Preliminary and incomplete.

FORECAST EVALUATION OF SMALL NESTED MODEL SETS

Kirstin Hubrich
European Central Bank

Kenneth D. West
University of Wisconsin

November 2007

Abstract

We propose and compare procedures for inference about mean squared prediction error (MSPE) when comparing a benchmark model against a small number of alternatives that nest the benchmark. We evaluate two procedures that adjust MSPE differences in accordance with Clark and West (2007); one examines the maximum t-statistic, the other computes a chi-squared statistic. We also examine two procedures that do not adjust the MSPE differences: a chi-squared statistic, and White's (2000) reality check. In our simulations, the two statistics that adjust MSPE differences have most accurate size, and the procedure that looks at the maximum t-statistic has best power. We illustrate our procedures by comparing forecasts of different models for U.S. inflation.

Keywords: Out-of-sample, prediction, testing, multiple model comparisons, inflation forecasting

We thank Eleanora Granziera and Roberto Duncan for research assistance. West thanks the National Science Foundation for financial support. The views expressed here are not necessarily those of the European Central Bank.

1. INTRODUCTION

Forecast evaluation frequently involves comparison of a small set of models, one of which is a null model nested in some or all of the alternative models. There are two broad classes of applications. In one class, applicable to studies of asset returns, the null model is a martingale difference, perhaps with drift (i.e., a random walk or random walk with drift for the asset price). Examples include Hong and Lee (2003), who study exchange rates, and Sarno et al. (2005), who study interest rates; each paper compares a random walk to a half dozen or so other models. In the second class of applications, the null model sometimes relies on stochastic predictors, typically via a univariate autoregression. Examples include Billmeier (2004), who compares a univariate autoregression (AR) to four other models that include four different measures of the output gap, and Hubrich (2005) and Hendry and Hubrich (2006, 2007), who compare univariate forecasts of aggregate inflation to a couple of other forecast models that add disaggregate components of inflation to the univariate model. This class of applications is important at policy institutions or for policy observers where it is of interest to compare forecasts from different models in a suite of models built to account for different aspects of the economy.

Our aim in this paper is to propose and evaluate procedures for performing inference about equality of mean squared predictions errors (MSPEs) in applications, such as these, that involve a small number of models. We do not have a precise definition of "small." But, loosely, the idea is that the number of alternative models m is much less than the sample size T.

There are at least two existing procedures. Both use a $m \times 1$ vector whose elements consist of the difference between the MSPE of the null model and the MSPE of one of the alternative models. To test the null of equality of MSPEs across the models, one approach is to conduct the chi-squared test that is the straightforward generalization of the Diebold and Mariano (1995) and West (1996) (DMW) statistic that is used to compare a pair of models. This chi-squared statistic was used in West et al. (1993) and West and Cho (1995). It was also proposed by Giacomini and White (2006) in a context closely related

to ours. It is referenced in our paper as "χ^2 that does not adjust MSPE differences" or "χ^2 (unadj.);" the reason for the qualification "unadjusted" will become clear shortly. Under our null hypothesis, however, this statistic is flawed in terms of both size and power. In terms of size: under a reasonable set of technical assumptions, the statistic is unlikely to be well-approximated by a chi-squared, because the vector of MSPE differences is not centered at zero, even under the null. We explain this point in section 2 below. In terms of power: as argued by Ashley et al. (1980), the alternative in question is one-sided. So even if the statistic is adjusted so as to be centered at zero under the null, a large chi-squared value can reflect extreme behavior in either tail of the underlying distribution, and thus this statistic potentially has poor power.

A second procedure, or perhaps we should say class of procedures, is to obtain critical values on the distribution of the vector of MSPE differences via simulation. One such possibility is White's (2000) reality check. While White (2000) proposed his procedure in the context of applications with many ($m \approx T$) rather than a small ($m << T$) number of nested models, the technique has also been applied to small sets of nested models (Hong and Lee (2003)). A possible problem is that White's procedure might not accurately account for dependence of predictions on estimated regression parameters (a key aspect of the computational appeal of White's procedure is that it does not require reestimation of models during bootstrap repetitions). Alternatively, one could bootstrap in a fashion that includes reestimation of models (e.g., Rapach and Wohar (2006)). Such a bootstrap has been found to work well (Clark and McCracken (2006), Clark and West (2007)). Nevertheless, in our own applied work, and, we presume, in the applied work of some others, it will at times be desirable to have procedures that are not computationally intensive.

In this paper, we develop two closely related procedures, and compare them, via simulations, to the unadjusted chi-squared and White's (2000) reality check. The new procedures use Clark and West's (2007) *adjustment* of the $m \times 1$ vector of MSPE differences. This is intended to center the vector at zero,

under the null. Let model 0 denote the benchmark model, and number the alternative models 1 to *m*.

Our main proposal involves two steps: (a) use the adjusted MSPE differences to compute *m* Clark and West (2007) "MSPE-adjusted" t-statistics, one of which compares model 0 to model 1, the second of which compares model 0 to model 2, ..., the last of which compares model 0 to model *m*; and (b) conduct inference on the largest of the *m* t-statistics via the distribution of the maximum of correlated normals. In our tables, this is called "max t-stat (adj.)," where the qualifier "adj." signals use of adjusted MSPEs. Step (b) respects the one-sided nature of the alternative, and is intended to lead to good power. When there are only two alternative models in addition to the benchmark model, as in the simulations in the current draft of the paper, critical values for this test vary with a single parameter, namely, the correlation between the two t-statistics. We include a table that presents critical values for 10% and 5% tests, for a crude grid of possible correlations. We supply detailed critical values for a fine grid of correlations in a not-for-publication appendix.

Our second proposal is to compute a conventional $\chi^2(m)$ statistic from the $m \times 1$ vector of Clark and West (2007) MSPE-adjusted values. Since this procedure uses the adjusted differences, we conjecture that it will be well-sized. But since it uses both tails of the distribution, it is likely to have less power than does the procedure that looks to the maximum of the individual t-statistics. This procedure is denoted "χ^2 (adj.)" in our tables and is sometimes referenced in our text as "χ^2 statistics based on the adjusted MSPE differences."

In the limited set of simulations completed to date, we find the following. The max t-stat (adj.) statistic is slightly undersized, the χ^2 (adj.) statistic is slightly oversized. The χ^2 statistic used in West et al. (1993) and West and Cho (1995)–referenced as "χ^2 (unadj.) in our tables, because it is computed from the usual rather than from adjusted MSPE differences–is somewhat, and for small sample sizes grossly, oversized; the reality check statistic is somewhat, and for small sample sizes grossly, undersized. In terms of power (not adjusted for size), as expected, max t-stat (adj.) has higher power than the χ^2 (adj.) statistic

(although the differences are found not to be large); the χ^2 (adj.) statistic in turn has greater power than either the reality check or the χ^2 (unadj.) statistics (often substantially higher power, as it turns out).

Section 2 motivates our two new procedures. Section 3 gives a precise statement of the environment and the statistics we compute. While the statement is precise, the argument is informal: we do not prove any theorems, but instead refer the reader to other literature. Section 4 gives an overview of our simulation. Section 5 presents simulation results. Section 6 presents an empirical example.

2. OVERVIEW AND INTUITION

The starting point for our procedures is the following observation in Clark and West (2006, 2007). Suppose we are comparing a parsimonious model to a larger model. The parsimonious model is nested in the larger model. Under the null that the additional variables in the larger model have coefficients that in population are zero, the more parsimonious model has a strictly smaller out-of-sample mean squared prediction error (MSPE). This is because the attempt to estimate coefficients whose population values are zero inflates the variance of the prediction error of the larger model.

Figure 2.1 illustrates the logic spelled out in detail (and with algebra) in Clark and West (2006, 2007). The figure depicts some densities of the difference between the MSPE from null model and the MSPE from an alternative, larger model, or, in self evident notation, $\hat{\sigma}_0^2 - \hat{\sigma}_1^2$. The alternative model estimates coefficients whose population values are zero. The densities were obtained by smoothing MSPE differences that were generated across 1,000 simulations, using the data generating process described in the simulations below. The top panel (Figure 2.1A) is one in which the number of predictions *P* used to construct MSPEs was held constant at 100; the number of observations *R* used in the rolling sample to compute predictions varied from 40 to 400. All the densities are centered below zero. This is because, on average, the null model has a strictly smaller sample MSPE than does the alternative model. As the regression sample size *R* increase, the densities shift towards zero. This is

because a larger sample typically delivers estimates of coefficients closer to their population values of zero. Hence the inflation of the MSPE in the alternative model diminishes as R increases.

