Economics Bulletin

Volume 34, Issue 4

European equity fund managers: luck or skill?!

Enareta Kurtbegu
University of Evry-Val d''Essonne, EPEE and TEPP-CNRS

Juliana Caicedo-llano
University of Evry-Val d''Essonne, EPEE, TEPP-CNRS and EONOS Investment Technologies

Abstract

Seeking persistent abnormal portfolio performance has been a key question for academics and practitioners. The main challenge in the construction of fund-of-funds is the ex-ante selection of "skilled" managers, ex-post outperforming the benchmark. This empirical study focused on European mutual funds, consists in using the False Discovery Rate selecting procedure. The standard tests to identify funds with non-zero alphas do not adequately account for the presence of "luck", while this becomes an important issue when one deals with multiple testing. Different pricing models are used and the performance of constructed fund-of-funds is analyzed in-sample and out-of-sample for different investment strategies.

The authors wish to thank participants of the European Financial Management 2013 "Merton H. Miller" doctoral Seminar (Reading, UK), the International Istanbul Finance Congress 2013 (Istanbul, Turkey), the International Finance and Banking Society Conference 2014 (Lisbon, Portugal) and the 29th Annual Congress of the European Economic Association 2014 (Toulouse, France) for the helpful comments on the earlier versions of this paper.
Citation: Enareta Kurtbegu and Juliana Caicedo-llano, (2014) ''European equity fund managers: luck or skill?!'', *Economics Bulletin*, Vol. 34 No. 4 pp. 2340-2350.
Contact: Enareta Kurtbegu - enareta.kurtbegu@univ-evry.fr, Juliana Caicedo-llano - juliana.caicedollano@univ-evry.fr.
Submitted: May 19, 2014. **Published:** November 06, 2014.

1 Introduction

Persistence of abnormal return and portfolio managers selection are major topics of interest for academics and financial practitioners. In particular, multi-managers face the difficult task of selecting a small number of funds with attractive return properties based on past track records. Significant *alphas* are frequently used as a risk-adjusted performance measure to select performing funds. The *alpha's* estimation and its associated hypothesis test when repeated for several funds (multiple testing) increases the number of false discoveries and the error level.

In order to face this multiple-testing problem, a cross-sectional bootstrap procedure was proposed, to distinguish between funds with differential performance (including the ones verifying the alternative hypothesis) and the *type I error* whose null hypothesis is rejected while being true. Funds verifying the latter characteristic are "lucky" enough to remain in the filter together with the high performing ones and are unable to be easily identified separately. The selection procedure consists on controlling the proportion of "lucky" fund managers in the selection procedure and in the best case, not to leave any of them pass through the "filter holes".

There is a wide branch of literature oriented towards fund performance devoted to detect first whether there is a significant *alpha* and secondly if that is the case, to determine whether one could *ex-ante* pre-define it. No analysis in European fund industry are yet done except the country specific study by Cuthbertson and Nitzsche (2010) based on German long-only Equity mutual fund database.

All the existing bootstrap methods used to measure false discoveries are based on the estimation of the *alpha* parameter and thus the asset pricing model used to estimate it plays a key role. Therefore, as robustness check, we consider three different pricing models. Barras et al. (2010) aim at using the False Discovery Rate (FDR) approach as a method to estimate the proportions of three fund categories such as the "skilled" funds, the zero-*alpha* and the "unskilled" ones. Following this approach, our study consists on measuring these proportions and comparing this fund performance for different subsamples belonging to distinct economic periods. Our goal goes further, first in using the FDR not only as a selection procedure but also in constructing FoF based on the occurred selection using various investment strategies. Secondly, a detailed comparison is done between in-sample and out-of-sample strategies for each sample.

2 Methodology

Using standard notation for hypothesis testing, the null hypothesis H_0 indicates the fund i achieves no significant performance whereas the alternative hypothesis H_1, if it is true, identifies funds with differential performance.

$$\begin{cases} H_0 : \alpha_i = 0 \\ H_1 : \alpha_i > 0 \text{ } or \text{ } \alpha_i < 0. \end{cases} \tag{1}$$

The multiple testing hypothesis consists on simultaneously testing a set of hypothesis. Let us suppose that there are N hypothesis to be tested and the probability that the null hypothesis is accepted at a certain significance level is q ($0 \leq q \leq 1$). On one hand, if multiple testing is taken into consideration then, there would be $(1 - q) * N$ true hypotheses to be rejected. On the other hand, if the multiple testing is not taken into account and each hypothesis is tested independently, then the probability that at least one true null hypothesis will be rejected is $1 - q^N$. As the number of multiple tests N increases and q is small enough, this probability goes to one and the number of true hypothesis rejected converges toward N.

