

A predictability test for a small number of nested models.

Eleonora Granziera, Kirstin Hubrich, Roger Moon

June 10, 2008

preliminary

Abstract

In this paper we introduce a new test to evaluate the null that a parsimonious model performs equally well as a small number of models which nest the benchmark. The size and the power performances of this likelihood-ratio type test are compared with the ones of two existing tests: a chi-square test, as described in Clark and West (2007) applied to multivariate comparison in Hubrich and West (2008), and the maximum of correlated normal test outlined in Hubrich and West (2008). The Monte Carlo experiments conducted in the paper reveal that the chi-square test performs poorly in terms of power as it disregards the one-sided nature of the test, while the ranking between the likelihood-ratio type test and the correlated normal test depends on the simulation settings. Further simulation experiments suggest that the relative performance of the tests is related to the rank of the variance covariance matrix of the MSPE differences.

1 Introduction

Evaluation of forecast accuracy usually requires to compare a parsimonious model with one or more models which might or not nest the benchmark model. The asymptotic distribution of the test statistics depend on whether the models to compare are nested or not. A substantial chunk of the literature on out-of-sample forecast accuracy has focused on the asymptotic and finite sample properties of tests for non-nested models. Diebold and Mariano (1995) make inference on a vector of moments of predictions or prediction errors proving that the distribution is asymptotically standard normal. Their framework accommodates various loss functions but assumes that the forecasts do not rely on regression estimates. West (1996) shows that, if the models are non-nested, when conducting inference about MSPE parameter estimation error is asymptotically irrelevant and the asymptotic distribution of the difference in sample mean square forecast errors is the same as the distribution of the population mean square forecast errors.

Asymptotic irrelevance fails in a number of cases and the asymptotic normality of the statistics derived under the assumptions in Diebold and Mariano (1985) and West (1996) do not apply to nested models because of a rank condition that is not satisfied in this framework. The problem with nested models arises because under the null of equal predictive accuracy, the errors of the different models are the same and hence the variance covariance matrix of the estimator is not full rank. Formal characterization of limiting distributions for the comparison of two nested models has been attained by McCracken (2004) and Clark and McCracken (2001, 2005) when the parameters are estimated through nonlinear least squares. In this environment the test statistic to evaluate the null of equal predictive ability is derived as functionals of Brownian motions and is asymptotically pivotal if certain additional conditions hold: the forecast horizon is one and the forecast errors are conditionally homoskedastic or if the larger model contains only one additional regressor. Clark and West (2006,2007), thereafter CW,

argues that for nested models the sample MSPE difference is positive and introduce an adjustment term to center the statistic around zero. They also provide Monte Carlo evidence to justify the assumption of normality of the adjusted MSPE difference.

The papers reviewed so far considered a two by two model comparison. There are two existing procedures to compare a small number of nested models, outlined in Hubrich and West (2008), thereafter HW. The first one is a direct extension of the work of Diebold and Mariano (1995) and West (1996), (DMW), and consists of a chi-squared statistic, the other examines the maximum of correlated normals. Both tests adjust the MSPE differences as advocated in CW.

We propose a new likelihood ratio type predictability test for the comparison of a small number of models nested to each other. A rigorous definition of 'small' is not provided, but as a practical rule we suggest that the number of models should be smaller than the size of the out-of sample. We distinguish among three different cases according to the structure of the alternative models considered: in the first case the models are nested within each other, in the second there is no nested relation between the models, in the third, more general case, the models can be grouped such that within each group the models are nested, but there is no nested relation among groups. We evaluate the size and power properties of the test via Monte Carlo simulations for 1-step ahead forecasts under different assumption on the limiting distributions of the statistics. The Monte Carlo investigation reveals that the chi-square test performs poorly in terms of power as it disregard the one-sided nature of the test, while the ranking between the likelihood-ratio type test and the correlated normal test depends on the simulation settings. Further simulation experiments suggest that the relative performance of the tests is related to the rank of the variance covariance matrix of the MSPE differences.

The outline of the paper is as follows: section 2 introduces the notation

and the forecasting environment and presents the test. Section 3 provides with inference on the test. In section 4 the Monte Carlo simulation experiment is described and the size and power properties of the test are commented. Section 5 concludes.

2 Basic Mathematical Framework

We evaluate the null that a parsimonious model performs equally well as a small number of models which nest the benchmark.

We refer to the work of HW for the description of the environment. We are interested in forecasting a scalar y_t through $M+1$ linear models estimated by least squares. The benchmark model, denoted as "0", and the alternative models, denoted as "m" with $m = 1, \dots, M$, can be written as:

$$\begin{aligned} y_t &= X'_{0t}\beta_0 + u_{0t} \\ &\vdots \\ y_t &= X'_{mt}\beta_m + u_{mt} \\ &\vdots \\ y_t &= X'_{Mt}\beta_M + u_{Mt}, \end{aligned}$$

where u_{it} are i.i.d random variables satisfying $E(u_{it}X_{it}) = 0$, and X_{0t}, \dots, X_{Mt} are vectors of regressors such that $X_{0t} = (x'_{0t})$ is of dimension $k_0 \times 1$ and $X_{mt} = (x'_{0t}, x'_{mt})$ is of dimension $k_m \times 1$ with $k_0 < k_m$. Under the null model 0 is the true model and hence each model m includes $k_m - k_0$ excess parameters: $\beta_m = (\beta'_0, 0'_{k_m - k_0})' \forall m = 1, \dots, M$; moreover the errors are identical $u_{0t} = u_{1t} = \dots = u_{Mt}$. Under the alternative, instead, one of the alternative models is the correctly specified model and hence the additional parameters estimated are non-zero in population. For simplicity we focus on one period ahead forecasts. Let $T+1$ be the total sample size, R the size of the sample used to generate the initial estimates and P the observations used for out of sample evaluation. The one period ahead forecasts,