The lower panel is one in which the regression sample size R is held fixed at 100, but the number of predictions P varies from 40 to 200. The difference in MSPEs stays centered at approximately the same value, but the distribution gets tighter and tighter around that value. This is because the law of large numbers causes the difference in MSPEs to pile up at the expected difference in MSPEs.

Clark and West (2006, 2007) proposed adjusting the difference in MSPEs between a pair of models to account for the inflation of the variance of the prediction error of the larger model. This adjustment centers the difference at zero, and is intended to produce a test statistic with good size. We will describe this adjustment in the next section.

We propose multivariate extensions of Clark and West (2006, 2007) applicable for testing a parsimonious benchmark model against a set of $m > 1$ other models, with the parsimonious model nested within all the other models. To achieve good size, we recommend adjusting the difference between the MSPE of the benchmark model and each of the m other models in the fashion proposed by Clark and West. As stated in the introduction, one can then take either of two steps. One can respect the fact that the alternative is one sided and, in a generalization of Clark and West (2006, 2007), look to the maximum of the m t-statistics produced after adjusting MSPE differences ("max t-stat (adj.)"). Critical values here are nonstandard, but can be obtained by simulation or, in simple cases, by reference to tabulations. Alternatively, one can simply construct a chi-squared statistic in straightforward fashion, computing a long run variance if predictions are multistep or for other reasons ("χ^2 (adj.)"). In this case, one looks to critical values of a $\chi^2(m)$ random variable.

3. ECONOMETRIC PROCEDURE

We suppose that there are $m + 1$ models under consideration. Each of the models is to be used to

predict a scalar y_t. For expositional clarity, we assume in this section that $m = 2$ and that the forecast horizon is one step ahead. (Generalization to arbitrary m is straightforward. As well, the procedures about to be described extend immediately to multistep forecasts using the direct method, though, as noted below, the theoretical justification for our procedure does not always extend.) Model 0 is a parsimonious benchmark model nested in alternative models 1 and 2. For example, model 0 might be a univariate autoregression in y_t, models 1 and 2 bivariate and trivariate vector autoregressions in which the right hand side variables include lags of y_t. Alternatively, model 1 might nest model 0 by adding lags of a second variable while model 2 adds lags of a third variable. Thus, while model 0 is nested in models 1 and 2, model 1 may or may not be nested in model 2 and model 2 may or may not be nested in model 1.

3.1 Mechanics

Write the null and two alternative models as

(3.1) $y_t = X_{0t}'\beta_0^* + e_{0t}$,

$y_t = X_{1t}'\beta_1^* + e_{1t}$,

$y_t = X_{2t}'\beta_2^* + e_{2t}$.

By assumption X_{0t} is a strict subset of X_{1t} and of X_{2t}. Our dating convention allows (indeed, presumes) that for each model, X_{it} is observed prior to period t. For example, we might have $X_{0t} = (1\ y_{t-1})'$, $X_{1t} = (1\ y_{t-1}\ y_{t-2})'$, $X_{2t} = (1\ y_{t-1}\ z_{t-1})'$ for some z that is observed in period t-1 (or $X_{2t} = (1\ y_{t-1}\ y_{t-2}\ y_{t-3}\ y_{t-4})'$–again, models 1 and 2 may or may not be nested in one another). It is possible that $X_{0t} \equiv 0$, i.e., that the null model presumes that y_t is white noise. The β^*'s are understood to be linear projections, with e_{it} by construction orthogonal to X_{it}. The assumption of linearity is for expositional convenience; methods such as nonlinear least squares are allowed by our test procedures.

Under the null, the coefficients on the additional regressors in X_{1t} and X_{2t} are zero. (In the example, just given, this means that the coefficients on y_{t-2} in X_{1t} and on z_{t-1} in X_{2t} are zero.) That is,

under the null, $X_{0t}'\beta_0^* = X_{1t}'\beta_1^* = X_{2t}'\beta_2^*$ and $e_{0t} = e_{1t} = e_{2t}$. Under the alternative, at least one of the additional regressors in X_{1t} and/or X_{2t} has a nonzero coefficient. For $i = 0,1,2$, let $\sigma_i^2 \equiv Ee_{it}^2$ denote the population variance of the forecast error.[1] We have

(3.2) H_0: $\sigma_0^2 - \sigma_1^2 = 0$, $\sigma_0^2 - \sigma_2^2 = 0$; H_A:$\max(\sigma_0^2 - \sigma_1^2, \sigma_0^2 - \sigma_2^2) > 0$.

Note that the alternative is one-sided. This is in accordance with Ashley et al. (1980) and a long list of subsequent studies. If, indeed, one or more of the coefficients in β_1^* or β_2^* are nonzero, then σ_1^2 or σ_2^2 must be less than σ_0^2.

Define the following notation, putting aside for the moment details such as whether a rolling or recursive scheme is used to generate prediction errors:

(3.3)(a) $\hat{\beta}_{it}$: an estimate of β_i^* computed using period t or earlier data, $i = 0,1,2$;

(b) $\hat{y}_{it+1}$: the one step ahead forecast from model i, ($i = 0,1,2$), $\hat{y}_{it+1} = X_{it+1}'\hat{\beta}_{it}$;

(c) $\hat{e}_{it+1}$:one step ahead prediction error from model i ($i = 0,1,2$), $\hat{e}_{it+1} \equiv y_{t+1} - \hat{y}_{it+1}$;

(d) P: the number of predictions and prediction errors;

(e) $\hat{\sigma}_i^2$: mean squared prediction error (MSPE) from model i ($i = 0,1,2$), $\hat{\sigma}_i^2 \equiv P^{-1}\sum_t \hat{e}_{it+1}^2$;

(f) $\hat{\sigma}_i^2$-adj: Clark and West's (2007) adjusted MSPE for models $i = 1, 2$,

$$\hat{\sigma}_i^2\text{-adj} = \hat{\sigma}_i^2 - P^{-1}\sum_t(\hat{y}_{0t+1} - \hat{y}_{it+1})^2;$$

(g) $\hat{f}_{it+1} \equiv \hat{e}_{0t+1}^2 - \hat{e}_{it+1}^2 + (\hat{y}_{0t+1} - \hat{y}_{it+1})^2$ ($i = 1,2$);

(h) $\bar{f}_i$: the adjusted difference in MSPEs between model i ($i = 1,2$) and model 0,

$$\bar{f}_i = \hat{\sigma}_0^2 - (\hat{\sigma}_i^2\text{-adj}) = P^{-1}\sum_t \hat{f}_{it+1};$$

(i) $\hat{v}_i$: an estimate of a long run variance computed using autocovariances of $\hat{f}_{it+1}$ ($i = 1,2$) (typically, $\hat{v}_i$ = sample variance of $\hat{f}_{it+1}$, though one might want to allow for non-zero autocovariances to guard against misspecification that leads to serial correlation);

(j) $P^{1/2}\bar{f}_i/\sqrt{\hat{v}_i}$: for $i = 1,2$, the *MSPE-adjusted* t-statistic.

Clark and West (2006, 2007) argue that for the purpose of comparing model 0 to model 1, one can compute the MSPE-adjusted t-statistic $P^{½}\bar{f}_1/\sqrt{\hat{v}_1}$ and use standard normal critical values, i.e., one can assume $P^{½}\bar{f}_1/\sqrt{\hat{v}_1} \sim_A N(0, 1)$; similarly, one can compare model 0 to model 2 via $P^{½}\bar{f}_2/\sqrt{\hat{v}_2} \sim_A N(0, 1)$. This motivates us to assume the following when we conduct inference:

$$(3.4) \qquad P^{½}\begin{pmatrix} \dfrac{\bar{f}_1}{\sqrt{\hat{v}_1}} \\ \dfrac{\bar{f}_2}{\sqrt{\hat{v}_2}} \end{pmatrix} \sim_A N(0, \Omega), \ \Omega \equiv \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

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

$$(3.5) \qquad \hat{z} = \max[P^{½}\bar{f}_1/\sqrt{\hat{v}_1}, P^{½}\bar{f}_2/\sqrt{\hat{v}_2}] \equiv \text{max t-stat (adj.)}.$$

Consider a test at the α level of significance. Let $g_z(z)$ denote the density of the larger of two standard normal random variables with correlation ρ. Let $c_\alpha(\rho)$ be such that $\int_{-\infty}^{c\alpha(\rho)} g_z(z)dz = 1 - \alpha$. We propose rejecting the null in favor of the alternative if $\hat{z} > c_\alpha(\hat{\rho})$, where $\hat{\rho}$ is the sample correlation between the two t-statistics $\bar{f}_1/\sqrt{\hat{v}_1}$ and $\bar{f}_2/\sqrt{\hat{v}_2}$.

