Large-Scale Multiple Testing without Data Snooping Bias: Methods and Applications

CHUNG-MING KUAN

Department of Finance & CRETA

National Taiwan University

May 2, 2011

Lecture Outline

Lecture Outline (cont'd)

5. Empirical Study I
   - Technical Rules and Performance Measures
   - Empirical Results
6. Empirical Study II
7. Improved Stepwise Procedure
   - Single-Step Control of k-FWER
   - Step-Down Control of k-FWER
   - The Step-SPA(k) Test
   - Asymptotic Properties
8. Simulations
9. Concluding Remarks

Data Snooping

- In economics and finance, it is common to test if a "new" model for some target variable (e.g., inflation or index return) has superior performance than the benchmark.
  - This is a multiple testing problem because many other related models have also been tested before.
  - There may be data snooping bias when those models are evaluated using the same data set and when the test results are ignored (Lo and MacKinglay, 1990; Brock, Lakonishok, and LeBaron, 1992).
- Data snooping is mainly due to data re-use; ignoring data snooping bias may yield very misleading conclusions.

Example 1: Predictive Power of Technical Trading Rules

The predictive power of technical analysis has been a long-debated issue in both industry and academia since Fama and Blume (1966). Recent supporting evidence includes Sweeney (1988), Blume, Easley, and O'Hara (1994), Neely, Weller, and Dittmar (1997), Brown, Goetzmann, and Kumar (1998), Gencay (1998), Lo, Mamaysky, and Wang (2000), and Savin, Weller, and Zvingelis (2007), among others.

> *"given enough computer time, we are sure that we can find a mechanical trading rule which 'works' on a table of random numbers ..." (Jensen and Benington,* J. of Finance, **25**, *p. 470).*

- Q1: Is the predictive ability of these rules real or due to chance?
- Q2: If real, what are they?

Example 1: Predictive Power of Technical Trading Rules

The predictive power of technical analysis has been a long-debated issue in both industry and academia since Fama and Blume (1966). Recent supporting evidence includes Sweeney (1988), Blume, Easley, and O'Hara (1994), Neely, Weller, and Dittmar (1997), Brown, Goetzmann, and Kumar (1998), Gencay (1998), Lo, Mamaysky, and Wang (2000), and Savin, Weller, and Zvingelis (2007), among others.

> *"given enough computer time, we are sure that we can find a mechanical trading rule which 'works' on a table of random numbers ..." (Jensen and Benington,* J. of Finance, **25**, *p. 470).*

- Q1: Is the predictive ability of these rules real or due to chance?
- Q2: If real, what are they?

Example 2: Performance of Mutual Funds

Mutual (hedge) funds are usually evaluated based on their performance relative to an index, in terms of mean returns and/or Sharpe ratios. For example, among 220 mutual funds in Taiwan, there are 153 funds with mean monthly returns higher than that of Taipei Weighted Index during 2002–2007. Also, 134 funds have higher Sharpe ratios during the same period.

- Q1: Did those funds really beat the market? Is the "superior" performance real or due to chance?
- Q2: If real, what are they?

Example 3: Predictive Ability of Term Spreads

It has been found that the term spreads between some short- and long-term interest rates have predictive ability for real GDP growth, e.g., Laurent (1988, 1989), Stock and Watson (1989), Estrella and Hardouvelis (1991), Estrella and Mishkin (1998), Hamilton and Kim (2002), Mody and Taylor (2004), and Bordo and Haubrich (2008b). Ang et al. (2006) find that models with certain short rates suffice, yet Bordo and Haubrich (2008b) show that combination of short rates and term spreads provides superior predictive power.

- Q1: Is the "superior" predictive power of a term spread model real or due to chance?
- Q2: If real, what are they?

Solutions to Data Snooping

1. Data approach: Testing different but comparable data sets (e.g. Lakonishok, Shleifer, and Vishny, 1994); validating a result using sub-samples (e.g. Brock, Lakonishok, and LeBaron, 1992).
2. Testing procedures:
   - Individual tests with the significance level controlled by the Bonferroni inequality (e.g. Lakonishok and Smidt, 1988).
   - One-step tests: Reality Check (RC) of White (2000); Superior Predictive Ability (SPA) test of Hansen (2005)
   - Stepwise tests: Step-RC of Romano and Wolf (2005); Step-SPA test of Hsu, Hsu, and Kuan (2010).
   - Generalization based on generalized familywise error rate: Lehmann and Romano (2005), Romano and Shaikh (2006a, b), Romano and Wolf (2007), and Donald, Hsu, and Kuan (2010).

Applications of Multiple Tests

1. Technical trading rule performance: Sullivan, Timmermann, and White (1999), White (2000), Hsu and Kuan (2005), Qi and Wu (2006), Hsu, Hsu, and Kuan (2010).
2. Calendar effect in stock returns: Sullivan, Timmermann, and White (2001), Hansen, Lunde, and Nason (2004), Coakley, Marzano, and Nankervis (2010).
3. Model comparison: Hansen (2005), Hansen and Lunde (2005), Kao, Kuan, and Chen (2010).
4. Fund performance: Romano and Wolf (2005), Chuang and Kuan (2010), Yen and Hsu (2010).

A Multiple Testing Problem

$d_{k,t}$, $k = 1, \ldots, m$ and $t = 1, \ldots, n$, are the performance measures (relative to a benchmark) of the k-th model at time t.

