On a Bootstrap-Based Diebold-Mariano Test for Forecast Evaluations

Marián Vávra
National Bank of Slovakia
E-mail: marian.vavra@nbs.sk

NBS Research Seminar, Bratislava, Slovakia, 2015

Introduction

- Monetary policy should be forward-looking in order to efficiently impact the economy.
- As a result, accurate forecasts of key economic variables (e.g. output and inflation) are of the fundamental importance for central banks (and many other institutions/firms).
- Correct forecast evaluation thus helps to select the perspective methods/models.

Policy questions...

- Can DSGE models beat VARs (or other models)?
- How to forecast the key exogenous variables such as the oil price? (surveys, futures, or AR models?)
- How to forecast the yield curve? (FX reserve management)

Motivation

- Often used approach for forecast evaluations - based on comparing MSFEs from models at hand - is simply useless! (see, e.g., Smets an Wouters (2004); Adolfson, Linde, and Villani (2007); Edge, Kiley, and Laforte (2010)).
- But it is fair to admit that appropriate forecast evaluation - using, for instance, the Diebold-Mariano (DM) test statistic - is by no means easy...

Diebold-Mariano (DM) test

- Basic assumptions: $\{(X_{1,t}, X_{2,t}) : t \in \mathbb{Z}\}$ is a pair of the covariance stationary correlated forecast errors coming from two alternative (non-nested) models.
- The hypothesis: Diebold and Mariano (1995) proposed a conceptually simple statistic for testing the equal forecast accuracy of the errors based on a mean squared error (MSE) measure: $H_0 : \mathbb{E}(X_1^2) = \mathbb{E}(X_2^2)$ against $H_1 : \mathbb{E}(X_1^2) \neq \mathbb{E}(X_2^2)$.
- The test statistic:

$$\mathcal{D} = \sqrt{n}\left(\frac{\bar{d}}{\hat{\sigma}}\right) \xrightarrow{d} N(0,1), \quad \text{as} \quad n \to \infty, \tag{1}$$

$\bar{d} = n^{-1}\sum_{t=1}^{n}(X_{1,t}^2 - X_{2,t}^2)$ stands for a sample average of the loss differentials and the asymptotic variance $\hat{\sigma}^2 = \hat{\gamma}_0 + \sum_{j=1}^{m} w(j/m)\hat{\gamma}_j$, where $w(\cdot)$ are the Bartlett weights, m is a real-valued bandwidth such that $m \to \infty$ and $m/n \to 0$ as $n \to \infty$, and $\hat{\gamma}_j = n^{-1}\sum_{t=j+1}^{n}(d_t - \bar{d})(d_{t-j} - \bar{d})$.

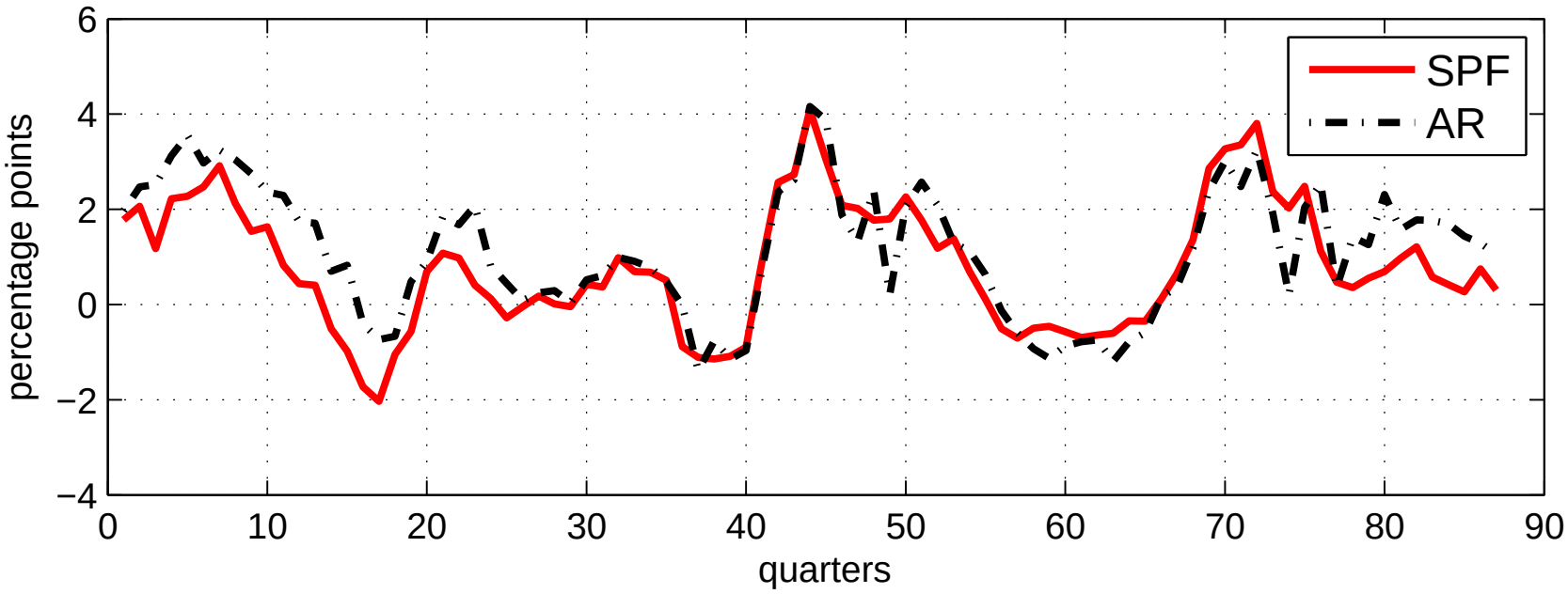
Complications with the DM test statistic