To use this result requires knowledge of the quantiles of $g_z(z)$. Let ϕ denote a standard normal density, Φ a standard normal CDF. Let Ω be the $2 \times \times 2$ variance-covariance matrix of a standard bivariate normal with off-diagonal element ρ. The density of the maximum of a bivariate $N(0, \Omega)$ random variable is (Cain (1994), Ker (2001))

$$(3.6) \qquad g_z(z) = 2\phi(z)\Phi(kz) \text{ where } k \text{ is the constant } k = (1 - \rho)/(1 - \rho^2)^{½} = [(1 - \rho)/(1 + \rho)]^{½}.$$

Table 3.1 contains 10% and 5% critical values for a range of values of ρ. These were computed for $\rho \neq -1$ by numerically solving for the value of c such that $\int_{-\infty}^{c} g_z(z)dz = 0.90$ or $\int_{-\infty}^{c} g_z(z)dz = 0.95$, for $\rho = -1$ using logic about to be described. The entries for positive ρ may also be found in Gupta et al. (1973). The entries for $\rho = -1$, $\rho = 1$ and $\rho = 0$ are intuitive or familiar. Let z_1 and z_2 denote two standard

normal variables. If $\rho = -1$, then $z_1 = -z_2$ and prob[max(z_1, z_2) > c] = prob[$z_1 > c$] + prob[$z_1 < -c$], so a 10% critical value is $c = 1.645$ (since prob[$z_1 > 1.645$] + prob[$z_1 < -1.645$] = .10). If $\rho = 1$, then $z_1 = z_2$ and prob[max(z_1, z_2) > c] = prob($z_1 > c$) so a 10% critical value is $c = 1.282$. If $\rho = 0$, familiar results on order statistics from independent observations tell us that the 10% critical value satisfies $\Phi(c)^2 = .90$, yielding the value of $c = 1.632$ given in the table. The critical values fall monotonically as ρ rises, initially with little change, but with an accelerating decline as ρ nears 1.

We look at three other statistics, in addition to the maximum of the MSPE-adjusted statistics.

•χ^2 using adjusted MSPEs. Define:

(3.7)(a) $\hat{f}_{t+1} \equiv (\hat{f}_{1t+1}, \hat{f}_{2t+1})' \equiv (\hat{e}^2_{0t+1} - \hat{e}^2_{1t+1} + (\hat{y}_{0t+1} - \hat{y}_{1t+1})^2, \hat{e}^2_{0t+1} - \hat{e}^2_{2t+1} + (\hat{y}_{0t+1} - \hat{y}_{2t+1})^2)'$,

(b) $\hat{V} \equiv P^{-1}\sum_t (\hat{f}_{t+1} - \bar{f})(\hat{f}_{t+1} - \bar{f})'$.

The diagonal elements of $\hat{V}$ are $\hat{v}_1$ and $\hat{v}_2$, defined in (3.3(i)). Then

(3.8) χ^2 (adj.) $\equiv P\bar{f}'\hat{V}^{-1}\bar{f}$.

We evaluate (3.8) using $\chi^2(2)$ critical values.

•χ^2 using unadjusted MSPEs: Use "~" on top of a quantity to define MSPE differences that are not adjusted as in Clark and West (2007):

(3.9)(a) $\tilde{f}_{it+1}$: $\hat{e}^2_{0t+1} - \hat{e}^2_{it+1}$ ($i = 1,2$);

(b) $\bar{\tilde{f}}_i = \hat{\sigma}^2_0 - \hat{\sigma}^2_i = P^{-1}\sum_t \tilde{f}_{it+1}$ ($i = 1,2$);

(c) $\tilde{f}_{t+1} \equiv (\tilde{f}_{1t+1}, \tilde{f}_{2t+1})' \equiv (\hat{e}^2_{0t+1} - \hat{e}^2_{1t+1}, \hat{e}^2_{0t+1} - \hat{e}^2_{2t+1})'$;

(d) $\bar{\tilde{f}} \equiv (\bar{\tilde{f}}_1, \bar{\tilde{f}}_2)' = P^{-1}\sum_t \tilde{f}_{t+1}$;

(e) $\tilde{V} \equiv P^{-1}\sum_t (\tilde{f}_{t+1} - \bar{\tilde{f}})(\tilde{f}_{t+1} - \bar{\tilde{f}})'$;

(f) χ^2 (unadj.): $P\bar{\tilde{f}}'\tilde{V}^{-1}\bar{\tilde{f}}$.

We evaluate (3.9(f)) using $\chi^2(2)$ critical values. For clarity, we observe that the adjusted and unadjusted MSPE differences are related via

(3.10) *i*'th adjusted MSPE difference = *i*'th unadjusted MSPE difference + $P^{-1}\sum_t(\hat{y}_{0t+1} - \hat{y}_{it+1})^2$

•**White's (2000) reality check**. See discussion of simulations below.

3.2 Mechanics, more complex settings

The grid supplied in Table 3.1 for the case of $m = 2$ alternative models is coarse. For this case, values across steps in ρ of 0.01 are available on request. More generally, for any *m*, one can compute the p-value of a "max MSPE-adj. t-statistic" by a simple simulation. One computes *m* MSPE-adjusted t-statistics, and constructs and $m \times m$ matrix $\hat{\Omega}$; here, the *i, j* element of $\hat{\Omega}$ is the sample correlation between $\hat{\sigma}_0^2$ - ($\hat{\sigma}_i^2$-adj) and $\hat{\sigma}_0^2$ - ($\hat{\sigma}_j^2$-adj). One then does a series of draws (say, 50,000 draws) on a N(0, $\hat{\Omega}$) random vector, and, for each draw, saves the largest of the *m* elements of that draw's random vector. The p-value for sample maximum MSPE-adjusted statistic is computed from the distribution of maxima from the simulation.

The statistics defined in (3.8) and (3.9(f)) generalize immediately to an environment with $m > 2$.

All these statistics also generalize immediately to multistep forecasts executed using the direct method. $\hat{V}$ (3.7(b)) and $\tilde{V}$ (3.9(e)) become estimates of a long run variance; the diagonal elements of $\hat{V}$ are used in the denominator of the MSPE-adjusted statistics, and the off-diagonal correlations in $\hat{\Omega}$ are compute from the off–diagonal elements of the long run variance estimate $\hat{V}$. With these changes, the formulas above are applicable.

3.3 Theoretical justification

As a formal matter, the max t-stat (adj.) and χ^2 (adj.) procedures require that $m \times 1$ vector $P^{½}\bar{f}$ be asymptotically normal with a variance that can be estimated in standard fashion. Under technical conditions such as in Giacomini and White (2006), it is straightforward to show that this holds when

(a) the null model posits that y_t is a martingale difference (i.e., $X_{0t} \equiv 0$, $\hat{y}_{0t+1} \equiv 0$ for all t), and (b) rolling samples are used to generate the regression estimates.[2] Under the conditions just stated, asymptotic normality also follows for multistep prediction errors if predictions are made using the direct method.[3]

Alternatively, under the technical conditions of Clark and McCracken (2001), asymptotic normality follows if the number of predictions P is very small relative to the number of observations R used in the first regression sample used to estimate the β^*'s. The precise requirement is that $P/R \to 0$ as the total sample size grows. This result holds for both recursive and rolling samples, and does not require that the null model be a martingale difference. It does require one-step ahead predictions. Extension to multistep predictions has been worked out only in special cases (Clark and McCracken (2005)).

The conditions of the previous two paragraphs do not by any means span the environment of applications that compare small sets of nested models. But the argument of Clark and West (2007) suggests that the quantiles of the right tail of the t-statistics described above will be approximately those of a standard normal in a wide range of environments. Hence the max t-stat (adj.) procedure should yield tests that are approximately accurately sized. In particular, using numerical methods, Clark and McCracken (2001) have tabulated critical values for the adjusted t-statistic, which they call "enc-t". These critical values assume that $P, R \to \infty$. The critical values depend on the limiting value of P/R, on the regression scheme (rolling vs. recursive) and on the number of extra regressors in the larger model (i.e., on the difference between the dimension of X_{1t} and X_{0t} or between X_{2t} and X_{0t}). But apart from a handful of exceptions, for all tabulated values of P/R and the number of extra regressors, the critical values obey the following inequalities:

(3.11) .90 quantile $\leq$ 1.282 $\leq$ .95 quantile $\leq$ 1.645 $\leq$ .99 quantile.

