Evaluating Multiple Interval Forecasts

Christian Brownlees
Andre B.M. Souza

upf.

Universitat Pompeu Fabra & Barcelona GSE

upf.

Introduction

Introduction

- Interval forecasts are extensively used in Finance and Economics.
- Often, practitioners are faced with the task of producing interval forecasts for multiple time series
- *i.e* Value-at-Risk forecasts for multiple trading desks, M4 Competition.

  Berkowitz, Christoffersen, and Pelletier (2011), Makridakis, Spiliotis, and Assimakopoulos (2018)
- Most of the backtesting methodology is designed for univariate series

  Christoffersen (1998), Engle and Manganelli (2004)
- Little work has been done on the evaluation of multiple interval forecasts. Length has been used, but properties are unclear.

  Askanazi, Diebold, Shin, and Schorfheide (2018), Makridakis et al. (2018)

upf.

In this work...

- We develop a methodology to evaluate multiple interval forecasts
- We assume a forecaster has M collections of interval forecasts for a panel of time series and must provide a ranking.
  All forecasts are assumed to have correct coverage for all series
- We propose as an optimality criteria to select the method that minimizes a measure of dependence across cross-sectional violations
  Everything equal, forecasters prefers methods that minimize the probability of simultaneous interval forecast violations.
- We develop a data-driven selection procedure and establish its consistency.

upf.

Empirical Illustration

- We apply the proposed methodology to evaluate common interval forecasting methods for all S&P 500 components
- VaR forecasts: GARCH, TARCH, Factor GARCH and Rolling Window
- Our methodology selects the collection generated by the F-GARCH
- Financial crisis: 78% reduction in simultaneous hits when compared to RW and 20% when compared to TARCH.

upf.

Related Literature

1. Absolute evaluation of interval forecasts
   - Christoffersen (1998), Engle and Manganelli (2004).
2. Relative evaluation of interval forecasts
   - Askanazi et al. (2018), Winkler (1972), Giacomini and Komunjer (2005), Gneiting and Raftery (2007).
3. Evaluation of vectors of forecasts
   - Sinclair, Stekler, and Carnow (2015), Hendry and Martinez (2017).
4. Risk Management
   - Berkowitz et al. (2011), Escanciano and Olmo (2011), Escanciano and Hualde (2017)

upf.

Methodology

Setup

- Let Y_t denote a n-dimensional vector of time series observed from $t = 1, \ldots, T$.

- An interval forecast for series i at time t with coverage $1 - \alpha$ is

$$PI_{i\,t} = [PI^{L}_{i\,t}, PI^{U}_{i\,t}] \quad \text{s.t} \quad \mathbb{P}(Y_{i\,t} \in PI_{i\,t}) = 1 - \alpha$$

- There are $m = 1, \ldots, M$ procedures to construct interval forecasts with nominal coverage $1 - \alpha$ and we denote by

$$\mathcal{P}_m = \{PI_{m\,i,t}\}_{i=1,t=1}^{n,T}$$

  the collection of such forecasts generated by method m.

- We are interested in ranking the collections $\mathcal{P}_m$ for $m = 1, \ldots, M$.

upf.

Remarks

- We **do not** consider prediction regions with uniform coverage $(1-\alpha)$.
  Already considered in Christoffersen (1998).
- Without loss of generality, we focus on 1-step ahead interval forecasts rather than *h*-step ahead forecasts
- One may want to combine collections. For simplicity, we do not consider this.
- We assume all intervals are based on the same quantiles for all series

upf.

Methodology

- Let $H_{m\,it} = \mathbf{1}\{Y_{i\,t} \notin PI_{m\,it}\}$ be the hit variable for series i at time t, generated by method m.
- It is standard to evaluate interval forecasts by the properties of the univariate series of hits.
- We assume methods have correct (unc.) coverage i.e. $\mathbb{E}[H_{m\,it}] = \alpha$
- Our evaluation methodology is based on the cross-sectional properties of $\mathbf{H}_t = (H_{m\,1t}, \ldots, H_{m\,nt})'$
- We assume $\mathbf{H}_t$ is covariance stationary.

upf.

Efficiency Criteria

We propose to rank collections according to their dependence properties.