- The finite sample properties of the DM test are not convincing – the magnitude of a size distortion makes the DM statistic unreliable for empirical applications (see Table 1).
- Why? Although consistency of $\hat{\sigma}^2$ is well established, the quantity is downward biased in "small" samples (e.g. $n < 100$) due to high persistence of the forecast errors (see Figure 1).

The main task of the paper

The main task is to modify the Diebold-Mariano test using an appropriate bootstrap technique.

An example of the 4Q ahead Treasure bill forecast errors



Assumptions about the stochastic process

Assumption 1 We consider a real-valued Wold representation for the bivariate forecast errors $\boldsymbol{x}_t = (X_{1,t}, X_{2,t})'$ given by

$$\boldsymbol{x}_t = \boldsymbol{\mu} + \sum_{j=1}^{\infty} \boldsymbol{\psi}_j \boldsymbol{\epsilon}_{t-j} + \boldsymbol{\epsilon}_t, \quad t \in \mathbb{Z}, \tag{2}$$

where $\boldsymbol{\mu} \in \mathbb{R}^2$ and the error sequence $\{\boldsymbol{\epsilon}_t : t \in \mathbb{Z}\}$ is assumed to be a strictly stationary and ergodic vector of innovations such that $\mathbb{E}(\boldsymbol{\epsilon}_t) = \mathbf{0}$, $\mathbb{E}(\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_t') = \boldsymbol{\Sigma}$, which is a symmetric and positive definite matrix, $\mathbb{E}(||\boldsymbol{\epsilon}_t||^2) < \infty$ and the density function $f(\boldsymbol{\epsilon}_t)$ is absolutely continuous on $\mathbb{R}^2$. Additionally, we assume the spectral density matrix of $\boldsymbol{x}_t$ fulfils the boundedness condition.

Assumptions about the stochastic process

Assumption 2 We consider that the loss differential function $d = g(X_1, X_2)$ is twice continuously differentiable and

$$\frac{\partial^2 d}{\partial X_1 \partial X_2}$$

satisfies a Lipschitz condition.

Note

Under an additional mild assumption the process in (2) can be written into a bivariate VAR(∞) model $\Rightarrow$ a VAR-sieve bootstrap.

VAR-sieve bootstrap

Algorithm 1

(i) Select an appropriate lag order p of a VAR model for a bivariate forecast error vector $\{\boldsymbol{x}_t : t = 1, \ldots, n\}$, using AIC.

(ii) Estimate the unknown VAR parameters by the multivariate least-squares (LS) method.