For a standard normal, the .90 quantile is of course 1.282 and the .95 quantile is 1.645. Hence t-tests using standard normal critical values will be somewhat undersized. Our presumption is that the same will

apply to the max t-stat (adj.) procedure.

Rationalization of χ^2 (adj.) requires that the quantiles of the left as well as the right tails of the MSPE-adj. statistics are approximately those of a standard normal. Michael McCracken has kindly supplied unpublished tables of .01, .05 and .10 quantiles for the left tail of the distribution of the MSPE-adj. t-statistics. When combined with the comparable values for the right tail, which are available on Todd Clark's web page, we find that apart from a handful of cases,

(3.12) $0.02 <$ prob [square of t statistic (adj.) $> 1.96^2] < 0.10$,

$0.06 <$ prob [square of t statistic (adj.) $> 1.645^2] < 0.15$.

The handful of exceptions to the above inequalities would all be eliminated were we to slightly increase the 1.645^2 in the second line to 1.66^2. Hence, were we to apply our χ^2 (adj.) statistic to an example with $m = 1$ (which we have not done), we expect the size of tests computed using the standard critical values for a $\chi^2(1)$ to be roughly right.

Under any of the conditions described above, χ^2 (unadj.), the statistic defined in (3.9(f)), will not be correctly sized. This is because of the miscentering depicted in Figure 2.1.

4. SIMULATION OVERVIEW

The data generating processes used in our simulations are motivated by the use of disaggregate data to forecast an aggregate (Hubrich (2005), Hendry and Hubrich (2006, 2007)) in the literature. There is an aggregate y_t that is the sum of three disaggregate series,

(4.1) $y_t = y_{1t} + y_{2t} + y_{3t}$.

Model 0 is a univariate autoregression in the aggregate y_t; model 1 adds a lag of y_{1t} as a right hand side variable; model 2 adds a lag of both y_{1t} and y_{2t} as right hand side variables:

(4.2) $y_t = \text{const.} + \beta^*_{01} y_{t-1} + e_{0t} \equiv X_{0t}'\beta^*_0 + e_{0t}$,

$y_t = \text{const.} + \beta^*_{11} y_{t-1} + \beta^*_{12} y_{1t-1} + e_{1t} \equiv X_{1t}'\beta^*_1 + e_{1t}$,

$y_t = \text{const.} + \beta^*_{21} y_{t-1} + \beta^*_{22} y_{1t-1} + \beta^*_{23} y_{2t-1} + e_{2t} \equiv X_{2t}'\beta^*_2 + e_{2t}$.

We specify the data generating processes in terms of the three disaggregates y_{1t}, y_{2t} and y_{3t}. We assume that $(y_{1t}, y_{2t}, y_{3t})'$ follows a VAR of order 1 with 3×3 matrix of autoregressive parameters Φ, and zero mean i.i.d. normal disturbances $U_t \equiv (u_{1t}\, u_{2t}\, u_{3t})'$,

(4.3) $Y_t \equiv (y_{1t}, y_{2t}, y_{3t})' = \mu + \Phi Y_{t-1} + U_t$, $EU_tU_t' = I_3$.

Throughout, the mean vector μ was set to $(1, 1, 1)'$.

When examining size properties, we ensure that the three models in (4.2) have equal MSPE by specifying Φ to be diagonal with common parameter ϕ on the diagonal.[4] That is, each disaggregate follows a univariate AR(1) with common parameter ϕ:

(4.4) $y_{it} = 1 + \phi y_{it-1} + u_{it}$, $|\phi| < 1$, $i = 1,2,3$

$\Rightarrow y_t = 3 + \phi y_{t-1} + e_t$, $e_t = u_{1t} + u_{2t} + u_{3t}$, $Ee_t^2 = 3$.

As indicated in (4.4), it follows that y_t also follows an AR(1) with parameter ϕ. The baseline simulations set $\phi = 0.5$. We call this DGP 1A. This process is motivated by empirical applications involving aggregate inflation and its disaggregate components.

In (4.3), the aggregate will be Granger caused by one of the disaggregates once we depart from the specification (4.4). In simulations reported below, for evaluation of power, we set

(4.5) $$\Phi = \begin{pmatrix} 0.5 & -0.6 & 0 \\ -0.4 & 0.3 & 0 \\ 0 & 0 & 0.5 \end{pmatrix}.$$

In such a setting, the univariate process for the aggregate y_t is an ARMA(3,2). The eigenvalues of Φ are 0.5, -0.1 and 0.9. We call this DGP 1B. Two components that depend on each other might be commodities and services inflation, while there is a third component (such as food or energy inflation) that shows less interdependence with the other two components on average.

In all simulations reported here, and as indicated in (4.2), only one lag of any variable is included on the right hand side. Thus in simulations under the null, correct specification is assumed (i.e., no unnecessary lags are included), while in simulations under the alternative, an incorrect specification is used.

Our simulation proceeds as follows. For each of 1000 replications, and for values of R and P given below:

1. We draw a 3×1 vector Y_0 from a N(0, I_3) distribution, and generate 100 presample observations using $Y_t = \Phi Y_{t-1} + U_t$, $U_t \sim$ N(0, I_3). Call the 100'th presample observation Y_1.

2. For $t = 1, ..., R + P$, set $Y_t = \Phi Y_{t-1} + U_t$, $U_t \sim$ N(0, I_3).

3. Estimate the models in (4.2) by both rolling and recursive schemes, predicting one step ahead and computing one step ahead prediction errors.

4. Compute the following statistics.

a. Max t-stat (adj.). We flag rejections by using the values such as those given in Table 3.1, first rounding the sample correlation $\hat{\rho}$ to the nearest tenth.

b.,c. χ^2 (adj.), χ^2 (unadj.): rejections flagged using standard $\chi^2(2)$ critical values, i.e., 4.61 for 10% tests.

d. Reality check: The procedure described in White (2000) was followed, with the parameter that he calls q set to 0.5. The number of bootstrap repetitions in each simulation sample was 1000.

We also analyzed the maximum t-statistic computed from unadjusted MSPE differences. We do not report these results because they were very similar to the $m = 1$ simulation results reported in Clark and West (2006, 2007): this statistic was quite undersized under the null, and displayed very poor power

under the alternative.

We report results for 12 combinations of R (rolling regression size / size of smallest recursive sample) and P (number of predictions): $R = 40$, 100, 200 and 400 / $P = 40$, 100 and 200. This is intended to capture sample sizes that are representative for macro applications. Quarterly samples are represented by the smaller values of R and P; monthly samples by the larger values.

5. SIMULATION RESULTS

Table 5.1 has the results of the simulation, for nominal .10 tests. Results with .05 tests were similar. Consider size first, in panel A. Max t-stat (adj.) is modestly undersized. This is consistent with the modest undersizing predicted by the Clark and McCracken (2005) univariate asymptotics (see the discussion below equation (3.11)) and the simulation results in Clark and West (2007). For rolling samples, size ranges from 0.057 to 0.108; for recursive samples, the range is 0.062 to 0.087. The median of the 24 values presented in the table for max t-stat (adj.) is 0.081.

χ^2 using the adjusted difference in MSPEs (i.e, "χ^2 (adj.)"): this statistic is modestly oversized. The range for rolling and recursive is similar, from about 0.10 to about 0.14. The median of the 24 values for χ^2 (adj.) is 0.125.

χ^2 using the unadjusted difference in MSPEs (i.e, "χ^2 (unadj.)"): this statistic is grossly oversized for small R, with the extreme example being a size of 0.470 for $R = 40$ and $P = 200$, rolling scheme. For larger values of P and R, χ^2 (unadj.) is slightly oversized relative to χ^2 (adj.).[5] This is consistent with Figure 2.1A: for rolling samples, the mis-sizing is worse for bigger P, holding R fixed. The median of the 24 values of χ^2 (unadj.) is 0.147.

The reality check is grossly undersized. The median size is 0.014. The largest size (for $R = 400$, $P = 40$, rolling scheme) is 0.045. For $R = 40$, $P = 200$, rolling scheme, not a single one of the 1000 simulation samples rejected the null. The fact that the reality check works better for small values of P/R

is consistent with the technical conditions in White (2000), which include the requirement that $P/R \to 0$ at a certain rate. That the reality check is undersized is consistent with the simulations in Hansen (2005) and Clark and McCracken (2006).

