A note on jointly backtesting models for multiple assets and horizons[☆]

David Ardia[a,b,*], Anas Guerrouaz[b], Lennart F. Hoogerheide[c]

[a] *Institute of Financial Analysis, University of Neuchâtel, Switzerland*
[b] *Département de finance, assurance et immobilier, Université Laval, Québec, Canada*
[c] *Department of Econometrics, Vrije Universiteit Amsterdam, The Netherlands*

Abstract

We propose a simulation-based methodology which allows us to test the performance of multi-level and/or multi-horizon Value-at-Risk forecasts.

Keywords: Bootstrap test, GARCH, dependent time series, multiple testing, Value-at-Risk

1. Introduction

The simulation framework proposed by Ardia et al. (2014) is useful when one wants to evaluate the validity of a set of models for dependent univariate time series and summarize the test results in one joint test statistic and p-value. This approach leads to higher power than a test for a univariate time series, which is especially useful for short time series. The bootstrap approach can be used in conjunction with tests aiming at evaluating the validity of interval forecasts such as Christoffersen (1998)'s unconditional coverage (UC) test which relies on LR statistics. In our case, the interval forecast of interest will be the Value-at-Risk.

If we consider the sum of the individual UC test statistics (LR statistics) as the joint test statistic, the new statistic does not have a known parametric distribution under the null hypothesis (i.e., all the marginals are correctly specified) when the time series' independence is violated (it would be χ^2-distributed if the series were independent). The p-value of the test must then be computed by bootstrapping the test statistic's distribution, through simulations of the time series under the null hypothesis and the corresponding simulations of the UC test statistic.

This note makes use of the bootstrap approach to test multiple VaR series for unconditional coverage. We extend the original article in the fact that we seek to test jointly the VaR estimates of multiple assets, for multiple nominal coverage levels and/or multiple horizons. We will illustrate this by testing three different VaR levels for four assets (large international indices), for three different horizons.

The bootstrap approach is even more crucial here in the sense that the series are dependent because of cross-asset dependence but also by construction (since we consider multiple VaRs levels and horizons per asset). Thus we expect the joint statistic's departure from its asymptotic distribution (under the independence hypothesis) to be even more pronounced than the joint testing of one VaR series per asset, as is the case in the original article. The LR statistic for an individual time series does not even have a $\chi^2(1)$ distribution under the null hypothesis in this case.

[☆] All analyses have been performed in the `R` statistical language (R Core Team, 2015) with the package **rugarch** (Ghalanos, 2015) and are available from the authors upon request. Any remaining errors or shortcomings are the authors' responsibility.

[*] Corresponding author.

Email addresses: `david.ardia@unine.ch` (David Ardia), `Anas.Guerrouaz@gmail.com` (Anas Guerrouaz), `l.f.hoogerheide@vu.nl` (Lennart F. Hoogerheide)

Electronic copy available at: https://ssrn.com/abstract=2732069

2. Methodology

Given marginal models for the mean and variance of the asset log-returns, we proceed to a daily estimation of the models' parameters (using a fixed-length estimation sample) and we simulate the future log-returns using the estimated parameters.

At this point our null hypothesis H_0 is that all marginal models are correct and several specification tests could be applied to assess the validity of this hypothesis.

In our case we focus on the left tail of the distribution, which is arguably of bigger concern than the distribution's interior to financial institutions. To do so, we will apply an unconditional coverage test to our VaR estimates so we need a sequence of VaR violations for each VaR series, which can be obtained by comparing the vector of realized log returns to the vector containing the relevant conditional VaR estimate. The same result can be achieved by comparing the Probability Integral Transform of the realized return to the nominal VaR level.

The rest of our methodology focuses on the PITs as they contain the information about VaR violations for any level (indeed, a PIT below $\alpha\%$ is synonymous with a violation of the $\alpha\%$ conditional VaR).

The PITs are calculated as empirical cumulative probabilities, using the empirical conditional distribution of future log-returns obtained through daily simulation. At each estimation step, we simulate 10,000 paths of which the length equals the maximum VaR horizon; we hence obtain cumulated log-return distributions for each VaR horizon (and the PITs corresponding to the realized cumulated returns; we now have one PIT series per asset, per horizon). The transformation of the PITs into conditional VaR violation sequences is trivial so we can readily apply the relevant test to the desired violation sequences, thus obtaining one test statistic per violation sequence. The calculation of the realized joint test statistic is quite straightforward, as we sum the LR statistics from the unconditional coverage test (one LR statistic per asset, per VaR level, per horizon).

As stated before, this new test statistic does not have a known parametric distribution since the tested series are dependent. We will hence simulate its distribution through our bootstrap approach. To take the dependence between the PIT series into account, we compute the p-value by simulating arrays of PITs (one series per column) under H_0 then calculate the test statistic for each of these simulated arrays in order to determine the statistic's distribution.

We first compute the array R, of which each column contains the ranking numbers of the corresponding column of the original PIT array.

The simulation of the PIT array under H_0 is in two steps:

1. We first simulate the PIT ranks by selecting randomly (according to a uniform distribution) rows among the actual PITs' rank array (same row numbers for all horizons). By selecting a whole row, we preserve the rank-wise dependence structure between columns (assets); by selecting the same row for all horizons, we preserve the rank-wise dependence structure between different return horizons. Note that in the case of multi-day return horizons we select blocks of consecutive rows of which the length equals the return horizon of interest.