Definition Efficiency

Assume we have $m = 1, \ldots, M$ collections of prediction intervals. Let

$$\bar{\tau}_m = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{i-1} \mathbb{P}(H_{m\,it} \cap H_{m\,jt}) \ .$$

Then we say a collection m^* is efficient if

$$m^* = \arg\min_{m \in M} \left| \bar{\tau}_m - \alpha^2 \right| \ .$$

Reminiscent to mixing, mutual information, KL divergences, etc...

upf.

Remarks on Criteria

- Univariate Setting: Length vs Coverage
  1. Shorter intervals are presumably conditioning on more valuable information sets. Granger, White, and Kamstra (1989)
- Multivariate: Length vs Coverage vs Dependence
  1. Given correct unconditional coverage, we assume forecasters prefer collections of intervals where the dependence between violations is minimized.
  2. Other optimality criteria (or combinations of such) may be chosen.
  3. Askanazi et al. (2018): Lengths are not directly comparable across series with different scales.
- Correct (Unc.) Coverage is empirically hard to reject. Berkowitz (2001)

upf.

Example

Example

Consider a simple linear model:

$$Y_{i\,t} = X_t + Z_{i\,t} \qquad X_t \sim N(0,1), Z_{i\,t} \sim N(0,1)$$

and the following PIs:

$$PI_{1\,it} = \left[\sqrt{2}\Phi^{-1}\left(\frac{\alpha}{2}\right), \sqrt{2}\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right]$$
$$PI_{2\,it} = \left[X_t + \Phi^{-1}\left(\frac{\alpha}{2}\right), X_t + \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right]$$

For $i = 1, \ldots, n$, $t = 1, \ldots, T$ and where Φ is the standard normal cdf.

upf.

Example

- Both methods provide correct unconditional coverage
- Simultaneous hits are more likely under method 1 than method 2
- In this example we know that X_t drives the factor structure in the panel, so we can easily test for optimality wrt X_t
- If we dont know what drives the panel, we have a model selection issue.

upf.

Evaluation

The following lemma makes our definition of efficiency operational.

Lemma

Let

$$\overline{H}_{m\,t} = \frac{1}{n}\sum_{i=1}^{n} H_{m\,it} \qquad \text{and} \qquad \bar{\tau}_m = \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{i-1}\mathbb{P}(H_{m\,it} \cap H_{m\,jt})$$

Then, under the previously stated assumptions and assuming additionally $\bar{\tau}_m \geqslant \alpha^2$ *for all* $m = 1, \ldots, M$*, we have that*

$$\arg\min_{m\in M} \mathbb{E}\left[(\overline{H}_{m\,t} - \alpha)^2\right] = \arg\min_{m\in M} \left|\bar{\tau}_m - \alpha^2\right| \ .$$

upf.

Comments

- We restrict $\overline{\tau}_m$ to be at least that obtained under cross-sectional independence.
- Empirically justified in Finance and Economics.
- We note that $\overline{\tau}_m \geqslant \alpha^2 - \frac{\alpha(1-\alpha)}{n-1}$, so the assumption $\overline{\tau}_m \geqslant \alpha^2$ is not so restrictive, for n large.

upf.

Empirical Evaluation

- We evaluate collections by the empirical analog of $\mathbb{E}\left[(\overline{H}_{m\,t} - \alpha)^2\right]$:

$$L_m = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{n}\sum_{i=1}^{N} H_{m\,it} - \alpha\right)^2$$

- We assume $\mathbf{H}_t = (H_{m\,1t}, \ldots, H_{m\,nt})'$ is strong mixing with coefficient ψ that satisfies:

$$\psi(l) \leqslant \exp\left(-Al^{\beta}\right)$$

for some $A > 0$ and $\beta > 1$.

upf.

Empirical Evaluation

Lemma

Let m^* be the efficient collection and $\widehat{m}^* = \arg\min_{m \in M} L_m$ Then,

$$\mathbb{P}\left(\widehat{m}^* = m^*\right) \geqslant 1 - C \exp\left\{\log(M) - \frac{(n-1)}{n}\left(\overline{\tau}_{m'} - \overline{\tau}_{m^*}\right) T^{\frac{\beta}{\beta+1}}\right\}$$

where $m' = \arg\min_{m \in M \setminus m^*} \mathbb{E}[(\overline{H}_{m\,t} - \alpha)^2]$ and for a constant C.