Table 5.1B has results for power. This is raw, and *not* size adjusted power. As expected, max t-stat (adj.) has greater power than does χ^2 (adj.) Since max t-stat (adj.) is undersized and χ^2 (adj.) is oversized (see panel A), the discrepancy in power would be greater had we reported size-adjusted power. For smaller P or R, χ^2 (unadj.) and the reality check have considerably less power than do the other two statistics. For example, for $R = 40$, rolling estimation scheme, power for χ^2 (unadj.) and the reality check is in each case less than half that for max t-stat (adj.) and χ^2 (adj.). Poor power for the reality check was also found in Hansen (2005).

We conclude that from the perspective of size, max t-stat (adj.) and χ^2 (adj.) are about comparable, with one being slightly undersized, the other slightly oversized. From the point of view of power, max t-stat (adj.) is slightly preferable.

6. FORECASTING AGGREGATE U. S. INFLATION

In this section, we analyze empirically different methods of forecast accuracy evaluation for a small set of models that nest the benchmark, including the tests we proposed.

The series we forecast is aggregate inflation. In the first two applications, we investigate whether to include disaggregate inflation components in the aggregate model does improve over forecasting the aggregate only using past aggregate information. The third application also includes models with activity variables.

We focus on one-step ahead forecasts for US CPI inflation. The estimation period starts in 1960 and the forecast evaluation periods are the pre- and post-1984 periods (see e.g. Atkeson and Ohanian (2001), Stock and Watson (2007) and Hendry and Hubrich (2007) for recent contributions to

predictability of US inflation). The data are described in detail in Section 6.1.

We evaluate whether disaggregate information and/or macroeconomic variables do help predicting aggregate US inflation and whether disaggregates with different stochastic properties help to a different degree. Furthermore, we investigate whether there is a difference in the predictive content of disaggregates and macroeconomic variables for aggregate inflation in a low and a high inflation regime.

We compare forecast accuracy of the different models using rolling estimation samples based on the test procedures we propose and those previously suggested in the literature.[6] We relate the findings to our simulation results.

The remainder of the section is organized as follows: Section 6.1 describes the data, while Section 6.2 describes the forecast methods employed. That section also presents details on the transformations used for building the forecast models and for forecast evaluation. Finally, the results of the pseudo out-of-sample forecast experiment based on a rolling estimation sample are discussed.

6.1 Data

The data employed in this study include all items US consumer price index as well as its breakdown into four subcomponents: food (p^f), commodities less food and energy commodities (p^c), energy (p^e) and services less energy services prices (p^s). We employ monthly, seasonally adjusted CPI data CPI-U (source: the Bureau of Labor Statistics). In one of our applications we also employ two other macroeconomic variables that are available on a monthly frequency. We use industrial production to approximate output growth and unemployment as predictors.

We consider a sample period for inflation from 1960(1) to 2004(12), where earlier data from 1959(1) onwards are used for the transformation of the price level. As observed by other authors before, there has been a substantial change in the mean and the volatility of aggregate inflation (see also e.g. Stock and Watson (2007)) as well as in the disaggregates (see Hendry and Hubrich (2007)) between the two samples. Figure 6.1 depicts aggregate year-on-year US inflation for all items CPI and its four

subcomponents. It shows that also the disaggregate components exhibit a substantial change in mean and volatility. Aggregate as well as component inflation all exhibit high and volatile inflation until the beginning or mid 80s and lower, more stable inflation rates afterwards.

In Table 6.1 we show the substantial reduction in the mean for the disaggregate component inflation from the first to the second sample. The mean inflation rate has been reduced from 3.8-5.9% to 1.4-3.9%, while the standard deviation is reduced from between 2.9-8.2% to ranges of 1.0-8.3%. Thus, also the standard deviation has been reduced substantially, except for energy prices.

In section 6.2 we present results of an out-of-sample experiment for two different forecast evaluation periods: 1970(1) - 1983(12) and 1984(1) - 2004(12). The date 1984 for splitting the sample coincides with estimates of the beginning of the great moderation and is in line with what is chosen in Atkeson and Ohanian (2001) and Stock and Watson (2007) . We use the same split sample for comparability of our results to those studies in terms of aggregate inflation forecasts.

The out-of-sample forecast evaluation period includes therefore 14 and 21 years for forecast evaluation, respectively. Hendry and Hubrich (2007) have carried out simple ADF unit root tests for aggregate and disaggregate inflation, for different samples. The results are mixed. For the purpose of illustrating the application of our proposed test procedures, we present empirical results for the level of inflation.

6.2 Forecast methods and test results

The pseudo out-of-sample forecast results we present and discuss in the following are based on a rolling estimation window. Note that model selection and estimation is carried out for each rolling sample. The models selected are based on the AIC criterion due to the overall favorable forecast accuracy for US inflation (see Stock and Watson (2007) and Hendry and Hubrich (2007)). The forecast evaluation results presented are based on models formulated in first differences and forecast accuracy is evaluated based on year-on-year forecasts. Hendry and Hubrich (2007) find that formulating the model in terms of

month-on-month inflation improves forecast accuracy over formulating the model in year-on-year differences directly.

The one month ahead forecast is based on the following model:

$$(6.1) \qquad \pi^a_{t+1} = \text{const.} + \alpha_1 \pi^a_t + \sum_{i=1}^{n} \alpha_{i2} \pi^i_t + e_{t+1},$$

where aggregate inflation π^a_t as growth in prices $(P^a_t - P^a_{t-1})/P^a_{t-1}$ and π^i_t (also specified as growth of prices) are the *i* subcomponents of inflation (or other macroeconomic variables in the third application) included in the forecasting model.[7] The forecast evaluation is based on a transformation of the resulting forecasts to year-on-year inflation $(\hat{P}^a_{t+1} - P^a_{t+1-12})/P^a_{t+1-12}$. We estimate VARs, but since we only present one month ahead forecasts in the following, we have presented only the equation for the aggregate from the VAR in (6.1).

We present three empirical applications. In each empirical application we applied our proposed test procedures, i.e. the test based on the maximum of correlated normal random variables (max t-test adjusted) and the adjusted chi-squared statistic, using here and throughout a 10 percent level of significance. For the 3-model comparison we also present the estimated correlations between the squared forecast error differentials of the benchmark and two alternative models. We also present the respective critical value for the different tests in each of the applications. We compare the results for the pairwise model comparison based on the t-statistic, again adjusted in line with Clark and West (2006, 2007), with the other tests that compare all the models simultaneously. Additional test results displayed in the table include the unadjusted chi-squared statistic and White's reality check that we have also analyzed in the simulation study. We also present the absolute root mean squared prediction error (RMSPE) for the $AR_{(p)}$ model and the relative RMSPE for the alternative models. The lag length for the AR and VAR models in the current tables is chosen by AIC.

3-model comparison The first application closely resembles the simulation set-up and compares 3

different models: An $AR_{(p)}$ as a benchmark model for aggregate inflation is compared with two vector autoregressive models (VARs) with different disaggregate variables as predictors. One VAR includes services price inflation in addition to lagged aggregate inflation, $VAR^{a,s}_{(p)}$. The other VAR includes services as well as commodities price inflation in addition to lagged aggregate inflation, $VAR^{a,s,c}_{(p)}$, i.e. in this model we include the two disaggregate components that constitute the so-called 'core' inflation rate. In this case the alternative models are nested within each other as in the simulation experiments.

The upper panel of Table 6.2 presents the results for this application on forecasting US aggregate inflation for the sample period of high and volatile inflation 1970(1) to 1983(12). We find that in the pairwise model comparison equal predictive ability is rejected for both comparisons, i.e. the comparison of the benchmark AR model versus the smaller model, $VAR^{a,s}_{(p)}$, as well as for the comparison with the larger model, $VAR^{a,s,c}_{(p)}$. If we compare both alternative models to the benchmark simultaneously using the higher critical value for the maximum t-statistic (see Table 3.1 for the critical values for the maximum of two correlated standard normal variables), we still find a rejection of equal predictive ability. Note that the correlation between the two forecast error differentials of the alternative models and the benchmark is relatively small. Therefore, the critical value is clearly higher than the one for the adjusted t-statistic. The unadjusted chi-squared test and the reality check do not reject. That can be explained by the substantially lower power that we found for these tests in our simulations.

