FORECAST EVALUATION OF SMALL NESTED MODEL SETS

Kirstin Hubrich and Kenneth West

European Central Bank / Federal Reserve Board and University of Wisconsin

January 2009

NBER Working Paper 14601, ECB WP No. 1030, 2009

Motivation

- Forecast evaluation often compares a small set of models
- Criterion: usually mean squared prediction error (MSPE)
- Usually: Sequence of pairwise model comparisons is carried out, using Diebold-Mariano-West statistic for non-nested models or Clark- McCracken and Clark-West statistics for nested models
- Our proposal: Comparing models simultaneously; two statistics easy to compute for simultaneously comparing a parsimonious benchmark model to m alternative models that nest the benchmark
- We take into account potential correlation of model forecasts

Conclusions from pairwise model comparison might be misleading

Motivation

- Forecast evaluation often compares a small set of models
- Criterion: usually mean squared prediction error (MSPE)
- Usually: Sequence of pairwise model comparisons is carried out, using Diebold-Mariano-West statistic for non-nested models or Clark- McCracken and Clark-West statistics for nested models
- Our proposal: Comparing models simultaneously; two statistics easy to compute for simultaneously comparing a parsimonious benchmark model to m alternative models that nest the benchmark
- We take into account potential correlation of model forecasts

Conclusions from pairwise model comparison might be misleading

Motivation

- Forecast evaluation often compares a small set of models
- Criterion: usually mean squared prediction error (MSPE)
- Usually: Sequence of pairwise model comparisons is carried out, using Diebold-Mariano-West statistic for non-nested models or Clark- McCracken and Clark-West statistics for nested models
- Our proposal: Comparing models simultaneously; two statistics easy to compute for simultaneously comparing a parsimonious benchmark model to m alternative models that nest the benchmark
- We take into account potential correlation of model forecasts

Conclusions from pairwise model comparison might be misleading

Motivation

- Forecast evaluation often compares a small set of models
- Criterion: usually mean squared prediction error (MSPE)
- Usually: Sequence of pairwise model comparisons is carried out, using Diebold-Mariano-West statistic for non-nested models or Clark- McCracken and Clark-West statistics for nested models
- Our proposal: Comparing models simultaneously; two statistics easy to compute for simultaneously comparing a parsimonious benchmark model to m alternative models that nest the benchmark
- We take into account potential correlation of model forecasts

Conclusions from pairwise model comparison might be misleading

Motivation

- Forecast evaluation often compares a small set of models
- Criterion: usually mean squared prediction error (MSPE)
- Usually: Sequence of pairwise model comparisons is carried out, using Diebold-Mariano-West statistic for non-nested models or Clark- McCracken and Clark-West statistics for nested models
- Our proposal: Comparing models simultaneously; two statistics easy to compute for simultaneously comparing a parsimonious benchmark model to m alternative models that nest the benchmark
- We take into account potential correlation of model forecasts

Conclusions from pairwise model comparison might be misleading

Motivation

- Forecast evaluation often compares a small set of models
- Criterion: usually mean squared prediction error (MSPE)
- Usually: Sequence of pairwise model comparisons is carried out, using Diebold-Mariano-West statistic for non-nested models or Clark- McCracken and Clark-West statistics for nested models
- Our proposal: Comparing models simultaneously; two statistics easy to compute for simultaneously comparing a parsimonious benchmark model to m alternative models that nest the benchmark
- We take into account potential correlation of model forecasts

Conclusions from pairwise model comparison might be misleading

Recent literature: Pairwise Model Comparison

- Non-nested model comparisons: Diebold-Mariano-West tests
  - null hypothesis of equal predictive accuracy (in population)
  - Diebold and Mariano (1995): allow for wide variety of forecast accuracy measures and general assumptions on forecast errors
  - West (1996): allows for estimation uncertainty
  - provide conditions for t-type statistics $\sim_A N(0,1)$
- Nested model comparisons: McCracken (2007), Clark and McCracken (2001, 2005), Clark and West (2006, 2007)
  - nested models: under null of equal predictive accuracy the variance of the forecast error differential is zero
  - derive standard and non-standard limiting distributions
  - Clark and West (2006, 2007): test statistic adjusted for estimation uncertainty is approximately normal

Recent literature: Pairwise Model Comparison

- Non-nested model comparisons: Diebold-Mariano-West tests
  - null hypothesis of equal predictive accuracy (in population)
  - Diebold and Mariano (1995): allow for wide variety of forecast accuracy measures and general assumptions on forecast errors
  - West (1996): allows for estimation uncertainty
  - provide conditions for t-type statistics $\sim_A N(0, 1)$
- Nested model comparisons: McCracken (2007), Clark and McCracken (2001, 2005), Clark and West (2006, 2007)
  - nested models: under null of equal predictive accuracy the variance of the forecast error differential is zero
  - derive standard and non-standard limiting distributions
  - Clark and West (2006, 2007): test statistic adjusted for estimation uncertainty is approximately normal

Recent literature: Pairwise Model Comparison

- Non-nested model comparisons: Diebold-Mariano-West tests
  - null hypothesis of equal predictive accuracy (in population)
  - Diebold and Mariano (1995): allow for wide variety of forecast accuracy measures and general assumptions on forecast errors
  - West (1996): allows for estimation uncertainty
  - provide conditions for t-type statistics $\sim_A N(0,1)$
- Nested model comparisons: McCracken (2007), Clark and McCracken (2001, 2005), Clark and West (2006, 2007)
  - nested models: under null of equal predictive accuracy the variance of the forecast error differential is zero
  - derive standard and non-standard limiting distributions
  - Clark and West (2006, 2007): test statistic adjusted for estimation uncertainty is approximately normal

Recent literature: Pairwise Model Comparison

- Non-nested model comparisons: Diebold-Mariano-West tests
  - null hypothesis of equal predictive accuracy (in population)
  - Diebold and Mariano (1995): allow for wide variety of forecast accuracy measures and general assumptions on forecast errors
  - West (1996): allows for estimation uncertainty
  - provide conditions for t-type statistics $\sim_A N(0,1)$
- Nested model comparisons: McCracken (2007), Clark and McCracken (2001, 2005), Clark and West (2006, 2007)
  - nested models: under null of equal predictive accuracy the variance of the forecast error differential is zero
  - derive standard and non-standard limiting distributions
  - Clark and West (2006, 2007): test statistic adjusted for estimation uncertainty is approximately normal

Recent literature: Pairwise Model Comparison

- Non-nested model comparisons: Diebold-Mariano-West tests
  - null hypothesis of equal predictive accuracy (in population)
  - Diebold and Mariano (1995): allow for wide variety of forecast accuracy measures and general assumptions on forecast errors
  - West (1996): allows for estimation uncertainty
  - provide conditions for t-type statistics $\sim_A N(0,1)$
- Nested model comparisons: McCracken (2007), Clark and McCracken (2001, 2005), Clark and West (2006, 2007)
  - nested models: under null of equal predictive accuracy the variance of the forecast error differential is zero
  - derive standard and non-standard limiting distributions
  - Clark and West (2006, 2007): test statistic adjusted for estimation uncertainty is approximately normal

Recent literature: Pairwise Model Comparison

- Non-nested model comparisons: Diebold-Mariano-West tests
  - null hypothesis of equal predictive accuracy (in population)
  - Diebold and Mariano (1995): allow for wide variety of forecast accuracy measures and general assumptions on forecast errors
  - West (1996): allows for estimation uncertainty
  - provide conditions for t-type statistics $\sim_A N(0,1)$
- Nested model comparisons: McCracken (2007), Clark and McCracken (2001, 2005), Clark and West (2006, 2007)
  - nested models: under null of equal predictive accuracy the variance of the forecast error differential is zero
  - derive standard and non-standard limiting distributions
  - Clark and West (2006, 2007): test statistic adjusted for estimation uncertainty is approximately normal

Recent literature: Pairwise Model Comparison

- Non-nested model comparisons: Diebold-Mariano-West tests
  - null hypothesis of equal predictive accuracy (in population)
  - Diebold and Mariano (1995): allow for wide variety of forecast accuracy measures and general assumptions on forecast errors
  - West (1996): allows for estimation uncertainty
  - provide conditions for t-type statistics $\sim_A N(0,1)$
- Nested model comparisons: McCracken (2007), Clark and McCracken (2001, 2005), Clark and West (2006, 2007)
  - nested models: under null of equal predictive accuracy the variance of the forecast error differential is zero
  - derive standard and non-standard limiting distributions
  - Clark and West (2006, 2007): test statistic adjusted for estimation uncertainty is approximately normal

Recent literature: Pairwise Model Comparison

- Non-nested model comparisons: Diebold-Mariano-West tests
  - null hypothesis of equal predictive accuracy (in population)
  - Diebold and Mariano (1995): allow for wide variety of forecast accuracy measures and general assumptions on forecast errors
  - West (1996): allows for estimation uncertainty
  - provide conditions for t-type statistics $\sim_A N(0,1)$
- Nested model comparisons: McCracken (2007), Clark and McCracken (2001, 2005), Clark and West (2006, 2007)
  - nested models: under null of equal predictive accuracy the variance of the forecast error differential is zero
  - derive standard and non-standard limiting distributions
  - Clark and West (2006, 2007): test statistic adjusted for estimation uncertainty is approximately normal

Recent literature: Pairwise Model Comparison

- Non-nested model comparisons: Diebold-Mariano-West tests
  - null hypothesis of equal predictive accuracy (in population)
  - Diebold and Mariano (1995): allow for wide variety of forecast accuracy measures and general assumptions on forecast errors
  - West (1996): allows for estimation uncertainty
  - provide conditions for t-type statistics $\sim_A N(0,1)$
- Nested model comparisons: McCracken (2007), Clark and McCracken (2001, 2005), Clark and West (2006, 2007)
  - nested models: under null of equal predictive accuracy the variance of the forecast error differential is zero
  - derive standard and non-standard limiting distributions
  - Clark and West (2006, 2007): test statistic adjusted for estimation uncertainty is approximately normal

Motivation

Example: Our empirical application considers forecasts of U.S. CPI all items inflation, comparing:

- univariate AR forecast ("model 0") vs. m=4 other models
- the m=4 other models are bivariate VARs with CPI inflation and
  - one the component of the CPI (food, energy, commodities or services inflation)
  - output growth, unemployment, commodities or services inflation

⇒ see e.g. Hubrich (2005) and Hendry and Hubrich (2009) for further empirical results and for analytical and simulation results on the use of disaggregate information in forecasting the aggregate

Motivation

Example: Our empirical application considers forecasts of U.S. CPI all items inflation, comparing:

- univariate AR forecast ("model 0") vs. m=4 other models
- the m=4 other models are bivariate VARs with CPI inflation and
  - one the component of the CPI (food, energy, commodities or services inflation)
  - output growth, unemployment, commodities or services inflation

⇒ see e.g. Hubrich (2005) and Hendry and Hubrich (2009) for further empirical results and for analytical and simulation results on the use of disaggregate information in forecasting the aggregate

Motivation

Example: Our empirical application considers forecasts of U.S. CPI all items inflation, comparing:

- univariate AR forecast ("model 0") vs. m=4 other models
- the m=4 other models are bivariate VARs with CPI inflation and
  - one the component of the CPI (food, energy, commodities or services inflation)
  - output growth, unemployment, commodities or services inflation

⇒ see e.g. Hubrich (2005) and Hendry and Hubrich (2009) for further empirical results and for analytical and simulation results on the use of disaggregate information in forecasting the aggregate

Motivation

Example: Our empirical application considers forecasts of U.S. CPI all items inflation, comparing:

- univariate AR forecast ("model 0") vs. m=4 other models
- the m=4 other models are bivariate VARs with CPI inflation and
  - one the component of the CPI (food, energy, commodities or services inflation)
  - output growth, unemployment, commodities or services inflation

⇒ see e.g. Hubrich (2005) and Hendry and Hubrich (2009) for further empirical results and for analytical and simulation results on the use of disaggregate information in forecasting the aggregate

