Cross-validation as an alternative to out-of-sample inference under instability

Helle Bunzel and Gray Calhoun
Iowa State University, Econ Dept

ISNPS Annual Conference
19 June 2012

Introduction

Overview

- Out-of-sample (OOS) comparisons are popular in Macroeconomics and Finance and other fields
  - Eg: Meese and Rogoff (1983), Stock and Watson (2003), Goyal and Welch (2008)
- Basic procedure:
  - Construct a series of pseudo forecast errors for the last 30 years (say) for each model
  - Use those forecast errors to calculate average performance
- In Economics applications:
  - Usually not interested in forecasting per se
  - Want to determine whether a theoretically-motivated candidate model (or models) outperforms a simple benchmark
  - Usually concerned about
    - **inference**
    - Population models
  - Eg: Diebold and Mariano (1995), West (1996), Clark and McCracken (2001), Giacomini and White (2006)

Overview

- But it's not clear that OOS tests are necessary for inference about population models
  - Full-sample tests control size and have higher power (Inoue and Kilian, 2004, 2006)
  - OOS and full-sample tests require similar assumptions (moments, dependence, etc.)
- Counterargument: maybe OOS tests are more robust, but the theory isn't there yet
  - Robust to problems in the DGP
  - Robust to problems in the models
  - One example: unmodeled instability
- In this paper, we look at properties of OOS tests when there is unmodeled instability
  - Find that the test sample must be quite small
  - Propose a new test statistic based on cross-validation

Basic setup:

- y_{t+1} is the variable we want to predict
- x_{1t} and x_{2t} are regressors for two different models
  1. $y_{t+1} = x_{1t}'\beta_1 + \varepsilon_{1,t+1}$ (benchmark model)
  2. $y_{t+1} = x_{2t}'\beta_2 + \varepsilon_{2,t+1}$ (alternative model)
- Simplifications for this talk:
  - I'll present results for linear models and OLS
    - The results hold for many semi-parametric and non-parametric estimators as well, but the notation becomes more complicated
  - Strictly stationary regressors and innovations
- T total observations, $T = R + P + 1$
  - *Training sample* is the first R observations
  - *Test sample* is the last P observations

Construction of the OOS test statistic

- The test statistic is based on the difference in average forecasting performance over the test sample (DMW test):

$$\bar{f} \equiv \frac{1}{P} \sum_{t=R+1}^{T-1} f_t(\hat{\beta}_t)$$

  - $f_t(\beta) = L(y_{t+1} - x'_{1t}\beta_1) - L(y_{t+1} - x'_{2t}\hat{\beta}_2)$
    - L is a known and smooth loss function
  - $\hat{\beta}_{it} \equiv (\sum_{s=1}^{t-1} x_{is}x'_{is})^{-1} \sum_{s=1}^{t-1} x_{is}y_{i,s+1}$
- Researcher will use $\sqrt{P}\frac{\bar{f}}{\hat{\sigma}}$ for a hypothesis test
  - $\hat{\sigma}^2$ is an OOS estimator of the asymptotic variance of $\bar{f}$

Theoretical development of OOS inference (much omitted)

- Diebold and Mariano (1995): derive the asymptotic distribution for known β
  - OOS t-test is asymptotically normal
- West (1996): allows for estimated β by expanding $f_t(\hat{\beta}_t)$ around $f_t(\beta)$
  - Asymptotically normal
  - Under some conditions, the variance of $\hat{\beta}_t$ affects the variance of the OOS statistic
- Clark and McCracken (2001): show that the statistic is degenerate when the models give the same forecasts in the limit
  - Not normal, so Clark and McCracken derive critical values
- All test the null that $\mathrm{E}\, f_t(\beta) = 0$
  - β is the pseudo-true parameter value
- Many other subsequent papers

Previous research on OOS inference under instability

- Key contribution of our paper is to look at a null hypothesis that accomodates breaks
- Many other papers have focus on rejecting under instability
  - Clark and McCracken (2005), Inoue and Kilian (2004, 2006), Giacomini and Rossi (2009, 2010), Rossi and Sekhposyan (2011),
- Tests for optimality of break-robust forecasting methods
  - Giacomini and White (2006)
- Choice of estimation window under breaks
  - Pesaran and Timmermann (2007), Hansen and Timmermann (2012), Inoue and Rossi (2012),

