Forecast evaluation in panel data: some recent advancements

Andrea Viselli

University of Milan
andrea.viselli@unimi.it

Milan Econometrics Reading Group

October 30, 2024

Disclaimer: Any errors or omissions are solely the responsibility of Andrea Viselli and do not reflect upon the original authors.

Panels of forecasts

Panels of forecasts are increasingly available and becoming large:

- International institutions and private companies – for example, the IMF, World Bank, and firms in the financial industry such as banks – forecast a number of economic variables;
- Forecasts are collected through surveys or directly produced using a large number of models/methods/procedures/assessments;
- Forecasts of the same variables – e.g. the inflation rate and gross domestic product – for several countries generate over time large datasets of historical predictions.

Today's questions:

1. **Given two forecasts, which one is more accurate?**
2. **How this translates to panel data / cross-sections?**

Short review of the literature

Some refresh of the existing literature:

- comparing predictive accuracy with time series has been a prolific field (no space for all references);
- Pesaran et al. (2009) propose a version of the Diebold-Mariano test for panel data;
- recently, Akgun et al. (2023) propose a variety of tests for panel data.

Comparing predictive ability over time is **challenging** because:

- the sample size may be very small, so the statistical power may be low;
- if instead the sample is very long, non-stationarities may affect inference.

Today's presentation

We consider the work of Qu, Timmermann and Zhu (2023, 2024):

- 2023:
  - panel data setting, namely n countries and T time series;
  - cross-sectional dependence among the clusters.
- 2024:
  - cross-section setting, where only n is large;
  - the forecast error has a factor structure.

Why is this interesting compared to a time series setting?

- tests for differences in predictive ability utilizing the cross-sectional dimension – possibly very large – translate to more power;
- they allow to evaluate more timely the forecasts at the end of the sample, by fixing the time dimension.

The (standard) Diebold-Mariano test (1)

Description of the **environment** (Diebold and Mariano, 1995):

- Denote with $\hat{y}_{t|t-h,1}$ and $\hat{y}_{t|t-h,2}$ two competing h-step ahead forecasts of the variable y_t, made at time $t-h$.
- Denote $e_{t|t-h,m} = y_t - \hat{y}_{t|t-h,m}$ as the **forecast error**, for $m = 1, 2$.
- For evaluation, denote the **loss** associated to forecast m as $L(e_{t|t-h,m})$ for $m = 1, 2$. For example, $L(e_{t|t-h,m}) = e^2_{t|t-h,m}$ in MSE terms.
- Denote $d_{t|t-h} = L(e_{t|t-h,1}) - L(e_{t|t-h,2})$ as the **loss differential**.

We are interested in testing the null hypothesis of (unconditional) equal predictive accuracy:

$$H_0 : E(d_{t|t-h}) = 0, \qquad \text{for } t = 1, \ldots, T.$$

The (standard) Diebold-Mariano test (2)

Diebold and Mariano (1995) propose to use the time average

$$\bar{d} = \frac{1}{t} \sum_{t=1}^{T} d_{t|t-h},$$

and consider the test statistic

$$J_{DM} = \sqrt{T} \frac{\bar{d}}{\hat{\sigma}(\bar{d})},$$

where $\hat{\sigma}^2(\bar{d})$ is a consistent estimate of the long run variance $\sigma^2(\bar{d}) = Var(\bar{d})$.

Under H_0 and regularity conditions, as $T \to \infty$ it follows that

$$J_{DM} \xrightarrow{d} \mathcal{N}(0, 1).$$

Diebold-Mariano test for panel data

Now, we shift to a **panel data** setting:

- The target variable is y_{it}, for unit $i = 1, \dots, n$ and time $t = 1, \dots, T$.
- The h-step ahead forecast is denoted as $\hat{y}_{it|t-h,m}$, for the forecaster or model indexed by $m = 1, \dots, M$.
- Let $e_{it|t-h,m} = y_t - \hat{y}_{i,t|t-h,m}$ be the forecast error, then under quadratic loss function we have $L(e_{it|t-h,m}) = e^2_{it|t-h,m}$, for $m = 1, 2$.
- Suppose $M = 2$. Then $d_{it|t-h} = e^2_{it|t-h,1} - e^2_{it|t-h,2}$ is the loss differential for forecasts $m = 1, 2$.