(iii) Construct a sequence of the estimated residuals $\{\hat{\boldsymbol{\epsilon}}_t : t = p+1, \ldots, n\}$ by the recursion

$$\hat{\boldsymbol{\epsilon}}_t = \boldsymbol{x}_t - \hat{\boldsymbol{c}} - \sum_{j=1}^{p} \hat{\boldsymbol{\phi}}_j \boldsymbol{x}_{t-j}.$$

VAR-sieve bootstrap

Algorithm 1

(iv) Draw a random vector $\{\hat{\boldsymbol{\epsilon}}_t^* : t = 1, \ldots, n + 100\}$ from a bivariate empirical distribution function given by $\hat{F}_n(\boldsymbol{u}) = \frac{1}{n-p} \sum_{t=p+1}^{n} \mathbb{I}(\hat{\boldsymbol{\epsilon}}_t \leq \boldsymbol{u})$, where $\mathbb{I}(\cdot)$ denotes a standard indicator function and $\boldsymbol{u} \in \mathbb{R}^2$.

(v) Generate bootstrap replicates $\{\boldsymbol{x}_t^* : t = 1, \ldots, n + 100\}$ by the recursion

$$\boldsymbol{x}_t^* = \hat{\boldsymbol{c}} + \sum_{j=1}^{p} \hat{\boldsymbol{\phi}}_j \boldsymbol{x}_{t-j}^* + \hat{\boldsymbol{\epsilon}}_t^*.$$

where the process is initiated by a vector of sample averages $\bar{\boldsymbol{x}} = n^{-1} \sum_{t=1}^{n} \boldsymbol{x}_t$: $(\boldsymbol{x}_{-p+1}^*, \ldots, \boldsymbol{x}_0^*) = (\bar{\boldsymbol{x}}, \ldots, \bar{\boldsymbol{x}})$. The first 100 data points are then discarded in order to eliminate start-up effects and the remaining n data points are used.

VAR-sieve bootstrap

Algorithm 1

(vi) Consistently with the null hypothesis (i.e. the equality of mean squared forecast errors: $\mathbb{E}(X_{1,t}^2) = \mathbb{E}(X_{2,t}^2)$), generate the normalized bootstrap vector $\boldsymbol{z}_t^* = (Z_{1,t}^*, Z_{2,t}^*)'$ according to

$$Z_{1,t}^* = X_{1,t}^* \sqrt{(\omega_1^2 + \omega_2^2)/2\omega_1^2},$$
$$Z_{2,t}^* = X_{2,t}^* \sqrt{(\omega_1^2 + \omega_2^2)/2\omega_2^2},$$

where $\omega_i^2 = n^{-1} \sum_{t=1}^{n} X_{i,t}^2$ denotes the sample second raw moment.

(vii) Construct a bootstrap analogy of the DM test statistic $\mathcal{B}^*$ calculated from the normalized bootstrap samples $\{\boldsymbol{z}_t^* : t = 1, \ldots, n\}$.

VAR-sieve bootstrap

Algorithm 1

(viii) Repeat steps (iv)–(vi) independently B times to get a sample of the bootstrap DM statistics $\{\mathcal{B}_j^* : j = 1, \ldots, B\}$. Then, the sampling distribution of the $\mathcal{B}$ test statistic is approximated by the empirical distribution function associated with $\{\mathcal{B}_j^* : j = 1, \ldots, B\}$: $H^*(u) = B^{-1} \sum_{j=1}^{B} I(\mathcal{B}_j^* \leq u)$, where $u \in \mathbb{R}$. Finally, a bootstrap test of the nominal level α rejects the null hypothesis if

$$|\mathcal{B}| > \inf\{u : H^*(u) \geq (1 - \alpha/2)\},$$

where $\mathcal{B}$ is the DM test statistic obtained from the observed samples $\{\boldsymbol{x}_t : t = 1, \ldots, n\}$.

Monte Carlo setup

The finite-sample properties of the DM and BDM tests are assessed using the following DGPs:

$$X_{i,t} = c_i + \phi_i X_{i,t-1} + \kappa_i \epsilon_{i,t}, \quad \text{for} \quad i \in \{1, 2\}. \tag{3}$$

The configuration of individual parameters is as follows:

M1: $c_1 = c_2 = 0.2$, $\phi_1 = \phi_2 = 0.5$, $\kappa_1 = \kappa_2 = 1.0$;

M2: $c_1 = c_2 = 0.2$, $\phi_1 = \phi_2 = 0.8$, $\kappa_1 = \kappa_2 = 1.0$;

M3: $c_1 = 0.4$, $c_2 = 0.2$, $\phi_1 = \phi_2 = 0.8$, $\kappa_1 = \kappa_2 = 1.0$;

M4: $c_1 = c_2 = 0.2$, $\phi_1 = 0.8$, $\phi_2 = 0.5$, $\kappa_1 = \kappa_2 = 1.0$;

M5: $c_1 = c_2 = 0.2$, $\phi_1 = \phi_2 = 0.8$, $\kappa_1 = \sqrt{2.0}$, $\kappa_2 = 1.0$;

Two values of the pairwise correlation between model innovations $\rho = \mathrm{Corr}(\epsilon_{1,t}, \epsilon_{2,t}) \in \{0.25, 0.75\}$ are considered for the Monte Carlo experiments.

