Comparing predictive ability in presence of instability over a very short time

Fabrizio Iacone[a,b] Luca Rossini[a,c] Andrea Viselli[a,d]

[a]University of Milan, Italy [b]University of York, United Kingdom
[c]Fondazione Eni Enrico Mattei [d]University of Pavia, Italy

European Association of Young Economists (EAYE)
2024 Annual Meeting

Paris, May 24, 2024

Motivation

- The Covid-19 pandemic caused a large, unexpected macroeconomic shock, posing a great challenge both for forecasting and evaluation (Foroni et al. 2022);
- It spans a **very short period of time** and is a moment of **extreme instability**;
- Several procedures have been proposed to address, or improve, forecasting and nowcasting in this extreme scenario (Schorfheide and Song, 2021; Lenza and Primiceri, 2022)
- In comparison, **the issue of forecast evaluation has received less attention**. As we move forward, it becomes the most relevant challenge.

$\longrightarrow$ **How should we treat the performance of forecasting/nowcasting models during the Covid-19 period?**

Our contributions

- We show that tests like the Diebold and Mariano (1995, DM) for equal forecasting ability or the Giacomini and Rossi (2010) Fluctuation test (Fl) have **no power**:

    1. when differences in predictive ability only span a very short period;
    2. even when one forecast is notably superior over the whole evaluation period if the shock is particularly large;

- These situations may be detected using **non-parametric diagnostics for local breaks or extreme values**, such as the Andrews' (2003) test and MAX procedure of Harvey et al. (2021);

- We illustrate these results in a **Monte Carlo** exercise, and a **nowcasting exercise** using the U.S. Survey of Professional Forecasters (SPF).

Empirical application

- We consider the current quarter median forecast (i.e. the nowcast) of the U.S. nominal GDP growth from the SPF over the period $Q1 : 2000 - Q3 : 2020$;
- We compare it against a naïve benchmark, namely zero nominal GDP growth, which corresponds to nowcasting GDP with the last available observation;
- For comparison, we consider the two-sided DM and Fl test with Bartlett's estimate of the long-run variance. The 5% critical values are 2.032 for $|DM|$ (fixed smoothing, Coroneo and Iacone, 2020) and 3.012 for $|Fl|$.

Nowcast evaluation $Q1 : 2000 - Q3 : 2020$ (1)

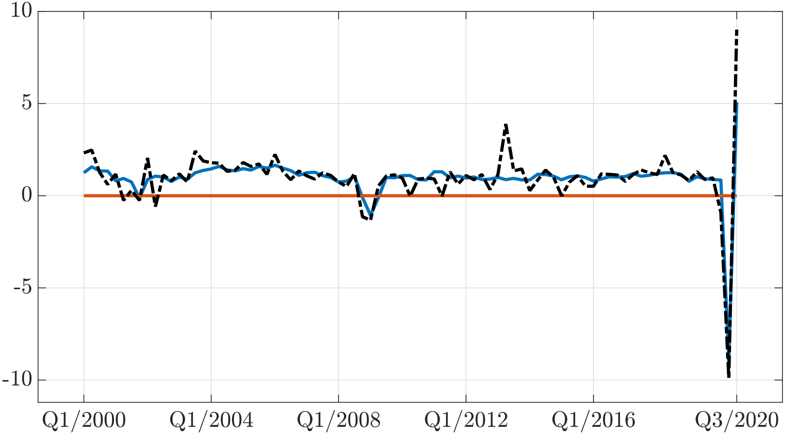

Figure 1: U.S. nominal GDP growth (black dotted line), GDP nowcast from the SPF (blue line), and naïve nowcast (red line) over the period 2000:Q1 to 2020:Q3.

Nowcast evaluation $Q1:2000 - Q3:2020$ (2)

Table 1: Nowcast evaluation of the SPF and the naïve benchmark using the RMSE, DM, and Fluctuation test over the periods 2000:Q1 – 2019:Q4 and 2000:Q1 – 2020:Q3.