Some hypothesis of interest:

$$H_0 : E(d_{it|t-h}) = 0.$$

- Which **subset** of the i's? Which of the t's?

Pesaran et al. (2009, IJF):

Which forecast is more accurate for **all** time periods t and units i?

Test for the pooled average (1)

Test for differences in predictive accuracy **across all** i **and** t, that is

$$H_0^P : E(d_{it|t-h}) = 0, \quad \text{for } t = 1, \dots, T \text{ and } i = 1, \dots, n.$$

To test this null, Pesaran et al. (2009) propose to use the global average

$$\bar{d}_{n,T} = \frac{1}{nT} \sum_{t=1}^{T} \sum_{i=1}^{n} d_{it|t-h},$$

and consider the test statistic

$$J_P = (nT)^{-\frac{1}{2}} \frac{\bar{d}_{n,T}}{\hat{\sigma}(\bar{d}_{n,T})},$$

where $\hat{\sigma}(\bar{d}_{n,T})$ is a consistent estimate of $\sigma(\bar{d}_{n,T}) = \sqrt{Var(\bar{d}_{n,T})}$.

Test for the pooled average (2)

In practice, consider the sequence of cross-sectional average loss differentials

$$\bar{d}_{t|t-h} = \frac{1}{n}\sum_{i=1}^{n} d_{it|t-h}, \qquad \text{for } t = 1, \ldots, T,$$

and retrieve the Diebold-Mariano test statistic, namely

$$J_P = \sqrt{T}\frac{\sum_{t=1}^{T}(\sqrt{n}\,\bar{d}_{t|t-h})}{\hat{\sigma}(\bar{d}_{t|t-h})},$$

where $\hat{\sigma}(\bar{d}_{t|t-h})$ is a Newey and West (1987) estimate of the long run variance (e.g. the Bartlett kernel).

Test for the pooled average (3)

Theorem 1

Suppose that:

1. $\max_{1\leq t\leq T} E|R_{t|t-h|^r}$ is bounded with $r > 2$;
2. $\{R_{t|t-h}\}_{t=1}^{T}$ is α-mixing of size $-r/(r-2)$;
3. $\hat{\sigma}(\bar{d}_{t|t-h}) = \bar{\sigma}_{n,T} + o_p(1)$ and $\bar{\sigma}_{n,T}$ is bounded away from zero, where $\bar{\sigma}_{n,T} = (nT)^{-\frac{1}{2}} \sum_{t=1}^{T} \sum_{i=1}^{n} \Delta L_{it|t-h}$.

Then, under H_0^P, as $T \to \infty$,

$$J_P \xrightarrow{d} \mathcal{N}(0,1).$$

Qu et al. (2024, IJF):

Which forecast is more accurate for **some** time periods t and units i?

For example:

- Normal and extraordinary times (e.g. periods of recession);
- Advanced and developing countries.

Test for time clusters (1)

Partition the panel data along the time series dimension into K mutually exclusive clusters, or subperiods, whom set is denoted as $\mathcal{T}_k$.

For each cluster k:

- the time series length is the cardinality of $\mathcal{T}_k$.
- the average loss differential is $\bar{d}_{t|t-h,k} = \frac{1}{T_k} \sum_{t \in \mathcal{T}_k} d_{t|t-h}$.

The null hypothesis of interest is whether two forecasts are equally accurate **within each of the time clusters**:

$$H_0^{TC} : E(\bar{d}_{t|t-h,1}) = \ldots = E(\bar{d}_{t|t-h,K}) = 0.$$

Test for time clusters (2)

Let $\bar{d} = \frac{1}{K}\sum_{j=1}^{K} \bar{d}_{t|t-h,k}$. Consider the test statistic

$$J_{TC} = \frac{\sqrt{K}\bar{d}}{\sqrt{(K-1)^{-1}\sum_{j=1}^{K}(\bar{d}_{t|t-h,k} - \bar{d})^2}}.$$

Assumption 1 (Ibragimov and Müller, 2010):

Let $\bar{d}_{(n)} = (\bar{d}_{t|t-h,1}, \ldots, \bar{d}_{t|t-h,K})' \in \mathbb{R}^K$. Then, $\bar{d}_{(n)} - E(d_{(n)}) \xrightarrow{d} \mathcal{N}(0, \Omega)$ as $n \to \infty$, where Ω is diagonal.