Empirical Example: Forecasting All Items CPI Inflation

	1984-2004					
method	RMSPE (altern)/ RMSPE (bench)	t-stat adj.	max t-stat adj.	χ^2 adj	χ^2 unadj	Reality check
$AR_{(AIC)}$ (bench)	**0.187**					
Test AR						
vs $VAR^{a,f}_{(AIC)}$	0.999	1.860*				
vs $VAR^{a,e}_{(AIC)}$	1.097	–0.027				
vs $VAR^{a,c}_{(AIC)}$	1.048	0.290				
vs $VAR^{a,s}_{(AIC)}$	1.027	–0.463				
vs 4 models			1.860	3.905	11.926*	0.0007
critical value		1.282	1.919	7.78	7.78	0.059

Motivation

- Illustrative of class of applications relevant for our procedures: MSPE is measure of forecast performance, null model nested in a "small" number of other models
  - "small" number of models m is greater than 1 but much smaller than sample size
  - relevant applications include ones that conduct evaluations of this sort simultaneously for several data sets
- **Relevant asset pricing applications:** e.g. Hong and Lee (REStat, 2003), Goyal and Welch (working paper, 2004), Cheung et al. (JIMF, 2005), McCracken and Sapp (JMCB, 2005), Sarno et al. (JMCB, 2005), Rapach and Wohar (JEmpFin, 2006)
- **Relevant inflation applications:** e.g. Atkeson and Ohanian (FRBMinnQR, 2001), Billmeier (working paper, 2004), Hubrich (IJF, 2005), Cristadoro et al. (JMCB, 2005), D'Agostino et al. (2006), Hendry and Hubrich (JBES, 2009)

Motivation

- Illustrative of class of applications relevant for our procedures: MSPE is measure of forecast performance,
  null model nested in a "small" number of other models
  - "small" number of models m is greater than 1 but much smaller than sample size
  - relevant applications include ones that conduct evaluations of this sort simultaneously for several data sets
- **Relevant asset pricing applications:** e.g. Hong and Lee (REStat, 2003), Goyal and Welch (working paper, 2004), Cheung et al. (JIMF, 2005), McCracken and Sapp (JMCB, 2005), Sarno et al. (JMCB, 2005), Rapach and Wohar (JEmpFin, 2006)
- **Relevant inflation applications:** e.g. Atkeson and Ohanian (FRBMinnQR, 2001), Billmeier (working paper, 2004), Hubrich (IJF, 2005), Cristadoro et al. (JMCB, 2005), D'Agostino et al. (2006), Hendry and Hubrich (JBES, 2009)

Motivation

- Illustrative of class of applications relevant for our procedures: MSPE is measure of forecast performance, null model nested in a "small" number of other models
  - "small" number of models m is greater than 1 but much smaller than sample size
  - relevant applications include ones that conduct evaluations of this sort simultaneously for several data sets
- **Relevant asset pricing applications:** e.g. Hong and Lee (REStat, 2003), Goyal and Welch (working paper, 2004), Cheung et al. (JIMF, 2005), McCracken and Sapp (JMCB, 2005), Sarno et al. (JMCB, 2005), Rapach and Wohar (JEmpFin, 2006)
- **Relevant inflation applications:** e.g. Atkeson and Ohanian (FRBMinnQR, 2001), Billmeier (working paper, 2004), Hubrich (IJF, 2005), Cristadoro et al. (JMCB, 2005), D'Agostino et al. (2006), Hendry and Hubrich (JBES, 2009)

Motivation

- Illustrative of class of applications relevant for our procedures: MSPE is measure of forecast performance, null model nested in a "small" number of other models
  - "small" number of models m is greater than 1 but much smaller than sample size
  - relevant applications include ones that conduct evaluations of this sort simultaneously for several data sets
- **Relevant asset pricing applications:** e.g. Hong and Lee (REStat, 2003), Goyal and Welch (working paper, 2004), Cheung et al. (JIMF, 2005), McCracken and Sapp (JMCB, 2005), Sarno et al. (JMCB, 2005), Rapach and Wohar (JEmpFin, 2006)
- **Relevant inflation applications:** e.g. Atkeson and Ohanian (FRBMinnQR, 2001), Billmeier (working paper, 2004), Hubrich (IJF, 2005), Cristadoro et al. (JMCB, 2005), D'Agostino et al. (2006), Hendry and Hubrich (JBES, 2009)

Existing procedures

Compute χ^2 statistic (multivariate version of DMW):

- Construct $m \times 1$ vector of MSPE differences; compute long run variance of differences; compute the usual quadratic form - "χ^2(unadj.)" in our tables (West et al. (1993))
- This statistic has possible problems with size and power:
  - **size:** under our null, the vector of MSPE differences is not centered at zero
  - **power:** the alternative is one sided
    $\Rightarrow$ even if the vector is recentered appropriately, to deliver correct size under the null, a large chi-squared value can come from the wrong tail of the distribution

Existing procedures

Compute χ^2 statistic (multivariate version of DMW):

- Construct $m \times 1$ vector of MSPE differences; compute long run variance of differences; compute the usual quadratic form - "χ^2(unadj.)" in our tables (West et al. (1993))
- This statistic has possible problems with size and power:
  - **size:** under our null, the vector of MSPE differences is not centered at zero
  - **power:** the alternative is one sided
    $\Rightarrow$ even if the vector is recentered appropriately, to deliver correct size under the null, a large chi-squared value can come from the wrong tail of the distribution

Existing procedures

Compute χ^2 statistic (multivariate version of DMW):

- Construct $m \times 1$ vector of MSPE differences; compute long run variance of differences; compute the usual quadratic form - "χ^2(unadj.)" in our tables (West et al. (1993))
- This statistic has possible problems with size and power:
  - **size:** under our null, the vector of MSPE differences is not centered at zero
  - **power:** the alternative is one sided
    $\Rightarrow$ even if the vector is recentered appropriately, to deliver correct size under the null, a large chi-squared value can come from the wrong tail of the distribution

Existing procedures

Compute χ^2 statistic (multivariate version of DMW):

- Construct $m \times 1$ vector of MSPE differences; compute long run variance of differences; compute the usual quadratic form - "χ^2(unadj.)" in our tables (West et al. (1993))
- This statistic has possible problems with size and power:
  - **size:** under our null, the vector of MSPE differences is not centered at zero
  - **power:** the alternative is one sided
    $\Rightarrow$ even if the vector is recentered appropriately, to deliver correct size under the null, a large chi-squared value can come from the wrong tail of the distribution

Existing procedures (continued)

Simulation / Bootstrap:

- Reality check (White (2000), proposed for $m \sim T$), Test for superior predictive ability (Hansen (2005))
  - problem: might not account for dependence of predictions on estimated regression parameters under our null
- Simulate, including reestimation of forecasting models (Rapach and Wohar (2006))
  - possible problem: time intensive

Further alternative: Construct a set of pairwise comparisons, adjust via Bonferroni or related procedure: low power

Existing procedures (continued)

Simulation / Bootstrap:

- Reality check (White (2000), proposed for $m \sim T$), Test for superior predictive ability (Hansen (2005))
  - problem: might not account for dependence of predictions on estimated regression parameters under our null
- Simulate, including reestimation of forecasting models (Rapach and Wohar (2006))
  - possible problem: time intensive

Further alternative: Construct a set of pairwise comparisons, adjust via Bonferroni or related procedure: low power

Existing procedures (continued)

Simulation / Bootstrap:

- Reality check (White (2000), proposed for $m \sim T$), Test for superior predictive ability (Hansen (2005))
  - problem: might not account for dependence of predictions on estimated regression parameters under our null
- Simulate, including reestimation of forecasting models (Rapach and Wohar (2006))
  - possible problem: time intensive

Further alternative: Construct a set of pairwise comparisons, adjust via Bonferroni or related procedure: low power

Existing procedures (continued)

Simulation / Bootstrap:

- Reality check (White (2000), proposed for $m \sim T$),
  Test for superior predictive ability (Hansen (2005))
  - problem: might not account for dependence of predictions on estimated regression parameters under our null
- Simulate, including reestimation of forecasting models (Rapach and Wohar (2006))
  - possible problem: time intensive

Further alternative: Construct a set of pairwise comparisons, adjust via Bonferroni or related procedure: low power

Existing procedures (continued)

Simulation / Bootstrap:

- Reality check (White (2000), proposed for $m \sim T$), Test for superior predictive ability (Hansen (2005))
  - problem: might not account for dependence of predictions on estimated regression parameters under our null
- Simulate, including reestimation of forecasting models (Rapach and Wohar (2006))
  - possible problem: time intensive

Further alternative: Construct a set of pairwise comparisons, adjust via Bonferroni or related procedure: low power

Existing procedures (continued)

Simulation / Bootstrap:

- Reality check (White (2000), proposed for $m \sim T$), Test for superior predictive ability (Hansen (2005))
  - problem: might not account for dependence of predictions on estimated regression parameters under our null
- Simulate, including reestimation of forecasting models (Rapach and Wohar (2006))
  - possible problem: time intensive

Further alternative: Construct a set of pairwise comparisons, adjust via Bonferroni or related procedure: low power

Our proposals

- construct a vector of **MSPE differences adjusted** as in Clark and West (2006, 2007) to center vector at zero under the null
- compute a variance-covariance matrix for the vector of MSPE differences
- conduct inference via either of the following two options
  1. "max t-stat (adj.)": inference on the largest of the m adjusted t-statistics that compare null model one by one to each of the m larger models **via the distribution of the maximum of correlated normals**;
  2. "χ^2 (adj.)": inference via the usual χ^2 statistic

Key features:

- we take estimation uncertainty into account
- we use standard or easily computed critical values

Our proposals

- construct a vector of **MSPE differences adjusted** as in Clark and West (2006, 2007) to center vector at zero under the null
- compute a variance-covariance matrix for the vector of MSPE differences
- conduct inference via either of the following two options
  1. "max t-stat (adj.)": inference on the largest of the m adjusted t-statistics that compare null model one by one to each of the m larger models **via the distribution of the maximum of correlated normals**;
  2. "χ^2 (adj.)": inference via the usual χ^2 statistic

Key features:

- we take estimation uncertainty into account
- we use standard or easily computed critical values

Our proposals

- construct a vector of **MSPE differences adjusted** as in Clark and West (2006, 2007) to center vector at zero under the null
- compute a variance-covariance matrix for the vector of MSPE differences
- conduct inference via either of the following two options
  1. "max t-stat (adj.)": inference on the largest of the m adjusted t-statistics that compare null model one by one to each of the m larger models **via the distribution of the maximum of correlated normals**;
  2. "χ^2 (adj.)": inference via the usual χ^2 statistic

Key features:

- we take estimation uncertainty into account
- we use standard or easily computed critical values

Our proposals

- construct a vector of **MSPE differences adjusted** as in Clark and West (2006, 2007) to center vector at zero under the null
- compute a variance-covariance matrix for the vector of MSPE differences
- conduct inference via either of the following two options
  1. "max t-stat (adj.)": inference on the largest of the m adjusted t-statistics that compare null model one by one to each of the m larger models **via the distribution of the maximum of correlated normals**;
  2. "χ^2 (adj.)": inference via the usual χ^2 statistic

Key features:

- we take estimation uncertainty into account
- we use standard or easily computed critical values

Our proposals

- construct a vector of **MSPE differences adjusted** as in Clark and West (2006, 2007) to center vector at zero under the null
- compute a variance-covariance matrix for the vector of MSPE differences
- conduct inference via either of the following two options
  1. "max t-stat (adj.)": inference on the largest of the m adjusted t-statistics that compare null model one by one to each of the m larger models **via the distribution of the maximum of correlated normals**;
  2. "χ^2 (adj.)": inference via the usual χ^2 statistic

Key features:

- we take estimation uncertainty into account
- we use standard or easily computed critical values

Our proposals

- construct a vector of **MSPE differences adjusted** as in Clark and West (2006, 2007) to center vector at zero under the null
- compute a variance-covariance matrix for the vector of MSPE differences
- conduct inference via either of the following two options
  1. "max t-stat (adj.)": inference on the largest of the m adjusted t-statistics that compare null model one by one to each of the m larger models **via the distribution of the maximum of correlated normals**;
  2. "χ^2 (adj.)": inference via the usual χ^2 statistic

Key features:

- we take estimation uncertainty into account
- we use standard or easily computed critical values