$\hat{y}_{0t+1}, \dots, \hat{y}_{mt+1}, \dots, \hat{y}_{Mt+1}$, are obtained through either the expanding window or the rolling scheme for $t=R, \dots, T$. In the expanding window scheme the size of the estimation sample grows while in the rolling scheme the size of the estimation sample stays constant.¹ Following CW we denote as f_{mt+1} the difference of the loss functions (MSPE) between the benchmark and alternative model m : $f_{m,t+1} = \sigma_0^2 - \sigma_m^2$, with $\sigma_i^2 \equiv E(u_{it}^2)$ being the population variance of the forecast error, which is assumed to be a stationary process. Collect the MSPE differences in the vector f_{t+1} :

$$f_{t+1} = (f_{1,t+1}, \dots, f_{M,t+1})'$$

Let $\hat{u}_{i,t+1} = y_{t+1} - \hat{y}_{i,t+1}$ be the 1-step ahead forecast error from model i with $i = 0, \dots, M$. The sample analogous of $f_{m,t+1}$, denoted by $\hat{f}_{m,t+1}$, is given by:

$$\hat{f}_{m,t+1} = (y_{t+1} - \hat{y}_{0t+1})^2 - (y_{t+1} - \hat{y}_{mt+1})^2 = (\hat{u}_{0t+1})^2 - (\hat{u}_{mt+1})^2$$

Then μ is the expected value of \hat{f}_{t+1} , the vector obtained by stacking together the sample MSPE differences:

$$\hat{f}_{t+1} = (\hat{f}_{1t+1}, \dots, \hat{f}_{Mt+1})'$$

$$\mu = E(\hat{f}_{t+1})$$

Then the sample counterpart of μ is given by:

$$\bar{f} = P^{-1} \left(\sum_{t=R}^T \hat{f}_{1t+1}, \dots, \sum_{t=R}^T \hat{f}_{mt+1}, \dots, \sum_{t=R}^T \hat{f}_{Mt+1} \right)'$$

¹Consider a sequence P of estimates for $t=R, \dots, T$ for the regression model $Y_t = \beta X_t + \varepsilon_t$ with Y_t and X_t scalar random variables. In the recursive scheme $\hat{\beta}_t^{rec} = \sum_{s=1}^t (X_s Y_s) / \sum_{s=1}^t (X_s^2)$ while in the rolling scheme $\hat{\beta}_t^{rol} = \sum_{s=t-R+1}^t (X_s Y_s) / \sum_{s=t-R+1}^t (X_s^2)$. For the recursive scheme the size of the estimation sample is R for the first sample and $T=R+P-1$ for the last sample, while for the rolling scheme it is R for any sample.

However, CW show that under the null that model 0 is the correctly specified model the sample MSPE from the parsimonious model will be generally lower than the MSPE from the alternative model, so it will be the case that $P^{-1} \sum_{t=R}^T (\hat{f}_{mt+1}) < 0$. Hence, they suggest the following adjustment to center the MSPE difference around zero:

$$\begin{aligned} \hat{f}_{mt+1}^{adj} &= (y_{t+1} - \hat{y}_{0t+1})^2 - \left[(y_{t+1} - \hat{y}_{mt+1})^2 - (\hat{y}_{0t+1} - \hat{y}_{mt+1})^2 \right] \\ &= (\hat{u}_{0t+1})^2 - \left[(\hat{u}_{mt+1})^2 - (\hat{y}_{0t+1} - \hat{y}_{mt+1})^2 \right]. \end{aligned}$$

Analogous quantities defined above for \hat{f}_{mt+1} can be derived from \hat{f}_{mt+1}^{adj} :

$$\hat{f}_{t+1}^{adj} = \left(\hat{f}_{1t+1}^{adj}, \dots, \hat{f}_{Mt+1}^{adj} \right)'$$

$$\mu^{adj} = E \left(\hat{f}_{t+1}^{adj} \right).$$

$$\bar{f}^{adj} = P^{-1} \left(\sum_{t=R}^T \hat{f}_{1t+1}^{adj}, \dots, \sum_{t=R}^T \hat{f}_{mt+1}^{adj}, \dots, \sum_{t=R}^T \hat{f}_{Mt+1}^{adj} \right)'$$

We will specify the null hypothesis as $H_0 : \mu^{adj} = 0$ while the specification of the alternative hypothesis will depend on the assumptions on the structure of the alternative models. We will distinguish between three cases: in the first one the models are nested within each other, in the second there is no nesting relation between the alternative models, in the last one the models are nested within groups.

2.1 Special Case 1: When the Alternative Models are Nested With Each Other

We characterize the case in which each model $m-1$ is nested in model m by imposing that model m includes $k_m - k_{m-1}$ additional regressors: $X_{m,t} = (x'_{m-1,t}, x'_{m,t})$ so that $k_0 < \dots < k_m < \dots < k_M$.

Given the structure of the problem, when considering the unadjusted MSPE we know that if model m is true it will hold that, for models $m, m - 1, \dots, 1$, $\sigma_m^2 < \sigma_{m-1}^2 < \dots < \sigma_1^2$ and hence $0 < \mu_1 = \hat{\sigma}_0^2 - \hat{\sigma}_1^2 < \dots < \hat{\sigma}_0^2 - \hat{\sigma}_{m-1}^2 = \mu_{m-1} < \hat{\sigma}_0^2 - \hat{\sigma}_m^2 = \mu_m$ while for models $m + 1, \dots, M$, $\sigma_m^2 = \sigma_{m+1}^2 = \dots = \sigma_M^2$. This ordering is invariant to the introduction of the CW adjustment² Then the null and the alternative hypotheses can be expressed as:

$$\begin{aligned} H_0 & : \mu^{adj} = 0. \\ H_1 & : 0 \leq \mu_1^{adj} \leq \mu_2^{adj} \leq \dots \leq \mu_M^{adj}, \mu^{adj} \neq 0. \end{aligned}$$

Hence we test equal forecast accuracy versus the alternative that at least one of the models performs better than the benchmark. We consider a one-side alternative as first suggested by Ashley, Granger and Schmalensee (1980) and subsequently assumed in many studies (CW, HW).

The test we propose to evaluate the null of equal predictive ability is a likelihood-ratio type test of the form:

$$\begin{aligned} \mathcal{T}_{LRT} & = P \bar{f}^{adj}' \hat{v}^{-1} \bar{f}^{adj} - \min_{\mu^{adj} \geq 0} P \left(\bar{f}^{adj} - \mu^{adj} \right)' \hat{v}^{-1} \left(\bar{f}^{adj} - \mu^{adj} \right) \\ & = P \tilde{\mu}^{adj}' \hat{v}^{-1} \tilde{\mu}^{adj}. \end{aligned}$$

where $\tilde{\mu} = \arg \min \left(\bar{f}^{adj} - \mu^{adj} \right)' \hat{v}^{-1} \left(\bar{f}^{adj} - \mu^{adj} \right)$, \hat{v} is a consistent estimator of v :

$$\hat{v} = P^{-1} \sum_{t=R}^T \left(\hat{f}_{t+1}^{adj} - \bar{f}^{adj} \right)^2 \quad (1)$$

and

$$D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}.$$

²See Appendix A for a formal proof.

is picked such that the parameter set implied by the null and the alternative can be expressed as

$$H_1 : D\mu^{adj} \geq 0.$$

i.e. the alternative models follow the structure: $\mu_1^{adj} \leq \mu_2^{adj} \leq \dots \leq \mu_M^{adj}$.

