An Overview of Recent Developments in Forecast Evaluation

Todd E. Clark[1]

[1]Federal Reserve Bank of Kansas City

October 25, 2007

Introduction: organization

I. Evaluation in simple world: no model estimation, and no nested models

II. What if my forecast models have estimated parameters?

III. What if my models are nested?

- nested: null model is a restricted version of alternative

IV. What if my forecasts are real time?

Introduction: scope of overview

(1) tests of equal MSE applied to point forecasts

- pair of models

(2) tests of encompassing applied to point forecasts

- encompassing: is there information in the alternative forecast not incorporated in the null forecast?
- pair of models

(3) tests of accuracy of density forecasts

- single model and pair of models

omit much in the interest of time

- many tests: Corradi and Swanson (2006c), West (2006)
- multiple models: ditto
- Bayesian methods: Geweke and Whiteman (2006)

I. No model estimation, no nested models

(1) testing for equal MSE

- Diebold and Mariano (1995) t-test
- loss_{t+h} = $e^2_{1,t+h}$ - $e^2_{2,t+h}$
- DM test = mean of loss/standard error
- DM test $\rightarrow^d$ N(0,1)

(2) testing for forecast encompassing

- Harvey, Leybourne, and Newbold (1998) t-test
- loss_{t+h} = $e_{1,t+h}(e_{1,t+h} - e_{2,t+h})$
- HLN test = mean of loss/standard error
- HLN test $\rightarrow^d$ N(0,1)

I. No model estimation, no nested models

(3) density forecasts: single model

- Diebold, Gunther, and Tay (1998): examine probability integral transform (PIT)
  - PIT = prob($y \leq y_{t+h}$)
  - $h = 1$: PIT is iid uniform
  - $h > 1$: PIT is still uniform, but not independent
- DGT: graphically investigate departures of PIT from uniformity
  - histogram of PIT series
  - ECDF against 45 degree line (uniform CDF)
- others: formally test by comparing the maximum distance between the uniform CDF and the ECDF against the Kolmogorov distribution
  - test only valid for $h = 1$

II. What if my models have estimated parameters?

answer depends in part on whether the estimation scheme is recursive or rolling

- recursive: sample expands
- rolling: constant sample size, rolled forward
- notation: R = # of observations used to estimate model generating first forecast, P = # of forecasts

West (1996) asymptotics: $P, R \to \infty$ at the same rate

- allows use of recursive and rolling schemes
- each coefficient estimate $\to_p$ population value

Giacomini and White (2006) asymptotics: R fixed, $P \to \infty$

- allows only rolling scheme
- coefficients always include parameter estimation error
- GW results are conditional on the forecasts (estimated parameters, methods, etc.), rather than unconditional

II. What if my models have estimated parameters?

(1) testing for equal MSE: DM test can be compared against the N(0,1) distribution.

- West (1996): Asymptotically, impacts of parameter estimation error (PEE) on the standard error cancel out.
  - cancellation occurs because estimation and loss functions are the same
  - applies under recursive and rolling schemes
- Giacomini and White (2006): Despite PEE, conventional DM test is N(0,1).
  - applies only under rolling scheme

challenge: DM test rejects too often for $h > 1$

- adjust standard error for bias (HLN (1997)); not asymptotically justified for estimated models
- moving block bootstrap of forecast errors (White (2000))