The lower panel in Table 6.2 exhibits the results for the sample period 1984(1)-2004(12). For this period we do not find a rejection of equal predictive ability for any of the tests (except for one), no matter whether we consider a pairwise or a 3-model comparison. The only rejection we get is for the unadjusted chi-squared test, a test that we find to be oversized in our simulations, also for this kind of sample size. Therefore, our results are overall in line with previous findings that during the recent period of low and relatively stable inflation it is difficult for any model to outperform simple benchmark models as the random walk or the autoregressive model due to a lack of variability in aggregate inflation and a

lack of predictive content of most explanatory variables.

5-model comparison: Disaggregate predictors The second and third applications present a five model comparison. In the second application, we compare the $AR_{(p)}$ benchmark model against four different VAR models, where each of the alternative models includes a different disaggregate predictor in addition to lagged aggregate inflation: $VAR^{a,f}_{(p)}$, $VAR^{a,e}_{(p)}$, $VAR^{a,c}_{(p)}$, and $VAR^{a,s}_{(p)}$. The different models therefore include disaggregate regressors with very different properties. Energy and food inflation are much more volatile and difficult to forecast than commodities and services inflation (see Figure 6.1 and Table 6.1). The four alternative models in this second example are not nested within one another, only the benchmark model is nested in both alternative models.

The results of the second empirical application are presented in Table 6.3. As mentioned, in this example the four alternative models include disaggregate inflation rates as predictors that have quite different properties. When we carry out a pairwise model forecast evaluation using the adjusted t-statistic, we find a rejection of equal forecast accuracy of the benchmark AR model and two alternative models with commodities and services inflation for the high inflation period. If we compare the five models simultaneously using the appropriate simulated critical value for correlated normals, we still find that the null of equal forecast accuracy is rejected. The χ^2 (adj.) statistic does not reject but the statistic is close to the critical value. The χ^2 (unadj.) statistic and the reality check do not reject, perhaps because of low power.

Notably, for the low and stable inflation period - where it is usually difficult to improve over a simple AR model - the tests for the pairwise model comparisons presented in the lower panel of Table 6.3 indicate predictive content of food inflation for aggregate inflation, but no predictive content of the other disaggregate components. However, the result for the maximum t-test based on the higher critical value simulated for this test statistic based on the maximum of correlated normals does not reject equal forecast accuracy of all models. Here we get a different test result once we take into account that we are

comparing 5 models simultaneously. As in the 3 model comparison, only the χ^2 (unadj.) rejects.

5-model comparison: Disaggregate and other macroeconomic predictors In the third empirical application, we consider two models with disaggregate predictors, i.e. with services and commodity inflation, and two models that include other macroeconomic variables and compare those models to the benchmark. One out of those four models is a Phillips curve type model including the change in unemployment as a predictor (in this context the change in unemployment provided lower RMSPE than the level of unemployment). The other is a model with output growth capturing economic activity (see Orphanides and van Norden (2005), who suggest that using output growth instead of an output gap measure might be useful for forecasting in real time).

In this empirical application, reported in Table 6.4, all pairwise model comparisons reject equal forecast accuracy for the sample 1970-1983. Furthermore, also our proposed test procedures for the 5-model comparison both reject. Only the unadjusted chi-squared statistic and the reality check do not reject, which is likely due to the very low power of those procedures. Overall we conclude that for this sample period at least one of the alternative models, in particular the one with the largest test statistic, i.e. including unemployment changes, has higher forecast accuracy than the benchmark.

For the sample period 1984-2004 we do find predictive content of unemployment changes for aggregate inflation from the pairwise model comparison, but not for the 5-model comparison. From all tests applied for the 5-model comparison, again only the unadjusted chi-squared test rejects. We have greater confidence in the results of the other tests on the basis of our simulation results. Therefore, we conclude that equal forecast accuracy of those five models is not rejected, indicating–in line with previous literature–no predictive content of those disaggregates and macroeconomic variables employed here over the information contained in lags of aggregate inflation. This is due to a lack of variability of aggregate inflation to be explained and a lack of predictive content of most explanatory variables.

To conclude, these applications demonstrate that one might draw wrong conclusions on the basis

of pairwise model forecast evaluation tests. This is particularly the case if the correlations between the forecast error differentials vis-à-vis the benchmark are quite low and the critical value of the maximum t-test is therefore high. Also, this might occur in times of low inflation where the differences in terms of forecast accuracy of alternative models in comparison to the benchmark are rather small.

FOOTNOTES

1. The assumption of constant second moments (i.e., the fact that σ_i^2 is not subscripted by *t*) is for expositional convenience. We can accommodate moment drift at the expense of complications in notation.

2. To prevent confusion, we note that we reference the technical conditions, and not the procedures in Giacomini and White (2006). The procedure they propose is essentially what we call χ^2 using the *unadjusted* MSPE differences–and that is not a procedure that we endorse in the present context.

3. Suppose the null model relies on estimated regression parameters to predict. Then under the conditions of this paragraph, and for either single or multistep predictions, the vector of adjusted MSPE differences will be asymptotically normal, but possibly not centered at zero. In this case we expect some missizing; see the discussion around equation (4.4) in Clark and West (2007).

4. Hendry and Hubrich (207) show that in this case of equal eigenvalues, slope misspecification is minimized and estimation uncertainty differences will dominate forecast accuracy comparisons.

5. In a first round of simulations, we tried computing a robust HAC covariance matrix, as recommended by Giacomini and White (2006). The behavior of χ^2 (unadj.) was similar to what is reported in the table.

6. Results with recursive samples are similar, and are omitted to save space.

7. To prevent confusion, we note that π_t^a places the role of the variable called y_t in previous sections.

REFERENCES

Ashley. R., Granger, Clive .W.J. and Richard Schmalensee, 1980, "Advertising and Aggregate Consumption: An Analysis of Causality," *Econometrica* 48, 1149-1168.

Atkeson, Andrew and Ohanian, Lee E. (2001), "Are Phillips Curves Useful for Forecasting Inflation?," Federal Reserve Bank of Minneapolis Quarterly Review 25(1): 2–11.

Billmeier, Andreas, 2004, "Ghostbusting: Which Output Gap Measure Really Works?", IMF Working paper WP/04/146.

Cain, Michael, 1994, "The Moment Generating Function of the Minimum of Bivariate Normal Random Variables," *The American Statistician*, 1994, 48:2, 124-125.

Clark, Todd E. and Michael W. McCracken, 2001, "Tests of Equal Forecast Accuracy and Encompassing for Nested Models," *Journal of Econometrics* 105, 85-110.

Clark, Todd E. and Michael W. McCracken, 2005, "Evaluating Direct Multistep Forecasts," *Econometric Reviews*, 369-404.

Clark, Todd E. and Michael W. McCracken, 2006, "Reality Checks and Nested Forecast Model Comparisons," manuscript, Board of Governors of the Federal Reserve.

Clark, Todd E. and Kenneth D. West, 2006, "Using Out-of-Sample Mean Squared Prediction Errors to Test the Martingale Difference Hypothesis," *Journal of Econometrics* 135 (1-2) (2006), 155-186.

Clark, Todd E. and Kenneth D. West, 2007, "Approximately Normal Tests for Equal Predictive Accuracy in Nested Models," *Journal of Econometrics*, 138(1) (2007), 291-311.

D'Agostino, Antonello, Domenico Giannone and Paolo Surico, 2006, "(Un)Predictability and Macroeconomic Stability," European Central Bank Working Paper No. 605.

Diebold, Francis X. and Robert S. Mariano, 1995, "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics* 13, 253-263.

Giacomini, Rafaella and Halbert White, 2006, "Tests of Conditional Predictive Ability," *Econometrica* 74, 1545-78.

Granger, C.W.J and Paul Newbold, 1977, *Forecasting Economic Time Series*, New York: Academic Press.

Gupta, Shanti S., Klaus Nagel and S. Panchapakesan, 1973, "On the Order Statistics from Equally Correlated Normal Random Variables," *Biometrika* 60, 403-413.

Hansen, Peter Reinhard, 2005, "A Test for Superior Predictive Ability," *Journal of Business and Economic Statistics*, 23, 365-380.

Hendry, David F. and Kirstin Hubrich, 2006, "Forecasting Aggregates by Disaggregates," European Central Bank Working Paper 589.

Hendry, David F. and Kirstin Hubrich, 2007, "Combining Disaggregate Forecasts or Combining Disaggregate Information to Forecast an Aggregate," manuscript, European Central Bank.