2.2 Special Case 2: When the Alternative Models are Non-nested

In this case there is no nesting relation between the alternative models, but still each of them nests the benchmark. We test for:

$$H_0 : \mu^{adj} = 0$$

against

$$H_0 \cup H_1 : \mu_1^{adj} \geq 0, \dots, \text{ or } \mu_M^{adj} \geq 0.$$

Denote by \mathcal{A} the region such that

$$\begin{aligned} \mathcal{A} &= \left\{ \mu^{adj} : \mu_1^{adj} \geq 0, \dots, \text{ or } \mu_M^{adj} \geq 0 \right\} \\ \mathcal{A}_m &= \left\{ \mu^{adj} : \mu_m^{adj} \geq 0 \right\} \end{aligned}$$

In this case, the likelihood ratio is given by:

$$\begin{aligned} \mathcal{T}_{LRT} = & \max \left\{ P \bar{f}^{adj'} \hat{v}^{adj-1} \bar{f}^{adj} - \min_{\mu^{adj} \in \mathcal{A}_1} \left(\bar{f}^{adj} - \mu^{adj} \right)' \hat{v}^{adj-1} \left(\bar{f}^{adj} - \mu^{adj} \right), \dots \right. \\ & \left. \dots, P \bar{f}^{adj'} \hat{v}^{adj-1} \bar{f}^{adj} - \min_{\mu^{adj} \in \mathcal{A}_M} \left(\bar{f}^{adj} - \mu^{adj} \right)' \hat{v}^{adj-1} \left(\bar{f}^{adj} - \mu^{adj} \right) \right\} \end{aligned}$$

2.3 General Case: When the Alternative Models are Nested within Groups

Now we consider a general case. Suppose that the alternative models can be grouped according to the nested relations such that within each group the models are nested however across different groups, the models are not nested. In particular, consider K groups such that within each group \mathcal{G}_k :

$\mu_{k1} \leq \mu_{k2} \leq \dots \leq \mu_{kM_k}$. Then the null and the alternative can be rewritten as:

$$\begin{aligned} H_0 & : \mu^{adj} = 0. \\ H_1 & : \mu_{11}^{adj} \leq \dots \leq \mu_{M_1}^{adj}, \dots, \mu_{1K}^{adj} \leq \dots \leq \mu_{M_K}^{adj} \quad \mu^{adj} \neq 0. \end{aligned}$$

For this case we propose a likelihood-ratio type test that combines the two tests outlined in the previous sections.

$$\begin{aligned} \mathcal{T}_{LRT} = \max & \left\{ P \bar{f}^{adj'} \hat{v}^{adj-1} \bar{f}^{adj} - \min_{\mu^{adj} \in \mathcal{A}_1} \left(\bar{f}^{adj} - \mu^{adj} \right)' \hat{v}^{adj-1} \left(\bar{f}^{adj} - \mu^{adj} \right), \dots \right. \\ & \left. \dots, P \bar{f}^{adj'} \hat{v}^{adj-1} \bar{f}^{adj} - \min_{\mu^{adj} \in \mathcal{A}_M} \left(\bar{f}^{adj} - \mu^{adj} \right)' \hat{v}^{adj-1} \left(\bar{f}^{adj} - \mu^{adj} \right) \right\} \end{aligned}$$

where now:

$$\mathcal{A}_k = \left\{ \mu^{adj} \in \mathbb{R}^M : \mathfrak{D}_k \mu^{adj} \geq 0 \right\} \text{ for } k = 1, \dots, K$$

$$\mathfrak{D}_k = \left[\begin{array}{c|c|c} 0_{(\max(M_1, \dots, M_K) \times (\sum_{i=1}^k M_i))} & D_k, M_k \times M_k & 0_{(\max(M_1, \dots, M_K) \times (M_K - (\sum_{i=1}^k M_i)))} \end{array} \right]$$

$$D_k = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}.$$

with \mathfrak{D}_k being a matrix of dimension $(\max(M_1, \dots, M_K)) \times (M)$, and D_k of dimension $M_k \times M_k$.

2.4 Alternative Tests

We consider two alternative forecast accuracy tests for multi-model comparison: a chi-square test, which has been originally designed for a bi-model

comparison in CW and the correlated normals test, proposed by HW. CW considers a Wald-type test involving the statistic $\mathcal{T}_{chi^2} = P \bar{f}_m^{adj} \hat{v}_m^{-1} \bar{f}_m^{adj}$. As discussed above it focuses on the adjusted MSPE in order to center the statistic around zero. HW exploits the one-sided nature of the test and select as test statistic: $\mathcal{T}_{\max t} = \max[\sqrt{P} \bar{f}_m^{adj} / \sqrt{\hat{v}_m}] \equiv \max \text{t-stat}(\text{adj.})$, with \hat{v}_m the sample variance of \hat{f}_m^{adj} .

3 Evaluation of MSPE differences.

The properties of the tests described above rely on the distribution of the adjusted MSPE. Observe that the adjusted MSPE can be rewritten for each model m as:

$$\bar{f}_m^{adj} = 2P^{-1} \sum_{t=R}^T \hat{u}_{0,t+1} (\hat{u}_{0,t+1} - \hat{u}_{m,t+1}).^3$$

which is analogous to the statistic considered in Clark and McCracken (2001) who show that if $\lim_{P,R \rightarrow \infty} P/R = \pi$, $\pi > 0$, i.e. if the size of the estimation sample grows at the same rate than the out-of-sample, the limiting distribution of $\bar{f}_m^{adj} / \sqrt{\hat{v}_m}$ is non normal under the null when the models are nested. They derive the asymptotic distribution of such statistic for one step ahead forecasts in conditionally homoskedastic environment as a functional of Brownian motion which depends on the excess parameters in model m , k_m , the large sample limit of the ratio between the in-sample and out-of-sample size, π , and the estimation scheme used (expanding window or rolling). Simulation experiments show that for one-step ahead forecasts and homoskedastic prediction errors, applying standard normal inference to $\bar{f}_m^{adj} / \sqrt{\hat{v}_m}$ leads to slightly undersized test results. The standard normal approximation performs reasonably well also in heteroskedastic environment when the number of additional regressors, k_m is equal to one. This finding is