1. Introduction
2. Proposed procedures: max t-stat (adjusted), adjusted χ^2
3. Simulation results
4. Empirical example
5. Conclusions

MSPE-adjusted t-stats: Intuition

Intuition

(for pairwise comparison, generalises to comparison of model sets)

- Consider first a comparison of a parsimonious model to a single larger model (m=1, in our terminology); Example:
  "model 0": $y_t = \beta_0 + \beta_1 y_{t-1} + e_{0t}$
  "model 1": $y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 z_{t-1} + e_{1t}$
- $H_0 : \sigma_0^2 - \sigma_1^2 = 0$, $H_A = \sigma_0^2 - \sigma_1^2 > 0$
- under the null of equal forecast accuracy attempt to estimate parameters whose population values are zero
  $\Rightarrow$ will inflate variance of larger model
  $\Rightarrow$ MSPE of the null model will be strictly smaller than that of the larger model

MSPE-adjusted t-stats: Intuition

Intuition

(for pairwise comparison, generalises to comparison of model sets)

- Consider first a comparison of a parsimonious model to a single larger model (m=1, in our terminology); Example:
  "model 0": $y_t = \beta_0 + \beta_1 y_{t-1} + e_{0t}$
  "model 1": $y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 z_{t-1} + e_{1t}$
- $H_0 : \sigma_0^2 - \sigma_1^2 = 0$, $H_A = \sigma_0^2 - \sigma_1^2 > 0$
- under the null of equal forecast accuracy attempt to estimate parameters whose population values are zero
  $\Rightarrow$ will inflate variance of larger model
  $\Rightarrow$ MSPE of the null model will be strictly smaller than that of the larger model

MSPE-adjusted t-stats: Intuition

Intuition

(for pairwise comparison, generalises to comparison of model sets)

- Consider first a comparison of a parsimonious model to a single larger model (m=1, in our terminology); Example:
  "model 0": $y_t = \beta_0 + \beta_1 y_{t-1} + e_{0t}$
  "model 1": $y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 z_{t-1} + e_{1t}$
- $H_0 : \sigma_0^2 - \sigma_1^2 = 0$, $H_A = \sigma_0^2 - \sigma_1^2 > 0$
- under the null of equal forecast accuracy attempt to estimate parameters whose population values are zero
  $\Rightarrow$ will inflate variance of larger model
  $\Rightarrow$ MSPE of the null model will be strictly smaller than that of the larger model

MSPE-adjusted t-stats: Intuition

Intuition

(for pairwise comparison, generalises to comparison of model sets)

- Consider first a comparison of a parsimonious model to a single larger model (m=1, in our terminology); Example:
  "model 0": $y_t = \beta_0 + \beta_1 y_{t-1} + e_{0t}$
  "model 1": $y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 z_{t-1} + e_{1t}$
- $H_0 : \sigma_0^2 - \sigma_1^2 = 0$, $H_A = \sigma_0^2 - \sigma_1^2 > 0$
- under the null of equal forecast accuracy attempt to estimate parameters whose population values are zero
  $\Rightarrow$ will inflate variance of larger model
  $\Rightarrow$ MSPE of the null model will be strictly smaller than that of the larger model

MSPE-adjusted t-stats: Intuition

Intuition

(for pairwise comparison, generalises to comparison of model sets)

- Consider first a comparison of a parsimonious model to a single larger model (m=1, in our terminology); Example:
  "model 0": $y_t = \beta_0 + \beta_1 y_{t-1} + e_{0t}$
  "model 1": $y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 z_{t-1} + e_{1t}$
- $H_0 : \sigma_0^2 - \sigma_1^2 = 0$, $H_A = \sigma_0^2 - \sigma_1^2 > 0$
- under the null of equal forecast accuracy attempt to estimate parameters whose population values are zero
  $\Rightarrow$ will inflate variance of larger model
  $\Rightarrow$ MSPE of the null model will be strictly smaller than that of the larger model

MSPE-adjusted t-stats: Intuition

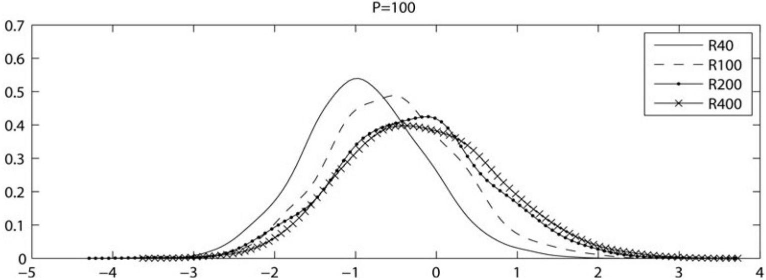
Intuition

(for pairwise comparison, generalises to comparison of model sets)

- Consider first a comparison of a parsimonious model to a single larger model (m=1, in our terminology); Example:
  "model 0": $y_t = \beta_0 + \beta_1 y_{t-1} + e_{0t}$
  "model 1": $y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 z_{t-1} + e_{1t}$
- $H_0 : \sigma_0^2 - \sigma_1^2 = 0,\ H_A = \sigma_0^2 - \sigma_1^2 > 0$
- under the null of equal forecast accuracy attempt to estimate parameters whose population values are zero
  $\Rightarrow$ will inflate variance of larger model
  $\Rightarrow$ MSPE of the null model will be strictly smaller than that of the larger model

MSPE-adjusted t-stats: Intuition

P = number of predictions and prediction errors
R = size of rolling regression sample used to estimate parameter
Smoothed density estimates of
sample MSPE(null) - sample MSPE(alternative) ($\hat{\sigma}_0^2$ - $\hat{\sigma}_1^2$)
across 1000 simulations:



Adjustment Clark and West, Pairwise model comparison

Clark and West (2006, 2007)

- magnitude of the downward shift is estimable; larger the larger number of extraneous regressors in the alternative model
- when comparing a parsimonious model to one other model
  - adjust difference in MSPEs by estimated downward shift
  - after adjustment, conduct inference in standard Diebold-Mariano-West (DMW) fashion
- Denote Clark-West adjusted MSPE as $\bar{f}_1 = \hat{\sigma}_0^2 - (\hat{\sigma}_1^2 - adj.) = \hat{\sigma}_0^2 - \hat{\sigma}_1^2 + P^{-1}\Sigma_t(\hat{y}_{0,t+1} - \hat{y}_{1,t+1})^2$
  - $\hat{\sigma}_0^2$, $\hat{\sigma}_1^2$ = sample MSPE from null and alternative model
  - $\hat{y}_{0,t+1}$, $\hat{y}_{1,t+1}$ = one step ahead forecasts
  - Corresponding t-statistic is $P^{1/2}\bar{f}_1/\sqrt{\hat{\nu}_1}$
- Clark and West: conduct inference via $P^{1/2}\bar{f}_1/\sqrt{\hat{\nu}_1} \sim N(0,1)$

Adjustment is intended to produce test statistics with good size

Adjustment Clark and West, Pairwise model comparison

Clark and West (2006, 2007)

- magnitude of the downward shift is estimable; larger the larger number of extraneous regressors in the alternative model
- when comparing a parsimonious model to one other model
  - adjust difference in MSPEs by estimated downward shift
  - after adjustment, conduct inference in standard Diebold-Mariano-West (DMW) fashion
- Denote Clark-West adjusted MSPE as $\bar{f}_1 = \hat{\sigma}_0^2 - (\hat{\sigma}_1^2 - adj.) = \hat{\sigma}_0^2 - \hat{\sigma}_1^2 + P^{-1}\Sigma_t(\hat{y}_{0,t+1} - \hat{y}_{1,t+1})^2$
  - $\hat{\sigma}_0^2$, $\hat{\sigma}_1^2$ = sample MSPE from null and alternative model
  - $\hat{y}_{0,t+1}$, $\hat{y}_{1,t+1}$ = one step ahead forecasts
  - Corresponding t-statistic is $P^{1/2}\bar{f}_1/\sqrt{\hat{\nu}_1}$
- Clark and West: conduct inference via $P^{1/2}\bar{f}_1/\sqrt{\hat{\nu}_1} \sim N(0,1)$

Adjustment is intended to produce test statistics with good size

Adjustment Clark and West, Pairwise model comparison

Clark and West (2006, 2007)

- magnitude of the downward shift is estimable; larger the larger number of extraneous regressors in the alternative model
- when comparing a parsimonious model to one other model
  - adjust difference in MSPEs by estimated downward shift
  - after adjustment, conduct inference in standard Diebold-Mariano-West (DMW) fashion
- Denote Clark-West adjusted MSPE as $\bar{f}_1 = \hat{\sigma}_0^2 - (\hat{\sigma}_1^2 - adj.) = \hat{\sigma}_0^2 - \hat{\sigma}_1^2 + P^{-1}\Sigma_t(\hat{y}_{0,t+1} - \hat{y}_{1,t+1})^2$
  - $\hat{\sigma}_0^2$, $\hat{\sigma}_1^2$ = sample MSPE from null and alternative model
  - $\hat{y}_{0,t+1}$, $\hat{y}_{1,t+1}$ = one step ahead forecasts
  - Corresponding t-statistic is $P^{1/2}\bar{f}_1/\sqrt{\hat{\nu}_1}$
- Clark and West: conduct inference via $P^{1/2}\bar{f}_1/\sqrt{\hat{\nu}_1} \sim N(0,1)$

Adjustment is intended to produce test statistics with good size

Adjustment Clark and West, Pairwise model comparison

Clark and West (2006, 2007)

- magnitude of the downward shift is estimable; larger the larger number of extraneous regressors in the alternative model
- when comparing a parsimonious model to one other model
  - adjust difference in MSPEs by estimated downward shift
  - after adjustment, conduct inference in standard Diebold-Mariano-West (DMW) fashion
- Denote Clark-West adjusted MSPE as $\bar{f}_1 = \hat{\sigma}_0^2 - (\hat{\sigma}_1^2 - adj.) = \hat{\sigma}_0^2 - \hat{\sigma}_1^2 + P^{-1}\Sigma_t(\hat{y}_{0,t+1} - \hat{y}_{1,t+1})^2$
  - $\hat{\sigma}_0^2$, $\hat{\sigma}_1^2$ = sample MSPE from null and alternative model
  - $\hat{y}_{0,t+1}$, $\hat{y}_{1,t+1}$ = one step ahead forecasts
  - Corresponding t-statistic is $P^{1/2}\bar{f}_1/\sqrt{\hat{v}_1}$
- Clark and West: conduct inference via $P^{1/2}\bar{f}_1/\sqrt{\hat{v}_1} \sim N(0,1)$

Adjustment is intended to produce test statistics with good size

Adjustment Clark and West, Pairwise model comparison

Clark and West (2006, 2007)

- magnitude of the downward shift is estimable; larger the larger number of extraneous regressors in the alternative model
- when comparing a parsimonious model to one other model
  - adjust difference in MSPEs by estimated downward shift
  - after adjustment, conduct inference in standard Diebold-Mariano-West (DMW) fashion
- Denote Clark-West adjusted MSPE as $\bar{f}_1 = \hat{\sigma}_0^2 - (\hat{\sigma}_1^2 - adj.) = \hat{\sigma}_0^2 - \hat{\sigma}_1^2 + P^{-1}\Sigma_t(\hat{y}_{0,t+1} - \hat{y}_{1,t+1})^2$
  - $\hat{\sigma}_0^2$, $\hat{\sigma}_1^2$ = sample MSPE from null and alternative model
  - $\hat{y}_{0,t+1}$, $\hat{y}_{1,t+1}$ = one step ahead forecasts
  - Corresponding t-statistic is $P^{1/2}\bar{f}_1/\sqrt{\hat{\nu}_1}$
- Clark and West: conduct inference via $P^{1/2}\bar{f}_1/\sqrt{\hat{\nu}_1} \sim N(0,1)$

Adjustment is intended to produce test statistics with good size

Adjustment Clark and West, Pairwise model comparison

Clark and West (2006, 2007)