II. What if my models have estimated parameters?

(2) testing for encompassing: depends on asymptotics

- West (2001): PEE affects the variance of the encompassing statistic.
  - standard error in HLN test must be adjusted, as in West (2001, 2006)
  - specific adjustment depends on whether scheme is recursive or rolling
  - the bigger P/R, the more important is PEE adjustment
- Giacomini and White (2006): Despite PEE, HLN encompassing test is N(0,1).
  - applies only under rolling scheme

challenge: HLN test rejects too often for $h > 1$

- same as with DM test

II. What if my models have estimated parameters?

(3) density forecasts: single model – alternative tests

- PIT/Kolmogorov-based approach: test uniformity of PIT, making some adjustment for impact of PEE
  - applicable for h =1, not $h > 1$
- (a) Bai (2003): a test statistic corrected for the impact of PEE has a simple asymptotic distribution
  - correct the test statistic directly
  - Bai's published results only apply to in-sample analysis.
  - Corradi and Swanson (2006c) show the Bai result carries over to densities of out-of-sample forecasts.
- (b) Corradi and Swanson (2006a): use conventional test statistic (max distance between ECDF and 45 degree line), and bootstrap the resulting non-standard distribution
  - moving block bootstrap of data, with adjustment for recursive estimation

II. What if my models have estimated parameters?

(4) density forecasts: pair of models – alternative tests based on different asymptotics

- (a) Amisano and Giacomini (2007): likelihood ratio test based on log predictive density score
  - based on Giacomini and White (2006) asymptotics, in which PEE is a permanent component of forecast error
  - applicable to $h \geq 1$ and many model estimation approaches
  - but models must be estimated with rolling scheme
- (b) Corradi and Swanson (2006b): test based on distributional analog of MSE (average squared distance between densities)
  - based on West (1996)-type asymptotics
  - non-standard asymptotic distribution which incorporates impact of PEE; use moving block bootstrap with adjustment for critical values
  - applicable to $h \geq 1$
  - applies under recursive and rolling schemes

III. What if my models are nested?

West (1996) requires a variance matrix of moment conditions to have full rank. But with nested models, the forecast errors are perfectly correlated, making the variance matrix singular.

- H_0: small model true $\Rightarrow e_{1,t+h} = e_{2,t+h} \; \forall \; t$
- $\Rightarrow$ The population difference in MSE is exactly 0, with 0 variance.

Nested model tests based on West-type asymptotics are tests of H_0: small model true — i.e., whether some coefficients = 0.

- perfectly valid, but whether a coefficient = 0 may not be what we want to know
- We may want to know if forecasts are equally accurate in a finite sample.

Giacomini and White (2006) asymptotics provide a way of testing whether forecasts are equally accurate.

III. What if my models are nested?

(1) testing for equal MSE: depends on asymptotics

- McCracken (2007) and Clark and McCracken (2005): DM test has a non-standard asymptotic distribution.
  - distribution comes entirely from PEE
  - applies under recursive and rolling schemes
  - $h = 1$: asymptotic critical values from McCracken (2007)
  - $h \geq 1$: MC simulations or parametric bootstrap of null model
- Under H_0: small model true, the MCM asymptotic distribution is shifted well to the left of the N(0,1).
  - Why? Estimation error associated with extraneous parameters in large model makes MSE(large) $>$ MSE(small).
  - Using N(0,1) critical values, the DM test rarely rejects, whether or not the null model is true.
- McCracken (2007) and Clark and McCracken (2005): An F-type test of equal MSE has much better power.

III. What if my models are nested?

(1) testing for equal MSE: depends on asymptotics

- Giacomini and White (2006): DM test is N(0,1).
  - H_0: conditioned on the forecasts, the forecasts are equally accurate
  - applies only under rolling scheme

III. What if my models are nested?

(2) testing for forecast encompassing: depends on asymptotics

- Clark and McCracken (2001, 2005): HLN test has a non-standard asymptotic distribution.
  - distribution comes entirely from PEE
  - applies under recursive and rolling schemes
  - $h = 1$: asymptotic critical values from CM (2001)
  - $h \geq 1$: MC simulations or parametric bootstrap of null model
- Clark and West (2007): N(0,1) critical values reasonably approximate the CM critical values.
  - for $h > 1$, better to use parametric bootstrap of null model
- Giacomini and White (2006): HLN test is N(0,1).
  - H_0: conditioned on the forecasts, encompassing
  - applies only under rolling scheme

