested models.

$$\begin{aligned} \mathrm{M}_0 &: \quad \widehat{y}_{0,t+1} = P\left(Y | X_0 \right) \\ \mathrm{M}_i &: \quad \widehat{y}_{i,t+1} = P\left(Y | X_0, X_i \right) \end{aligned}$$

If X_i does not have predictive power for Y then (a) the forecasts from M_0 encompass those from M_i and (b) the two models have same predictive accuracy.

Testing equal predictive ability and testing forecast encompassing ...

- ... is equivalent for the case of nested models.

$$\begin{aligned} \mathrm{M}_0 &: \quad \widehat{y}_{0,t+1} = P\left(Y \middle| X_0 \right) \\ \mathrm{M}_i &: \quad \widehat{y}_{i,t+1} = P\left(Y \middle| X_0, X_i \right) \end{aligned}$$

  If X_i does not have predictive power for Y then (a) the forecasts from M_0 encompass those from M_i and (b) the two models have same predictive accuracy.

- *Definition of FE*: $\widehat{y}_{0,t+1}$ encompasses $\widehat{y}_{i,t+1}$ if there is no gain from combining them into a composite forecast

$$\widehat{y}_{c,t+1} = (1-\lambda)\,\widehat{y}_{0,t+1} + \lambda \widehat{y}_{i,t+1},$$

  for some $\lambda > 0$.

Testing equal predictive ability and testing forecast encompassing ...

- ... is equivalent for the case of nested models.

$$\begin{aligned} \mathrm{M}_0 &: \quad \widehat{y}_{0,t+1} = P\left(Y|\, X_0\right) \\ \mathrm{M}_i &: \quad \widehat{y}_{i,t+1} = P\left(Y|\, X_0, X_i\right) \end{aligned}$$

  If X_i does not have predictive power for Y then (a) the forecasts from M_0 encompass those from M_i and (b) the two models have same predictive accuracy.

- *Definition of FE*: $\widehat{y}_{0,t+1}$ encompasses $\widehat{y}_{i,t+1}$ if there is no gain from combining them into a composite forecast

$$\widehat{y}_{c,t+1} = (1-\lambda)\,\widehat{y}_{0,t+1} + \lambda \widehat{y}_{i,t+1},$$

  for some $\lambda > 0$.

- In fact, as recalled in the paper, the test of EPA based on the adjusted MSPE's is equivalent to a test of FE ($H_0 : \lambda = 0$).

Thus, one possibility is to extend this framework to other FE tests ...

- ... that perhaps may lead to power improvements. In particular, based on Clark and Mc Cracken (2001), can construct the "max t-test"

Thus, one possibility is to extend this framework to other FE tests ...

- ... that perhaps may lead to power improvements. In particular, based on Clark and Mc Cracken (2001), can construct the "max t-test"
- (a) using the correct limiting distribution of each t-statistic $P^{1/2}\overline{f}_i / \sqrt{\widehat{\nu}_i}$

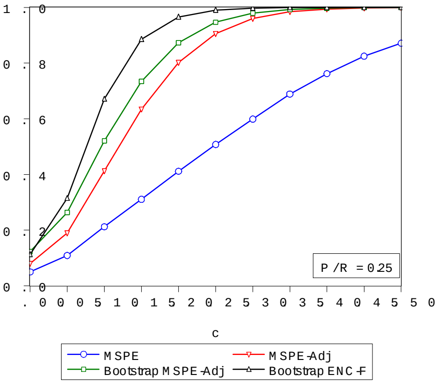
Thus, one possibility is to extend this framework to other FE tests ...

- ... that perhaps may lead to power improvements. In particular, based on Clark and Mc Cracken (2001), can construct the "max t-test"
- (a) using the correct limiting distribution of each t-statistic $P^{1/2}\overline{f}_i/\sqrt{\widehat{\nu}_i}$
- (b) using the alternative individual statistics $P\overline{f}_i/\widehat{\sigma}_i^2$, where $\widehat{\sigma}_i^2 = P^{-1}\sum_{s=1}^{P}\widehat{e}_{i,t+s}^2$

Thus, one possibility is to extend this framework to other FE tests ...

- ... that perhaps may lead to power improvements. In particular, based on Clark and Mc Cracken (2001), can construct the "max t-test"
- (a) using the correct limiting distribution of each t-statistic $P^{1/2}\overline{f}_i/\sqrt{\widehat{\nu}_i}$
- (b) using the alternative individual statistics $P\overline{f}_i/\widehat{\sigma}_i^2$, where $\widehat{\sigma}_i^2 = P^{-1}\sum_{s=1}^{P}\widehat{e}_{i,t+s}^2$
- Getting critical values would be more complicated as cannot simply simulate from a multivariate normals with an estimated correlation structure. But a bootstrap approximation should go through, like in Hansen (2005). An idea of the order of magnitude of the power gain can be obtained looking at the simulated power functions computed in Busetti, Marcucci and Veronese (2009) for $m = 1$

Power functions of the MSPE, MSPE-adj and other FE tests (m=1)



How about non-nested models?

- The "joint FE test" of a benchmark model against several alternatives provided here should be useful also for the case of non-nested model. Asymptotic Gaussianity OK.

How about non-nested models?