- magnitude of the downward shift is estimable; larger the larger number of extraneous regressors in the alternative model
- when comparing a parsimonious model to one other model
  - adjust difference in MSPEs by estimated downward shift
  - after adjustment, conduct inference in standard Diebold-Mariano-West (DMW) fashion
- Denote Clark-West adjusted MSPE as
  $\bar{f}_1 = \hat{\sigma}_0^2 - (\hat{\sigma}_1^2 - adj.) = \hat{\sigma}_0^2 - \hat{\sigma}_1^2 + P^{-1}\Sigma_t(\hat{y}_{0,t+1} - \hat{y}_{1,t+1})^2$
  - $\hat{\sigma}_0^2$, $\hat{\sigma}_1^2$ = sample MSPE from null and alternative model
  - $\hat{y}_{0,t+1}$, $\hat{y}_{1,t+1}$ = one step ahead forecasts
  - Corresponding t-statistic is $P^{1/2}\bar{f}_1/\sqrt{\hat{\nu}_1}$
- Clark and West: conduct inference via $P^{1/2}\bar{f}_1/\sqrt{\hat{\nu}_1} \sim N(0,1)$

Adjustment is intended to produce test statistics with good size

Adjustment Clark and West, Pairwise model comparison

Clark and West (2006, 2007)

- magnitude of the downward shift is estimable; larger the larger number of extraneous regressors in the alternative model
- when comparing a parsimonious model to one other model
  - adjust difference in MSPEs by estimated downward shift
  - after adjustment, conduct inference in standard Diebold-Mariano-West (DMW) fashion
- Denote Clark-West adjusted MSPE as
  $\bar{f}_1 = \hat{\sigma}_0^2 - (\hat{\sigma}_1^2 - adj.) = \hat{\sigma}_0^2 - \hat{\sigma}_1^2 + P^{-1}\Sigma_t(\hat{y}_{0,t+1} - \hat{y}_{1,t+1})^2$
  - $\hat{\sigma}_0^2$, $\hat{\sigma}_1^2$ = sample MSPE from null and alternative model
  - $\hat{y}_{0,t+1}$, $\hat{y}_{1,t+1}$ = one step ahead forecasts
  - Corresponding t-statistic is $P^{1/2}\bar{f}_1/\sqrt{\hat{v}_1}$
- Clark and West: conduct inference via $P^{1/2}\bar{f}_1/\sqrt{\hat{v}_1} \sim N(0,1)$

Adjustment is intended to produce test statistics with good size

Adjustment Clark and West, Pairwise model comparison

Clark and West (2006, 2007)

- magnitude of the downward shift is estimable; larger the larger number of extraneous regressors in the alternative model
- when comparing a parsimonious model to one other model
  - adjust difference in MSPEs by estimated downward shift
  - after adjustment, conduct inference in standard Diebold-Mariano-West (DMW) fashion
- Denote Clark-West adjusted MSPE as
  $\bar{f}_1 = \hat{\sigma}_0^2 - (\hat{\sigma}_1^2 - adj.) = \hat{\sigma}_0^2 - \hat{\sigma}_1^2 + P^{-1}\Sigma_t(\hat{y}_{0,t+1} - \hat{y}_{1,t+1})^2$
  - $\hat{\sigma}_0^2$, $\hat{\sigma}_1^2$ = sample MSPE from null and alternative model
  - $\hat{y}_{0,t+1}$, $\hat{y}_{1,t+1}$ = one step ahead forecasts
  - Corresponding t-statistic is $P^{1/2}\bar{f}_1/\sqrt{\hat{\nu}_1}$
- Clark and West: conduct inference via $P^{1/2}\bar{f}_1/\sqrt{\hat{\nu}_1} \sim N(0,1)$

Adjustment is intended to produce test statistics with good size

Adjustment Clark and West, Pairwise model comparison

Clark and West (2006, 2007)

- magnitude of the downward shift is estimable; larger the larger number of extraneous regressors in the alternative model
- when comparing a parsimonious model to one other model
  - adjust difference in MSPEs by estimated downward shift
  - after adjustment, conduct inference in standard Diebold-Mariano-West (DMW) fashion
- Denote Clark-West adjusted MSPE as
  $\bar{f}_1 = \hat{\sigma}_0^2 - (\hat{\sigma}_1^2 - adj.) = \hat{\sigma}_0^2 - \hat{\sigma}_1^2 + P^{-1}\Sigma_t(\hat{y}_{0,t+1} - \hat{y}_{1,t+1})^2$
  - $\hat{\sigma}_0^2$, $\hat{\sigma}_1^2$ = sample MSPE from null and alternative model
  - $\hat{y}_{0,t+1}$, $\hat{y}_{1,t+1}$ = one step ahead forecasts
  - Corresponding t-statistic is $P^{1/2}\bar{f}_1/\sqrt{\hat{\nu}_1}$
- Clark and West: conduct inference via $P^{1/2}\bar{f}_1/\sqrt{\hat{\nu}_1} \sim N(0,1)$

Adjustment is intended to produce test statistics with good size

Adjustment Clark and West, Pairwise model comparison

Clark and West (2006, 2007)

- magnitude of the downward shift is estimable; larger the larger number of extraneous regressors in the alternative model
- when comparing a parsimonious model to one other model
  - adjust difference in MSPEs by estimated downward shift
  - after adjustment, conduct inference in standard Diebold-Mariano-West (DMW) fashion
- Denote Clark-West adjusted MSPE as
  $\bar{f}_1 = \hat{\sigma}_0^2 - (\hat{\sigma}_1^2 - adj.) = \hat{\sigma}_0^2 - \hat{\sigma}_1^2 + P^{-1}\Sigma_t(\hat{y}_{0,t+1} - \hat{y}_{1,t+1})^2$
  - $\hat{\sigma}_0^2$, $\hat{\sigma}_1^2$ = sample MSPE from null and alternative model
  - $\hat{y}_{0,t+1}$, $\hat{y}_{1,t+1}$ = one step ahead forecasts
  - Corresponding t-statistic is $P^{1/2}\bar{f}_1/\sqrt{\hat{\nu}_1}$
- Clark and West: conduct inference via $P^{1/2}\bar{f}_1/\sqrt{\hat{\nu}_1} \sim N(0,1)$

Adjustment is intended to produce test statistics with good size

Adjustment Clark and West, Pairwise model comparison

Clark and West (2006, 2007)

- magnitude of the downward shift is estimable; larger the larger number of extraneous regressors in the alternative model
- when comparing a parsimonious model to one other model
  - adjust difference in MSPEs by estimated downward shift
  - after adjustment, conduct inference in standard Diebold-Mariano-West (DMW) fashion
- Denote Clark-West adjusted MSPE as
  $\bar{f}_1 = \hat{\sigma}_0^2 - (\hat{\sigma}_1^2 - adj.) = \hat{\sigma}_0^2 - \hat{\sigma}_1^2 + P^{-1}\Sigma_t(\hat{y}_{0,t+1} - \hat{y}_{1,t+1})^2$
  - $\hat{\sigma}_0^2$, $\hat{\sigma}_1^2$ = sample MSPE from null and alternative model
  - $\hat{y}_{0,t+1}$, $\hat{y}_{1,t+1}$ = one step ahead forecasts
  - Corresponding t-statistic is $P^{1/2}\bar{f}_1/\sqrt{\hat{\nu}_1}$
- Clark and West: conduct inference via $P^{1/2}\bar{f}_1/\sqrt{\hat{\nu}_1} \sim N(0,1)$

Adjustment is intended to produce test statistics with good size

Adjustment Clark and West, Pairwise model comparison

Clark and West (2006, 2007)

- magnitude of the downward shift is estimable; larger the larger number of extraneous regressors in the alternative model
- when comparing a parsimonious model to one other model
  - adjust difference in MSPEs by estimated downward shift
  - after adjustment, conduct inference in standard Diebold-Mariano-West (DMW) fashion
- Denote Clark-West adjusted MSPE as
  $\bar{f}_1 = \hat{\sigma}_0^2 - (\hat{\sigma}_1^2 - adj.) = \hat{\sigma}_0^2 - \hat{\sigma}_1^2 + P^{-1}\Sigma_t(\hat{y}_{0,t+1} - \hat{y}_{1,t+1})^2$
  - $\hat{\sigma}_0^2$, $\hat{\sigma}_1^2$ = sample MSPE from null and alternative model
  - $\hat{y}_{0,t+1}$, $\hat{y}_{1,t+1}$ = one step ahead forecasts
  - Corresponding t-statistic is $P^{1/2}\bar{f}_1/\sqrt{\hat{\nu}_1}$
- Clark and West: conduct inference via $P^{1/2}\bar{f}_1/\sqrt{\hat{\nu}_1} \sim N(0,1)$

Adjustment is intended to produce test statistics with good size

"Max t-stat (adjusted)" statistic

Our first proposal for comparison of model sets

- Compute adjusted t-statistic for each of the m model comparisons; Suppose $m = 2$ alternative models for simplicity. We propose basing inference on

$$P^{1/2}\begin{pmatrix}\frac{\bar{f}_1}{\sqrt{\hat{\nu}_1}}\\ \frac{\bar{f}_2}{\sqrt{\hat{\nu}_2}}\end{pmatrix} \sim_A N(0,\Omega),\ \Omega = \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}.$$

- Here, ρ is the correlation between the two t-statistics.

"Max t-stat (adjusted)" statistic

Our first proposal for comparison of model sets

- Compute adjusted t-statistic for each of the m model comparisons; Suppose $m = 2$ alternative models for simplicity. We propose basing inference on

$$P^{1/2}\begin{pmatrix} \frac{\bar{f}_1}{\sqrt{\hat{\nu}_1}} \\ \frac{\bar{f}_2}{\sqrt{\hat{\nu}_2}} \end{pmatrix} \sim_A N(0,\Omega),\ \Omega = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

- Here, ρ is the correlation between the two t-statistics.

"Max t-stat (adjusted)" statistic

Our first proposal for comparison of model sets

- Compute adjusted t-statistic for each of the m model comparisons; Suppose $m = 2$ alternative models for simplicity. We propose basing inference on

$$P^{1/2}\begin{pmatrix} \frac{\bar{f}_1}{\sqrt{\hat{\nu}_1}} \\ \frac{\bar{f}_2}{\sqrt{\hat{\nu}_2}} \end{pmatrix} \sim_A N(0, \Omega),\ \Omega = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

- Here, ρ is the correlation between the two t-statistics.

"Max t-stat (adjusted)" statistic

Inference via the distribution of the maximum of correlated normals

- Let $\hat{z}$ be the larger of the two t-statistics

$$\hat{z} = max(P^{1/2}\bar{f}_1/\sqrt{\hat{\nu}_1},\ \ \bar{f}_2/\sqrt{\hat{\nu}_2}) = \text{max t-stat (adj.)}.$$

- Reject the null when $\hat{z}$ is sufficiently large (one-tailed test).
- Critical values for m=2:

	ρ							
	1	0.8	0.6	0.4	0.2	0	-0.2	-1
size=5 %	1.645	1.846	1.900	1.929	1.946	1.955	1.959	1.960
size=10 %	1.282	1.493	1.556	1.594	1.617	1.632	1.640	1.645

"Max t-stat (adjusted)" statistic

Inference via the distribution of the maximum of correlated normals

- Let $\hat{z}$ be the larger of the two t-statistics

$$\hat{z} = max(P^{1/2}\bar{f}_1/\sqrt{\hat{\nu}_1},\ \ \bar{f}_2/\sqrt{\hat{\nu}_2}) = \text{max t-stat (adj.)}.$$

- Reject the null when $\hat{z}$ is sufficiently large (one-tailed test).
- Critical values for m=2:

	ρ							
	1	0.8	0.6	0.4	0.2	0	-0.2	-1
size=5 %	1.645	1.846	1.900	1.929	1.946	1.955	1.959	1.960
size=10 %	1.282	1.493	1.556	1.594	1.617	1.632	1.640	1.645

"Max t-stat (adjusted)" statistic

Inference via the distribution of the maximum of correlated normals

- Let $\hat{z}$ be the larger of the two t-statistics

$$\hat{z} = max(P^{1/2}\bar{f}_1/\sqrt{\hat{\nu}_1},\ \ \bar{f}_2/\sqrt{\hat{\nu}_2}) = \text{max t-stat (adj.)}.$$

- Reject the null when $\hat{z}$ is sufficiently large (one-tailed test).
- Critical values for m=2:

	ρ							
	1	0.8	0.6	0.4	0.2	0	-0.2	-1
size=5 %	1.645	1.846	1.900	1.929	1.946	1.955	1.959	1.960
size=10 %	1.282	1.493	1.556	1.594	1.617	1.632	1.640	1.645

"Max t-stat (adjusted)" statistic

Inference via the distribution of the maximum of correlated normals

- Let $\hat{z}$ be the larger of the two t-statistics

$$\hat{z} = max(P^{1/2}\bar{f}_1/\sqrt{\hat{\nu}_1}, \ \ \bar{f}_2/\sqrt{\hat{\nu}_2}) = \text{max t-stat (adj.)}.$$

- Reject the null when $\hat{z}$ is sufficiently large (one-tailed test).
- Critical values for m=2:

	ρ							
	1	0.8	0.6	0.4	0.2	0	-0.2	-1
size=5 %	1.645	1.846	1.900	1.929	1.946	1.955	1.959	1.960
size=10 %	1.282	1.493	1.556	1.594	1.617	1.632	1.640	1.645

"Max t-stat (adjusted)" statistic

Inference via the distribution of the maximum of correlated normals

- Let $\hat{z}$ be the larger of the two t-statistics

$$\hat{z} = max(P^{1/2}\bar{f}_1/\sqrt{\hat{\nu}_1},\;\; \bar{f}_2/\sqrt{\hat{\nu}_2}) = \text{max t-stat (adj.)}.$$

- Reject the null when $\hat{z}$ is sufficiently large (one-tailed test).
- Critical values for m=2:

	ρ							
	1	0.8	0.6	0.4	0.2	0	-0.2	-1
size=5 %	1.645	1.846	1.900	1.929	1.946	1.955	1.959	1.960
size=10 %	1.282	1.493	1.556	1.594	1.617	1.632	1.640	1.645

"Max t-stat (adjusted)" statistic: critical values

- critical values in tables obtained by numerically integrating the relevant density
- More generally, for arbitrary $m > 1$, one can obtain a p-value from a set of draws from a normal distribution:
  - compute m MSPE-adjusted t-statistics, each of which compares benchmark model to one of the m larger models
  - compute $\hat{\Omega}$: mxm sample correlation matrix of m t-statistics
  - do (say) 50,000 draws on $mx1$ vector $\sim N(0, \hat{\Omega})$, saving the maximum value from each draw
  - use the distribution of 50,000 maxima to compute the p-value of the largest of the m MSPE-adjusted t-statistics computed from the actual data
- Whatever the method of obtaining the critical value, we call this procedure "max t-stat (adj.)"

"Max t-stat (adjusted)" statistic: critical values

- critical values in tables obtained by numerically integrating the relevant density
- More generally, for arbitrary $m > 1$, one can obtain a p-value from a set of draws from a normal distribution:
  - compute m MSPE-adjusted t-statistics, each of which compares benchmark model to one of the m larger models
  - compute $\hat{\Omega}$: mxm sample correlation matrix of m t-statistics
  - do (say) 50,000 draws on $mx1$ vector $\sim N(0, \hat{\Omega})$, saving the maximum value from each draw
  - use the distribution of 50,000 maxima to compute the p-value of the largest of the m MSPE-adjusted t-statistics computed from the actual data
- Whatever the method of obtaining the critical value, we call this procedure "max t-stat (adj.)"

"Max t-stat (adjusted)" statistic: critical values

- critical values in tables obtained by numerically integrating the relevant density
- More generally, for arbitrary $m > 1$, one can obtain a p-value from a set of draws from a normal distribution:
  - compute m MSPE-adjusted t-statistics, each of which compares benchmark model to one of the m larger models
  - compute $\hat{\Omega}$: mxm sample correlation matrix of m t-statistics
  - do (say) 50,000 draws on $mx1$ vector $\sim N(0, \hat{\Omega})$, saving the maximum value from each draw
  - use the distribution of 50,000 maxima to compute the p-value of the largest of the m MSPE-adjusted t-statistics computed from the actual data
- Whatever the method of obtaining the critical value, we call this procedure "max t-stat (adj.)"

"Max t-stat (adjusted)" statistic: critical values

- critical values in tables obtained by numerically integrating the relevant density
- More generally, for arbitrary $m > 1$, one can obtain a p-value from a set of draws from a normal distribution:
  - compute m MSPE-adjusted t-statistics, each of which compares benchmark model to one of the m larger models
  - compute $\hat{\Omega}$: mxm sample correlation matrix of m t-statistics
  - do (say) 50,000 draws on $mx1$ vector $\sim N(0, \hat{\Omega})$, saving the maximum value from each draw
  - use the distribution of 50,000 maxima to compute the p-value of the largest of the m MSPE-adjusted t-statistics computed from the actual data
- Whatever the method of obtaining the critical value, we call this procedure "max t-stat (adj.)"

χ^2 statistic (adjusted)

Second proposal for comparison of model sets

- compute m adjusted differences in MSPEs
- compute variance or long run variance of $mx1$ vector of adjusted differences
- compute usual quadratic form, use critical values from $\chi^2(m)$
- "χ^2 (adj.)" $= P\bar{f}'\hat{V}^{-1}\bar{f}$

with $\bar{f} = \hat{\sigma}_0^2 - (\hat{\sigma}_i^2 - (adj.))$ for $i = 1, ..., m$

$\Rightarrow$ possible sacrifice in power given the one-sided nature of the test

χ^2 statistic (adjusted)

Second proposal for comparison of model sets

- compute m adjusted differences in MSPEs
- compute variance or long run variance of $mx1$ vector of adjusted differences
- compute usual quadratic form, use critical values from $\chi^2(m)$
- "χ^2 (adj.)" $= P\bar{f}'\hat{V}^{-1}\bar{f}$

with $\bar{f} = \hat{\sigma}_0^2 - (\hat{\sigma}_i^2 - (adj.))$ for $i = 1, ..., m$

$\Rightarrow$ possible sacrifice in power given the one-sided nature of the test

χ^2 statistic (adjusted)

Second proposal for comparison of model sets

- compute m adjusted differences in MSPEs
- compute variance or long run variance of $mx1$ vector of adjusted differences
- compute usual quadratic form, use critical values from $\chi^2(m)$
- "χ^2 (adj.)" $= P\bar{f}'\hat{V}^{-1}\bar{f}$

with $\bar{f} = \hat{\sigma}_0^2 - (\hat{\sigma}_i^2 - (adj.))$ for $i = 1, ..., m$

$\Rightarrow$ possible sacrifice in power given the one-sided nature of the test

χ^2 statistic (adjusted)

Second proposal for comparison of model sets

- compute m adjusted differences in MSPEs
- compute variance or long run variance of $mx1$ vector of adjusted differences
- compute usual quadratic form, use critical values from $\chi^2(m)$
- "χ^2 (adj.)" $= P\bar{f}'\hat{V}^{-1}\bar{f}$

  with $\bar{f} = \hat{\sigma}_0^2 - (\hat{\sigma}_i^2 - (adj.))$ for $i = 1, ..., m$

  $\Rightarrow$ possible sacrifice in power given the one-sided nature of the test

χ^2 statistic (adjusted)

Second proposal for comparison of model sets

- compute m adjusted differences in MSPEs
- compute variance or long run variance of $mx1$ vector of adjusted differences
- compute usual quadratic form, use critical values from $\chi^2(m)$
- "χ^2 (adj.)" $= P\bar{f}'\hat{V}^{-1}\bar{f}$

with $\bar{f} = \hat{\sigma}_0^2 - (\hat{\sigma}_i^2 - (adj.))$ for $i = 1, ..., m$

$\Rightarrow$ possible sacrifice in power given the one-sided nature of the test

χ^2 statistic (adjusted)

Second proposal for comparison of model sets

- compute m adjusted differences in MSPEs
- compute variance or long run variance of $mx1$ vector of adjusted differences
- compute usual quadratic form, use critical values from $\chi^2(m)$
- "χ^2 (adj.)" $= P\bar{f}'\hat{V}^{-1}\bar{f}$

  with $\bar{f} = \hat{\sigma}_0^2 - (\hat{\sigma}_i^2 - (adj.))$ for $i = 1, ..., m$

  $\Rightarrow$ possible sacrifice in power given the one-sided nature of the test

Theoretical justification for our procedures

- asymptotic validity of use of normal distribution follows from conditions such as in Giacomini and White (2007):
  $\Rightarrow$ rolling windows (asymptotics: R fixed, P grows); null model: white noise; also for multistep forecasts if direct method used
- approximate asymptotic validity of normal distribution for pairwise model comparison (m=1) follows from Clark and McCracken (2001, 2005) asymptotics:
  $\Rightarrow$ rolling and recursive windows (asymptotics: P grows at slower rate than R, i.e. $P/R \rightarrow 0$ as T grows); forecasts from null model rely on estimated regression parameters; generally for one-step ahead forecasts, for multistep: depends on DGP
  - we conjecture approximate asymptotic validity for $m > 1$ (see paper for more details)

Theoretical justification for our procedures

- asymptotic validity of use of normal distribution follows from conditions such as in Giacomini and White (2007):
  $\Rightarrow$ rolling windows (asymptotics: R fixed, P grows); null model: white noise; also for multistep forecasts if direct method used
- approximate asymptotic validity of normal distribution for pairwise model comparison (m=1) follows from Clark and McCracken (2001, 2005) asymptotics:
  $\Rightarrow$ rolling and recursive windows (asymptotics: P grows at slower rate than R, i.e. $P/R \rightarrow 0$ as T grows); forecasts from null model rely on estimated regression parameters; generally for one-step ahead forecasts, for multistep: depends on DGP
  - we conjecture approximate asymptotic validity for $m > 1$ (see paper for more details)

Theoretical justification for our procedures

- asymptotic validity of use of normal distribution follows from conditions such as in Giacomini and White (2007): $\Rightarrow$ rolling windows (asymptotics: R fixed, P grows); null model: white noise; also for multistep forecasts if direct method used
- approximate asymptotic validity of normal distribution for pairwise model comparison (m=1) follows from Clark and McCracken (2001, 2005) asymptotics:
  $\Rightarrow$ rolling and recursive windows (asymptotics: P grows at slower rate than R, i.e. $P/R \rightarrow 0$ as T grows); forecasts from null model rely on estimated regression parameters; generally for one-step ahead forecasts, for multistep: depends on DGP
  - we conjecture approximate asymptotic validity for $m > 1$ (see paper for more details)

Theoretical justification for our procedures

- asymptotic validity of use of normal distribution follows from conditions such as in Giacomini and White (2007):
  $\Rightarrow$ rolling windows (asymptotics: R fixed, P grows); null model: white noise; also for multistep forecasts if direct method used
- approximate asymptotic validity of normal distribution for pairwise model comparison (m=1) follows from Clark and McCracken (2001, 2005) asymptotics:
  $\Rightarrow$ rolling and recursive windows (asymptotics: P grows at slower rate than R, i.e. $P/R \rightarrow 0$ as T grows); forecasts from null model rely on estimated regression parameters; generally for one-step ahead forecasts, for multistep: depends on DGP
  - we conjecture approximate asymptotic validity for $m > 1$ (see paper for more details)

Theoretical justification for our procedures

- asymptotic validity of use of normal distribution follows from conditions such as in Giacomini and White (2007):
  $\Rightarrow$ rolling windows (asymptotics: R fixed, P grows); null model: white noise; also for multistep forecasts if direct method used
- approximate asymptotic validity of normal distribution for pairwise model comparison (m=1) follows from Clark and McCracken (2001, 2005) asymptotics:
  $\Rightarrow$ rolling and recursive windows (asymptotics: P grows at slower rate than R, i.e. $P/R \rightarrow 0$ as T grows); forecasts from null model rely on estimated regression parameters; generally for one-step ahead forecasts, for multistep: depends on DGP
  - we conjecture approximate asymptotic validity for $m > 1$ (see paper for more details)