- (Rate) Stronger time series dependence leads to slower selection
- (# of Collections) M can grow as a power of T.

upf.

Empirical Application

Empirical Application

- We apply our framework to evaluate Value-at-Risk forecasting methods for the individual components of S&P500.
- We consider log returns for the S&P 500 components, from 01/01/2000 to 01/01/2019.
- VaR forecasts are constructed using Rolling Window, GARCH, TARCH and Factor GARCH.
- We take 25% of the available observations for each stock as the in sample period. We re-estimate the volatility parameters once per year out of sample.
- We construct quantile forecasts by means of Filtered Historical Simulation (Barone-Adesi, Engle, and Mancini (2008))

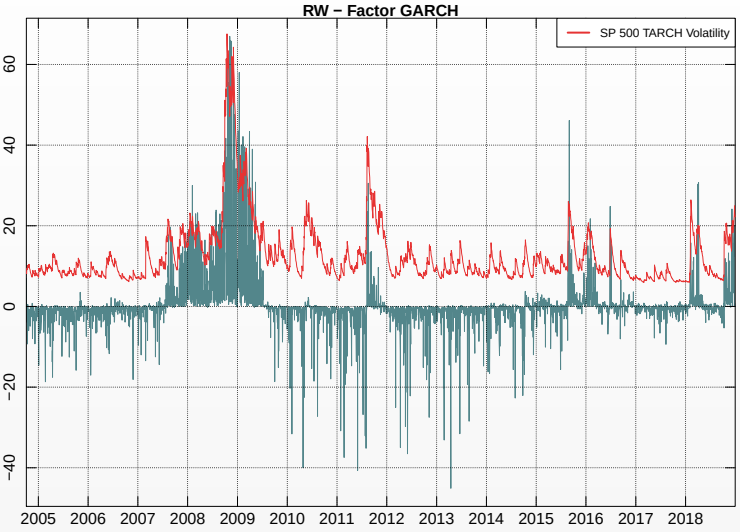
upf.

Empirical Application: Results

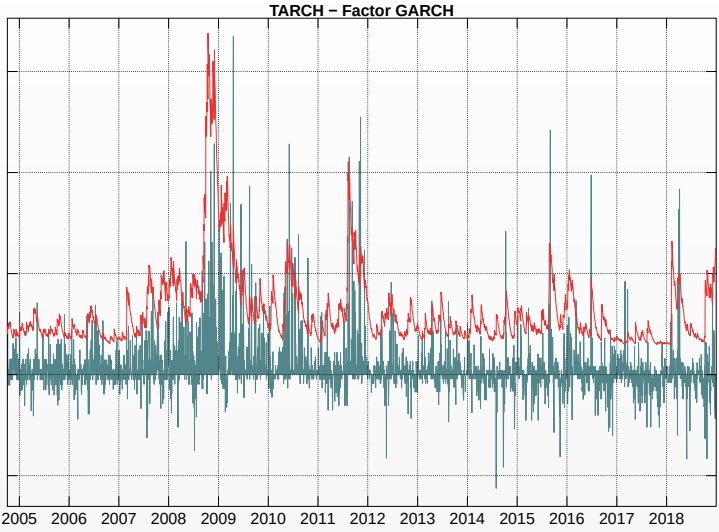
	Factor GARCH	TARCH	GARCH	Rolling Window
Average Hits(%)	4.796	5.015	4.954	5.138
Average Length	2.982	2.898	2.914	3.099
UC(1%)	93.8%	98.4%	98.2%	89.4%
CC(1%)	85.2%	93.1%	87.4%	20.4%
DQ(1%)	74.5%	75.8%	63.4%	7.30%
$\overline{\tau} \times 100$ (Note: $\alpha^2 \times 100 = 0.25$)	0.807	0.876	0.901	1.212
Loss	17.604	19.152	19.484	27.852

Since $\mathbb{E}L_m \leqslant \alpha(1-\alpha)$, we report $Loss_m = \frac{L_m}{\alpha(1-\alpha)} \times 100$

Empirical Application: Figures



Empirical Application: Figures



Conclusion

upf.

Summary

- We introduce a criteria to evaluate multiple interval forecasts
- We propose to rank methods according to the dependence properties of the collection of hit series
- We apply our framework to evaluate methods to construct VaR forecasts for each of the S&P500 stocks
- Methods that incorporate the factor structure of volatility reduce the dependence across VaR hits, performing better according to the proposed metric.

upf.