The aim of multiple testing is to keep under control the number of funds that satisfy the H_0 but are rejected by accident. We implement a cross-sectional bootstrap procedure as in Wermers et al. (2006) to face the problem of testing multiple H_0 simultaneously and to control for false discoveries. Therefore, if no adjustment of the p-values is made, there is a quite high probability that some of the true H_0 will be rejected. Bonferroni's method Hochberg (1988) was one of the first classical approaches proposed whereas other methods were proposed afterwords such as: the ones based on marginal p-values, the stepwise methods (Wolf and Wunderli (2011)), the generalized error rates and the re-sampling methods (Benjamini and Hochberg (1995), Barras et al. (2010), Cuthbertson et al. (2008)).

As FDR approach is based on the fund manager's *alpha*, we consider three distinct factor pricing models for the estimation: the standard Capital Asset Pricing Model ($CAPM$), the Fama-French three factors model (FF) and the four factors model introduced by Carhart (CA)[1]. The FDR is defined as the expected proportion of false discoveries to the significant funds, corresponding to each tail of the cross-sectional *alpha* distribution, based on a given pre-defined significance level denoted by *delta*. We consider a fund to have significant estimated *alpha* if its p-value is smaller than the threshold *delta*. Three portfolio construction strategies are implemented: equal-weights (EW), minimum-variance (MV) and equal-risk (ER) portfolio. We analyze each of these strategies including funds from the FDR selection using different risk and return measures.

3 Data description

We consider 89 European equity mutual funds for the period from January 2005 to February 2012 observed at a daily basis.

[1]In this paper we will focus more in details on CAPM and Fama-French models. Nevertheless some main results provided by using Carhart model are shown and the rest is available under request.

Our database corresponds to open-end funds domiciled in different European countries mainly in Luxembourg, France, Germany and Spain (70% of the funds) [2] invested in medium and large companies from European countries and valued in Euro currency. Several funds correspond to a Value style management but their benchmarks are European Indexes mainly $EUROSTOXX$ 50, $STOXX$ Europe 50 and $MSCI$ Europe. We used historic fund values from Datastream and additional information from Thomson Reuters to complete the database. We focused on the period $2005-2012$ since particularly during crisis time the fund selection is one of the biggest investment challenges usually resulting in lower success rates. Therefore, the analyzed period used will serve as a robustness check of the FDR selection procedure. Moreover, this set of 89 funds is representative of the target investment universe for this particular time period.

We use the series of the $EURIBOR$ one month as risk-free returns and the returns of $MSCI$ Europe Index as market index returns. For the size factor (SMB), we calculated the excess performance of the $MSCI$ Small Cap Index over the performance of an Index composed with only 50 biggest companies from the European region. As for the style factor (HML), we computed the extra performance of the $MSCI$ Value Index over the $MSCI$ Growth Index. For the momentum factor, we used the same construction technique as Carhart (1997), thus we build this factor based on one month and one year past performance using the top and worst 30% performers. The Fama-French factors' descriptive statistics are expressed on daily basis in Table I.

Table I: Descriptive Statistics MSCI Index and Fama-French factors

						Correlation		
	TotR	AvgR	Vol	IR	MDD	MSCI	SMB	HML
MSCI Europe	−8.6	−1.217	26.1	0	−96.3	1	−0.4614	0.5242
SMB	86.3	8.374	12.2	0.7	−24.4	−0.4614	1	−0.3943
HML	−18.3	−2.719	8.6	−0.3	−32.7	0.5242	−0.3943	1

Note:Period: [Jan 2005;Feb 2012]; Performance indicators: {$TotR$; $AvgR$; Vol; IR; MDD} [3]; Computations by the authors.

The funds present an average return for the whole period of 1.4% and an average annualized volatility of 21.7%. This database is interesting because it contains two sub-samples. The first one corresponds to a bull market period while the second one to the 2008 global financial crisis which is mainly a bear market over the period 2008 – 2012. The average annual return drops from nearly 4% to −1% and the volatility increases from

[2]The rest of funds (30%) are issued in Austria, Finland, Ireland, Italy, Netherlands, Portugal and Switzerland.