- For each k, $\mathbb{E}(d_{k,t}) = \mu_k$ for all t; for each t, $d_{k,t}$ may be dependent across k.
- Example: For a given asset with return r_t, let $d_{k,t} = \delta_{k,t-1} r_t$ denote its realized return based on the k-th trading rule, where $\delta_{k,t-1}$ is the trading signal of the k-th rule. Clearly, $d_{k,t}$ involve the same r_t and hence are dependent across k.

Null: No model has positive performance measure

$$H_0^k : \mu_k \leq 0, \quad k = 1, ..., m.$$

Individual Tests

- Letting $\mathcal{I}$ be the set of indices of true hypotheses,

$$\text{familywise error rate (FWER)} = \mathbb{P}(\text{Reject at least one } H_0^k,\ k \in \mathcal{I}).$$

- Assuming independence among the tests of testing H_0^k at 5% level, the FWER of is:

m	1	2	5	10	50
FWER	.05	.10	.23	.40	.92

- Bonferroni: Setting each significance level to α/m we have

$$\text{FWER} \le \sum_{k \in \mathcal{I}} \mathbb{P}(\text{Reject } H_0^k) \le \sum_{k \in \mathcal{I}} \frac{\alpha}{m} \le \alpha.$$

These inequalities are very loose and hence yield a very conservative test. This method is not practically useful when m is large.

Joint Test

- We may construct a joint test of $\boldsymbol{\mu}$ based on the asymptotic normality:

$$\sqrt{n}(\bar{\mathbf{d}}_n - \boldsymbol{\mu}) \xrightarrow{D} \mathcal{N}(\mathbf{0},\ \boldsymbol{\Omega}),$$

  where $\bar{\mathbf{d}}_n = n^{-1}\sum_{t=1}^{n} \mathbf{d}_t$ with $\mathbf{d}_t = (d_{1,t}, \ldots, d_{m,t})^\top$, $\boldsymbol{\mu} = \mathbb{E}(\mathbf{d}_t)$, and $\boldsymbol{\Omega}$ is the asymptotic covariance matrix.

- More difficulties:
  - Implementing this test is not easy when m is large. For example, consistent estimation of $\boldsymbol{\Omega}$ for a large m would be practically cumbersome.
  - It is not clear how the null distribution should be determined under inequality hypothesis.

Multiple Testing vs. Joint Testing

Multiple testing is concerned with drawing individual inferences about m hypotheses (considering equality hypothesis for now):

$$H_0^k : \mu_k = 0, \quad \text{vs.} \quad H_a^k : \mu_k \neq 0, \ \text{for } k = 1, ..., m.$$

Join testing is concerned with testing the single hypothesis:

$$H_0 : \mu_k = 0 \ \forall k, \quad \text{vs.} \quad H_a : \mu_k \neq 0 \ \text{ for some } \ k.$$

One may conduct multiple testing based on a joint test.

- As shown in Romano and Wolf (2005), a join test for a multiple testing problem is sub-optimal.
- A rejection of the joint hypothesis H_0 does not necessarily lead to the rejection of one of the individual hypotheses H_0^k.

Least Favorable Configuration

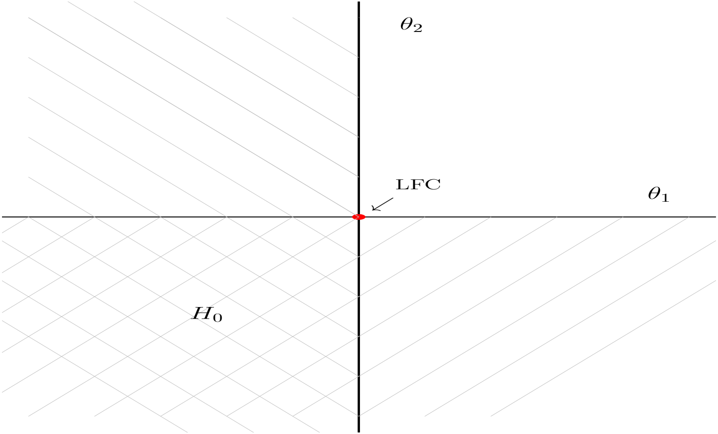

Figure: The configuration under the null ($\theta_1 \leq 0$ nad $\theta_2 \leq 0$) that is least favorable to the alternative.

White's Reality Check

- White's RC determines the null distribution by the LFC: $\boldsymbol{\mu} = \mathbf{0}$:

$$\mathrm{RC}_n = \max_{k=1,\ldots,m} \sqrt{n}\bar{d}_k \xrightarrow{D} \max_{i=1,\ldots,m} Z_i,$$

where $(Z_1, \ldots, Z_m) \sim \mathcal{N}(\mathbf{0}, \ \boldsymbol{\Omega})$.

- The inference is based on the bootstrapped null distribution:

$$\max_{k=1,\ldots,m} \sqrt{n}\big(\bar{d}_k^*(b) - \bar{d}_k\big), \quad b = 1, \ldots, B,$$

where $\bar{d}_k^*(b)$ is the sample average of the b-th bootstrapped sample. *Note*: Bootstrap $\mathbf{d}_t = (d_{1,t}, d_{2,t}, \ldots, d_{m,t})^\top$, $t = 1, \ldots, n$, to preserve dependence across models.