Monte Carlo results

Table : Rejection frequencies of the BDM and DM test statistics at 0.10 nominal level

	$n = 50$				$n = 100$				$n = 200$			
	$\rho = 0.25$		$\rho = 0.75$		$\rho = 0.25$		$\rho = 0.75$		$\rho = 0.25$		$\rho = 0.75$	
DGP	$\mathcal{B}$	$\mathcal{D}$	$\mathcal{B}$	$\mathcal{D}$	$\mathcal{B}$	$\mathcal{D}$	$\mathcal{B}$	$\mathcal{D}$	$\mathcal{B}$	$\mathcal{D}$	$\mathcal{B}$	$\mathcal{D}$
M1	0.08	0.19	0.07	0.16	0.10	0.16	0.09	0.15	0.09	0.11	0.10	0.14
M2	0.10	0.32	0.09	0.29	0.09	0.24	0.10	0.25	0.09	0.19	0.11	0.21
M3	0.22	0.53	0.39	0.68	0.38	0.60	0.61	0.82	0.59	0.76	0.88	0.95
M4	0.30	0.64	0.46	0.80	0.64	0.87	0.78	0.96	0.93	0.98	0.99	1.00
M5	0.20	0.47	0.32	0.61	0.32	0.54	0.57	0.76	0.53	0.70	0.82	0.92

Can professionals beat AR models?

- We test the null hypothesis $H_0 : MSFE(SPF) = MSFE(AR)$ against $H_0 : MSFE(SPF) \neq MSFE(AR)$;
- The following set of economic variables is considered – the 3-month Treasury Bill rate (TBILL), the AAA Corporate Bond yield (AAA), the real Gross Domestic Product growth rate (RGDP), the GDP deflator growth rate (PGDP), the Industrial Production growth rate (IP), the Unemployment rate (UR), and the Housing Starts (HOUS);
- Forecasts over 1 and 4 quarters ahead for each variable;
- Due to an institutional break in the SPF survey – the Federal Reserve Bank of Philadelphia took over the survey from the National Bureau of Economic Research in 1990 – we conduct our analysis over two different sub-periods: (i) 1983 Q1 – 2012 Q3 (i.e 119 obs.); and (ii) 1991 Q1 – 2012 Q3 (i.e. 87 obs.).

Can professionals beat AR models?

Table : *P*-values of the BDM and DM Test Statistics

		1983 Q1 – 2012 Q3			1991 Q1 – 2012 Q3		
variables	horizon	$\mathcal{D}$	$\mathcal{B}$	result($\mathcal{B}$)	$\mathcal{D}$	$\mathcal{B}$	result($\mathcal{B}$)
AAA	1	0.056	0.486		0.011	0.031	SPF
	4	0.166	0.723		0.232	0.396	
TBILL	1	0.012	0.089	SPF	0.008	0.026	SPF
	4	0.151	0.330		0.084	0.172	
PGDP	1	0.003	0.042	SPF	0.077	0.116	
	4	0.027	0.146		0.232	0.494	
GDP	1	0.076	0.132		0.150	0.216	
	4	0.230	0.307		0.807	0.841	
IP	1	0.067	0.211		0.177	0.266	
	4	0.090	0.150		0.241	0.372	
UR	1	0.001	0.012	SPF	0.015	0.030	SPF
	4	0.007	0.029	SPF	0.064	0.145	
HOUS	1	0.020	0.089	SPF	0.052	0.188	
	4	0.030	0.117		0.047	0.224	

What next...

- A short paper analyzing financial variables from the Consensus Forecast dataset (joint work with Peter Tóth (NBS)).
- A multivariate extension of the BDM test and comparison of the forecast accuracy of DSGE and VAR models (joint work with Ron Smith, Zacharias Psaradakis (both UoL), and Stanislav Tvrz (NBS)).

References

Diebold, F. and R. Mariano (1995). Comparing predictive accuracy. *Journal of Business & economic statistics 13*.

Thanks

Thank you for attention.