³as $f_{m,t} = \hat{u}_{0,t+1}^2 - \hat{u}_{m,t+1}^2 + (\hat{y}_{0,t+1} - \hat{y}_{m,t+1})^2 = \hat{u}_{0,t+1}^2 - \hat{u}_{m,t+1}^2 + (\hat{u}_{m,t+1} - \hat{u}_{0,t+1})^2 = 2\hat{u}_{0,t+1}^2 - 2\hat{u}_{0,t+1}\hat{u}_{m,t+1}$

confirmed by simulations in CW which finds an empirical size between 0.05 and 0.1 for a 10% nominal size for both heteroskedastic and homoskedastic forecast errors, for both the expanding window and rolling estimation scheme and for values of π ranging between one third and six. They also compare the performance of the test when using simulated or bootstrapped critical values rather than asymptotic normal critical values and they do not find substantial size or power improvements. This is taken as justification for the assumption of normality of $\bar{f}_m^{adj}/\sqrt{\hat{v}_m}$ in HW which extend the work of CW for a multi-model comparison setting.

In this paper we will investigate the size and power properties of the proposed likelihood-ratio statistic, of the CW and of the HW statistics using critical values derived both under the asymptotic normality assumption and through simulation of the limiting distribution.

When $\mu \sim N(0, V)$ the limiting distribution of \mathcal{T}_{LRT} under the null is known to be a mixture of independent chi-square: $\mathcal{T}_{LRT} \Longrightarrow \omega_1 \chi_1^2 + \dots + \omega_M \chi_M^2$, where M , the number of alternative models considered, is the size of the vector μ^{adj} and $\omega_m = \omega_m(M, \nu)$ is the probability that exactly m of the M components are strictly positive (see Perlman 1969). For a given significance level α , the test rejects the null when $\mathcal{T}_{LRT} > c_\alpha$ where c_α is such that $\alpha = \Pr [\omega_1 \chi_1^2 + \dots + \omega_M \chi_M^2 \geq c_\alpha]$. Critical values c_α , necessary to evaluate \mathcal{T}_{LRT} , can be derived through Monte Carlo Simulations once a consistent estimator for V is obtained. In presence of homoskedastic and uncorrelated forecast errors, as it is the case for one-step ahead forecasts, a consistent estimator for V is given simply by the sample covariance $\hat{v} = P^{-1} \sum_{t=R}^T \left(\hat{f}_{t+1}^{adj} - \bar{f}^{adj} \right)^2$.

For the \mathcal{T}_{chi^2} statistic the asymptotic normality assumption implies that inference can be conducted through critical values from a χ_M^2 , while for Hubrich and West the critical values to evaluate $\mathcal{T}_{max t}$ for a given nominal size α can be found solving $\int_{-\infty}^{c_\alpha(\rho)} g_z(z) dz = 1 - \alpha$ where $g_z(z)$ denotes the density of the larger of m standard normal variables with correlation ρ .

4 Monte Carlo Simulation

We now outline in detail the experimental design for the Monte Carlo simulation and the procedures used to obtain the critical values. We present two sets of experiments, one simple design which has a general formulation and can be applied to many empirical studies and one more specific, suited to the evaluation of forecasts for inflation. The evaluation of the tests is implemented with asymptotic critical values derived through simulations for three settings: first assuming normality of the adjusted MSPE, simulating the limiting distribution of the test statistics and through resampling of the data.

4.1 Experimental Design

The first part of the simulation exercise requires the design of the DGP process for the size and the power experiment.

In the first design the process chosen as DGP for the size experiment is an order one autoregressive process of the form:

$$y_t = c + \rho y_{t-1} + \varepsilon_t \quad (2)$$

with $c = 1$, $\rho = 0.2$ and $\varepsilon_t \sim N(0, 1)$.

The process chosen as DGP for the power experiment is the following:

$$y_t = c + \rho y_{t-1} + \gamma x_{1t} + \varepsilon_t \quad (3)$$

with $c = 1$, $\rho = 0.2$, $\varepsilon_t \sim N(0, 1)$. The experiment is repeated for 3 different values of $\gamma = (0.2, 0.1, 0.05)$. The exogenous variable x_t is determined by:

$$x_{1t} = a + u_{1t} \quad (4)$$

with $a = 1$, $u_t \sim N(0, 1)$ and $u_{1t} \perp \varepsilon_t$.

Next we need to select the regression models. The model used for benchmark:

$$M0 : \quad y_t = \hat{c} + \hat{\rho} y_{t-1} \quad (5)$$

There are $M=2$ alternative models, each one of the form:

$$M_m : \quad y_t = \hat{c}_m + \hat{\rho}_m y_{t-1} + \hat{\varphi}_m x_{mt} \quad m = 1, 2. \quad (6)$$

where the extra regressor x_{2t} is generated from:

$$x_{2t} = a + u_{2t} \quad (7)$$

with $u_{2t} \perp \varepsilon_t$ and $u_{1t} \perp u_{2t}$.

In the design of the second experiment we follow HW and we assume the series y_t is an aggregate variable obtained as sum of a small number of components: $y_t = \sum_{l=1}^L x_{l,t}$. The disaggregate variables follow a VAR(1) process:

$$\mathbf{x}_t = \mathbf{a} + \Phi \mathbf{x}_{t-1} + \varepsilon_t$$

with $\mathbf{x}_t = (x_{1,t} \dots x_{L,t})'$, \mathbf{a} an $L \times 1$ vector of constants and $\varepsilon_t \sim N(0, \mathbf{I})$. For both the size and power experiment the aggregate is the sum of $L=3$ disaggregate series, and \mathbf{a} is a vector of ones. The VAR(1) regression coefficient varies when considering the size or the power experiment; for the size experiment:

$$\Phi = \begin{bmatrix} 0.5 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.5 \end{bmatrix},$$

while for the power experiment

$$\Phi = \begin{bmatrix} 0.5 & -0.6 & 0 \\ -0.4 & 0.3 & 0 \\ 0 & 0 & 0.5 \end{bmatrix},$$

Note that in the size experiment, the aggregate process is an AR(1) process with autoregressive parameter $\rho = 0.5$ and constant $c = 3$, while in the power experiment the aggregate is an ARMA(3,2) process. Next we need to select the regression models. The model used for benchmark is:

$$M0 : \quad y_t = \hat{c} + \hat{\rho} y_{t-1}. \quad (8)$$

There are $M=2$ alternative models, each one of the form:

$$M_m : y_t = \hat{c}_m + \hat{\rho}_m y_{t-1} + \hat{\varphi}_m x_{mt-1} \quad m = 1, 2 \quad (9)$$

The estimates are carried out through OLS, with rolling scheme, such that each estimation sample has the same size $R=\{40,100,200,400\}$. The forecasts are produced for horizon $h=1$ and the size of the out-of-sample is $P=\{40,100,200,400\}$.