- **Note:** Statistic based on adjusted difference in MSPEs is algebraically identical with the encompassing statistic:

$$\begin{aligned}
\hat{f}_{i,t+1} &= \hat{e}^2_{0,t+1} - (\hat{e}^2_{i,t+1} - (\hat{y}_{0,t+1} - \hat{y}_{i,t+1})^2) \\
&= \hat{e}^2_{0,t+1} - \hat{e}^2_{i,t+1} + (\hat{y}_{0,t+1} - \hat{y}_{i,t+1})^2 \\
&= \hat{e}^2_0 - \hat{e}^2_i + (\hat{y}_{0,t+1} + y_{t+1} - y_{t+1} - \hat{y}_{i,t+1})^2 \\
&= 2\hat{e}^2_{0,t+1} - 2\hat{e}_{0,t+1}\hat{e}_{i,t+1} \\
&= 2\hat{e}_{0,t+1}(\hat{e}_{0,t+1} - \hat{e}_{i,t+1})
\end{aligned}$$

$\Rightarrow$ we provide an encompassing test for small model sets, extending pairwise encompassing test literature; we explicitly allow for estimation uncertainty in contrast to Harvey and Newbold (2000);

- **Note:** Statistic based on adjusted difference in MSPEs is algebraically identical with the encompassing statistic:

$$
\begin{aligned}
\hat{f}_{i,t+1} &= \hat{e}_{0,t+1}^2 - (\hat{e}_{i,t+1}^2 - (\hat{y}_{0,t+1} - \hat{y}_{i,t+1})^2) \\
&= \hat{e}_{0,t+1}^2 - \hat{e}_{i,t+1}^2 + (\hat{y}_{0,t+1} - \hat{y}_{i,t+1})^2 \\
&= \hat{e}_0^2 - \hat{e}_i^2 + (\hat{y}_{0,t+1} + y_{t+1} - y_{t+1} - \hat{y}_{i,t+1})^2 \\
&= 2\hat{e}_{0,t+1}^2 - 2\hat{e}_{0,t+1}\hat{e}_{i,t+1} \\
&= 2\hat{e}_{0,t+1}(\hat{e}_{0,t+1} - \hat{e}_{i,t+1})
\end{aligned}
$$

$\Rightarrow$ we provide an encompassing test for small model sets, extending pairwise encompassing test literature; we explicitly allow for estimation uncertainty in contrast to Harvey and Newbold (2000);

- **Note:** Statistic based on adjusted difference in MSPEs is algebraically identical with the encompassing statistic:

$$\begin{aligned}
\hat{f}_{i,t+1} &= \hat{e}^2_{0,t+1} - (\hat{e}^2_{i,t+1} - (\hat{y}_{0,t+1} - \hat{y}_{i,t+1})^2) \\
&= \hat{e}^2_{0,t+1} - \hat{e}^2_{i,t+1} + (\hat{y}_{0,t+1} - \hat{y}_{i,t+1})^2 \\
&= \hat{e}^2_0 - \hat{e}^2_i + (\hat{y}_{0,t+1} + y_{t+1} - y_{t+1} - \hat{y}_{i,t+1})^2 \\
&= 2\hat{e}^2_{0,t+1} - 2\hat{e}_{0,t+1}\hat{e}_{i,t+1} \\
&= 2\hat{e}_{0,t+1}(\hat{e}_{0,t+1} - \hat{e}_{i,t+1})
\end{aligned}$$

$\Rightarrow$ we provide an encompassing test for small model sets, extending pairwise encompassing test literature;
we explicitly allow for estimation uncertainty in contrast to Harvey and Newbold (2000);

Simulation design

Our simulations compare "max t-stat (adj.)" and "$\chi^2(adj.)$" with

- "$\chi^2(unadj.)$" proceeds as does "$\chi^2(adj.)$", but uses raw rather than adjusted differences in MSPEs
- White's (2000) bootstrap reality check

Simulation set-up (macro and finance DGPs):

- Macro DGP: $(y_{1t}, y_{2t}, ..., y_{it})'$ follows a VAR(1)
- Null model $\Rightarrow$ univariate AR(1) in y_t
- m=2 and m=4 alternative models, each of which uses a constant + one lag of y_t + one lag of other variables
- rolling (and recursive) samples; nominal size 10% (5%)

Simulation design

Our simulations compare "max t-stat (adj.)" and "$\chi^2(adj.)$" with

- "$\chi^2(unadj.)$" proceeds as does "$\chi^2(adj.)$", but uses raw rather than adjusted differences in MSPEs
- White's (2000) bootstrap reality check

Simulation set-up (macro and finance DGPs):

- Macro DGP: $(y_{1t}, y_{2t}, ..., y_{it})'$ follows a VAR(1)
- Null model $\Rightarrow$ univariate AR(1) in y_t
- m=2 and m=4 alternative models, each of which uses a constant + one lag of y_t + one lag of other variables
- rolling (and recursive) samples; nominal size 10% (5%)

Simulation design

Our simulations compare "max t-stat (adj.)" and "$\chi^2(adj.)$" with

- "$\chi^2(unadj.)$" proceeds as does "$\chi^2(adj.)$", but uses raw rather than adjusted differences in MSPEs
- White's (2000) bootstrap reality check

Simulation set-up (macro and finance DGPs):

- Macro DGP: $(y_{1t}, y_{2t}, ..., y_{it})'$ follows a VAR(1)
- Null model $\Rightarrow$ univariate AR(1) in y_t
- m=2 and m=4 alternative models, each of which uses a constant + one lag of y_t + one lag of other variables
- rolling (and recursive) samples; nominal size 10% (5%)

Simulation design

Our simulations compare "max t-stat (adj.)" and "$\chi^2(adj.)$" with

- "$\chi^2(unadj.)$" proceeds as does "$\chi^2(adj.)$", but uses raw rather than adjusted differences in MSPEs
- White's (2000) bootstrap reality check

Simulation set-up (macro and finance DGPs):

- Macro DGP: $(y_{1t}, y_{2t}, ..., y_{it})'$ follows a VAR(1)
- Null model $\Rightarrow$ univariate AR(1) in y_t
- m=2 and m=4 alternative models, each of which uses a constant + one lag of y_t + one lag of other variables
- rolling (and recursive) samples; nominal size 10% (5%)

Simulation design

Our simulations compare "max t-stat (adj.)" and "$\chi^2(adj.)$" with

- "$\chi^2(unadj.)$" proceeds as does "$\chi^2(adj.)$", but uses raw rather than adjusted differences in MSPEs
- White's (2000) bootstrap reality check

Simulation set-up (macro and finance DGPs):

- Macro DGP: $(y_{1t}, y_{2t}, ..., y_{it})'$ follows a VAR(1)
- Null model $\Rightarrow$ univariate AR(1) in y_t
- m=2 and m=4 alternative models, each of which uses a constant + one lag of y_t + one lag of other variables
- rolling (and recursive) samples; nominal size 10% (5%)

Simulation design

Our simulations compare "max t-stat (adj.)" and "$\chi^2(adj.)$" with

- "$\chi^2(unadj.)$" proceeds as does "$\chi^2(adj.)$", but uses raw rather than adjusted differences in MSPEs
- White's (2000) bootstrap reality check

Simulation set-up (macro and finance DGPs):

- Macro DGP: $(y_{1t}, y_{2t}, ..., y_{it})'$ follows a VAR(1)
- Null model $\Rightarrow$ univariate AR(1) in y_t
- m=2 and m=4 alternative models, each of which uses a constant + one lag of y_t + one lag of other variables
- rolling (and recursive) samples; nominal size 10% (5%)

Simulation design

Our simulations compare "max t-stat (adj.)" and "$\chi^2(adj.)$" with

- "$\chi^2(unadj.)$" proceeds as does "$\chi^2(adj.)$", but uses raw rather than adjusted differences in MSPEs
- White's (2000) bootstrap reality check

Simulation set-up (macro and finance DGPs):

- Macro DGP: $(y_{1t}, y_{2t}, ..., y_{it})'$ follows a VAR(1)
- Null model $\Rightarrow$ univariate AR(1) in y_t
- m=2 and m=4 alternative models, each of which uses a constant + one lag of y_t + one lag of other variables
- rolling (and recursive) samples; nominal size 10% (5%)

Simulation design

Our simulations compare "max t-stat (adj.)" and "$\chi^2(adj.)$" with

- "$\chi^2(unadj.)$" proceeds as does "$\chi^2(adj.)$", but uses raw rather than adjusted differences in MSPEs
- White's (2000) bootstrap reality check

Simulation set-up (macro and finance DGPs):

- Macro DGP: $(y_{1t}, y_{2t}, ..., y_{it})'$ follows a VAR(1)
- Null model $\Rightarrow$ univariate AR(1) in y_t
- m=2 and m=4 alternative models, each of which uses a constant + one lag of y_t + one lag of other variables
- rolling (and recursive) samples; nominal size 10% (5%)

Simulation results: Size of Tests

Empirical Size of Nominal .10 Tests, 1 Step

		-----	*m*=2		-----
P		*R*=40	*R*=100	*R*=200	*R*=400
40	Max t-stat (adj.)	0.081	0.082	0.085	0.084
	χ^2 (adj.)	0.119	0.138	0.134	0.109
	χ^2 (unadj.)	0.157	0.134	0.137	0.116
	Reality check	0.019	0.039	0.066	0.072
100	Max t-stat (adj.)	0.073	0.058	0.080	0.065
	χ^2 (adj.)	0.112	0.109	0.125	0.129
	χ^2 (unadj.)	0.241	0.147	0.147	0.137
	Reality check	0.001	0.011	0.036	0.047
200	Max t-stat (adj.)	0.100	0.069	0.060	0.043
	χ^2 (adj.)	0.134	0.114	0.098	0.101
	χ^2 (unadj.)	0.416	0.200	0.122	0.127
	Reality check	0.000	0.005	0.018	0.024

Simulation results: Size of Tests

Empirical Size of Nominal .10 Tests, 1 Step Ahead Predictions

		-----	*m*=2		-----	-----	*m*=4		-----
P		*R*=40	*R*=100	*R*=200	*R*=400	*R*=40	*R*=100	*R*=200	*R*=400
40	Max t-stat (adj.)	0.081	0.082	0.085	0.084	0.065	0.072	0.085	0.076
	χ^2 (adj.)	0.119	0.138	0.134	0.109	0.148	0.161	0.185	0.162
	χ^2 (unadj.)	0.157	0.134	0.137	0.116	0.191	0.175	0.187	0.177
	Reality check	0.019	0.039	0.066	0.072	0.013	0.038	0.063	0.068
100	Max t-stat (adj.)	0.073	0.058	0.080	0.065	0.069	0.075	0.063	0.064
	χ^2 (adj.)	0.112	0.109	0.125	0.129	0.114	0.113	0.112	0.133
	χ^2 (unadj.)	0.241	0.147	0.147	0.137	0.299	0.162	0.135	0.134
	Reality check	0.001	0.011	0.036	0.047	0.000	0.017	0.031	0.056
200	Max t-stat (adj.)	0.100	0.069	0.060	0.043	0.084	0.055	0.057	0.062
	χ^2 (adj.)	0.134	0.114	0.098	0.101	0.117	0.091	0.120	0.144
	χ^2 (unadj.)	0.416	0.200	0.122	0.127	0.505	0.210	0.158	0.168
	Reality check	0.000	0.005	0.018	0.024	0.000	0.001	0.011	0.035

Simulation results: Power of Tests

Power

P		----- m=2			-----
		R=40	R=100	R=200	R=400
40	Max t-stat (adj.)	0.648	0.767	0.809	0.832
	χ^2 (adj.)	0.584	0.651	0.703	0.708
	χ^2 (unadj.)	0.177	0.252	0.301	0.298
	Reality check	0.230	0.408	0.478	0.522
100	Max t-stat (adj.)	0.885	0.983	0.987	0.991
	χ^2 (adj.)	0.851	0.954	0.966	0.971
	χ^2 (unadj.)	0.268	0.430	0.519	0.564
	Reality check	0.314	0.658	0.753	0.766
200	Max t-stat (adj.)	0.989	0.997	0.999	1.000
	χ^2 (adj.)	0.986	0.998	0.997	1.000
	χ^2 (unadj.)	0.465	0.743	0.790	0.814
	Reality check	0.483	0.900	0.933	0.944

Simulation results: Power of Tests

Power

P		-----	m=2		-----	-----	m=4		-----
		R=40	R=100	R=200	R=400	R=40	R=100	R=200	R=400
40	Max t-stat (adj.)	0.648	0.767	0.809	0.832	0.422	0.502	0.559	0.567
	χ^2 (adj.)	0.584	0.651	0.703	0.708	0.394	0.474	0.513	0.530
	χ^2 (unadj.)	0.177	0.252	0.301	0.298	0.198	0.244	0.254	0.280
	Reality check	0.230	0.408	0.478	0.522	0.140	0.256	0.342	0.364
100	Max t-stat (adj.)	0.885	0.983	0.987	0.991	0.672	0.831	0.856	0.876
	χ^2 (adj.)	0.851	0.954	0.966	0.971	0.603	0.781	0.803	0.841
	χ^2 (unadj.)	0.268	0.430	0.519	0.564	0.244	0.297	0.359	0.437
	Reality check	0.314	0.658	0.753	0.766	0.171	0.425	0.532	0.578
200	Max t-stat (adj.)	0.989	0.997	0.999	1.000	0.891	0.962	0.981	0.989
	χ^2 (adj.)	0.986	0.998	0.997	1.000	0.851	0.953	0.984	0.988
	χ^2 (unadj.)	0.465	0.743	0.790	0.814	0.424	0.484	0.554	0.604
	Reality check	0.483	0.900	0.933	0.944	0.269	0.664	0.766	0.799

Simulation results: Summary

Size:

- "max t-stat (adj.)" and "χ^2 (adj.)" have comparable size, one slightly undersized, one slightly oversized
- previous proposals have size problems: "χ^2 (unadj.)" slightly to grossly oversized, Reality check grossly undersized

Power:

- "max t-stat (adj.)" best, also better than "χ^2 (adj.)"
- previous procedures rank third (Reality check) and fourth ("χ^2 (unadj.)")