Theorem 2 with condition (1) only (Qu et al., 2024):

Suppose that Assumption 1 holds and $K \geq 2$ and $\alpha \leq 0.08326$.
Then, under H_0^{TC},

$$\limsup_{n\to\infty} P(|J_{TC}| > t_{K-1,1-\alpha/2}) \leq \alpha,$$

where $t_{K-1,1-\alpha/2}$ denotes the $1-\alpha$ quantile of the Student-t distribution with $K-1$ degrees of freedom.

Test for cross-sectional clusters

Now, cluster the data at the unit level into K mutually exclusive sets denoted as $\mathcal{H}_j$, for $j = 1, \ldots, K$.

For each cluster k:

- denote the cardinality as H_j, such that $\sum_{j=1}^{K} H_j = n$.
- the average loss differential is $\bar{d}_k = \frac{1}{TH_j} \sum_{i \in \mathcal{H}_j} \sum_{t=1}^{T} d_{it|t-h}$.

The null hypothesis of interest is whether two forecasts are equally accurate **within each of the cross-sectional clusters**:

$$H_0^{CC} : E(\bar{d}_1) = \ldots = E(\bar{d}_K) = 0.$$

The test follows by the same arguments of the test for time clusters (in particular, through Assumption 2 and Theorem 4 of Qu et al., (2024))

Remarks

1. The test for time clusters does **not** test whether two forecasts are equally accurate at different periods of time, but whether the two forecasts are **jointly** equally accurate **within** each time cluster;

- For example, if two forecasts are jointly equally accurate within times of recession and expansion.

2. The test for cross-sectional clusters does **not** test whether two forecasts are equally accurate at different periods of time, but whether the two forecasts are **jointly** equally accurate **within** each cross-sectional cluster;

- For example, if two forecasts are jointly equally accurate within two subgroups of advanced and developing countries.

3. These tests are agnostic about the nature of cross-sectional dependence.

4. When Ω in Assumptions 1,2 is not diagonal, Qu et al. (2024) propose to use a factor model to decorrelate the forecast errors.

Empirical example

Data on the IMF WEO, Consensus Economics (CE), and AR(1) forecasts:

- comparison forecasts with origin in spring and fall every year, and horizon $h = 0$ and $h = 1$ (a total of four horizons);
- comparison for 85 (real output growth) or 86 (inflation) countries.

Findings:

- the pooled test fails to be significant for any of the individual horizons (both WEO vs CE and WEO vs AR1);
- the test where three time clusters are built around the GFC fails to be significant for $h = 0$, yet it is for $h = 1$ (IMF WEO is significantly more accurate than AR1);
- similar results for the test with cross-sectional clusters (macro regions are considered as clusters), where IMF WEO is significantly more accurate than AR1.

Qu et al. (2023, JoE):

- a cross-section of forecast (forecasts of n variables);
- cross-sectional correlation with a factor structure.

Factor structure

Suppose again that forecasts are originated at $t - h$ (suppressed for convenience in the notation) and there are **2** forecasters.

To capture cross-sectional dependence, decompose the forecast error as

$$e_{itm} = \lambda'_{im} f_t + u_{itm}, \qquad \text{for } m = 1, 2,$$

where:

- f_t is the common factors;
- λ_{im} the loadings to the factors;
- u_{itm} the idiosyncratic component.

The common component f_t does **not** vanish asymptotically even as $n \to \infty$.

Idea: "control" for the common component, then test the equality of the loss differential of idiosyncratic error variances.