[3]TotR: total return, AvgR: average return, Vol:Volatility, IR: Information Ratio, MDD: Max Draw-down

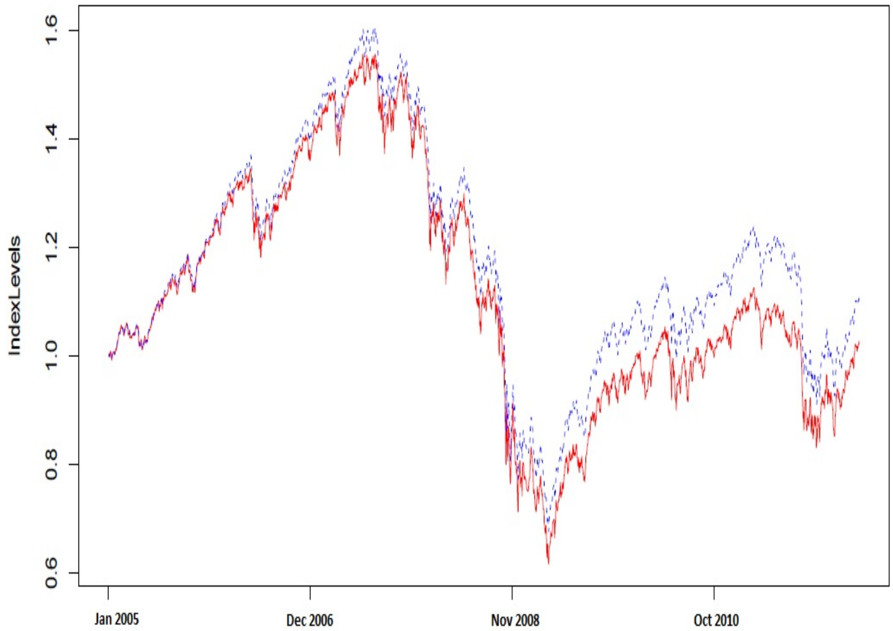
15% to 26% (see Table II).

Table II: Descriptive Statistics Mutual Funds

	Mean	St. Dev	Skew	Kurt	JB-pvalue
All data	1.39	21.71	0	0	0
$S1$	3.75	15.21	0	0	0
$S2$	-0.99	26.38	0	0	0

Note: Descriptive statistics related to three samples studied; Period S: [Jan 2005–Feb 2012]; Period $S1$: [Jan 2005–Aug 2008]; Period $S2$: [Aug 2008–Feb 2012]; Computations by the authors.

Figure 1: Benchmark ($MSCI$ Index) and EW portfolio.



Note: Benchmark portfolio (in red), EW portfolio among all funds (in blue); Computations by the authors.

Considering the features of this database (see graph 1), we test our method based on three different samples: the whole sample (S), the bullish period ($S1$) and the market crash followed by a recovery period ($S2$). We estimated three different factor models individually for each fund in the database using the whole sample. For the $CAPM$ model the average estimated *alpha* is close to zero but not significant on average and the beta coefficient for the market factor is equal to 0.63. The average R^2 coefficients for these regression is equal to 0.57. We compare these results to the three factors estimation where the average sensitivity to the market factor is 0.66, and the coefficients for the size factor and the style factor are equal to 0.26 and 0.21, the adjusted R^2 being equal to 0.6. The main average statistics of the estimated coefficients for these regressions are presented in Table III.

Table III: Average regression estimates

	Estimate	t-stat	$R^2/\hat{R}^2$
alpha-CAPM	0.00	0.21	0.57
MKT	0.63	78.78	0.57
alpha-3 FF	-0.00	-0.05	0.61
MKT	0.66	66.83	0.60
SMB	0.26	7.84	
HML	0.21	5.29	

Note: Average estimated *alphas* are non-significantly zero; Sample: S; Pricing model: $CAPM$ (1 factor) and FF (3 factor); Calculations by the authors.