We try 2 additional experiments for the power:

$$\Phi = \begin{bmatrix} 0.5 & -0.06 & 0 \\ -0.04 & 0.3 & 0 \\ 0 & 0 & 0.5 \end{bmatrix}, \quad (10)$$

and

$$\Phi = \begin{bmatrix} 0.5 & 0 & 0 \\ 0 & 0.3 & 0 \\ 0 & 0 & 0.5 \end{bmatrix}, \quad (11)$$

to investigate the sensitiveness of the test performances to the structure of the interdependence between the disaggregate series.

4.2 Evaluation of the tests

To proceed with the forecast evaluation the quantity \hat{f}_{mt+1}^{adj} is computed for each model and stacked in the vector $\hat{f}_{t+1}^{adj} = \left(\hat{f}_{1t+1}^{adj}, \dots, \hat{f}_{mt+1}^{adj}, \dots, \hat{f}_{Mt+1}^{adj} \right)'$. The sample average of \hat{f}_{t+1}^{adj} is denoted as \bar{f}^{adj} .

The forecast evaluation is carried out for three different tests: adjusted chi-squared, correlated normal and likelihood-ratio.

The chi-square test considers as statistic:

$$\mathcal{T}_{chi^2} = P \bar{f}^{adj'} \hat{v}^{-1} \bar{f}^{adj} \quad (12)$$

where \hat{v} is the estimated long run variance, defined as $\hat{v} = P^{-1} \sum_{t=R}^T \left(\hat{f}_{t+1}^{adj} - \bar{f}^{adj} \right)^2$.

The statistic for the correlated normal test is given by:

$$\mathcal{T}_{\max t} = \max \left[P^{1/2} \bar{f}_1^{adj} / \sqrt{\hat{v}_1}, \dots, P^{1/2} \bar{f}_M^{adj} / \sqrt{\hat{v}_M} \right] \quad (13)$$

with \hat{v}_i defined as above.

The statistic for the likelihood ratio test is given by:

$$\mathcal{T}_{LRT} = P \tilde{f}' \hat{v}^{-1} \tilde{f} \quad (14)$$

where \tilde{f} is the solution to the optimization problem: $\min_{\Pi f \geq 0} (\bar{f}^{adj} - f)' \hat{v}^{-1} (\bar{f}^{adj} - f)$ where Π can be the matrix D , \mathbb{D} or the identity matrix, depending on the structure of the alternative models.

As discussed above the adjusted MSPE does not have a normal limiting distribution. However, following Clark and McCracken (2001), CW and HW we conduct three sets of experiments: in the first one we embrace the normality assumption for the MSPE adjusted and derive critical values under this assumption; in the second we simulate the exact distribution of the statistic; the third one is based on resampling.

Under the normality assumption, the critical values for \mathcal{T}_{chi^2} are given by a χ_M^2 distribution. For the $\mathcal{T}_{\max t}$ the critical values depend on the correlation matrix of the forecast errors and need to be simulated. The procedure works as follows: first the sample correlation matrix of the f_{-adj} vector, $\hat{\Omega}$, is estimated. Then a vector of size M from a multivariate normal with zero mean and sample correlation $\hat{\Omega}$ is drawn. Next, the maximum element from this vector is selected. The experiment is repeated d times. The 1-alpha percentile of the simulated distribution is the $\alpha - th$ percent critical value. For the \mathcal{T}_{LRT} the critical values are generated as follows: first a vector Z of size $P \times 1$ is generated from a multivariate normal distribution with mean zero and variance-covariance matrix $\hat{v} : Z \sim N(0, \hat{v})$, with \hat{v} as defined above being the sample covariance matrix of the $M \times P$ matrix of prediction errors from the models of interest. Then \tilde{f}^{adj} is obtained by solving $\min_{Df \geq 0} (\bar{f}^{adj} - f)' \hat{v}^{-1} (\bar{f}^{adj} - f)$ (we label this as LRT_D),

when imposing a nested structure for the alternative models, or $\min_{If \geq 0} (\bar{f}^{adj} - f)' \hat{v}^{-1} (\bar{f}^{adj} - f)$ (LRT_I), when no structure is imposed for the alternative models. Next, the statistic $\mathcal{T}_{LRT} = f \overset{\sim adj'}{\hat{v}^{-1}} \overset{\sim adj}{f}$ is computed and the value of the test statistic is stored. The procedure is repeated d times and the critical values c_α is derived as the $(1 - \alpha)$ -th quantile of the simulated distribution.

The evaluation according to the 'true' limiting distribution is carried out in the following fashion: the statistics \mathcal{T}_{chi^2} , $\mathcal{T}_{max t}$ and \mathcal{T}_{LRT} are computed and stored d times with $d=10000$. Then the critical value used to evaluate the power is represented by the $(1 - \alpha) - th$ percentile of the simulated distributions under the null.