Hong, Yongmiao and T. H. Lee, 2003, "Inference on Predictability of Foreign Exchange Rates via Generalized Spectrum and Nonlinear Time Series Models". *Review of Economics and Statistics* 85, 1048–1062.

Hubrich, Kirstin, 2005, "Forecasting Euro Area Inflation: Does Aggregating Forecasts by HICP Component Improve Forecast Accuracy?," *International Journal of Forecasting* 21, 119-36.

Ker, Alan, 2001, "On the Maximum of Bivariate Normal Variables", *Extremes*, 4:2, 185-186.

Lütkepohl, Helmet, 1984, "Forecasting contemporaneously aggregated vector ARMA processes," *Journal of Business & Economic Statistics* 2(3), 201–214.

Lütkepohl, Helmet, 1987, *Forecasting Aggregated Vector ARMA Processes*, Springer-Verlag.

Orphanides, Athanasios and Simonvan Norden, 2005), "The Reliability of Inflation Forecast Based on Output Gap Estimates in Real Time," *Journal of Money, Credit, and Banking* 37, 583-600.

Rapach, David E. and Mark E. Wohar, 2006, "In-Sample vs. Out-of-Sample Tests of Stock Return Predictability in the Context of Data Mining," *Journal of Empirical Finance*, 13(2), 231-247.

Sarno, Lucio, Daniel L. Thornton and Giorgio Valente, 2005, "Federal Funds Rate Prediction," *Journal of Money, Credit and Banking* 37, 449-472.

Stock, James H. and Mark W. Watson, 2007, "Why has U.S. inflation become harder to forecast?," *Journal of Money, Credit and Banking*.

West, Kenneth D., 1996, "Asymptotic Inference About Predictive Ability," *Econometrica* 64 , 1067-1084.

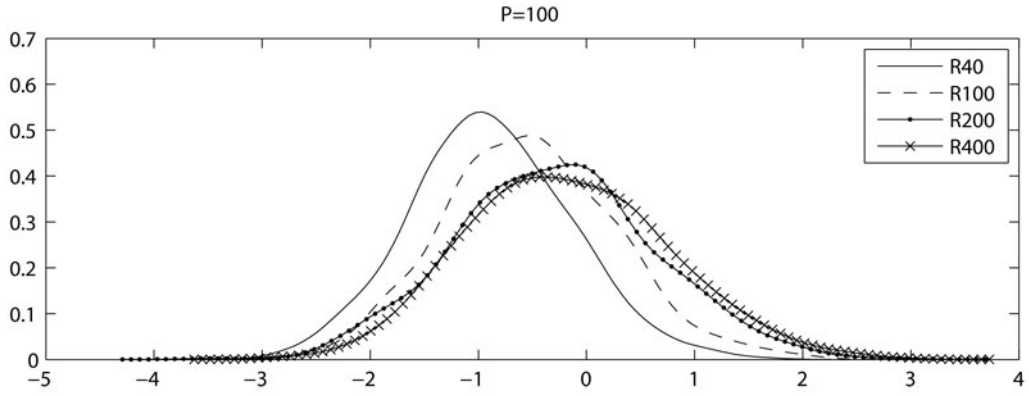
West, Kenneth D. and Dongchul Cho, 1995, "The Predictive Ability of Several Models of Exchange Rate Volatility," *Journal of Econometrics* 69, 367-391.

West, Kenneth D., Hali J. Edison and Dongchul Cho, 1993, "A Utility Based Comparison of Some Models of Exchange Rate Volatility," *Journal of International Economics* 35, 23-46.

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

Figure 2.1
Density of MSPE Differences Under the Null, DGP 1A

A. *P*=100, *R* varying



B. *R*=100, *P* varying

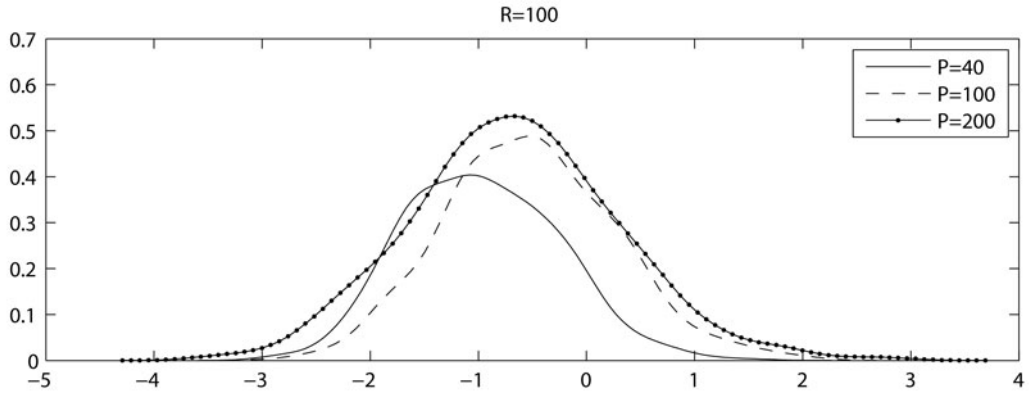


Table 3.1

Critical Values for the Maximum of Two Correlated Standard Normals

	ρ										
	1	0.8	0.6	0.4	0.2	0	-0.2	-0.4	-0.6	-0.8	-1
size=5%	1.645	1.846	1.900	1.929	1.946	1.955	1.959	1.960	1.960	1.960	1.960
size=10%	1.282	1.493	1.556	1.594	1.617	1.632	1.640	1.644	1.645	1.645	1.645

Notes:

1. Let z_1 and z_2 be standard normal variables, with correlation ρ. The table presents the .95 and .90 quantiles for the random variable: $z \equiv \max(z_1, z_2)$.

Table 5.1
Empirical Size and Power, 1-Step Ahead Forecasts, Nominal Size = 10%

A. Size: DGP 1A

		1. Rolling				2. Recursive			
P		*R*=40	*R*=100	*R*=200	*R*=400	*R*=40	*R*=100	*R*=200	*R*=400
40	Max t-stat (adj.)	0.082	0.067	0.091	0.087	0.083	0.074	0.084	0.087
	χ^2 (adj.)	0.125	0.144	0.138	0.117	0.125	0.139	0.141	0.130
	χ^2 (unadj.)	0.172	0.139	0.144	0.124	0.153	0.137	0.146	0.131
	Reality check	0.007	0.022	0.044	0.043	0.011	0.033	0.036	0.045
100	Max t-stat (adj.)	0.086	0.065	0.090	0.081	0.069	0.076	0.084	0.077
	χ^2 (adj.)	0.107	0.097	0.126	0.128	0.108	0.126	0.141	0.133
	χ^2 (unadj.)	0.248	0.147	0.139	0.130	0.167	0.162	0.156	0.134
	Reality check	0.001	0.005	0.022	0.032	0.001	0.015	0.023	0.038
200	Max t-stat (adj.)	0.108	0.071	0.057	0.058	0.083	0.063	0.062	0.064
	χ^2 (adj.)	0.120	0.114	0.106	0.099	0.121	0.116	0.136	0.118
	χ^2 (unadj.)	0.470	0.223	0.142	0.120	0.195	0.178	0.160	0.141
	Reality check	0.000	0.002	0.007	0.014	0.001	0.009	0.014	0.020

B. Power: DGP 1B

		1. Rolling				2. Recursive			
P		*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	0.722	0.777	0.812	0.830
	χ^2 (adj.)	0.584	0.651	0.703	0.708	0.627	0.658	0.701	0.711
	χ^2 (unadj.)	0.177	0.252	0.301	0.298	0.222	0.271	0.305	0.297
	Reality check	0.230	0.408	0.478	0.522	0.331	0.424	0.483	0.520
100	Max t-stat (adj.)	0.885	0.983	0.987	0.991	0.959	0.989	0.990	0.992
	χ^2 (adj.)	0.851	0.954	0.966	0.971	0.933	0.971	0.967	0.973
	χ^2 (unadj.)	0.268	0.430	0.519	0.564	0.415	0.498	0.540	0.572
	Reality check	0.314	0.658	0.753	0.766	0.615	0.714	0.768	0.775
200	Max t-stat (adj.)	0.989	0.997	0.999	1.000	1.000	1.000	1.000	1.000
	χ^2 (adj.)	0.986	0.998	0.997	1.000	0.999	0.999	0.998	1.000
	χ^2 (unadj.)	0.465	0.743	0.790	0.814	0.736	0.816	0.809	0.819
	Reality check	0.483	0.900	0.933	0.944	0.900	0.938	0.938	0.942

See notes on next page.

