The Test for Superior Predictive Ability (SPA): A Detailed Explanation

1 Motivation

In evaluating predictive models—such as volatility forecasts from different GARCH models—we often compare a large set of models to a benchmark (e.g., GARCH(1,1)).

The Problem

When many models are compared, at least one may appear to outperform the benchmark just by chance, due to sampling variability. This issue is known as the **data snooping problem** or **multiple testing bias**. To avoid false positives (Type I errors), we require a statistical test that:

- Accounts for all models being compared.
- Adjusts for the selection bias inherent in choosing the best model.
- Provides valid inference on whether any model has *superior predictive ability* (SPA).

2 Objective of the SPA Test

The SPA test evaluates the null hypothesis:

$$\begin{aligned} H_0 &: \text{None of the alternative models perform better (in expectation) than the benchmark,} \\ H_1 &: \text{At least one model has strictly better (lower expected) loss than the benchmark.} \end{aligned}$$

3 Notation and Setup

Let K be the number of alternative models compared to a benchmark (model 0). Define:

- $L_t^{(k)}$: Loss at time t from model k.
- $d_{k,t} = L_t^{(0)} - L_t^{(k)}$: Loss differential.

Interpretation:

- $d_{k,t} > 0$ indicates model k performs better than the benchmark.
- $\mu_k = \mathbb{E}[d_{k,t}]$: Expected loss differential.

Define the sample average:

$$\bar{d}_k = \frac{1}{n}\sum_{t=1}^{n} d_{k,t}.$$

The null hypothesis becomes:

$$H_0 : \max_{1\le k\le K} \mu_k \le 0,$$

with the alternative:

$$H_1 : \exists k \text{ such that } \mu_k > 0.$$

4 SPA Test Statistic

The SPA test statistic is:

$$T_n = \max_{1\le k\le K} \left(\sqrt{n}\cdot \bar{d}_k\right)_+,$$

where $x_+ = \max(x, 0)$.

5 Bootstrap Inference

Step 1: Centering

Under H_0, center the loss differentials:

$$\tilde{d}_{k,t} = d_{k,t} - \bar{d}_k.$$

Step 2: Stationary Bootstrap

Apply the stationary bootstrap (Politis and Romano, 1994) to generate resampled sequences $\tilde{d}^*_{k,t}$. Compute:

$$\bar{d}^*_k = \frac{1}{n}\sum_{t=1}^{n} \tilde{d}^*_{k,t}, \quad T^*_n = \max_{1\le k\le K} \left(\sqrt{n}\cdot \bar{d}^*_k\right)_+.$$

Step 3: p-value Computation

With B bootstrap replications, the SPA p-value is estimated by:

$$\text{p-value} = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}(T^{*(b)}_n > T_n).$$

Reject H_0 if the p-value $< \alpha$.

6 Comparison with Reality Check (RC)

Feature	Reality Check (RC)	SPA Test
Centering	Global	Selective
Studentization	No	Yes
Power	Low	High
Bootstrap	Block/Stationary	Stationary

7 Example: Hansen and Lunde (2005)

In their paper, 330 GARCH-type models were compared to GARCH(1,1) using the SPA test. The results:

- For IBM equity data: GARCH(1,1) was significantly outperformed (p-value < 0.05).
- For DM/USD exchange rate: GARCH(1,1) was not outperformed (p-value > 0.05).

8 Summary

Let:

- $d_{k,t} = L_t^{(0)} - L_t^{(k)}$: Loss difference.
- $\bar{d}_k = \frac{1}{n}\sum_t d_{k,t}$: Average difference.
- $T_n = \max_k(\sqrt{n} \cdot \bar{d}_k)_+$: SPA statistic.
- T_n^*: Bootstrap replicate.

Then:

$$\text{SPA p-value} = \overset{*}{\Pr}(T_n^* > T_n \mid H_0).$$

Reject H_0 if p-value $< \alpha$.