2. Once the new ranks have been simulated, we simulate the corresponding PITs: under H_0 (i.e., all conditional marginal models are correct) the PITs should be uniformly distributed on the [0,1] interval. Then the kth-order PIT (i.e., that of rank k) has a Beta distribution. Given a rank k out of n, the simulated kth-order PIT is a realization from the Beta-distribution with parameters k and $(n-k+1)$.

We repeat the simulation of the PIT array under H_0 multiple times (we use 500 bootstrap simulations in our article) and compute the test statistic (i.e., the sum of N LR statistics) for each bootstrapped PIT array to obtain a bootstrapped distribution of the test statistic, from which we compute the test's p-value.

Electronic copy available at: https://ssrn.com/abstract=2732069

A scheme of the procedure is presented in Figure 1.

[Insert Figure 1 about here.]

3. Illustrations

We illustrate our bootstrapped methodology through the joint testing of the VaR estimates for four assets, three nominal VaR levels and three forecast horizons.

Our raw data are comprised of 2000 daily log-returns of four international indices (S&P 500, Dow Jones Industrial Average, FTSE 100 and Nikkei 225) between 2000/01/04 and 2008/06/04.

Our marginal models for the mean and variance of the asset log-returns are an AR(1) and four GARCH-type specifications commonly used by finance practitioners:

- GARCH (Bollerslev, 1986) with normal innovations.
- GJR (Glosten et al., 1993) with normal innovations.
- GJR with student-t innovations.
- GJR with skewed student-t innovations.

The models' parameters are estimated using 1000 data points and used to simulate the future log-returns. We then compute the out-of-sample PITs by comparing the realized log-return to the simulated distribution of log-returns. The model is re-estimated daily, leading to 1000 out-of-sample PITs per asset, per forecast horizon.

Using these PITs, we apply our bootstrap methodology to conduct a joint UC test, of which the p-values are reported in Table 1.

[Insert Table 1 about here.]

Using a 5% significance level, we can see that the rate of rejection gradually diminishes with each variance model, in the above order. This progression could be expected, as each new model better takes into account stylized facts about the log-returns:

- In the GJR model, the impact of past innovations on the variance is allowed to be different for returns with positive or negative sign.
- Introducing student-t innovations allows for fat tails of the distribution.
- Introducing skew student-t innovations allows for asymmetry as well as fat tails of the distribution.

Our joint test procedure allows any combination and any number of horizons and VaR nominal levels to be jointly tested, thus adapting to the needs of different users.

4. Possible extensions

The approach presented here could be extended in several ways. While we have chosen to illustrate the approach by applying the same test (i.e., unconditional coverage) to all VaR series, we could combine different tests in a joint test statistic (e.g., a test for the independence of VaR exceedances and a test for the uniformity of loss quantiles). We would construct the joint statistic and bootstrap its distribution the same way we did in the single-test case. Another possible extension would be unequal weighting of the individual test statistics within the joint statistic, thus allowing to put the focus on specific VaR series and/or tests.

References

Ardia, D., Gatarek, L., Hoogerheide, L., 2014. A new bootstrap test for multiple assets joint risk testing. Working paper.
URL http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2312007

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics 31 (3), 307–327.

Christoffersen, P., 1998. Evaluating interval forecasts. Journal of Finance 39 (4), 841–862.

Ghalanos, A., 2015. rugarch: Univariate GARCH models in R. R package version 1.3-6.
URL https://cran.r-project.org/web/packages/rugarch/index.html

Glosten, L. R., Jagannathan, R., Runkle, D. E., 1993. On the relation between the expected value and the volatility of the nominal excess return on stocks. Journal of Finance 48 (4), 1779–1801.

R Core Team, 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL http://www.R-project.org/

Table 1: Backtesting results. For each model, forecasting horizon and VaR risk level, the table presents the p-value associated with the joint UC test. Values below the 5% significance level are reported in bold.

		VaR level			
Model	Horizon	99%	95%	90%	All
GARCH-N	d	**0.00**	0.62	0.88	**0.02**
	w	**0.00**	0.19	0.61	**0.05**
	m	**0.00**	**0.04**	0.30	**0.03**
	d+w	**0.00**	0.22	0.71	**0.02**
	d+w+m	**0.00**	0.06	0.42	**0.03**
GJR-N	1	**0.00**	0.88	0.38	**0.01**
	w	0.18	0.87	0.87	0.61
	m	0.48	0.44	0.50	0.51
	d+w	**0.02**	0.94	0.77	0.27
	d+w+m	0.09	0.64	0.64	0.46
GJR-S	1	**0.05**	0.71	0.82	0.32
	w	0.15	0.74	0.97	0.56
	m	0.19	0.26	0.41	0.31
	d+w	0.10	0.80	0.99	0.45
	d+w+m	0.13	0.39	0.61	0.37
GJR-SS	1	0.80	0.94	0.20	0.65
	w	0.34	0.89	0.75	0.72
	m	0.19	0.50	0.49	0.42
	d+w	0.53	0.97	0.56	0.79
	d+w+m	0.34	0.72	0.54	0.54

Figure 1: Scheme of the bootstrap test. Within each matrix the rows are the periods and the columns are the assets. h_1, h_2 and h_3 refer to the forecasting horizons.