Drawbacks of Reality Check

- When LFC fails (some $\mu_j = 0$ and some $\mu_i < 0$), $\mathrm{RC}_n \xrightarrow{D} \max\{\mathcal{N}(\mathbf{0},\ \mathbf{\Omega}_0)\}$, which does not depend on the "poor" models with negative mean. Yet, this distribution is stochastically dominated by the distribution under LFC: $\max\{\mathcal{N}(\mathbf{0},\ \mathbf{\Omega})\}$.
- The power of RC is thus adversely affected because the bootstrapped p-value is artificially increased (or the bootstrapped critical value is larger than it should be).
- The power of RC deteriorates when more models with $\mu < 0$ are included in the test (power could be driven to zero by including many poor models).

Hansen's SPA Test

Hansen (2005): $\text{SPA}_n = \max\big(\max_{k=1,\ldots,m} \sqrt{n}\bar{d}_k,\ 0\big)$, with a re-centered, bootstrapped distribution:

$$\max_{k=1,\ldots,m} \sqrt{n}\big(\bar{d}_k^*(b) - \bar{d}_k + \hat{\mu}_k\big), \quad b = 1, \ldots, B,$$

where $\hat{\mu}_k = \bar{d}_k \mathbf{1}\big(\sqrt{n}\bar{d}_k \leq A_{n,k}\big)$ and $A_{n,k} = -\hat{\sigma}_k \sqrt{2 \log \log n}$.

- When $\mu_k = 0$, $\hat{\mu}_k = 0$ almost surely.
- When $\mu_k < 0$, $n^{1/2}\bar{d}_k \leq A_{n,k}$ with probability approaching one, so that $\hat{\mu}_k \xrightarrow{\mathbb{P}} \mu_k$.
- We may replace $\log \log n$ by some $a_n \to \infty$ and $a_n/n \to 0$.

Note: Re-centering leads to a better approximation to $\max\{\mathcal{N}(\mathbf{0},\ \mathbf{\Omega}_0)\}$ and hence a more powerful test.

Digression: Holm's Procedure

- Step-down procedure: Set $\alpha_1 \le \alpha_2 \le \cdots \le \alpha_m$ and denote the ordered p-values for individual tests as $\hat{p}_{(1)} \le \hat{p}_{(2)} \le \cdots \le \hat{p}_{(m)}$.
  - If $\hat{p}_{(1)} > \alpha_1$, no hypothesis is rejected; otherwise reject $H_{(1)}$.
  - If $\hat{p}_{(1)} \le \alpha_1, \ldots, \hat{p}_{(r)} \le \alpha_r$, reject $H_0^{(1)}, \ldots, H_0^{(r)}$.
- Holm (1979) and its generalization:
  - Control FWER $\le \alpha$ by setting $\alpha_j = \alpha/(m-j+1)$.
  - Control k-FWER $\le \alpha$ by setting

$$\alpha_j = \begin{cases} k\alpha/m, & \text{if } j \le k; \\ k\alpha/(m-j+k), & \text{otherwise.} \end{cases}$$

- In Holm's procedure, α_j increases with j, so that the test is able to reject more hypothesis than does the Bonferroni method where $\alpha_j = \alpha/m$ or $\alpha_j = k\alpha/m$ is a constant.

Stepwise RC and SPA Tests

Romano and Wolf (2005) and Hsu et al. (2010): Identify as many outperforming models as possible using a stepwise procedure, while controlling

$$\text{FWER} = \mathbb{P}_{H_0}(\text{Reject at least one } H_0^i,\ i \in \mathcal{I}),$$

where $\mathcal{I}$ is the set of indices of true hypotheses.

1. Reject model k when $n^{1/2}\bar{d}_k$ is greater than the bootstrapped RC (or SPA) critical value.
2. Remove $\bar{d}_k$ of the rejected models from the data and re-bootstrap the critical value using the remaining data.
3. Repeating (1) and (2) based on the newly bootstrapped critical value.

The procedure stops when no model can be rejected.

Asymptotic Properties

Step-SPA Test (Hsu, Hsu, and Kuan, 2010)

1. (Exact FWER) Given level α_0, the Step-SPA test has FWER $= \alpha_0$ when n tends to infinity if and only if there is at least one $\mu_k = 0$.
2. (Consistency) H_0^k with $\mu_k > 0$ will be rejected by the Step-SPA test with probability approaching 1 when n tends to infinity.
3. (Power) The Step-SPA test is more powerful than the Step-RC test under the notions of power defined in Romano and Wolf (2005).

- The FWER of the Step-RC test $\leq \alpha_0$.
- If there is no $\mu_k = 0$, the FWER would be zero asymptotically, so that no null hypothesis will be incorrectly rejected.

The power notions in Romano and Wolf (2005):

- Minimal power: Probability of rejecting at least one false null hypothesis.
- Global power: Probability of rejecting all false null hypotheses.
- Average power: The average of the individual probabilities of rejecting each false null hypothesis. This is equivalent to the expected number of false hypotheses that will be rejected.
- The expected proportion of false hypotheses that will be rejected.
- The probability of rejecting at least $\gamma \times 100\%$ of the false null hypotheses, where γ is a user-chosen parameter.

Simulations

Returns: $x_{i,t} = c_i + \gamma x_{i,t-1} + \epsilon_{i,t}$, $i = 1, \ldots, m$ and $t = 1, \ldots, n$, where $\epsilon_{i,t}$ are i.i.d. $\mathcal{N}(0, \sigma^2)$, $c_i = a, 0, -a$ for, respectively, $i = 1, \ldots, m_1$, $i = m_1 + 1, \ldots, m_1 + m_2$, and $c_i = -a$ for $i = m_1 + m_2 + 1, \ldots, m$, $\bar{d}_k = \bar{x}_k$, the average of the k-th return series.