- The "joint FE test" of a benchmark model against several alternatives provided here should be useful also for the case of non-nested model. Asymptotic Gaussianity OK.
- Of course, EPA and FE are no longer equivalent. In particular a forecasting model can contain useful information even when its predictive accuracy is relatively bad (but FE by M_0 of M_1 implies that $MSPE_0 \leq MSPE_1$**).**

How about non-nested models?

- The "joint FE test" of a benchmark model against several alternatives provided here should be useful also for the case of non-nested model. Asymptotic Gaussianity OK.
- Of course, EPA and FE are no longer equivalent. In particular a forecasting model can contain useful information even when its predictive accuracy is relatively bad (but FE by M_0 of M_1 implies that $MSPE_0 \leq MSPE_1$**)**.
- It is also interesting that FE tests retain some advantage over the standard EPA tests for out-of-sample model selection. I have this example, taken from Busetti, Marcucci and Veronese (2009):

The set-up

- $$\begin{aligned} y_t &= \mu_y + \phi_y y_{t-1} + \beta x_{t-1} + \varepsilon_t, & \varepsilon_t &\sim IN\left(0,1\right) \\ x_t &= \mu_x + \phi_x x_{t-1} + u_{x,t} & u_{x,t} &\sim IN\left(0, q_x^2\right) \\ w_t &= x_t + u_{w,t} & u_{w,t} &\sim IN\left(0, q_w^2 \sigma_x^2\right) \end{aligned}$$

  So w_t and x_t are positively correlated with $\rho_{xw} = 1/\left(1 + q_w^2\right)$.

The set-up

- $$\begin{aligned} y_t &= \mu_y + \phi_y y_{t-1} + \beta x_{t-1} + \varepsilon_t, & \varepsilon_t &\sim IN(0,1) \\ x_t &= \mu_x + \phi_x x_{t-1} + u_{x,t} & u_{x,t} &\sim IN\left(0, q_x^2\right) \\ w_t &= x_t + u_{w,t} & u_{w,t} &\sim IN\left(0, q_w^2 \sigma_x^2\right) \end{aligned}$$

  So w_t and x_t are positively correlated with $\rho_{xw} = 1/\left(1 + q_w^2\right)$.
- Let M_X be the true model and M_W be a misspecified one.

  $$\begin{aligned} \mathrm{M}_X &: \quad P(Y|\, 1, Y_{-1}, X) \\ \mathrm{M}_W &: \quad P(Y|\, 1, Y_{-1}, W) \end{aligned}$$

  Assume that $\beta \neq 0$. The models are non nested (although, if $\rho_{xw} \to 1$ the two forecasts coincide)

FE tests can be very helpful for rejecting wrong models

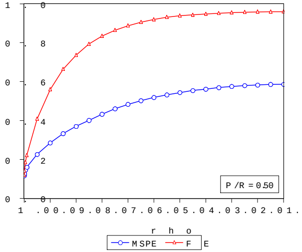
- EPA: H_0 : M_X and M_W have same predictive ability; H_A : M_X is better

FE tests can be very helpful for rejecting wrong models

- EPA: H_0 : M_X and M_W have same predictive ability; H_A : M_X is better
- FE: H_0 : M_W encompasses M_X; H_A : M_X helps forecasting (and thus should at least be included in a Forecast Combination)

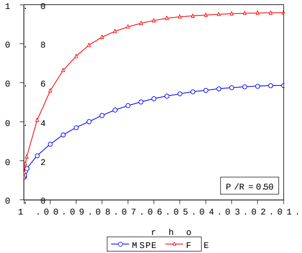
FE tests can be very helpful for rejecting wrong models

- EPA: H_0 : M_X and M_W have same predictive ability; H_A : M_X is better
- FE: H_0 : M_W encompasses M_X; H_A : M_X helps forecasting (and thus should at least be included in a Forecast Combination)
- The rejection frequencies of H_0 are reported in the graph against the value of ρ_{xw} (if $\rho_{xw} \rightarrow 1$ rejection frequencies$\rightarrow$size).



FE tests can be very helpful for rejecting wrong models

- EPA: H_0 : M_X and M_W have same predictive ability; H_A : M_X is better
- FE: H_0 : M_W encompasses M_X; H_A : M_X helps forecasting (and thus should at least be included in a Forecast Combination)
- The rejection frequencies of H_0 are reported in the graph against the value of ρ_{xw} (if $\rho_{xw} \to 1$ rejection frequencies$\to$size).



- In practice, we may have an "economic" model that provides (slightly) worse predictions than others. The FE test can help discriminate whether the worse performance is just due to randomness or not.

Conclusions

- The paper provides an important contribution towards evaluating the out-of-sample performance of several (nested) models. The tests are neat and easily applicable!

Conclusions

- The paper provides an important contribution towards evaluating the out-of-sample performance of several (nested) models. The tests are neat and easily applicable!
- Perhaps one might design more powerful alternative tests but a cost of a substantial complication which could inhibit the actual use of them

Conclusions

- The paper provides an important contribution towards evaluating the out-of-sample performance of several (nested) models. The tests are neat and easily applicable!
- Perhaps one might design more powerful alternative tests but a cost of a substantial complication which could inhibit the actual use of them
- The idea of a joint FE test of a benchmark model against various alternatives should be kept in mind also in the context of non nested model comparisons