Simulation results: Summary

Size:

- "max t-stat (adj.)" and "χ^2 (adj.)" have comparable size, one slightly undersized, one slightly oversized
- previous proposals have size problems: "χ^2 (unadj.)" slightly to grossly oversized, Reality check grossly undersized

Power:

- "max t-stat (adj.)" best, also better than "χ^2 (adj.)"
- previous procedures rank third (Reality check) and fourth ("χ^2 (unadj.)")

Simulation results: Summary

Size:

- "max t-stat (adj.)" and "χ^2 (adj.)" have comparable size, one slightly undersized, one slightly oversized
- previous proposals have size problems: "χ^2 (unadj.)" slightly to grossly oversized, Reality check grossly undersized

Power:

- "max t-stat (adj.)" best, also better than "χ^2 (adj.)"
- previous procedures rank third (Reality check) and fourth ("χ^2 (unadj.)")

Simulation results: Summary

Size:

- "max t-stat (adj.)" and "χ^2 (adj.)" have comparable size, one slightly undersized, one slightly oversized
- previous proposals have size problems: "χ^2 (unadj.)" slightly to grossly oversized, Reality check grossly undersized

Power:

- "max t-stat (adj.)" best, also better than "χ^2 (adj.)"
- previous procedures rank third (Reality check) and fourth ("χ^2 (unadj.)")

Simulation results: Summary

Size:

- "max t-stat (adj.)" and "χ^2 (adj.)" have comparable size, one slightly undersized, one slightly oversized
- previous proposals have size problems: "χ^2 (unadj.)" slightly to grossly oversized, Reality check grossly undersized

Power:

- "max t-stat (adj.)" best, also better than "χ^2 (adj.)"
- previous procedures rank third (Reality check) and fourth ("χ^2 (unadj.)")

Simulation results: Summary

Size:

- "max t-stat (adj.)" and "χ^2 (adj.)" have comparable size, one slightly undersized, one slightly oversized
- previous proposals have size problems: "χ^2 (unadj.)" slightly to grossly oversized, Reality check grossly undersized

Power:

- "max t-stat (adj.)" best, also better than "χ^2 (adj.)"
- previous procedures rank third (Reality check) and fourth ("χ^2 (unadj.)")

Empirical example: Forecasting Inflation

- predictand = annual U.S. CPI inflation, one month ahead
- two sample periods 1960-1983 and 1960-2004
  - prediction period =1970-1983; R=120, P=168
  - prediction period = 1984-2004; R=288, P=252
- rolling samples (also recursive)
- null model = univariate AR, lag length selected by AIC
- alternatives are bivariate VARs, lag lengths again selected by AIC. Second predictor (in addition to CPI inflation) is
  - component of inflation (food, energy, commodities, services)
  - output growth, change in unemployment

(see Hubrich (2005) and Hendry and Hubrich (2009) for further empirical results on use of disaggregate information in forecasting the aggregate)

Empirical example: Forecasting Inflation

- predictand = annual U.S. CPI inflation, one month ahead
- two sample periods 1960-1983 and 1960-2004
  - prediction period =1970-1983; R=120, P=168
  - prediction period = 1984-2004; R=288, P=252
- rolling samples (also recursive)
- null model = univariate AR, lag length selected by AIC
- alternatives are bivariate VARs, lag lengths again selected by AIC. Second predictor (in addition to CPI inflation) is
  - component of inflation (food, energy, commodities, services)
  - output growth, change in unemployment

(see Hubrich (2005) and Hendry and Hubrich (2009) for further empirical results on use of disaggregate information in forecasting the aggregate)

Empirical example: Forecasting Inflation

- predictand = annual U.S. CPI inflation, one month ahead
- two sample periods 1960-1983 and 1960-2004
    - prediction period =1970-1983; R=120, P=168
    - prediction period = 1984-2004; R=288, P=252
- rolling samples (also recursive)
- null model = univariate AR, lag length selected by AIC
- alternatives are bivariate VARs, lag lengths again selected by AIC. Second predictor (in addition to CPI inflation) is
    - component of inflation (food, energy, commodities, services)
    - output growth, change in unemployment

(see Hubrich (2005) and Hendry and Hubrich (2009) for further empirical results on use of disaggregate information in forecasting the aggregate)

Empirical example: Forecasting Inflation

- predictand = annual U.S. CPI inflation, one month ahead
- two sample periods 1960-1983 and 1960-2004
  - prediction period =1970-1983; R=120, P=168
  - prediction period = 1984-2004; R=288, P=252
- rolling samples (also recursive)
- null model = univariate AR, lag length selected by AIC
- alternatives are bivariate VARs, lag lengths again selected by AIC. Second predictor (in addition to CPI inflation) is
  - component of inflation (food, energy, commodities, services)
  - output growth, change in unemployment

(see Hubrich (2005) and Hendry and Hubrich (2009) for further empirical results on use of disaggregate information in forecasting the aggregate)

Empirical example: Forecasting Inflation

- predictand = annual U.S. CPI inflation, one month ahead
- two sample periods 1960-1983 and 1960-2004
    - prediction period =1970-1983; R=120, P=168
    - prediction period = 1984-2004; R=288, P=252
- rolling samples (also recursive)
- null model = univariate AR, lag length selected by AIC
- alternatives are bivariate VARs, lag lengths again selected by AIC. Second predictor (in addition to CPI inflation) is
    - component of inflation (food, energy, commodities, services)
    - output growth, change in unemployment

(see Hubrich (2005) and Hendry and Hubrich (2009) for further empirical results on use of disaggregate information in forecasting the aggregate)

Empirical example: Forecasting Inflation

- predictand = annual U.S. CPI inflation, one month ahead
- two sample periods 1960-1983 and 1960-2004
  - prediction period =1970-1983; R=120, P=168
  - prediction period = 1984-2004; R=288, P=252
- rolling samples (also recursive)
- null model = univariate AR, lag length selected by AIC
- alternatives are bivariate VARs, lag lengths again selected by AIC. Second predictor (in addition to CPI inflation) is
  - component of inflation (food, energy, commodities, services)
  - output growth, change in unemployment

(see Hubrich (2005) and Hendry and Hubrich (2009) for further empirical results on use of disaggregate information in forecasting the aggregate)

Empirical example: Forecasting Inflation

- predictand = annual U.S. CPI inflation, one month ahead
- two sample periods 1960-1983 and 1960-2004
  - prediction period =1970-1983; R=120, P=168
  - prediction period = 1984-2004; R=288, P=252
- rolling samples (also recursive)
- null model = univariate AR, lag length selected by AIC
- alternatives are bivariate VARs, lag lengths again selected by AIC. Second predictor (in addition to CPI inflation) is
  - component of inflation (food, energy, commodities, services)
  - output growth, change in unemployment

(see Hubrich (2005) and Hendry and Hubrich (2009) for further empirical results on use of disaggregate information in forecasting the aggregate)

Empirical example: Forecasting Inflation

- predictand = annual U.S. CPI inflation, one month ahead
- two sample periods 1960-1983 and 1960-2004
  - prediction period =1970-1983; R=120, P=168
  - prediction period = 1984-2004; R=288, P=252
- rolling samples (also recursive)
- null model = univariate AR, lag length selected by AIC
- alternatives are bivariate VARs, lag lengths again selected by AIC. Second predictor (in addition to CPI inflation) is
  - component of inflation (food, energy, commodities, services)
  - output growth, change in unemployment

(see Hubrich (2005) and Hendry and Hubrich (2009) for further empirical results on use of disaggregate information in forecasting the aggregate)

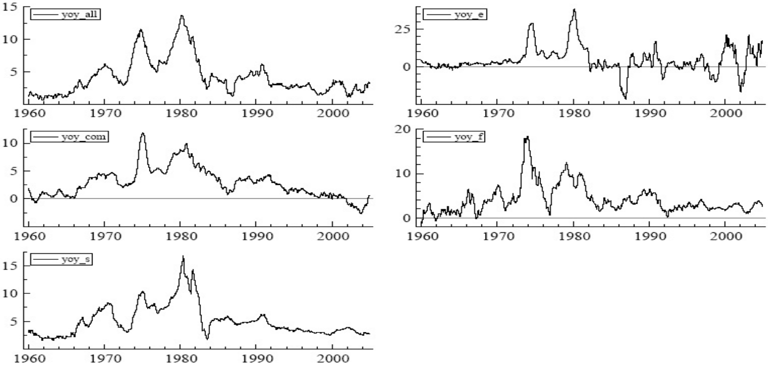
Empirical example: Forecasting Inflation

- predictand = annual U.S. CPI inflation, one month ahead
- two sample periods 1960-1983 and 1960-2004
  - prediction period =1970-1983; R=120, P=168
  - prediction period = 1984-2004; R=288, P=252
- rolling samples (also recursive)
- null model = univariate AR, lag length selected by AIC
- alternatives are bivariate VARs, lag lengths again selected by AIC. Second predictor (in addition to CPI inflation) is
  - component of inflation (food, energy, commodities, services)
  - output growth, change in unemployment

  (see Hubrich (2005) and Hendry and Hubrich (2009) for further empirical results on use of disaggregate information in forecasting the aggregate)

US CPI Inflation

US CPI year-on-year inflation: aggregate, energy, commodities, food and services



Forecasting US Inflation by components

Table 5: Tests of Equal Forecast Accuracy, US year-on-year inflation

	1970-1983					
method	RMSPE (altern)/ RMSPE (bench)	t-stat adj.	max t-stat. adj.	χ^2 adj	χ^2 unadj	Reality check
$AR_{(AIC)}$ (bench)	**0.307**					
Test AR						
vs $VAR^{a,f}_{(AIC)}$	1.039	0.666				
vs $VAR^{a,e}_{(AIC)}$	1.029	0.891				
vs $VAR^{a,c}_{(AIC)}$	1.016	1.743*				
vs $VAR^{a,s}_{(AIC)}$	0.986	2.311*				
vs 4 models			2.311*	7.743	7.207	0.032
critical value		1.282	1.902	7.78	7.78	0.118

- **Note 2:** It is possible to have the following seemingly paradoxical result:
  - sample MSPE from null model < sample MSPE from alternative model
  - we reject the null of equal population MSPE in favor of the alternative that the larger model has lower population MSPE
  - reason: adjustment larger than MSPE difference $\bar{f}_1 = \hat{\sigma}_0^2 - \hat{\sigma}_1^2 + P^{-1}\Sigma_t(\hat{y}_{0,t+1} - \hat{y}_{1,t+1})^2$
- Implication for future modeling: there is information in the larger model that is useful for forecasting, but we have not successfully exploited that information