The subsampling procedure follows White (2000), which resorts to the stationary bootstrap procedure by Politis and Romano (1994). The algorithm to obtain the resampled MSPE difference, $\left\{ \hat{f}_{p+1}^* \right\}_{p=1}^P$ or equivalently $\left\{ \hat{f}_{t+1}^* \right\}_{t=R}^T$, works as follows: first the number of resamples, L , and the smoothing parameter q are specified and $\left\{ \hat{f}_{t+1} \right\}_{t=R}^T$, \mathcal{T}_{chi^2} , $\mathcal{T}_{max t}$ and \mathcal{T}_{LRT} are computed; the resampling is based on an index $s(p)$ for $p=1, \dots, P$; for $p=1$ a number $s(1)$ is drawn independently and uniformly from $\{R, \dots, T\}$; denote as U a standard uniform random variable on $[0,1]$, then, if $U \geq q$, $s(p+1) = s(p) + 1$ while, if $U < q$, $s(p+1)$ is drawn independently and uniformly from $\{R, \dots, T\}$; repeat the procedure until you reach a resampled vector of size P ; for each model $i=0, \dots, M$, the parameters β_m are estimated through $\hat{\beta}_{m,s(p)} = \left(\sum_{t=s(p)-R}^{s(p)} X'_{m,t} X_{m,t} \right)^{-1} \sum_{t=s(p)-R}^{s(p)} X_{m,t} y_t$; the forecast $\hat{y}_{s(p)+1} = X'_{m,s(p)+1} \hat{\beta}_{m,s(p)}$ and the forecast error $\hat{u}_{m,s(p)+1} = \hat{y}_{s(p)+1} - y_{s(p)+1}$ are constructed; it follows that $\hat{f}_{m,p+1}^* = \hat{u}_{m,s(p)+1}^2 - \hat{u}_{0,s(p)+1}^2$; compute the statistic of interest from the resampled MSPE difference $\mathcal{T}_{chi^2}^*$, $\mathcal{T}_{max t}^*$ and \mathcal{T}_{LRT}^* and store them; repeat the procedure L times and evaluate \mathcal{T}_{chi^2} , $\mathcal{T}_{max t}$ and \mathcal{T}_{LRT} through the quantiles of $\mathcal{T}_{chi^2}^*$, $\mathcal{T}_{max t}^*$ and \mathcal{T}_{LRT}^* .

Some preliminary results are shown in table 1 through 5. Table 1 reports the size and the unadjusted power for experiment 2 under the assumption

of normality of the limiting distribution of the statistics. Table 2 through 5 provide with some preliminary results for the adjusted power experiment for a 10 percent significance level for exercise 1 and exercise 2 under different specification for Φ .

**Table 1. Size and unadjusted Power for Aggregation
Experiment.(a).**

		40		100		200		400		
		size	power	size	power	size	power	size	power	
R	40	chi2	0.168	0.194	0.207	0.319	0.374	0.562	0.675	0.865
		HW	0.016	0.177	0.001	0.264	0	0.426	0	0.706
		LRT_I	0.023	0.234	0.002	0.384	0	0.618	0	0.882
		LRT_D	0.016	0.287	0.003	0.464	0	0.699	0	0.915
	100	chi2	0.122	0.29	0.138	0.456	0.176	0.764	0.313	0.978
		HW	0.034	0.36	0.014	0.482	0.004	0.751	0.001	0.96
		LRT_I	0.035	0.395	0.014	0.622	0.005	0.85	0.001	0.995
		LRT_D	0.035	0.468	0.014	0.699	0.007	0.905	0.001	0.999
	200	chi2	0.137	0.319	0.128	0.516	0.144	0.822	0.188	0.987
		HW	0.057	0.402	0.024	0.595	0.006	0.848	0.002	0.988
		LRT_I	0.058	0.46	0.021	0.666	0.004	0.916	0.001	0.997
		LRT_D	0.05	0.537	0.017	0.739	0.004	0.944	0.003	0.999
	400	chi2	0.139	0.333	0.108	0.559	0.122	0.824	0.156	0.991
		HW	0.073	0.424	0.05	0.638	0.02	0.832	0.011	0.992
		LRT_I	0.065	0.459	0.05	0.7	0.022	0.915	0.013	1
		LRT_D	0.051	0.54	0.037	0.752	0.015	0.943	0.014	1

Nominal size 10%. Test statistics evaluated under the normality assumption of MSPE differences. Φ matrix as in HW.

Table 2. Exact Power Aggregation Experiment (a).

		P				
		40	100	200	400	
R	40	chi2	0.138	0.150	0.183	0.253
		HW	0.466	0.798	0.956	0.998
		LRT_I	0.517	0.844	0.970	0.999
		LRT_D	0.625	0.894	0.983	0.999
	100	chi2	0.235	0.382	0.619	0.889
		HW	0.519	0.831	0.978	1.000
		LRT_I	0.589	0.888	0.987	1.000
		LRT_D	0.684	0.934	0.993	1.000
	200	chi2	0.259	0.485	0.750	0.973
		HW	0.495	0.800	0.973	1.000
		LRT_I	0.557	0.872	0.991	1.000
		LRT_D	0.660	0.921	0.996	1.000
	400	chi2	0.273	0.514	0.799	0.981
		HW	0.484	0.759	0.944	0.999
		LRT_I	0.535	0.832	0.980	1.000
		LRT_D	0.638	0.895	0.991	1.000

Nominal size 10%. Φ matrix as in HW.

Table 3. Exact Power Aggregation Experiment (b)

		P				
		40	100	200	400	
R	40	chi2	0.093	0.071	0.057	0.039
		HW	0.141	0.199	0.245	0.331
		LRT_I	0.136	0.180	0.214	0.224
		LRT_D	0.154	0.210	0.249	0.194
	100	chi2	0.090	0.076	0.061	0.040
		HW	0.169	0.218	0.318	0.472
		LRT_I	0.168	0.206	0.282	0.425
		LRT_D	0.192	0.256	0.352	0.509
	200	chi2	0.099	0.087	0.070	0.048
		HW	0.175	0.252	0.361	0.533
		LRT_I	0.169	0.241	0.351	0.488
		LRT_D	0.203	0.293	0.413	0.579
	400	chi2	0.098	0.097	0.101	0.087
		HW	0.194	0.267	0.359	0.529
		LRT_I	0.193	0.260	0.347	0.515
		LRT_D	0.248	0.329	0.439	0.610

Nominal size 10%. Φ matrix is diagonal.

Table 4. Exact Power Aggregation Experiment (c)

		P				
		40	100	200	400	
R	40	chi2	0.090	0.065	0.050	0.030
		HW	0.160	0.235	0.306	0.430
		LRT_I	0.155	0.218	0.273	0.327
		LRT_D	0.181	0.267	0.334	0.306
	100	chi2	0.091	0.077	0.062	0.041
		HW	0.192	0.264	0.405	0.604
		LRT_I	0.190	0.253	0.365	0.564
		LRT_D	0.232	0.321	0.459	0.659
	200	chi2	0.102	0.097	0.086	0.069
		HW	0.195	0.300	0.444	0.664
		LRT_I	0.190	0.290	0.435	0.627
		LRT_D	0.240	0.363	0.524	0.723
	400	chi2	0.103	0.113	0.133	0.137
		HW	0.211	0.304	0.427	0.641
		LRT_I	0.211	0.301	0.421	0.635
		LRT_D	0.277	0.387	0.524	0.730

Nominal size 10%. Φ matrix as in (10).