- $a = 0.0008$ (8 basis points), $\gamma = 0.01$, and $\sigma = 0.005$.
- m_1 "outperforming" returns with positive mean 0.00081;
  m_2 "neutral" returns with a zero mean;
  $m - m_1 - m_2$ "poor" returns with a negative mean -0.00081.
- $m = 90, 900$ and 9,000; for each m, we consider $m_1 = m_2 = m/3$ (equal groups) and $m_1 = m_2 = m/9$ (unequal groups).
- $n = 1000$, $R = 500$, $B = 500$, $Q = 0.9$.

Table 1: Average rejection rates and FWERs of studentized tests.

	Equal group			Unequal group			All neutral
Test	AR rate (1-step)	AR rate (all-steps)	FWE rate	AR rate (1-step)	AR rate (all-steps)	FWE rate	FWE rate
	30 outperforming + 30 neutral + 30 poor			10 outperforming + 10 neutral + 70 poor			90 neutral
Step-SPA	96.6	98.0	4.8	98.8	99.4	4.8	3.2
Step-RC	95.3	96.4	2.2	95.4	95.6	0.8	3.2
	300 outperforming + 300 neutral + 300 poor			100 outperforming + 100 neutral + 700 poor			900 neutral
Step-SPA	86.1	89.4	3.0	92.0	94.3	3.4	1.8
Step-RC	83.8	85.9	1.2	84.0	84.6	0.4	1.8
	3000 outperforming + 3000 neutral + 3000 poor			1000 outperforming + 1000 neutral + 7000 poor			9000 neutral
Step-SPA	68.8	72.6	2.0	78.6	82.5	1.6	1.2
Step-RC	65.2	67.8	1.0	65.0	65.6	0.6	1.2

A Summary

- The stepwise procedure does identify more outperforming returns than the corresponding one-step test.
- In terms of the average rejection rate and FWER, the Step-SPA test performs better than the Step-RC test when there are "poor" models.
- The improvement of the Step-SPA test on the Step-RC test is more obvious when there are more models that have negative return (the case of unequal groups).
- The studentized tests perform slightly better than the non-studentized counterparts.

Empirical Study I: Trading Rule Performance

Hsu, Hsu, and Kuan (2010, JEF)

- We evaluate the predictive ability of technical trading rules based on the data of market indices and corresponding ETFs.
- ETFs have been powerful investment tools for arbitrageurs and hedge funds because they track market indices closely and can be conveniently traded at low transaction costs.
- Data: Global insight and Yahoo Finance
  - U.S. growth markets: S&P SmallCap 600/Citigroup Growth Index (SP600SG), Russell 2000 Index (RUT2000), NASDAQ Composite Index (NASDAQ), and the ETFs that track these indices.
  - Emerging markets: MSCI Emerging Markets Index, MSCI Brazil Index, MSCI South Korea Index, MSCI Malaysia Index, MSCI Mexico Index, MSCI Taiwan Index, and their ETFs.

Table 2: The pre- and post-ETF periods.

Market	Identifier	Index	Pre-ETF Period	Obs.	ETF Incept. date
U.S.	SP600SG	S&P SmallCap 600/Citigroup Growth Index	1/4/1989 – 12/31/1999	2779	July 24, 2000
Markets	RUT2000	Russell 2000 Index	1/3/1990 – 12/31/1999	2527	March 1, 2000
	NASDAQ	NASDAQ Composite Index	1/3/1990 – 12/31/1998	2275	Sept. 25, 2003
	Emerging	MSCI Emerging Markets Index	1/4/1993 – 12/31/2002	2601	April 7, 2003
	Brazil	MSCI Brazil Index	1/1/1990 – 12/31/1999	2610	July 10, 2000
Emerging	Korea	MSCI South Korea Index	1/2/1990 – 12/31/1999	2865	May 9, 2000
Markets	Malaysia	MSCI Malaysia Index	1/1/1988 – 12/29/1995	2086	March 12, 1996
	Mexico	MSCI Mexico Index	1/1/1988 – 12/29/1995	2086	March 12, 1996
	Taiwan	MSCI Taiwan Index	1/1/1990 – 12/31/1999	2610	June 20, 2000

Market	Ticker	ETF	Post-ETF Period	Obs.	ETF Incept. date
U.S.	IJT	S&P SmallCap 600 Growth Index Fund	7/28/2000 – 12/30/2005	1364	July 24, 2000
Markets	IWM	Russell 2000 Index Fund	5/30/2000 – 12/30/2005	1406	March 1, 2000
	ONEQ	NASDAQ Composite Index Tracking Fund	10/1/2003 – 12/30/2005	568	Sept. 25, 2003
	EEM	MSCI Emerging Markets Index Fund	10/2/2003 – 12/30/2005	566	April 7, 2003
	EWZ	MSCI Brazil Index Fund	7/14/2000 – 12/30/2005	1368	July 10, 2000
Emerging	EWY	MSCI South Korea Index Fund	6/1/2000 – 12/30/2005	1401	May 9, 2000
Markets	EWM	MSCI Malaysia Index Fund	4/1/1996 – 12/30/2005	2453	March 12, 1996
	EWW	MSCI Mexico Index Fund	4/1/1996 – 12/30/2005	2453	March 12, 1996
	EWT	MSCI Taiwan Index FUnd	6/26/2000 – 12/30/2005	1384	June 20, 2000

Technical Rules and Performance Measures

- There is a total of 16,380 rules: 9,120 moving averages (MA) rules and 7,260 filter rules (FR). These rules encompass 2,049 MA rules and 497 filter rules used in Brock et al. (1992) and Sullivan et al. (1999).
- The trading signals are generated from the technical rules operated on market indices.
- We evaluate whether technical rules have predictive power and, if they do, whether this power is affected by the introduction of ETF.
- Performance measures: Mean return, Sharpe ratio, x-statistic of Sweeney (1986, 1988) which is mean return adjusted for a proportion of risk premium, and studentized mean return. These measures take into account the risk-free rate and transaction cost.