Theoretical results for out-of-sample comparisons

Motivating our null hypothesis

- Under stationarity, testing $\mathrm{E}\, f_t(\beta) = 0$ in-sample is easy
- What is an interesting null hypothesis that can't be easily tested in-sample?
- Suppose there's a single break in period τ
  - For this talk, I'll assume that τ is **known** and that $\frac{\tau}{T} \to \mu$
  - The DGP is

$$y_{t+1} = \begin{cases} x_t'\dot{\alpha} + \varepsilon_{t+1} & t < \tau \\ x_t'\ddot{\alpha} + \varepsilon_{t+1} & t \geq \tau \end{cases}$$

    - For simplicity, suppose (x_t, ε_{t+1}) are stationary in this talk
    - x_t contains the elements of x_{1t} and x_{2t} without duplicates
    - ε_{t+1} has mean zero but does not need to be MDS
- Neither model accounts for that break (surprisingly common)
  1. $y_{t+1} = x_{1t}'\beta_1 + \varepsilon_{1,t+1}$ (benchmark model)
  2. $y_{t+1} = x_{2t}'\beta_2 + \varepsilon_{2,t+1}$ (alternative model)

Motivating our null hypothesis

- We want a null that is analagous to $\mathrm{E}\, f_t(\beta) = 0$
- One question of interest is, does the alternative model do better after the break?
- Formally, for $t \geq \tau$,

$$H_0 : \ \mathrm{E}\, f_t(\bar{\beta}) \leq 0$$

  - $\bar{\beta}$ is the pseudo-true full-sample parameter value
  - $\dot{\beta}$ and $\ddot{\beta}$ are pre- and post-break values of β
  - $\bar{\beta} \approx \mu\dot{\beta} + (1-\mu)\ddot{\beta}$ for OLS with stationary regressors
- The hope is that, for some V,

$$\sqrt{P}(\bar{f} - \mathrm{E}\, f_{\tau+1}(\bar{\beta})) \to^d N(0, V)$$

  Or that $\hat{V}^{-1/2}\sqrt{P}(\bar{f} - \mathrm{E}\, f_{\tau+1}(\bar{\beta}))$ is asymptotically pivotal

Basic assumptions

- We use essentially the same assumptions as West (1996) and McCracken (2000, 2007)
- The underlying series are strong mixing (i.e. a weak dependence assumption)
- Moment conditions that ensure the necessary CLTs hold
- The loss function is smooth (differentiable)
- $R,\ P \to \infty$ as $T \to \infty$

The OOS average is miscentered:

Theorem (Result 2.1)

Under standard assumptions and with $R > \tau$

$$\frac{1}{\sqrt{P}} \sum_{t=R+1}^{T-1} \left(f_t(\hat{\beta}_t) - \frac{1}{P} \sum_{s=R+1}^{T-1} \mathrm{E}\, f_s(\bar{\beta}_s) \right) \to^d N(0, \Omega^{oos})$$

- $\bar{\beta}_s = \frac{\tau}{s}\dot{\beta} + (1 - \frac{\tau}{s})\ddot{\beta}$
- Argument is similar to West's (1996) and McCracken's (2000)
- Ω^{oos} has known functional form and can be estimated
- Only if $\frac{P^{1.5}}{R} \to 0$ does

$$\frac{1}{\sqrt{P}} \sum_{s=R+1}^{T-1} \left(\mathrm{E}\, f_s(\bar{\beta}) - \mathrm{E}\, f_s(\bar{\beta}_s) \right) \to 0$$

  - The printed abstract has an error

The OOS average is correctly centered when P is small

$$\frac{1}{\sqrt{P}} \sum_{t=R+1}^{T-1} (\mathrm{E}\, f_t(\bar{\beta}_t) - \mathrm{E}\, f_t(\bar{\beta})) =$$

$$= \frac{1}{\sqrt{P}} \sum_{t=R+1}^{T-1} \mathrm{E}[\nabla f_t(\bar{\beta})'](\bar{\beta}_t - \bar{\beta}) + o(-)$$

$$= \frac{C}{\sqrt{P}} \sum_{t=R+1}^{T-1} \left(\frac{\tau}{t} - \frac{\tau}{T}\right) + o(-)$$

$$= \frac{C}{\sqrt{P}} \left(\tau \log(\frac{T}{R}) - P\mu\right) + o(-)$$

$$= C\mu\sqrt{P}\left(\frac{T}{R} - 1\right) + o(-)$$

$$= C\mu\frac{P^{3/2}}{R} + o(-)$$

(we assume $\frac{P}{T} \to 0$ in a few of the intermediate steps)