Period	RMSE_{SPF}	$\text{RMSE}_{\text{Naive}}$	Ratio	DM	Fl_l	Fl_u
Q1:2000 - Q4:2019	0.37	1.66	0.22	-7.27	-5.83	-2.04
Q1:2000 - Q3:2020	0.59	3.77	0.16	-1.92	-2.49	-0.21

Note: RMSE_{SPF} and $\text{RMSE}_{\text{Naïve}}$ are the average RMSE for the SPF and naïve benchmark, respectively. Ratio refers to the ratio $\text{RMSE}_{\text{SPF}}/\text{RMSE}_{\text{Naïve}}$. DM, Fl_l, and Fl_u are the DM, lower, and upper Fl test statistics. The 5% critical values for two-sided tests are 2.032 (fixed smoothing) and 3.012, respectively.

Asymptotic results for the DM / Fl tests (1)

Given regularity conditions as in Giacomini and White (2006), let us denote with d_s, for $s = 1, ..., T$, the loss differential and assume that

$$d_s = \delta_1 T^{-1/2} + \delta_2 T^a \mathbb{I}_s(\tau) + u_s,$$

where:

- u_s is a zero-mean process and $\mathbb{I}_s(\tau)$ is an indicator function, taking value 1 if $s = \lfloor \tau T \rfloor$ and 0 otherwise;
- the factor $\delta_2 T^a$ characterises the dimension of the change in the prediction differential concerning the sample size.

The long-run variance is estimated using the Bartlett kernel,

$$\hat{\sigma}_T^2 = c_0 + 2 \sum_{l=1}^{M} \frac{M-l}{M} c_l,$$

where c_l is the $l-$th sample covariance of d_s and M is a user-chosen bandwidth. The DM statistic is

$$t_{DM} = \sqrt{T} \frac{\bar{d}}{\hat{\sigma}_T}.$$

Asymptotic results for the DM / Fl tests (2)

Theorem 1: *Under the Giacomini and White (2006) regularity conditions,*

(i) if $a < 1/2$, then $t_{DM} \rightarrow_d Z + \frac{\delta_1}{\sigma}$;

(ii) if $a > 1/2$, then $|t_{DM}| \rightarrow_p 1$.

Discussion:

- if the change in the prediction differential is not very large, then the shock does not affect the DM test which ultimately may have non-trivial power;
- if the change in prediction differential is large, then the DM test has no power and cannot detect even differentials that would be otherwise significant;
- qualitatively similar results hold for the Fl test, as it uses a fraction of the sample size and the same estimate for the long-run variance.

Non-parametric detection of local instabilities (1)

Andrews (2003) proposes an end-of-sample instability test:

- used to detect changes in the prediction errors' distribution at the end of the sample, even for a very small number of post-change observations;
- because the number of post-change observations is small (in fact, as small as one), the F statistic does not converge to a χ_1^2;
- assuming the location of the shock is **known** in advance, the asymptotic distribution of the test statistic is estimated nonparametrically.
- we compare the test statistic $S = \tilde{u}^2_{\lfloor \tau T \rfloor}$ to the residuals $\{\hat{u}_s^2 : s = 1, \ldots, \lfloor \tau T \rfloor - 1, \lfloor \tau T \rfloor + 1, \ldots, T\}$ in the restricted model.
- for instabilities affecting k observations, Andrews (2003) proposes a statistic that accounts for residual autocorrelation, using the restricted residuals;
- however, it may be inconsistent in presence of local instabilities, so we consider a statistic that only use the unrestricted residuals.

Non-parametric detection of local instabilities (2)

Harvey et al. (2021) propose a the MAX monitoring procedure, where:

- the sample is split into a training and monitoring period, and the number of observations in each set determines the false positive rate (FPR) of the procedure;
- no instability is assumed to occur during the training period, but it may take place during the monitoring period;
- we compare the maximum residual in the training period with the one in the monitoring period, and detect an instability if

$$\max_{s=1,\dots,T^*} u_s^2 < \max_{s=T^*+1,\dots,E} u_s^2.$$

Non-parametric detection of the Covid-19 shock

Table 2: Andrews (2003) S test evaluated for three different values of Σ and MAX procedure over the period 2020:Q1 to 2020:Q3 (3 observations).