Conditional cross-sectional test (1)

When the loadings to the factors are **heterogeneous**, or $\lambda_{i1} \neq \lambda_{i2}$, the loss differential is

$$d_{it} = [(\lambda_{i1}' f_t)^2 - (\lambda_{i2}' f_t)^2] + [(u_{it1}^2 - u_{it2}^2) + 2(\lambda_{i1}' f_t u_{it1} - \lambda_{i2}' f_t u_{it2})].$$

Notice that:

- by a CLT, even as $n \to \infty$, $n^{-\frac{1}{2}} \sum_{i=1}^{n} [(\lambda_{i1}' f_t)^2 - (\lambda_{i2}' f_t)^2]$ is asymptotically normal only **conditional** on f_t;
- the information set at time t is $\mathcal{F} = \sigma(f_t, \{\lambda_{i1}, \lambda_{i2}\}_{i=1}^{n})$.

This suggests testing the null hypothesis of **conditional** equal predictive ability:

$$H_0^{con} : E(\bar{d}_t | \mathcal{F}) = 0.$$

Conditional cross-sectional test (2)

Assume that $E(u_{it1}|\mathcal{F}) = E(u_{it2}|\mathcal{F}) = 0$.

Let $\xi_{it} = \Delta L_{it|t-h} - E(\Delta L_{it|t-h})$, then

$$\xi_{it} = (u_{it1}^2 - u_{it2}^2) - E(u_{it1}^2 - u_{it2}^2|\mathcal{F}) + 2(\lambda_{i1} f_t u_{it1} - \lambda_{i2} f_t u_{it2})$$

whose variance $n^{-1}\sum_{i=1}^{n} \xi_{it}^2$ is unobservable.

The following test statistic is thus considered:

$$\widetilde{Q}_t = \frac{n^{\frac{1}{2}}\overline{\Delta L_t}}{\sqrt{n^{-1}\sum_{i=1}^{n}(\Delta L_{it|t-h} - \overline{\Delta L_t})^2}}.$$

.

Conditional cross-sectional test (3)

Assumption 2

1. Conditional on $\mathcal{F} = (f_t, \{\lambda_{i1}, \lambda_{i2}\}_{i=1}^n), \{(u_{it1}, u_{it2})\}_{i=1}^n$ is independent across i with mean zero and bounded $(4 + \delta)$ moments for some $\delta > 0$;
2. $\min_{1 \le i \le n} Var(\xi_{it} | \mathcal{F}) \ge c$ for some constant c.

Theorem 2

Suppose Assumption 2 holds. Then, under H_0^{con},

$$\limsup_{n \to \infty} P(|\widetilde{Q}_t > z_{1-\alpha/2}) \le \alpha,$$

where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)$ quantile of a $\mathcal{N}(0, 1)$ random variable.

Conditional cross-sectional test (4)

Now, apply the following bias-variance decomposition:

$$\underbrace{E(\bar{d}_t)}_{\frac{1}{n}\sum_{i=1}^{n} E(d_{it}|\mathcal{F})} = \underbrace{bias_t^2}_{\frac{1}{n}\sum_{i=1}^{n}[(\lambda'_{i1}f_t)^2-(\lambda'_{i2}f_t)^2]} + \underbrace{E(\Delta u_{it}^2|\mathcal{F})}_{\frac{1}{n}\sum_{i=1}^{n}(u_{it1}^2-u_{it2}^2|\mathcal{F})}$$

where the terms on right hand side are **not** observable. However,

$$\bar{d}_t - bias_t^2 = \frac{1}{n}\sum_{i=1}^{n}\Delta u_{it}^2 + \frac{2}{n}\sum_{i=1}^{n}(\lambda'_{i1}f_t u_{it1} - \lambda'_{i2}f_t u_{it2}),$$

is observable, in particular:

- if the last term on the right hand side is small, then $\bar{d}_t - bias_t^2$ is a good estimate of the loss differential of idiosyncratic error variances;
- a method to compute the $bias_t^2$ term is required.