Table 3: The numbers of outperforming rules in pre- and post-ETF periods.

Market	Index/ETF	Period	Outperforming rules			
			Mean return	Sharpe ratio	x-statistic	St. mean ret.
U.S. Indices	S&P600SG	pre-ETF	269	136	220	230
	RUT2000	pre-ETF	186	109	179	171
	NASDAQ	pre-ETF	33	1	5	7
U.S. ETFs	IJT	post-ETF	0	0	0	0
	IWM	post-ETF	0	0	0	0
	ONEQ	post-ETF	0	0	0	0
Emerging Market Indices	Emerging	pre-ETF	797	414	917	758
	Brazil	pre-ETF	117	88	0	143
	Korea	pre-ETF	0	0	0	0
	Malaysia	pre-ETF	81	2	70	68
	Mexico	pre-ETF	559	370	331	490
	Taiwan	pre-ETF	0	0	0	0
Emerging Market ETFs	EEM	post-ETF	0	0	0	0
	EWZ	post-ETF	0	0	0	0
	EWY	post-ETF	0	0	0	0
	EWM	post-ETF	55	0	66	0
	EWW	post-ETF	241	152	285	198
	EWT	post-ETF	0	0	0	0

Empirical Results

- The introduction of ETFs affects predictability.
  - US Markets: Technical rules are quite powerful in predicting U.S. indices in pre-ETF periods but not in post-ETF periods.
  - Emerging Markets: There are significant rules for 4 out of 6 emerging market indices in pre-ETF periods but only 2 in post-ETF periods.
- There are "thick" sets of outperforming rules which are strong evidence for return predictability (Timmermann and Granger, 2004).
- The predictive power is not a consequence of serial correlation in data.
  - The Step-SPA test identifies significant rules for MSCI Malaysia and Mexico Index Funds whose returns are serially uncorrelated.
  - The Step-SPA test does not find any outperforming rules for MSCI Taiwan Index Fund which has significant first-order autocorrelation.

Table 4: The identified best rules and their break-even transaction costs.

Market	Index/ETF	Best rule and break-even transaction cost								
		Best rule	Mean (p-value)	Break-even cost (bps)	Best rule	Sharpe (p-value)	Break-even cost (bps)	Best rule	x-stat (p-value)	Break-even cost (bps)
U.S. Indices	S&P600SG	MA	.16 (.00)	17	MA	.17 (.01)	12	MA	.15 (.00)	17
	RUT2000	MA	.15 (.00)	15	MA	.17 (.00)	11	MA	.14 (.00)	15
	NASDAQ	MA	.13 (.00)	13	MA	.11 (.00)	7	MA	.11 (.00)	9
U.S. ETFs	IJT	MA	.06 (.99)	N/A	FR	.06 (.94)	N/A	MA	.06 (.82)	N/A
	IWM	MA	.06 (.97)	N/A	FR	.05 (.96)	N/A	MA	.06 (.86)	N/A
	ONEQ	FR	.09 (.65)	N/A	FR	.09 (.89)	N/A	FR	.09 (.37)	N/A
Emerging Market Indices	Emerging	MA	.22 (.00)	27	MA	.19 (.00)	22	MA	.22 (.00)	28
	Brazil	FR	.30 (.00)	16	FR	.10 (.01)	12	FR	.18 (.58)	N/A
	Korea	MA	.14 (.38)	N/A	FR	.05 (.84)	N/A	MA	.14 (.37)	N/A
	Malaysia	MA	.15 (.00)	10	FR	.11 (.03)	6	MA	.15 (.00)	10
	Mexico	FR	.25 (.00)	28	FR	.13 (.00)	17	FR	.24 (.00)	22
	Taiwan	MA	.09 (.66)	N/A	FR	.05 (.75)	N/A	MA	.09 (.40)	N/A
Emerging Market ETFs	EEM	MA	.06 (.86)	N/A	FR	.08 (.88)	N/A	MA	.06 (.81)	N/A
	EWZ	FR	.17 (.44)	N/A	FR	.08 (.35)	N/A	FR	.17 (.32)	N/A
	EWY	FR	.14 (.77)	N/A	FR	.06 (.90)	N/A	FR	.13 (.42)	N/A
	EWM	FR	.21 (.04)	7	FR	.08 (.19)	N/A	FR	.21 (.03)	7
	EWW	FR	.24 (.00)	21	FR	.13 (.00)	14	FR	.24 (.00)	20
	EWT	MA	.09 (.99)	N/A	FR	.04 (1.00)	N/A	MA	.09 (.72)	N/A

Discussions

- Q: Why can technical rules predict the stock markets?
  - Due to serial correlations in the data (e.g., Fama and Blume, 1966).
  - Technical rules capture some information contained in the movements of prices, volumes, and order flows (Treynor and Ferguson, 1985; Brown and Jennings, 1990; Blume, Easley, and O'Hara, 1994; Kavajecz and Odders-White, 2004)
  - Market maturity matters (Ready, 2002; Hsu and Kuan, 2005); our results support this explanation.
- Q: Can the predictive power be transformed to profit?
  A: With good execution and low transaction cost, the potential profits from outperforming rules may exceed associated risk premia.