Requiring a small test sample matters

- Papers (e.g. Goyal and Welch, 2008) **often** cite instability as a reason for conducting an OOS comparison
- And they typically use a large test sample
  - e.g. the available data since 1970
- Unless the authors want to test hypotheses about $\frac{1}{T-\tau}\sum_{t=\tau}^{T-1}\mathrm{E}\, f_t(\bar{\beta}_t)$, this approach gives misleading conclusions
  - Think of τ here as being the date of the last large break
- $\mathrm{E}\, f_{\tau+1}(\bar{\beta})$ is our guess at a parameter that remains interesting under instability
  - But the general point remains even if researchers are interested in other parameters

A few remarks

- Since P must be so small, estimating the break date is unnecessary even if it is unknown.
- The general result continues to hold when the breaks are moderate
  - Moderate in the sense that $\dot{\beta} - \ddot{\beta} \sim \frac{1}{\sqrt{T}}$
  - In this setting, μ can not be estimated consistently
  - The restriction $\frac{P}{R} \to 0$ is still necessary
- The results also continue to hold when the OOS average is usually degenerate
  - $\frac{P}{R} \to 0$ resolves the degeneracy (Clark and McCracken, 2001)
- Parameter estimation error for $\bar{\beta}$ does not matter since P must be so small, so Ω^{cv} is especially simple (West, 1996)
- Conclusion: just do an OOS t-test at the very end of the sample

Theoretical results for cross-validation

Motivation for cross-validation

- The requirement that $\frac{P^{3/2}}{R} \to 0$ is very limiting
  - Power can be very low
  - CLT may be a poor approximation, distorting the test's size
- One solution is to replace $\hat{\beta}_t$ with estimators closer to $\bar{\beta}$
- We use post-break leave-one-out cross-validation:

$$\tilde{\beta}_{it} = \Big(\sum_{s \neq t} x_{is} x_{is}' \Big)^{-1} \sum_{s \neq t} x_{is} y_{s+1}$$

- Only look at the loss difference **after** the break. For $R > \tau$, the statistic becomes:

$$\frac{1}{\sqrt{P}\,\hat{\sigma}^2} \sum_{t=R+1}^{T-1} f_t(\tilde{\beta}_t)$$

  - $\hat{\sigma}^2$ is again an estimator of the asymptotic variance

Cross-validation asymptotic normality result

Theorem (Result 3.1)

If $R > \tau$, the asymptotic variance is positive, and other technical conditions hold, then

$$\frac{1}{\sqrt{P}} \sum_{t=R+1}^{T-1} \left[f_t(\tilde{\beta}_t) - \mathrm{E}\, f_t(\bar{\beta}) \right] \to^d N(0, \Omega^{cv})$$

- Similar proof to the previous result
- Holds even with $\lim \frac{P}{R} > 0$
- Ω^{cv} has known form and can be estimated
- Asymptotic variance reflects estimation error in $\tilde{\beta}_t$

Cross-validation asymptotic normality result

Consequently, under the previous assumptions,

$$\frac{1}{\sqrt{P\,\hat{\Omega}^{cv}}} \sum_{t=R+1}^{T-1} f_t(\tilde{\beta}_t) \rightarrow^d N(0,1)$$

under the null $\mathrm{E}\, f_t(\bar{\beta}) = 0$, if $\hat{\Omega}^{cv}$ is consistent for Ω^{cv}

Remarks about degeneracy

- To rule out degeneracy, we need a nonzero pseudo-true coefficient on at least one element unique to x_{1t} or x_{2t}
  - Otherwise the statistic can be non-Gaussian
- A little additional notation:
  - Let x_{0t} denote the regressors common to both models and z_{1t} and z_{2t} be the unique regressors
  - Let γ_0 be the coefficients on x_{0t} and let γ_1 and γ_2 be the coefficients on z_{1t} and z_{2t}
- Our recommended procedure for large breaks is very simple:
  1. Test $\bar{\gamma}_1 = 0$ and $\bar{\gamma}_2 = 0$ using the full dataset
  2. If the test rejects, test the null hypothesis $\mathrm{E}\, f_t(\bar{\beta}) \leq 0$ with

$$\frac{1}{\sqrt{P\,\hat{\Omega}^{cv}}} \sum_{t=R+1}^{T-1} f_t(\tilde{\beta}_t)$$