Table 5. Exact Power Experiment 1.

		P				
		40	100	200	400	
R	40	chi2	0.082	0.049	0.033	0.017
		HW	0.262	0.455	0.645	0.854
		LRT_I	0.258	0.437	0.605	0.659
		LRT_D	0.134	0.168	0.115	0.056
	100	chi2	0.106	0.103	0.089	0.101
		HW	0.282	0.486	0.737	0.934
		LRT_I	0.276	0.481	0.725	0.928
		LRT_D	0.128	0.168	0.224	0.252
	200	chi2	0.115	0.141	0.176	0.230
		HW	0.271	0.458	0.729	0.946
		LRT_I	0.264	0.443	0.722	0.943
		LRT_D	0.112	0.144	0.190	0.282
	400	chi2	0.131	0.175	0.246	0.392
		HW	0.264	0.441	0.667	0.918
		LRT_I	0.259	0.425	0.658	0.913
		LRT_D	0.099	0.120	0.155	0.213

Nominal size 10%. $\gamma = 0.2$

4.3 Simulation Results: Size

In the aggregation experiment under the normality assumption the HW and the LRT are highly undersized and the ranking of the tests depends on the particular combination of R and P considered. The chi-square test is always oversized, as found also in Hubrich and West. For all the tests the empirical size gets closer to the nominal size as the ratio P/R decreases. This is consistent with the theoretical framework in Clark and McCracken (2001): the Normal approximation holds asymptotically if the ratio P/R approaches zero as the total sample size increases and if the rolling estimation is used.

4.4 Simulation Results: Power

For all tests the power increases with the size of the out-of-sample. The effect of the increase in the size of the estimation sample depends on the test considered: for the chi-squared test the effect is positive, while for the other three tests it does not always hold. The performance of the chi-squared test is overall disappointing: for the aggregation experiment with critical values derived from the exact distribution the power of the chi-squared test is at most half of the power of the competitor tests; for the normality assumption the difference in the power is much lower, the chi-squared test ranks last. We do not provide with a theoretical explanation for the ranking of the correlated normal and LR test; however from an analysis of the Monte Carlo experiment results we find that when the eigenvalues of the estimated correlation matrix are close to zero, the HW test performs better, while for values of the eigenvalues far from zero the LRT ranks first. For experiment 1 imposing the additional constraint on the structure of the model penalizes the performance of the LRT_D test, while the correlated normal and the LRT_I are almost equivalent.

5 Conclusions

In this paper we introduced a likelihood ratio type predictability test for the comparison of a small number of models nested to each other. We distinguished among three cases according to the structure of the alternative models considered: a general case, in which the models can be grouped such that within each group the models are nested, but there is no nested relation among groups, and two extreme cases, one in which the models are nested within each other, one in which there is no nested relation between the models. We evaluated the size and power properties of the test via Monte Carlo simulations for 1-step ahead forecasts under different assumption on the limiting distributions of the statistics and for two simulation settings.

The Monte Carlo investigation reveals that the chi-square test performs poorly in terms of power as it disregards the one-sided nature of the test, while the ranking between the likelihood-ratio type test and the correlated normal test depends on the simulation frameworks. The normal approximation of the vector of MSPE differences, assumed in previous studies for multi-model comparison, proves to be reasonable for P/R going to zero, as found in CW and earlier on in Clark and McCracken (2001, 2005).

The relative performance of the LRT and HW test depends on the parametrization of the Monte Carlo experiment: when the eigenvalues of the correlation matrix of the adjusted MSPE differences are close to zero, i.e. when the variance covariance matrix of the adjusted MSPE differences is close to singular the LRT does worse than HW, while for bigger eigenvalues the LRT ranks better.

References

- [1] Ashley, R., C.W.J.Granger and R.Schmalense (1980), 'Advertising and Aggregate Consumption: an Analysis of Causality', *Econometrica*, vol.48 n.5.
- [2] Clark, T.E. and M. W. McCracken (2001), 'Tests of Equal Forecast Accuracy and Encompassing for Nested Models', *Journal of Econometrics*, vol.105, 85-110.
- [3] Clark, T.E. and K. West, (2006), 'Using out-of-sample Mean Squared Prediction Errors to test the Martingale Difference Hypothesis', *Journal of Econometrics*, vol.138, 291-311.
- [4] Clark, T.E. and K. West, (2007), 'Approximately Normal Tests for Equal Predictive Accuracy in Nested Models', *Journal of Econometrics*, vol.138, 291-311.
- [5] Diebold, F.X., and R.S. Mariano, (1995), 'Comparing Predictive Accuracy', *Journal of Business and Economic Statistics*, vol.13, 253-263.
- [6] Hubrich, K. and K. D. West, (2008), 'Forecast Evaluation of Small Nested Model Sets', mimeo.
- [7] Perlman, M.D., (1969), 'One-Sided Testing Problems in Multivariate Analysis', *The Annals of mathematical Statistics*, vol.40, 549-567.
- [8] Politis, D., and J. Romano, (1994), 'The Stationary Bootstrap', *Journal of the American Statistical Association*, vol.89, 1303-1313.
- [9] West, K.D., (1996) 'Asymptotic Inference about Predictive Ability', *Econometrica*, vol.64, 1067-1084.
- [10] White, H., (2000) 'A Reality Check for Data Snooping', *Econometrica*, vol.68, 1097-1126.

6 Appendix

In the following we will prove that the ranking of the MSPE differences is invariant to the introduction of the adjustment suggested by Clark and West. We will illustrate starting from an $M=1$ setting to derive a more general prove.

- There are two OLS fitted values: $\hat{Y}_1 = P_{X_1}Y$ and $\hat{Y}_2 = P_{X_2}Y$, where $X_2 = [X_{21}, X_{22}]$ and $X_1 = X_{21}$. Show that $\hat{Y}_1'\hat{Y}_1 \leq \hat{Y}_2'\hat{Y}_2$.

Proof. By definition $Y = \hat{Y}_k + \hat{U}_k$, for $k = 1, 2$. Also notice that $\hat{Y}_k'\hat{U}_k = 0$. Then, $Y'Y = \hat{Y}_1'\hat{Y}_1 + \hat{U}_1'\hat{U}_1 = \hat{Y}_2'\hat{Y}_2 + \hat{U}_2'\hat{U}_2$. By definition, $\hat{U}_2'\hat{U}_2 \leq \hat{U}_1'\hat{U}_1$. Therefore, $\hat{Y}_1'\hat{Y}_1 \leq \hat{Y}_2'\hat{Y}_2$.