Empirical Study II: Mutual Fund Performance

Chuang and Kuan (2010)

- Data: 220 mutual funds; 60 monthly data of 2002.11–2007.10
- Benchmarks: Taipei Weighted Index, MSCI Taiwan, TW50
- Performance measures: Mean return, Sharpe ratio, abnormal return (3-factor model)

Table 5: Number of funds that significantly outperform the benchmark

	Weighted Index			MSCI Taiwan			TW50		
Criterion	Num	t test	S-SPA	Num	t test	S-SPA	Num	t test	S-SPA
Mean	153	11	0	166	28	0	159	23	0
Sharpe	134	15	1	176	24	1	154	16	1
Abnormal	176	46	3	178	47	3	147	29	2

Generalized Familywise Error Rate

FWER is a stringent criterion because it is defined on one false rejection. If one is willing to tolerate more incorrect rejections, the resulting test would be able to identify more superior models (have better power).

- Lehmann and Romano (2005): For a given k, we control

$$k\text{-FWER} = \mathbb{P}_{H_0}(\text{Reject at least } k \text{ of } H_0^i,\ i \in \mathcal{I}).$$

- Instead of allowing for a fixed number of false rejections, one may allow for more false rejections by keeping the false discovery proportion (FDP) constant:

$$\text{FDP} = (\text{number of false rejections})/(\text{number of total rejections}).$$

For a given $\gamma \in (0, 1)$, one controls $\mathbb{P}(\text{FDP} > \gamma) < \alpha_o$.

Single-Step Control of k-FWER

- For $S \subset \mathcal{A} = \{1, ..., m\}$, define the following $(1-\alpha)$-th quantile:

$$c_S(1-\alpha, k, \mathbb{P}) := \inf\left\{ q : \ \mathbb{P}\left(k\text{-}\max_{i \in S} \sqrt{n}(\bar{d}_i - \mu_i) \leq q \right) \geq 1-\alpha \right\},$$

  where k-max denotes the k-th largest value.

- Setting $S = \mathcal{I}$, reject H_0^i if $\sqrt{n}\bar{d}_i > c_{\mathcal{I}}(1-\alpha, k, \mathbb{P})$, so that

$$k\text{-FWER} = \mathbb{P}\left\{ k\text{-}\max_{i \in \mathcal{I}} \sqrt{n}(\bar{d}_i - \mu_i) > c_{\mathcal{I}}(1-\alpha, k, \mathbb{P}) \right\} \leq \alpha,$$

  where $\mathcal{I}$ is the set of indices of true hypothesis.

- Replacing unknown $\mathcal{I}$ and $\mathbb{P}$ by, resp., $\mathcal{A}$ and the bootstrapped $\hat{\mathbb{P}}_n$, we reject H_0^i if $\sqrt{n}\bar{d}_i > c_{\mathcal{A}}(1-\alpha, k, \hat{\mathbb{P}}_n)$.

- All individual tests adopt the same criterion which is conservative because, as $\mathcal{I} \subset \mathcal{A}$, $c_{\mathcal{I}} \leq c_{\mathcal{A}}$.

Step-Down Control of k-FWER

Romano and Wolf (2007) (cf. Holm, 1979):

1. Let $\mathcal{A}_1 = \{1, \ldots, m\}$. Reject H_0^i if

$$\sqrt{n}\bar{d}_i > c_{\mathcal{A}_1}(1-\alpha, k, \hat{\mathbb{P}}_n).$$

2. Let $\mathcal{R}_1 = \{i : H_0^i \text{ is rejected at stage } 1\}$ and $\mathcal{A}_2 = \mathcal{A}_1 \backslash \mathcal{R}_1$.
   - The procedure stops if $|\mathcal{R}_1| < k$.
   - Reject H_0^i, $i \in \mathcal{A}_2$, if $\sqrt{n}\bar{d}_i > \hat{c}_{\mathcal{A}_2}(1-\alpha, k)$, where

$$\hat{c}_{\mathcal{A}_2}(1-\alpha, k) = \max_{M_1}\big\{c_{\mathcal{K}}(1-\alpha, k, \hat{\mathbb{P}}_n) : \mathcal{K} = M_1 \cup \mathcal{A}_2\big\},$$

with M_1 any subset of $\mathcal{R}_1$ such that $|M_1| = k-1$, i.e., a set of $k-1$ hypotheses that have been rejected at stage 1.

- Let $\mathcal{R}_j = \{i : H_0^i \text{ is rejected at stage } j\}$ and $\mathcal{A}_{j+1} = \mathcal{A}_j \backslash \mathcal{R}_j$, $j = 2, 3, \ldots$
  - The procedure stops if $|\mathcal{R}_j| < k$.
  - We reject H_0^i, $i \in \mathcal{A}_{j+1}$, if $\sqrt{n}\bar{d}_i > \hat{c}_{\mathcal{A}_{j+1}}(1-\alpha, k)$, where

$$\hat{c}_{\mathcal{A}_{j+1}}(1-\alpha, k) = \max_{M_j} \big\{ c_{\mathcal{K}}(1-\alpha, k, \hat{\mathbb{P}}_n) : \mathcal{K} = M_j \cup \mathcal{A}_{j+1} \big\},$$

with M_j any subset of $\cup_{i=1}^{j} \mathcal{R}_i$ such that $|M_j| = k-1$.