(3) testing for forecast encompassing: Corradi-Swanson approach

- Chao, Corradi and Swanson (2001): test of $E(e_{1,t+h}X_{2,t}) = 0$ has a standard distribution
  - West (1996)-type PEE correction of variance is necessary
  - applies under recursive and rolling schemes
  - more general, but power often lower than HLN test
- Corradi and Swanson (2002): CCS extended to allow for generic nonlinear alternative models has a non–standard distribution.
  - distribution reflects PEE
  - applies to recursive forecasts
  - moving block bootstrap of data, with adjustment for recursive estimation (Corradi and Swanson (2007))

III. What if my models are nested?

(4) density forecasts from pair of models: only one option

- Amisano and Giacomini (2007): likelihood ratio test based on log predictive density score
    - based on Giacomini and White (2006) asymptotics, which condition on the forecasts, making PEE a permanent component of forecast error
    - applicable to $h \geq 1$ and general set of model estimation approaches
    - but models must be estimated with rolling scheme

IV. What if my forecasts are real time?

Existing work makes stationarity assumptions that rule out real time data.

- Usual assumption: regression orthogonality conditions hold both in-sample and out-of-sample.
- But predictable data revisions will make the real-time forecast error correlated with the real-time predictors.

(1) testing for equal MSE, non-nested models

- Clark and McCracken (2007): DM test has a N(0,1) distribution, but only with a standard error corrected for the impact of predictability of revisions
    - Monte Carlo evidence: standard error correction is often helpful, but doesn't seem essential

IV. What if my forecasts are real time?

(2) testing for equal MSE, nested models

- Clark and McCracken (2007): DM test has a N(0,1) distribution, but only with a standard error based entirely on the impact of the predictability of revisions
  - sharp contrast to case of no revisions, in which distribution of t-test is non-standard
  - Monte Carlo evidence: taking account of real time revisions is helpful, but not necessarily essential
  - punchline: Consider both the F-type test of equal MSE that ignores the impact of predictable revisions and the t-test that does not.

Conclusions

I. Evaluation in simple world: no model estimation, and no nested models

- if only things were so simple....

II. What do I do if my forecast models have estimated parameters?

- MSE tests: simple DM
- encompassing test: depends on asymptotics, but probably a good idea to adjust HLN test for PEE as in West (2001, 2006)
- densities: take some care to account for PEE

Conclusions

III. What do I do if my models are nested?

- MSE tests: depends on asymptotics, but probably a good idea to use F-test and critical values of McCracken (2007) and Clark and McCracken (2005)
- encompassing test: reasonably safe to compare against standard critical values (short h)
- densities: only one option, which requires rolling scheme

IV. What if my forecasts are real time?

- good idea to make corrections for revision predictability in Clark and McCracken (2007), but also consider some uncorrected tests

Conclusions: other challenges

With typical sample sizes, it is not often the case that one forecast robustly beats another.

- result of instability?
- result of small sample imprecision and limited power?
  - Clark and McCracken (2006)

Predictability has declined sharply in some countries.

- now much harder for multivariate models or professional forecasts to beat univariate benchmarks
  - Campbell (2007), Clark and McCracken (2008), Faust and Wright (2007)
- outcome of better policy and optimal control?
  - Have unconditional forecast evaluations become less useful for model and forecast evaluations?
  - Should we consider conditional forecasts? If so, how do we evaluate them?

References

Amisano, G., and Giacomini, R. (2007), "Comparing Density Forecasts via Weighted Likelihood Ratio Tests," *Journal of Business and Economic Statistics* 25, 177-190.

Bai, J. (2003), "Testing Parametric Conditional Distributions of Dynamic Models," *Review of Economic and Statistics* 85, 531-549.

Campbell, S.D. (2007), "Macroeconomic Volatility, Predictability, and Uncertainty in the Great Moderation: Evidence from the Survey of Professional Forecasters,". *Journal of Business and Economic Statistics* 25, 191-200.