$S(I)$	$q_{S(I)}$	$S(\widetilde{\Sigma})$	$q_{S(\widetilde{\Sigma})}$	$S(\widehat{\Sigma})$	$q_{S(\widehat{\Sigma})}$	MAX	q_{MAX}
7576	10.9	0.21	1.92	3060	3.6	96.84^2	6.03^2

Note: $S(I)$, $S(\widetilde{\Sigma})$, and $S(\widehat{\Sigma})$ denote the S test statistics when the identity matrix, restricted residuals, and unrestricted residuals are used as weighting matrix, respectively, while $q_{S(I)}$, $q_{S(\widetilde{\Sigma})}$, and $q_{S(\widetilde{\Sigma})}$ denote the respective critical values. The theoretical size is 5%. MAX denotes the maximum procedure over the period 2020:Q1 to 2020:Q3 and q_{MAX} its MAX over the period 2000:Q1 to 2019:Q4. The false positive rate of the procedure is 3.6%.

Monte Carlo: DGP

We consider the data-generating process

$$
\begin{aligned}
y_t &= \beta x_t + \eta_t \\
x_t &= \rho_x x_{t-1} + \xi_t, \quad \xi_t \sim NID(o, \sigma_x^2), \quad |\rho_x| < 1, \\
\eta_t &= \rho_\eta \eta_{t-1} + \epsilon_t, \quad \epsilon_t \sim NID(o, \sigma_\eta^2), \quad |\rho_\eta| < 1,
\end{aligned}
$$

where we do not observe x_t, but

$$
\begin{aligned}
x_t^{(1)*} &= x_t + v_{1,t}, \quad v_t^{(1)} \sim NID(0, \sigma_1^2), \\
x_t^{(2)*} &= x_t + v_{2,t}, \quad v_t^{(2)} \sim NID(0, \sigma_2^2),
\end{aligned}
$$

and the forecasts are $\hat{y}_t^{(1)} = \hat{\beta}_t^{(1)} x_{1,t}^*$ and $\hat{y}_t^{(2)} = \hat{\beta}_t^{(2)} x_{2,t}^*$.

In our baseline experiment, we set:

$$
\beta = 1, \quad \rho_x = 0.75, \quad \sigma_x^2 = 1, \quad \rho_\eta^2 = 0.5, \quad \sigma_\eta^2 = 0.1, \quad \sigma_1^2 = 0.1, \quad \sigma_2^2 = 0.1\delta_s,
$$

where $\delta_s = 1$ yields equal unconditional predictive ability.

Monte Carlo: results

Table 3: Global and local equal predictive ability tests for different sizes and power. We report in the columns the DM, fluctuation, and S test and the MAX procedure using $T = 80$ and $R = 20$.

Size/Power	δ	DM	Fl	S	MAX
$\delta_s = \delta$ for all s	1	0.053	0.047	0.046	0.051
$\delta_s = \delta$ for all s	0.1	0.919	0.654	0.051	0.054
	2	0.925	0.658	0.049	0.051
	4	1.000	0.986	0.048	0.056
$\delta_s = \delta$ for $s > T - 20$	0.1	0.091	0.081	0.029	0.029
	2	0.160	0.249	0.106	0.150
	4	0.547	0.884	0.150	0.150
$\delta_s = \delta$ for $s = T$	0.1	0.053	0.048	0.024	0.044
	2	0.052	0.044	0.167	0.114
	4	0.047	0.034	0.426	0.335
	8	0.034	0.021	0.672	0.609

Note: the table exhibits the empirical size and power of the equal predictive ability tests. The theoretical size is set at 5% for all the tests.

Concluding remarks

- The DM and Fluctuation tests were not designed to capture very short-lived instabilities, and most importantly their power vanishes when the magnitude of the shock is very large;
- We consider two diagnostics (the S and MAX procedures), that are suitable in the presence of a very short instability, in a Monte Carlo experiment and in an empirical exercise for nowcasting the U.S. GDP growth using the SPF;
- We find strong evidence in favor of our claim both in the Monte Carlo and in the empirical application, due to the presence of the Covid-19 shock.