4 Main findings and results

4.1 In-sample results

The FDR selection method is based on a bootstrap procedure, therefore it requires specification of the number of bootstraps (B) needed to have stable estimates. After testing for different number of simulations, we established that the selection procedure becomes stable for $B = 2000$. Beyond this threshold, the sets of selected funds are always identical. The size of the time-series is crucial in deciding the number of bootstraps needed to have a convergent selection of funds. Table IV presents the number of selected funds chosen by the FDR procedure for different significance levels *delta*, pricing models and samples. As expected, an increase in *delta* increases the number of selected funds which varies between 4 and 19 funds for the $CAPM$ and between 1 and 4 for the FF model considering estimations done for sample S. We emphasize that the same group of funds is selected at each test for different confidence levels confirming the robustness of FDR selection.

Table IV: The variation of the number of FDR funds by *delta*

δ	CAPM-S	CAPM-S1	CAPM-S2	FF3-S	FF3-S1	FF3-S2
0.05	4	6	2	1	4	–
0.1	11	7	5	3	6	–
0.15	15	9	6	4	8	–
0.2	16	12	9	4	11	–
0.25	17	16	11	4	16	–
0.3	19	21	13	4	17	–

Note: Selection method: FDR; $B = 2000$; Pricing model: $CAPM$ and FF; Computations by the author.

During the bull period ($S1$) there are more funds with high *alpha* inducing a higher number of selected funds contrary to post-crisis sub-sample for each level of *delta*. The results for the long-term sample (S) stay in between the short-term ($S1$ and $S2$) in-sample results because of the crisis which decreases the value of *alpha*. Thus, the number of selected funds is higher in sub-sample $S1$ compared to sub-sample $S2$ for both pricing models (see Table IV). Furthermore, the number of selected funds by FF model is lower than the number of selected funds by $CAPM$. Comparing to the benchmark performance (see Table V), creating the FoF by selecting *ex-ante* a certain number of funds, induces out-performance, and a decrease in volatility from 21% to 18% on average. The maximum drawdown of the selected funds decreases as well compared to the benchmark portfolio results while the Sharpe ratio reaches 23% (10 times higher than the $MSCI$ Sharpe ratio). The increase in *delta* means higher probability of having "involuntary" selected funds with low *alphas* and increase in portfolio diversification.

Table V: $EW - FoF$ for different *deltas* (S) - $CAPM$ (1 Factor)

	MSCI	EW-all			FDR Portfolios			
δ			0.05	0.10	0.15	0.20	0.25	0.30
TotR	2.72	10.91	39.63	37.85	34.84	34.09	33.78	32.14
AvgR	0.36	1.39	4.49	4.31	4.02	3.94	3.91	3.75
Vol	21.28	15.15	19.43	18.67	18.91	18.67	18.48	18.63
IR	0.02	0.09	0.23	0.23	0.21	0.21	0.21	0.20
MDD	-83.93	-82.21	-77.13	-68.93	-69.38	-68.75	-69.11	-69.64

Note: Performance indicators: $\{TotR;\ AvgR;\ Vol;\ IR;\ MDD\}$; Sample: S; Pricing model: $CAPM$; Selection method: FDR; $B = 2000$; $\delta \in [0.05, 0.3]$; Computations by the authors.

Table VI: $EW - FoF$ for different *deltas* (S) - Fama-French (3 Factors)

	MSCI	EW-all			FDR Portfolios			
δ			0.05	0.10	0.15	0.20	0.25	0.30
TotR	2.72	10.91	60.65	58.65	54.47	54.47	54.47	54.47
AvgR	0.36	1.39	6.37	6.20	5.84	5.84	5.84	5.84
Vol	21.28	15.15	25.82	20.04	19.35	19.35	19.35	19.35
IR	0.02	0.09	0.22	0.30	0.29	0.29	0.29	0.29
MDD	-83.93	-82.21	-60.63	-67.93	-64.17	-64.17	-64.17	-64.17

Note: Performance indicators: $\{TotR;\ AvgR;\ Vol;\ IR;\ MDD\}$; Sample: S; Pricing model: FF; Selection method: FDR; $B = 2000$; $\delta \in [0.05, 0.3]$; Computations by the authors.

The performance for $EW-$portfolios based on FDR selections is shown in table VI. The results are completely in line with those obtained with $CAPM$ estimations. The best portfolio risk-adjusted performance is reached with a level of significance *delta* of

10%. Moreover, the increase in *delta* beyond 15% does not decrease the FoF performance both in terms of risk and return because of no extra selected funds.