- **Note 2:** It is possible to have the following seemingly paradoxical result:
  - sample MSPE from null model < sample MSPE from alternative model
  - we reject the null of equal population MSPE in favor of the alternative that the larger model has lower population MSPE
  - reason: adjustment larger than MSPE difference $\bar{f}_1 = \hat{\sigma}_0^2 - \hat{\sigma}_1^2 + P^{-1}\Sigma_t(\hat{y}_{0,t+1} - \hat{y}_{1,t+1})^2$
- Implication for future modeling: there is information in the larger model that is useful for forecasting, but we have not successfully exploited that information

- **Note 2:** It is possible to have the following seemingly paradoxical result:
  - sample MSPE from null model $<$ sample MSPE from alternative model
  - we reject the null of equal population MSPE in favor of the alternative that the larger model has lower population MSPE
  - reason: adjustment larger than MSPE difference $\bar{f}_1 = \hat{\sigma}_0^2 - \hat{\sigma}_1^2 + P^{-1}\Sigma_t(\hat{y}_{0,t+1} - \hat{y}_{1,t+1})^2$
- Implication for future modeling: there is information in the larger model that is useful for forecasting, but we have not successfully exploited that information

- **Note 2:** It is possible to have the following seemingly paradoxical result:
  - sample MSPE from null model < sample MSPE from alternative model
  - we reject the null of equal population MSPE in favor of the alternative that the larger model has lower population MSPE
  - reason: adjustment larger than MSPE difference $\bar{f}_1 = \hat{\sigma}_0^2 - \hat{\sigma}_1^2 + P^{-1}\Sigma_t(\hat{y}_{0,t+1} - \hat{y}_{1,t+1})^2$
- Implication for future modeling: there is information in the larger model that is useful for forecasting, but we have not successfully exploited that information

- **Note 2:** It is possible to have the following seemingly paradoxical result:
  - sample MSPE from null model < sample MSPE from alternative model
  - we reject the null of equal population MSPE in favor of the alternative that the larger model has lower population MSPE
  - reason: adjustment larger than MSPE difference $\bar{f}_1 = \hat{\sigma}_0^2 - \hat{\sigma}_1^2 + P^{-1}\Sigma_t(\hat{y}_{0,t+1} - \hat{y}_{1,t+1})^2$
- Implication for future modeling: there is information in the larger model that is useful for forecasting, but we have not successfully exploited that information

Forecasting US Inflation by components

	1984-2004					
method	RMSPE (altern)/ RMSPE (bench)	t-stat adj.	max t-stat adj.	χ^2 adj	χ^2 unadj	Reality check
$AR_{(AIC)}$ (bench)	**0.187**					
Test AR						
vs $VAR^{a,f}_{(AIC)}$	0.999	1.860*				
vs $VAR^{a,e}_{(AIC)}$	1.097	−0.027				
vs $VAR^{a,c}_{(AIC)}$	1.048	0.290				
vs $VAR^{a,s}_{(AIC)}$	1.027	−0.463				
vs 4 models			1.860	3.905	11.926*	0.0007
critical value		1.282	1.919	7.78	7.78	0.059

Forecasting US Inflation: components and real variables

Table 6: Tests of Equal Forecast Accuracy, US year-on-year inflation

	1970-1983					
method	RMSPE (altern)/ RMSPE (bench)	t-stat adj.	max t-stat adj.	χ^2 adj	χ^2 unadj	Reality check
$AR_{(AIC)}$ (bench)	**0.307**					
Test AR						
vs $VAR^{a,y}_{(AIC)}$	0.987	2.013*				
vs $VAR^{a,u}_{(AIC)}$	0.974	3.439*				
vs $VAR^{a,c}_{(AIC)}$	1.016	1.743*				
vs $VAR^{a,s}_{(AIC)}$	0.986	2.311*				
vs 4 models			3.439*	21.762*	2.432	0.061
critical value		1.282	1.917	7.78	7.78	0.146

Forecasting US Inflation: components and real variables

	1984-2004					
method	RMSPE (altern)/ RMSPE (bench)	t-stat adj.	max t-stat adj.	χ^2 adj	χ^2 unadj	Reality check
$AR_{(AIC)}$ (bench)	**0.187**					
Test AR						
vs $VAR^{a,y}_{(AIC)}$	1.046	–0.047				
vs $VAR^{a,u}_{(AIC)}$	1.024	1.867*				
vs $VAR^{a,c}_{(AIC)}$	1.048	0.290				
vs $VAR^{a,s}_{(AIC)}$	1.027	–0.463				
vs 4 models			1.867	4.605	12.680*	0.026
critical value		1.282	1.934	7.78	7.78	0.046

Empirical Example: Summary of Results

- in high and volatile inflation period 1970-1983 some disaggregate inflation rates or real variables can improve significantly, also according to the test of model sets

Important to apply appropiate test!

$\Rightarrow$ in low and stable inflation period 1984-2004 disaggregate food inflation and unemployment changes do improve forecast accuracy over simple AR model significantly with pairwise forecast accuracy test, but not with test of model sets
$\Rightarrow$ implication of both "max t-stat (adj.)" and "$\chi^2(adj.)$"

- "$\chi^2(unadj.)$" opposite result: can reject H0 1984-2004 but not 1970-1983; reality check - cannot reject H0 in either period; but: both tests have poor size and poor power

Empirical Example: Summary of Results

- in high and volatile inflation period 1970-1983 some disaggregate inflation rates or real variables can improve significantly, also according to the test of model sets

Important to apply appropiate test!

⇒ in low and stable inflation period 1984-2004 disaggregate food inflation and unemployment changes do improve forecast accuracy over simple AR model significantly with pairwise forecast accuracy test, but not with test of model sets
⇒ implication of both "max t-stat (adj.)" and "$\chi^2(adj.)$"

- "$\chi^2(unadj.)$" opposite result: can reject H0 1984-2004 but not 1970-1983; reality check - cannot reject H0 in either period; but: both tests have poor size and poor power

Empirical Example: Summary of Results

- in high and volatile inflation period 1970-1983 some disaggregate inflation rates or real variables can improve significantly, also according to the test of model sets

Important to apply appropiate test!

$\Rightarrow$ in low and stable inflation period 1984-2004 disaggregate food inflation and unemployment changes do improve forecast accuracy over simple AR model significantly with pairwise forecast accuracy test, but not with test of model sets
$\Rightarrow$ implication of both "max t-stat (adj.)" and "$\chi^2(adj.)$"

- "$\chi^2(unadj.)$" opposite result: can reject H0 1984-2004 but not 1970-1983; reality check - cannot reject H0 in either period; but: both tests have poor size and poor power

Conclusions

We have suggested **two tractable methods** for

Comparison of a null (benchmark) model to a small number of alternative models that nest the benchmark

- explicitly account for estimation error in parameters
- easily executed, do not require bootstrap procedures
- Simulation evidence suggests that our procedures have distinctly better size and power than existing procedures
- Empirical examples forecasting US inflation show
  - importance of simultaneously comparing small number of models with benchmark (correlations between forecasts)
  - sequence of pairwise model comparison can lead to wrong conclusions

Conclusions

We have suggested **two tractable methods** for

Comparison of a null (benchmark) model to a small number of alternative models that nest the benchmark

- explicitly account for estimation error in parameters
- easily executed, do not require bootstrap procedures
- Simulation evidence suggests that our procedures have distinctly better size and power than existing procedures
- Empirical examples forecasting US inflation show
  - importance of simultaneously comparing small number of models with benchmark (correlations between forecasts)
  - sequence of pairwise model comparison can lead to wrong conclusions

Conclusions

We have suggested **two tractable methods** for

Comparison of a null (benchmark) model to a small number of alternative models that nest the benchmark

- explicitly account for estimation error in parameters
- easily executed, do not require bootstrap procedures
- Simulation evidence suggests that our procedures have distinctly better size and power than existing procedures
- Empirical examples forecasting US inflation show
  - importance of simultaneously comparing small number of models with benchmark (correlations between forecasts)
  - sequence of pairwise model comparison can lead to wrong conclusions

Conclusions

We have suggested **two tractable methods** for

Comparison of a null (benchmark) model to a small number of alternative models that nest the benchmark

- explicitly account for estimation error in parameters
- easily executed, do not require bootstrap procedures
- Simulation evidence suggests that our procedures have distinctly better size and power than existing procedures
- Empirical examples forecasting US inflation show
  - importance of simultaneously comparing small number of models with benchmark (correlations between forecasts)
  - sequence of pairwise model comparison can lead to wrong conclusions

Conclusions

We have suggested **two tractable methods** for

Comparison of a null (benchmark) model to a small number of alternative models that nest the benchmark

- explicitly account for estimation error in parameters
- easily executed, do not require bootstrap procedures
- Simulation evidence suggests that our procedures have distinctly better size and power than existing procedures
- Empirical examples forecasting US inflation show
  - importance of simultaneously comparing small number of models with benchmark (correlations between forecasts)
  - sequence of pairwise model comparison can lead to wrong conclusions

Conclusions

We have suggested **two tractable methods** for

Comparison of a null (benchmark) model to a small number of alternative models that nest the benchmark

- explicitly account for estimation error in parameters
- easily executed, do not require bootstrap procedures
- Simulation evidence suggests that our procedures have distinctly better size and power than existing procedures
- Empirical examples forecasting US inflation show
  - importance of simultaneously comparing small number of models with benchmark (correlations between forecasts)
  - sequence of pairwise model comparison can lead to wrong conclusions

Conclusions

We have suggested **two tractable methods** for

Comparison of a null (benchmark) model to a small number of alternative models that nest the benchmark

- explicitly account for estimation error in parameters
- easily executed, do not require bootstrap procedures
- Simulation evidence suggests that our procedures have distinctly better size and power than existing procedures
- Empirical examples forecasting US inflation show
  - importance of simultaneously comparing small number of models with benchmark (correlations between forecasts)
  - sequence of pairwise model comparison can lead to wrong conclusions

Conclusions

We have suggested **two tractable methods** for

Comparison of a null (benchmark) model to a small number of alternative models that nest the benchmark

- explicitly account for estimation error in parameters
- easily executed, do not require bootstrap procedures
- Simulation evidence suggests that our procedures have distinctly better size and power than existing procedures
- Empirical examples forecasting US inflation show
  - importance of simultaneously comparing small number of models with benchmark (correlations between forecasts)
  - sequence of pairwise model comparison can lead to wrong conclusions

Further Research

1. use restrictions in accordance with the nesting properties of the alternative models for test procedure to improve power: Likelihood Ratio test (Granziera, Hubrich and Moon, 2008)
   - all alternative models are nested and nest the benchmark
   - all alternative models are non-nested, but nest the benchmark
   - some alternative models are nested, some are non-nested, all models nest the benchmark
2. extension to comparison of models that might or might not nest the benchmark (Hubrich, McCracken and West, 2009)
   - implications for asymptotic theory
   - implications for small sample performance
   - wider range of applications
3. "max t-stat (adj.)" might also be applicable to environments with number of models equal to sample size

Further Research

1. use restrictions in accordance with the nesting properties of the alternative models for test procedure to improve power: Likelihood Ratio test (Granziera, Hubrich and Moon, 2008)
   - all alternative models are nested and nest the benchmark
   - all alternative models are non-nested, but nest the benchmark
   - some alternative models are nested, some are non-nested, all models nest the benchmark
2. extension to comparison of models that might or might not nest the benchmark (Hubrich, McCracken and West, 2009)
   - implications for asymptotic theory
   - implications for small sample performance
   - wider range of applications
3. "max t-stat (adj.)" might also be applicable to environments with number of models equal to sample size

Further Research

1. use restrictions in accordance with the nesting properties of the alternative models for test procedure to improve power: Likelihood Ratio test (Granziera, Hubrich and Moon, 2008)
   - all alternative models are nested and nest the benchmark
   - all alternative models are non-nested, but nest the benchmark
   - some alternative models are nested, some are non-nested, all models nest the benchmark
2. extension to comparison of models that might or might not nest the benchmark (Hubrich, McCracken and West, 2009)
   - implications for asymptotic theory
   - implications for small sample performance
   - wider range of applications
3. "max t-stat (adj.)" might also be applicable to environments with number of models equal to sample size