Note: When $k = 1$, Step-SPA(k) simply reduces to the Step-SPA test of Hsu, Hsu, and Kuan (2010).

Remarks:

- To compute the critical value at each step, it is important to consider not only the hypotheses that have not been rejected ($\mathcal{A}_{j+1}$) but also those might have been incorrectly rejected in previous steps (M_j with $|M_j| = k - 1$). As the latter hypotheses are unknown to us, we take the largest possible critical value among those based on $M_j \cup \mathcal{A}_{j+1}$.
- Computing the critical values is computationally demanding, because we need to consider all possible $M_j \cup \mathcal{A}_{j+1}$.
- The step-down control of Romano and Wolf (2007) is based on the bootstrapped distribution without re-centering and hence ought to suffer from power loss, as shown in Hansen (2005) and Hsu, Hsu, and Kuan (2010).

The Step-SPA(k) Test

The Step-SPA(k) test is the stepwise SPA test that controls k-FWER.

- The (studentized) statistic is the same as the Step-SPA test.
- The critical values $\hat{q}_{A_j}(1-\alpha, k)$ are obtained from the re-centered, bootstrapped distribution of

$$k\text{-}\max_{j=1,\ldots,m}\sqrt{n}\big(\bar{d}_j^*(b) - \bar{d}_j + \hat{\mu}_j\big), \quad b = 1, \ldots, B,$$

  with $\hat{\mu}_j = \bar{d}_j \mathbf{1}\big(\sqrt{n}\bar{d}_j \leq -a_n\hat{\sigma}_j\big)$, where a_n diverges and $\lim_n a_n/\sqrt{n} = 0$.

  Note: a_n does not have to be $\log(\log n)$.

Asymptotic Properties

Step-SPA(k) Test

1. (k-FWER) Given level α_0, the Step-SPA(k) test has k-FWER $\leq \alpha_0$ when n tends to infinity.
2. (Consistency) The k-th model with $\mu_k > 0$ will be rejected by the Step-SPA(k) test with probability approaching 1 when n tends to infinity.
3. (Power) The Step-SPA(k) test is more powerful than the stepwise test of Romano and Wolf (2007), under any notions of power defined in Romano and Wolf (2005).

Simulations

- Data: Models with $\mathcal{N}(\mu, 1)$, each has $n = 250$ or 500 observations
  - $S = 125$: 100 of them with $\mu = 0$, 15 with $\mu = 0.25$, 10 with $\mu = 0.5$
  - $S = 125$: 50 with $\mu = 0$, 15 with $\mu = 0.25$, 10 with $\mu = 0.5$, 35 with $\mu = -0.25$, 15 with $\mu = -0.5$
  - $S = 250$: with and without negative means
  - $S = 500$: with and without negative means
- No. of bootstraps: $B = 500$; number of replications: $R = 1000$
- 4 Tests with $\alpha = 5\%$:
  - Step-RC of Romano and Wolf (2005)
  - Step-RC(3) of Romano and Wolf (2007) with 3-FWER
  - Step-SPA of Hsu, Hsu, and Kuan (2010)
  - Step-SPA(3) with 3-FWER

Table: Test performance: $n = 250$ and $S = 125$ without negative means.

	Model Correlation $\rho = 0$			
	Step-RC	Step-RC(3)	Step-SPA	Step-SPA(3)
Avg. Rej.	15.87	18.83	16.54	19.92
Avg. False Rej.	0.066	0.348	0.084	0.415
FWER	3.3%	12.3%	4.3%	12.6%
3-FWER	0.4%	4.2%	0.4%	4.4%
Avg. Power	63.2%	73.9%	65.8%	78.0%
	Model Correlation $\rho = 0.25$			
Avg. Rej.	18.58	21.28	19.38	22.31
Avg. False Rej.	0.067	0.341	0.082	0.449
FWER	3.1%	11.4%	4.1%	12.9%
3-FWER	0.4%	4.0%	0.4%	4.5%
Avg. Power	74.0%	83.8%	77.2%	87.4%

Table: Test performance: $n = 250$ and $S = 125$ with negative means.

	Model Correlation $\rho = 0$			
	Step-RC	Step-RC(3)	Step-SPA	Step-SPA(3)
Avg. Rej.	15.74	18.60	17.40	20.92
Avg. False Rej.	0.047	0.200	0.094	0.400
FWER	2.7%	8.0%	4.7%	13.9%
3-FWER	0.5%	2.1%	0.9%	4.1%
Avg. Power	62.8%	73.6%	69.2%	82.1%
	Model Correlation $\rho = 0.25$			
Avg. Rej.	18.35	21.00	20.23	23.10
Avg. False Rej.	0.042	0.195	0.095	0.413
FWER	2.6%	7.8%	4.7%	14.3%
3-FWER	0.4%	2.1%	0.8%	3.8%
Avg. Power	73.2%	83.2%	80.5%	90.8%

Note: There are only false rejections of models with mean zero.

Table: Test performance: $n = 250$ and $S = 250$ without negative means.