Main takeaway:

$\longrightarrow$ **the forecaster should not pool the sample, but exclude the short periods of high local instability from the evaluation exercise**.

Thank you very much for your attention

Scan the QR code to access the working paper.

References

Andrews, Donald W.K. *End-of-sample instability test*. Econometrica 71.6 (2003), pp. 1661–1694.

Coroneo, Laura, and Fabrizio Iacone. *Comparing predictive accuracy in small samples using fixed-smoothing asymptotics*. Journal of Applied Econometrics 35.4 (2020): 391-409.

Diebold, Francis X. and Mariano, Robert S. *Comparing Predictive Accuracy*. Journal of Business & Economic Statistics 20.1 (1995), pp. 134–144.

Foroni, Claudia, Marcellino, Massimiliano, and Stevanovic, Dalibor. *Forecasting the Covid-19 recession and recovery: Lessons from the financial crisis*. International Journal of Forecasting 38.2 (2022), pp. 596-612.

Giacomini, Raffaella, and Rossi, Barbara. *Forecast comparisons in unstable environments*. Journal of Applied Econometrics 25.4 (2010), pp. 595-620.

Giacomini, Raffaella, and White, Halbert. *Tests of conditional predictive ability*. Econometrica 74.6 (2006): 1545-1578.

Lenza, Michele, and Primiceri, Giorgio E. *How to estimate a vector autoregression after March 2020*. Journal of Applied Econometrics 37.4 (2022): 688-699.

Harvey, David I., et al. *Real-time detection of regimes of predictability in the US equity premium*. Journal of Applied Econometrics 36.1 (2021), pp. 45-70.

Rossi, Barbara. *Forecasting in the presence of instabilities: How we know whether models predict well and how to improve them*. Journal of Economic Literature 59.4 (2021), pp. 1135-1190.

Schorfheide, Frank, and Song, Dongho. *Real-time forecasting with a (standard) mixed-frequency VAR during a pandemic*. National Bureau of Economic Research (2021): No. w29535.

Giacomini and White (2006)

Denote the variable of interest (i.e. GDP growth) with y_t and a predictor with x_t. The observed vector is denoted by $w_t \equiv (y_t, x_t)'$.

Denote the two h-step-ahead forecasts obtained using two alternative methods with

$$\hat{y}_t^{(i)}(\hat{\delta}_{t-h,R_i}^{(i)}) = f^{(i)}(w_{t-h}, \ldots, w_{t-h-R_i+1}; \hat{\delta}_{t-h,R_i}^{(i)}), \quad \text{for } i = 1, 2,$$

where the semiparametric or nonparametric estimates $\hat{\delta}_{t-h,R_i}^{(i)}$ are based on a rolling window of size $R_i < \infty$, hence they are inconsistent.

Denote the forecast error by $\hat{e}_t^{(i)}(\hat{\delta}_{t-h,R_i}^{(i)}) = y_t - \hat{y}_t^{(i)}(\hat{\delta}_{t-h,R_i}^{(i)})$. For a loss function $L(\cdot)$, the loss differential is denoted as

$$d_t(\hat{\delta}_{t-h,R_1}^{(1)}, \hat{\delta}_{t-h,R_2}^{(2)}) = L(\hat{e}_t^{(1)}(\hat{\delta}_{t-h,R_1}^{(1)})) - L(\hat{e}_t^{(2)}(\hat{\delta}_{t-h,R_2}^{(2)}))$$

and the null hypothesis of equal predictive ability is

$$H_0 : E(d_s) = 0,$$

where $d_s = d_t(\hat{\delta}_{t-h,R_1}^{(1)}, \hat{\delta}_{t-h,R_2}^{(2)})$ and $s = t - (\max(R_1, R_2) + h) + 1$.

Giacomini and White (2006)

Theorem 1: As in Theorem 1 in Giacomini and White (2006), we assume that

(i) w_t is mixing with ϕ of size $r/(2r2)$, $r \geq 2$; or α of size $r/(r2)$, $r > 2$;

(ii) $E(|u_s|^{2r}) < \infty$ for all s;

(iii) $Var(\frac{1}{\sqrt{T}} \sum_{s=1}^{T} u_s) > 0$ for all T sufficiently large.