  3. Otherwise accept the null hypothesis

Extending to *estimated* breaks

- Suppose the break date is unknown but estimated with

$$\hat{\tau} = \arg\min_{\tau} \left\{ \sum_{t=1}^{\tau} \left(f_t(\tilde{\beta}_t) - \frac{1}{\tau} \sum_{s=1}^{\tau} f_s(\tilde{\beta}_s) \right)^2 + \sum_{t=\tau+1}^{T-1} \left(f_t(\tilde{\beta}_t) - \frac{1}{T-\tau-1} \sum_{s=\tau+1}^{T-1} f_s(\tilde{\beta}_s) \right)^2 \right\}$$

  and let $\hat{\mu} = \frac{\hat{\tau}}{T}$

- If $\sqrt{T}(\hat{\mu} - \mu) \to^d N(0, v)$ the previous results are unchanged
  - Happens if the break is large

Extending to *estimated* breaks

- If the break is moderate (i.e. $\sim \frac{1}{\sqrt{T}}$), it's more interesting
  - $\hat{\mu}$ is not consistent for μ
- For nondegenerate statistics, distribution becomes a function of Brownian Motion (Elliott and Mueller, 2007, 2010)
- Moderate breaks allow $\dot{\beta}$ and $\ddot{\beta}$ to be local to zero, introducing degeneracy
  - Distribution becomes a more complicated function of Brownian Motion
- Test statistic is non-pivotal and depends on parameters that can not be consistently estimated
- Worst-case critical values can be found by simulation (in progress)

Monte Carlo

Monte Carlo Setup

- DGP:

$$y_{t+1} = \begin{cases} 6x_t + \varepsilon_{t+1} & t = 1, \ldots, 110 \\ 2x_t + \varepsilon_{t+1} & t = 111, \ldots, 220 \quad \text{size simulations} \\ 2.25x_t + \varepsilon_{t+1} & t = 111, \ldots, 220 \quad \text{power simulations} \end{cases}$$

  - x_t and ε_{t+1} are both mutually independent Standard Normal
- Models:
  - $y_{1,t+1} = \beta_1 + u_{1,t+1}$
  - $y_{2,t+1} = \beta_1 + \beta_2 x_t + u_{2,t+1}$
- Squared-error loss: $L(y, \hat{y}) = (y - \hat{y})^2$
- The null hypothesis holds: for $t > 110$ under the "size" DGP,
  - $\mathrm{E}(y_{t+1} - \bar{\beta}_1)^2 = \mathrm{E}(2x_t + \varepsilon_{t+1})^2 = 4 + 1$
  - $\mathrm{E}(y_{t+1} - \bar{\beta}_1 - \bar{\beta}_2 x_t)^2 = \mathrm{E}((6-4)x_t + \varepsilon_{t+1})^2 = 4 + 1$

Monte Carlo Setup

- Use the following test statistics:
  - Our post-break CV
  - Our modified postbreak DMW test (adjusted standard error)
  - The naive DMW test (simple OOS-t test)
  - McCracken's (2007) test for nested models
  - Clark and West's (2005, 2006) test with a recursive window
- Current practice is to use Clark and West's or McCracken's test
  - These tests were not designed for this null
  - Again, this null seems to reflect the use of these tests in empirical research
- Assume that the break date is known
- Two choices of the test sample
  - 110 obs. (entire post-break sample)
  - 20 observations

Monte Carlo Results

	$P = 110$		$P = 20$	
	size (%)	power (%)	size (%)	power (%)
Cross-validation	9.1	55.2	9.7	25.5
New OOS	0.0	0.0	8.4	21.6
Naive DMW	0.0	0.0	9.9	25.9
McCracken	0.0	1.3	17.4	41.2
Clark & West	100.0	100.0	100.0	100.0

Nominal size is 10%

Conclusion

In Summary

- Introduced a null hypothesis that can hold under instability and looked at effectiveness of OOS comparisons
- OOS statistics have a limitation:
  - P must be very small for validity ($\frac{P^{3/2}}{R} \to 0$)
  - But the OOS test is robust to aspects of the instability
- We propose a new type of cross-validation that does well for this null
  - Leave-one-out over the post-break sample
- Caveats/future research
  - We only look at one form of instability
  - There are other justifications for OOS tests (e.g. overfit)
  - The proposed cross-validation estimator requires knowing the form of instability