Chao, J., Corradi, V., and Swanson, N.R. (2001), "An Out of Sample Test for Granger Causality," *Macroeconomic Dynamics* 5, 598-620.

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. (2006), "The Predictive Content of the Output Gap for Inflation: Resolving In–Sample and Out–of–Sample Evidence," *Journal of Money, Credit, and Banking* 38, 1127-1148.

Clark, T.E., and McCracken, M.W. (2007), "Tests of Equal Predictive Ability with Real-Time Data," Research Working Paper 07-06, Federal Reserve Bank of Kansas City.

Clark, T.E., and McCracken, M.W. (2008), "Forecasting with Small Macroeconomic VARs in the Presence of Instability," in *Forecasting in the Presence of Structural Breaks and Model Uncertainty*, Rapach D.E., and Wohar, M.E. (eds), Elsevier: Amsterdam (forthcoming).

Clark, T.E., and West, K.D. (2006), "Using Out-of-Sample Mean Squared Prediction Errors to Test the Martingale Difference Hypothesis," *Journal of Econometrics* 135, 155-186.

Clark, T.E., and West, K.D. (2007), "Approximately Normal Tests for Equal Predictive Accuracy in Nested Models," *Journal of Econometrics* 138, 291-311.

Corradi, V., and Swanson, N.R. (2002), "A Consistent Test for Nonlinear Out–of–Sample Predictive Accuracy," *Journal of Econometrics* 110, 353-381.

Corradi V., and Swanson, N.R. (2006a), "Bootstrap Conditional Distribution Tests In the Presence of Dynamic Misspecification," *Journal of Econometrics* 133, 779-806.

Corradi V., and Swanson, N.R. (2006b), "Predictive Density and Conditional Confidence Interval Accuracy Tests," *Journal of Econometrics* 135, 187-228.

Corradi V., and Swanson, N.R. (2006c), "Predictive Density Evaluation," in *Handbook of Economic Forecasting*, Elliott G., Granger C.W.J., and Timmermann A. (eds.), North Holland.

Corradi, V., and Swanson, N.R. (2007), "Nonparametric Bootstrap Procedures for Predictive Inference Based on Recursive Estimation Schemes," *International Economic Review* 48, 67-109.

Diebold, F.X., Gunther T.A., and Tay, A.S. (1998), "Evaluating Density Forecasts with Applications to Financial Risk Management," *International Economic Review* 39, 863-883.

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

Faust, J., and Wright, J.H. (2007), "Comparing Greenbook and Reduced Form Forecasts using a Large Realtime Dataset," manuscript, Johns Hopkin University.

Geweke, J., and Whiteman, C. (2006), "Bayesian Forecasting," in *Handbook of Economic Forecasting*, Elliott G., Granger C.W.J., and Timmermann A. (eds.), North Holland.

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

Harvey, D.I., Leybourne, S.J., and Newbold, P. (1997), "Testing the Equality of Prediction Mean Squared Errors," *International Journal of Forecasting* 13, 281-291.

Harvey, D.I., Leybourne, S.J., and Newbold, P. (1998), "Tests for Forecast Encompassing," *Journal of Business and Economic Statistics* 16, 254-259.

McCracken, M.W. (2007), "Asymptotics for Out–of–Sample Tests of Causality," *Journal of Econometrics* 140, 717-752.

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

West, K.D. (2001), "Tests for Forecast Encompassing When Forecasts Depend on Estimated Regression Parameters," *Journal of Business and Economic Statistics* 19, 29-33.

West, K.D. (2006), "Forecast Evaluation," in *Handbook of Economic Forecasting*, Elliott G., Granger C.W.J., and Timmermann A. (eds.), North Holland.

White, H. (2000), "A Reality Check for Data Snooping," *Econometrica* 68, 1097-1126.