	Model Correlation $\rho = 0$			
	Step-RC	Step-RC(3)	Step-SPA	Step-SPA(3)
Avg. Rej.	29.46	34.52	30.63	36.28
Avg. False Rej.	0.087	0.363	0.090	0.379
FWER	4.1%	11.2%	4.4%	12.3%
3-FWER	1.0%	4.1%	1.0%	4.3%
Avg. Power	58.7%	68.3%	61.1%	71.8%
	Model Correlation $\rho = 0.25$			
Avg. Rej.	34.63	39.57	36.14	41.74
Avg. False Rej.	0.079	0.374	0.081	0.385
FWER	4.0%	11.6%	4.2%	12.5%
3-FWER	0.8%	4.1%	0.8%	4.1%
Avg. Power	69.1%	78.4%	72.1%	82.7%

Table: Test performance: $n = 250$ and $S = 250$ with negative means.

	Model Correlation $\rho = 0$			
	Step-RC	Step-RC(3)	Step-SPA	Step-SPA(3)
Avg. Rej.	29.41	34.34	32.23	38.22
Avg. False Rej.	0.036	0.188	0.075	0.340
FWER	2.4%	9.0%	4.6%	13.7%
3-FWER	0.3%	2.3%	0.8%	3.6%
Avg. Power	58.7%	68.3%	64.3%	75.8%
	Model Correlation $\rho = 0.25$			
Avg. Rej.	34.56	39.42	37.87	43.47
Avg. False Rej.	0.035	0.195	0.083	0.354
FWER	2.4%	9.1%	4.8%	13.8%
3-FWER	0.3%	2.4%	0.9%	3.9%
Avg. Power	69.1%	78.5%	75.6%	86.2%

Note: There are only false rejections of models with mean zero.

Table: Test performance: $n = 500$ and $S = 250$ with negative means.

	Model Correlation $\rho = 0$			
	Step-RC	Step-RC(3)	Step-SPA	Step-SPA(3)
Avg. Rej.	42.43	45.83	45.56	48.73
Avg. False Rej.	0.047	0.202	0.085	0.393
FWER	2.4%	7.6%	4.3%	12.2%
3-FWER	0.6%	1.8%	0.7%	3.7%
Avg. Power	84.8%	91.3%	90.9%	96.7%
	Model Correlation $\rho = 0.25$			
Avg. Rej.	46.98	48.79	48.95	50/14
Avg. False Rej.	0.051	0.205	0.084	0.393
FWER	2.7%	7.6%	4.1%	12.1%
3-FWER	0.6%	1.8%	0.8%	3.7%
Avg. Power	93.9%	97.2%	97.7%	99.5%

Note: There are only false rejections of models with mean zero.

Table: Test performance: $n = 250$ and $S = 500$ without negative means.

	Model Correlation $\rho = 0$			
	Step-RC	Step-RC(3)	Step-SPA	Step-SPA(3)
Avg. Rej.	54.78	63.33	56.67	66.40
Avg. False Rej.	0.099	0.338	0.099	0.345
FWER	3.5%	8.2%	3.5%	8.9%
3-FWER	1.9%	3.4%	1.9%	3.4%
Avg. Power	54.8%	63.3%	56.7%	66.4%
	Model Correlation $\rho = 0.25$			
Avg. Rej.	65.07	74.66	67.77	78.33
Avg. False Rej.	0.083	0.301	0.083	0.317
FWER	2.8%	7.4%	2.8%	8.6%
3-FWER	1.6%	3.0%	1.6%	3.0%
Avg. Power	65.0%	74.4%	67.7%	78.0%

Table: Test performance: $n = 250$ and $S = 500$ with negative means.

	Model Correlation $\rho = 0$			
	Step-RC	Step-RC(3)	Step-SPA	Step-SPA(3)
Avg. Rej.	55.34	64.08	60.03	70.96
Avg. False Rej.	0.060	0.152	0.096	0.336
FWER	3.2%	5.2%	3.2%	110.4%
3-FWER	0.8%	1.6%	1.2%	3.2%
Avg. Power	55.3%	63.9%	59.9%	70.6%
	Model Correlation $\rho = 0.25$			
Avg. Rej.	65.44	74.85	71.36	83.24
Avg. False Rej.	0.068	0.160	0.100	0.352
FWER	3.2%	5.6%	3.2%	10.4%
3-FWER	0.8%	1.6%	1.6%	3.6%
Avg. Power	65.4%	74.7%	71.3%	81.9%

Note: There are only false rejections of models with mean zero.

A Summary

- Step-SPA(3) vs. Step-RC(3): More accurate 3-FWER and better power; power improvement is more obvious when there are models with negative means.
- Step-SPA(3) vs. Step-SPA: Better power
- For a given S, tests have better power when models are correlated.
- For a given n, all powers deteriorate when S increases, yet the power improvement of Step-SPA(3) over Step-RC(3) is roughly the same.
- When there are models with zero mean and negative mean, only those with zero mean may be incorrectly rejected. As such, allowing for more false rejections (k-FWER) is not very costly in practice.

Concluding Remarks

1. In large-scale, multiple testing problems, it is crucial to control a proper measure of familywise error, e.g. FWER, k-FWER, or FDP.
   - The choice of such error measure ought to be context dependent.
   - A procedure that controls FDP is possible when the rejection of a procedure that controls k-FWER is monotone in k, i.e., a hypothesis rejected by a k_1-FWER procedure must be rejected by a k_2-FWER if $k_1 < k_2$.
2. For testing inequality hypotheses, it would be better to avoid the least favorable configuration.
3. There are numerous potential applications of the new stepwise testing procedure, and they may result in different answers to empirical issues.