Notes to Table 5.1:

1. The mean squared prediction error (MSPE) from a first order univariate autoregression is compared to MSPEs from two other models, each of which add lags of other variables. The exact form and parameters for DGPs 1A and 1B are described in section 3 of the paper. In each simulation, and for each DGP, one step ahead forecasts of y_{t+1} are formed from each of the three models, using least squares regressions.

2. The number of simulations is 1000. R is the size of the rolling regression sample (panel A1 and B1), or the smallest regression sample (panel A2 and B2). P is the number of out-of-sample predictions.

3. The qualifier "(adj.)" means that the statistic is computed using MSPE differences adjusted as recommended in Clark and West (2007) and defined in equation 3.8(h); "(unadj.)" means that the usual equation 3.9(b) MSPE difference is used.

4. "Max t-stat" is the larger of the two Clark and West (2007) MSPE-adjusted t-statistics, and is defined in equation 3.5. The table reports the fraction of simulations in which each test statistic was greater than the critical value obtained by (a)rounding the sample correlation between the two MSPE-adjusted t-statistics to the nearest 0.1, and (b)using critical values obtained from numerically integrating the density given in (3.6). The χ^2 statistics are computed in standard fashion from the 2×1 vector of differences in MSPEs or adjusted difference in MSPEs, see (3.9(f)) and (3.8). For the reality check, White's (2000) bootstrap procedure was used, with 1000 bootstrap repetitions per simulation sample.

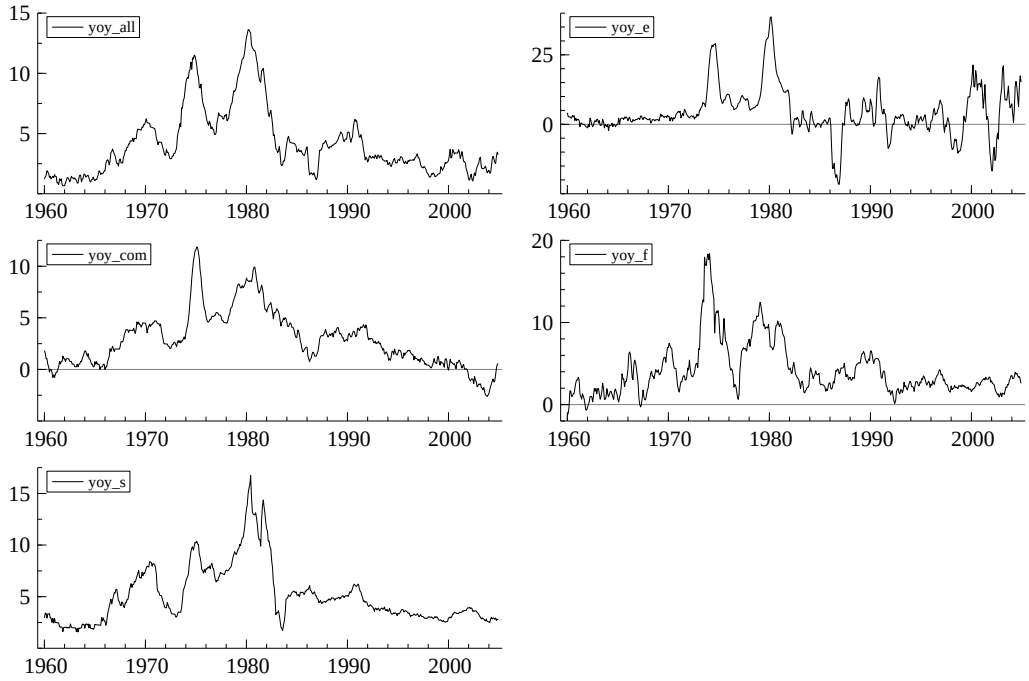


Figure 6.1: Year-on-year US CPI inflation rate, aggregate and subcomponents

Table 6.1: US, Descriptive Statistics, year-on-year CPI Inflation

1960-1983	all items	energy	commodities	food	services
Mean	4.86	5.91	3.80	4.75	5.81
Std Deviation	3.41	8.17	2.89	4.11	3.40
1984-2004	all items	energy	commodities	food	services
Mean	2.99	2.28	1.43	2.93	3.91
Std Deviation	1.06	8.26	1.65	1.26	0.99

Table 6.2: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 $VAR^{a,s}_{(AIC)}$	0.986	2.312^*				
vs $VAR^{a,s,c}_{(AIC)}$	1.035	1.554^*				
vs 2 models			2.312^*	6.311^*	2.963	0.032
est. correlation			0.261			
critical value		1.282	1.613	4.61	4.61	0.058
	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 $VAR^{a,s}_{(AIC)}$	1.027	-0.463				
vs $VAR^{a,s,c}_{(AIC)}$	1.055	0.923				
vs 2 models			0.923	1.331	6.576^*	-0.030
est. correlation			0.225			
critical value		1.282	1.617	4.61	4.61	0.019

Note: Forecast evaluation for 1 month ahead forecasts; actual RMSPE (non annualised) for $AR_{(AIC)}$ benchmark model in percentage points, for other models RMSPE relative to AR (RMSPE (altern)/RMSPE (bench)); rolling estimation window; rolling estimation samples 1960(1) to 1970(1),...,1983(12) (i.e. R=120 and P=168) and 1960(1) to 1984(1),...,2004(12), (i.e. R=288 and P=252); maximum number of lags: $p = 13$; Subscripts indicate model selection procedure, AIC: Akaike criterion, superscripts indicate model, $VAR^{a,c}$: VAR with lags of aggregate and commodities inflation, $VAR^{a,s}$: VAR with aggregate and services inflation; model specification in terms of month-on-month inflation; forecast evaluation for year-on-year inflation; estimated correlation between $f_i = \hat{e}_0 - \hat{e}_i$, for comparing model i=1,2 to the benchmark;critical value of respective test statistic;* indicates significance on a 10% nominal significance level

Table 6.3: 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
	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

Note: Forecast evaluation for 1 month ahead forecasts; actual RMSPE (non annualised) for $AR_{(AIC)}$ benchmark model in percentage points, for other models RMSPE relative to AR (RMSPE (altern)/RMSPE (bench)); rolling estimation window; rolling estimation samples 1960(1) to 1970(1),...,1983(12) (i.e. R=120 and P=168) and 1960(1) to 1984(1),...,2004(12), (i.e. R=288 and P=252); maximum number of lags: $p = 13$; Subscripts indicate model selection procedure, AIC: Akaike criterion, superscripts indicate model, $VAR^{a,f}$: VAR with lags of aggregate and food inflation, $VAR^{a,e}$: VAR with lags of aggregate and energy inflation, $VAR^{a,c}$: VAR with lags of aggregate and commodities inflation, $VAR^{a,s}$: VAR with aggregate and services inflation; model specification in terms of month-on-month inflation; forecast evaluation for year-on-year inflation; estimated correlation between $f_i = \hat{e}_0 - \hat{e}_i$, for comparing model i=1,...,4 to the benchmark;critical value of respective test statistic (simulated for max t-stat adj.);* indicates significance on a 10% nominal significance level

Table 6.4: 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
	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

Note: Forecast evaluation for 1 month ahead forecasts; actual RMSPE (non annualised) for $AR_{(AIC)}$ benchmark model in percentage points, for other models RMSPE relative to AR (RMSPE (altern)/RMSPE (bench)); rolling estimation window; rolling estimation samples 1960(1) to 1970(1),...,1983(12) (i.e. R=120 and P=168) and 1960(1) to 1984(1),...,2004(12), (i.e. R=288 and P=252); maximum number of lags: $p = 13$; Subscripts indicate model selection procedure, AIC: Akaike criterion, superscripts indicate model, $VAR^{a,y}$: VAR with lags of aggregate inflation and output growth, $VAR^{a,u}$: VAR with lags of aggregate inflation and change in unemployment, $VAR^{a,c}$: VAR with lags of aggregate and commodities inflation, $VAR^{a,s}$: VAR with aggregate and services inflation; model specification in terms of month-on-month inflation; forecast evaluation for year-on-year inflation; estimated correlation between $f_i = \hat{e}_0 - \hat{e}_i$, for comparing model i=1,...,4 to the benchmark;critical value of respective test statistic (simulated for max t-stat adj.);* indicates significance on a 10% nominal significance level