- There are three OLS fitted values: $\hat{Y}_0 = P_{X_0}Y$, $\hat{Y}_1 = P_{X_1}Y$, and $\hat{Y}_2 = P_{X_2}Y$, where $X_1 = [X_{11}, X_{12}]$ and $X_{11} = X_0$, and $X_2 = [X_{21}, X_{22}]$ and $X_{21} = X_1$. Show that $(\hat{Y}_1 - \hat{Y}_0)'(\hat{Y}_1 - \hat{Y}_0) \leq (\hat{Y}_2 - \hat{Y}_0)'(\hat{Y}_2 - \hat{Y}_0)$.

Proof.

$$\begin{aligned}
& (\hat{Y}_2 - \hat{Y}_0)'(\hat{Y}_2 - \hat{Y}_0) - (\hat{Y}_1 - \hat{Y}_0)'(\hat{Y}_1 - \hat{Y}_0) \\
&= Y'(P_{X_2} - P_{X_0})(P_{X_2} - P_{X_0})Y - Y'(P_{X_1} - P_{X_0})(P_{X_1} - P_{X_0})Y \\
&= Y'[(P_{X_2} - P_{X_0})(P_{X_2} - P_{X_0}) - (P_{X_1} - P_{X_0})(P_{X_1} - P_{X_0})]Y \\
&= Y'[(P_{X_2} - P_{X_0})(P_{X_2} - P_{X_0}) - (P_{X_1} - P_{X_0})(P_{X_1} - P_{X_0})]Y \\
&= Y'(P_{X_2} - P_{X_1})Y \\
&\geq 0
\end{aligned}$$

where the last equality holds since $P_{X_2}P_{X_0} = P_{X_1}P_{X_0} = P_{X_0}$ and the last inequality holds by Problem 1.

Now we can reproduce the same type of results in population with L_2 -linear projection. First define models as follows: for $m = 0, \dots, M$,

$$\begin{aligned} M_0 & : Y = \beta'_0 x_0 + U_0 = \beta'_0 X_0 + U_0, \text{ where } X_0 = x_0 \\ M_1 & : Y = \beta'_{10} x_0 + \beta'_{11} x_1 + U_1 = \beta'_1 X_1 + U_1, \text{ where } X_1 = [x'_0, x'_1]' \\ & \vdots \\ M_m & : Y = \beta'_{m0} x_0 + \beta'_{m1} x_1 + \dots + \beta'_{mm} x_m + U_m = \beta'_m X_m + U_m, \text{ where } X_m = [x'_0, x'_1, \dots, x'_m]' \end{aligned}$$

and

$$\beta_m = E(X_m X_m')^{-1} E(X_m Y).$$

Denote

$$Y_m = P_{X_m} Y = \beta'_m X_m.$$

For random variables Z_1 and Z_2 , $\|Z_1\| = (E(Z_1^2))^{1/2}$ signifies the L_2 norm and $\langle Z_1, Z_2 \rangle$ is the inner product defined by $E(Z_1 Z_2)$. Then, by definition

$$\langle Y_m, U_m \rangle = 0.$$

- Show that $P_{X_k} U_m = 0$ for any $k \leq m$.

Proof. Notice by definition,

$$P_{X_k} U_m = X'_k E(X_k X_k')^{-1} E(X_k U_m).$$

The required result follows because X_k is a subcomponent of X_m and

$$E(X_m U_m) = E(X_m Y) - E(X_m X_m') \beta_m = 0.$$

- Show that $\|U_m\|^2 \geq \|U_{m+1}\|^2$.

Proof. The required result follows since

$$\begin{aligned}
\|U_m\|^2 &= \|Y - P_{X_m}Y\|^2 \\
&= \|Y_{m+1} - P_{X_m}Y + U_{m+1}\|^2 \\
&= \|Y_{m+1} - P_{X_m}Y_{m+1} + U_{m+1}\|^2 \text{ since } P_{X_m}U_{m+1} = 0 \\
&= \|Y_{m+1} - P_{X_m}Y_{m+1}\|^2 + \|U_{m+1}\|^2 + 2\langle Y_{m+1} - P_{X_m}Y_{m+1}, U_{m+1} \rangle \\
&= \|Y_{m+1} - P_{X_m}Y_{m+1}\|^2 + \|U_{m+1}\|^2 \text{ since } \langle Y_{m+1}, U_{m+1} \rangle = 0, \langle P_{X_m}Y_{m+1}, U_{m+1} \rangle = 0.
\end{aligned}$$

- Show that $\|Y_m\|^2 \leq \|Y_{m+1}\|^2$.

Proof. By definition,

$$\|Y\|^2 = \|Y_m\|^2 + \|U_m\|^2 = \|Y_{m+1}\|^2 + \|U_{m+1}\|^2.$$

The required result follows since $\|U_m\|^2 \geq \|U_{m+1}\|^2$.

- Show that $\|Y_m - Y_k\|^2 = \|Y_m\|^2 - \|Y_k\|^2$ for any $k < m$.

Proof. By definition

$$\begin{aligned}
&\|Y_m - Y_k\|^2 \\
&= \|Y_m - P_{X_k}Y\|^2 \\
&= \|Y_m - P_{X_k}Y_m - P_{X_k}U_m\|^2 \\
&= \|Y_m - P_{X_k}Y_m\|^2 \text{ since } P_{X_k}U_m = 0 \\
&= \|Y_m\|^2 - \|P_{X_k}Y_m\|^2 \\
&= \|Y_m\|^2 - \|P_{X_k}Y\|^2 \\
&= \|Y_m\|^2 - \|Y_k\|^2.
\end{aligned}$$

- Show that $\mu_m = \|U_0\|^2 - \|U_m\|^2 + \|Y_m - Y_0\|^2 = 2\|U_0\|^2 - 2\|U_m\|^2$.

Proof. The result follows since

$$\begin{aligned}\mu_m &= \|U_0\|^2 - \|U_m\|^2 + \|Y_m - Y_0\|^2 \\ &= \|U_0\|^2 - \|U_m\|^2 + \|Y_m\|^2 - \|Y_0\|^2 \\ &= \|U_0\|^2 - \|U_m\|^2 + \|Y\|^2 - \|U_m\|^2 - \|Y\|^2 + \|U_0\|^2 \\ &= 2\|U_0\|^2 - 2\|U_m\|^2.\end{aligned}$$