Unconditional/Conditional cross-sectional test

Qu et al. (2023) propose to use three methods:

1. to consider clusters (known ex ante) where the factors loadings are homogeneous within the cluster;
2. methods for panel data

   2a. the common correlated effects model of Pesaran (2006);

   3b. principal component analysis (PCA).

Notice that:

- 1 determines a test of **unconditional** equal predictive accuracy, as the cluster-specific common component cancel out in the loss differential;
- depending on the method we use, we have a different expression of the test statistic (omitted from this discussion);
- one may want to use an unconditional test, provided that she/he pre-tests whether the factor loadings are homogeneous using a test for the equality of the bias term.

Remarks:

The unconditional and conditional tests differ in their interpretation:

- a conditional test is of interest if, for example during the Covid-19 pandemic, conditional on the economic shock to economic activity the performance of two alternative forecasts is equivalent;
- rejection of a conditional test (but not of an unconditional test) means that idiosyncratic aspects of the forecasts drive the performance;
- conversely, rejection of a unconditional test (but not of a conditional test) means that factor realizations drive the performance.

Empirical example (1)

Data from the Institutional Brokers Estimate System (IBES):

- Comparison of forecasts of quarterly earning per share (EPS) for four pairs of brokerage firms, namely Morgan Stanley vs. Goldman, Morgan Stanley vs. Merrill, Goldman vs. Merrill, and Lawrence (Deutsche Bank) vs. Merrill;
- the joint inspection of multiple test statistics–81 introduces a multiple hypothesis testing problem;
- for a sup-type bootstrap approach that evaluates the joint statistical significance of individual test statistics, see Qu et al. (2019).

Empirical example (2)

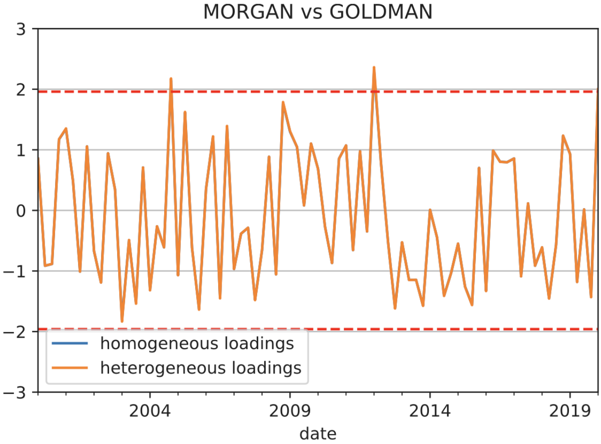

Figure: Cross-sectional test statistics for comparisons of the null of equal squared error loss. Positive values of the test statistics indicate that the second forecaster is more accurate than the first forecaster.

References

- Akgun, Oguzhan, et al. "Equal predictive ability tests based on panel data with applications to OECD and IMF forecasts." International Journal of Forecasting 40.1 (2024): 202-228.
- Diebold, Francis X., and Robert S. Mariano. "Comparing predictive accuracy." Journal of Business & economic statistics 20.1 (2002): 134-144.
- Ibragimov, R., & Müller, U. K. (2010). T-statistic based correlation and heterogeneity robust inference. Journal of Business & Economic Statistics, 28(4), 453–468.
- Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation-consistent covariance ma- trix. Econometrica, 55(3), 703–708.
- Pesaran, M. Hashem. "Estimation and inference in large heterogeneous panels with a multifactor error structure." Econometrica 74.4 (2006): 967-1012.
- Pesaran, M. H., Schuermann, T., & Smith, L. V. (2009). Forecasting economic and financial variables with global vars. International Journal of forecasting, 25(4), 642–675.
- Qu, Ritong, Allan Timmermann, and Yinchu Zhu. "Do any economists have superior forecasting skills?." (2019).
- Qu, Ritong, Allan Timmermann, and Yinchu Zhu. "Comparing forecasting performance in cross-sections." Journal of econometrics 237.2 (2023): 105186.
- Qu, Ritong, Allan Timmermann, and Yinchu Zhu. "Comparing forecasting performance with panel data." International journal of forecasting 40.3 (2024): 918-941.