Thanks!

upf.

Bibliography

Ross Askanazi, Francis X. Diebold, Minchul Shin, and Frank Schorfheide. On the comparison of interval forecasts. *Journal of Time Series Analysis*, 39:953–965, Sep 2018. doi: 10.1111/jtsa.12426.

Giovanni Barone-Adesi, Robert F. Engle, and Loriani Mancini. A garch option pricing model with filtered historical simulation. *Review of Financial Studies*, 21(3):1223–1258, May 2008. doi: 10.1093/rfs/hhn031.

Jeremy Berkowitz. Testing density forecasts, with applications to risk management. *Journal of Business & Economics Statistics*, 19: 465–474, 2001. doi: 10.1198/07350010152596718.

Jeremy Berkowitz, Peter Christoffersen, and Denis Pelletier. Evaluating value-at-risk models with desk-level data. *Management Science*, 12:2213–2227, 2011.

Peter Christoffersen. Evaluating interval forecasts. *International Economic Review*, 39(4):841–862, Nov 1998.

Robert F. Engle and Simone Manganelli. Caviar: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economics Statistics*, 22:367–381, Oct 2004. doi: 10.1198/073500104000000370.

J. Carlos Escanciano and Javier Hualde. Measuring asset market linkages: Nonlinear dependence and tail risk. *Working Paper*, 2017.

J. Carlos Escanciano and Jose Olmo. Robust backtesting tests for value-at-risk models. *Journal of Financial Econometrics*, 9:132–161, 2011.

Raffaella Giacomini and Ivana Komunjer. Evaluation and combination of conditional quantile forecasts. *Journal of Business & Economics Statistics*, 23(4):416–431, 2005.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378, 2007. doi: 10.1198/016214506000001437.

Clive W.G Granger, Halbert White, and Mark Kamstra. Interval forecasting: An analysis based upon arch-quantile estimators. *Journal of Econometrics*, 40:87–96, 1989.

David F. Hendry and Andrew B. Martinez. Evaluating multi-step system forecasts with relatively few forecast-error observations. *International Journal of Forecasting*, 33:359–372, 2017.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The m4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, 34:802–808, 2018.

Tara M. Sinclair, Herman O. Stekler, and Warren Carnow. Evaluating a vector of the fed's forecasts. *International Journal of Forecasting*, 31:157–164, 2015.

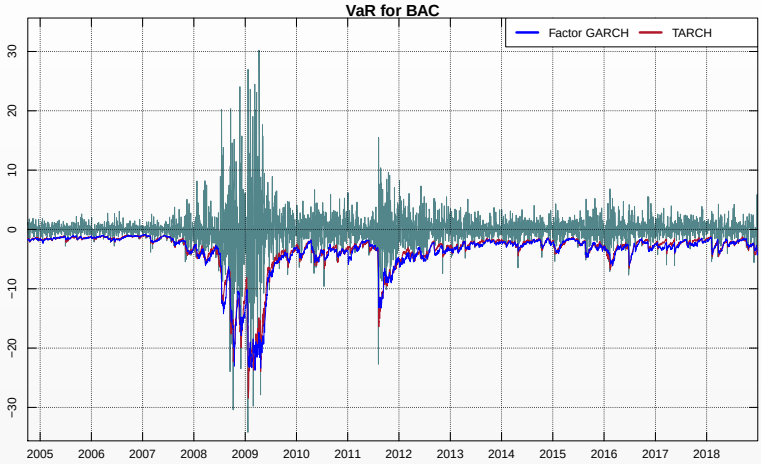
Robert L. Winkler. A decision theoretic approach to interval estimation. *Journal of the American Statistical Association*, 67:187–191, 1972. doi: 10.1080/01621459.1972.10481224.

upf.

Appendix

upf.

Empirical Application: Figures



Empirical Application: Additional Table

	Factor GARCH	TARCH	GARCH	Rolling Window
Loss	20.940	22.553	22.854	34.180
Average Hits(%)	5.005	5.186	5.116	5.662
$\overline{\tau}$	0.907	0.971	0.996	1.395
UC(10%)	1	0.997	1	0.914
CC(1%)	0.830	0.894	0.769	0.058
Length	2.920	2.846	2.862	3.030

upf.