Remark: Assumptions (i)–(iii) are the same as in Giacomini and White (2006), except for the fact that here (ii) and (iii) are expressed in terms of u_s instead of d_s, since the latter may diverge because of the drift $\delta_2 T^a I_s(\tau)$.

Giacomini and Rossi (2010)

Giacomini and Rossi (2010) propose a local test statistic,

$$Fl_{s,k} = \frac{1}{k\hat{\sigma}^2} \sum_{l=s-k/2}^{s+k/2-1} d_l$$

where $k = \lfloor \kappa T \rfloor$. Under H_0 and regularity conditions,

$$Fl_{s,k} \Rightarrow \frac{B(\rho + \kappa/2) - B(\rho - \kappa/2)}{\sqrt{\kappa}},$$

where $B(\cdot)$ is a standard Brownian motion and ρ is such that $s = \lfloor \rho T \rfloor$. The Fluctuation test statistic is thus

$$Fl_k = \max_s |Fl_{s,k}|.$$

Andrews (2003)

Denote with $d_t = e^2_{1,t} - e^2_{2,t}$ the quadratic loss differential associated with two alternative models, where $e^2_{i,t} = y_t - \hat{y}_t$ and $\hat{y}_t$ is the predicted value of y_t.

Consider the following test of hypotheses:

- $H_0 : E(d_t) = 0$, for $t = 1, \dots, \lfloor \tau T \rfloor$ and d_t is stationary and ergodic for $t > 1$,
- $H_1 : E(d_t) \neq 0$, for some $t = \lfloor \tau T \rfloor + 1, \dots, T$ and/or the distribution of $(d_{\lfloor \tau T \rfloor + 1}, \dots, d_T)$ is different from the one of $(d_t, \dots, d_{t+T-\lfloor \tau T \rfloor})$,

where the change point $\tau \in [0, 1]$ is known in advance, and the number of post-change observations $T - \lfloor \tau T \rfloor$ is small.

Use a non-parametric subsampling approach to estimate the e.d.f. of d_t:

- compare the least-squares residuals of the regression $d_s = \mu + \delta I_s(\tau) + u_s$, for $s = t, \dots, T - \lfloor \tau T \rfloor$, against the residuals of the unrestricted model $d_s = \mu + u_s$, denoted by $\hat{u}_s$;
- in practise, reject at the α significance level if $S = \tilde{u}^2_{\lfloor \tau T \rfloor}$ exceeds the $(1 - \alpha)$ sample quantile of $\{\hat{u}^2_s : s = 1, \dots, \lfloor \tau T \rfloor - 1, \lfloor \tau T \rfloor + 1, \dots, T\}$.

Harvey et al. (2021)

Denote with $\hat{u}_s$ the residuals from a restricted regression, as in Andrews (2003), where $\hat{u}_s = d_s - \hat{\mu}$ for $s = 1, \ldots, T$;

Consider a monitoring procedure where:

- the sample is split in a training period $s = 1, \ldots, T^*$, and a monitoring period $s = T^* + 1, \ldots, E$, where $E \leq T$, $T^* = \lfloor \lambda_1 T \rfloor$ and $E = \lfloor \lambda_2 T \rfloor$ are fractions of the sample size T, for $0 < \lambda_1 < \lambda_2 \leq 1$;
- compare $\max_{s=1,\ldots,T^*} u_s^2$, during the training period, to $\max_{s=T^*+1,\ldots,E} u_s^2$, during the monitoring period, and conclude that an instability occurred if $\max_{s=1,\ldots,T^*} u_s^2 < \max_{s=T^*+1,\ldots,E} u_s^2$.

In particular, Harvey et al. (2021) show that

$$\lim P\left(\max_{s=1,\ldots,T^*} \hat{u}_s^2 < \max_{s=T^*+1,\ldots,E} \hat{u}_s^2\right) = \frac{\lambda_2 - \lambda_1}{\lambda_2}$$

is the false positive rate (FPR) of the procedure.