Every portfolio selection based on the FDR procedure has therefore better risk-adjusted measures of performance compared to the benchmark and to the EW portfolio composed of all the funds in the database. This confirms the out-performance of FDR selection procedure independently of the sample characteristics.

Furthermore, equally-weighted portfolio constructed using FF three factor model has a Sharpe ratio of 30% for a significance level of 10% as the corresponding portfolio constructed with the estimated $CAPM$ one factor model has a Sharpe ratio of 23%. Therefore, one could conclude that unless the lower value of *alphas* and the lower number of selected funds, the increase in the number of factors in the pricing model increases the portfolio performance.

4.2 Backtesting results

As a continuation of the in-sample analysis we consider another similar in-sample exercise but with rolling estimations. First of all, we use the FDR procedure to select the "skilled" fund managers during the whole available period. The selected funds are used to create EW, ER and MV portfolios actualizing the respective weights every 3 months and using one year past observations to estimate the parameters needed for each portfolio construction. There is no difference between the EW-portfolios in-sample and any EW-backtesting portfolios but the $MV-FoF$ and $ER-FoF$ are different if the re-balancing period and the window of estimation changes (see Table VII). As the number of factors increases, the number of funds selected by the FDR method decreases but different FoF have stable performance. Moreover, the portfolio strategy which out-performs not only the benchmark but also all other constructed FoF, independently to the sample choice, is the $MV - FoF$.

Referring to $S1$ sub-sample, the $MV - FoF$ reaches its maximal Sharpe ratio and total return performance when $CAPM$ model is used to determine *alphas*. The CA model allows to construct a FoF with lower annual volatility reaching 8.38%. Because of the distinct economic characteristics there are differences in performances among samples using the same pricing model. A higher risk-adjusted performance is achieved in $S1$, followed by samples S and $S2$ consecutively.

Table VII: Performance of back-testing EW-FoF, ER-FoF and MV-FoF

	CAPM								
	S			S1			S2		
	EW	ER	MV	EW	ER	MV	EW	ER	MV
TotR	34.09	34.02	57.47	25.96	26.13	38.41	13.15	13.22	27.55
AvgR	3.94	3.94	6.10	6.20	6.24	8.74	3.33	3.34	6.55
Vol	18.67	18.64	15.07	12.89	13.00	10.93	23.34	23.26	19.69
IR	0.21	0.21	0.40	0.38	0.38	0.63	0.17	0.17	0.38
MDD	-68.75	-68.63	-56.10	-32.29	-31.89	-25.90	-47.78	-48.00	-35.68
	Fama-French						Carthart		
	S			S1			S1		
TotR	54.47	55.55	61.84	25.26	25.41	36.84	15.83	16.78	21.96
AvgR	5.84	5.94	6.47	6.05	6.09	8.43	3.95	4.17	5.34
Vol	19.35	19.34	17.16	12.98	12.98	10.89	9.53	8.87	8.38
IR	0.29	0.30	0.37	0.38	0.38	0.62	0.35	0.39	0.53
MDD	-64.17	-63.79	-58.11	-31.90	-31.30	-25.61	-32.16	-30.93	-26.91

Note: Performance indicators: $\{TotR;\ AvgR;\ Vol;\ IR;\ MDD\}$; In-sample portfolio strategies: EW, ER and MV; $B = 2000$; $\delta = 20\%$; Computations by the authors.

4.3 Out-of-sample results

We conduct an out-of-sample study, where we use observations from one year back to make a FDR selection, then invest during three upcoming months and analyze the performance of an EW strategy using $CAPM$ and FF model (see table VIII). No fee costs were taken into consideration.

Table VIII: Out-of-sample results

	CAPM20	FF3.20
TotR	6.75	9.10
AvgR	1.04	1.38
Vol	26.59	24.62
IR	0.04	0.06
MDD	-42.23	-34.59

Note: Performance indicators: $\{TotR;\ AvgR;\ Vol;\ IR;\ MDD\}$; Out-of-sample portfolio strategy; EW rolling window; Sample S; Pricing model: $CAPM$ and FF; Selection method: FDR; $B = 500$; $\delta = 20\%$; Computations by the authors.

On one hand, we observe that the FDR portfolio estimated from the $CAPM$ with a *delta* of 20% has a positive total return for the period of 6.7%, a risk adjusted performance of 4% and volatility of 26.6%. Low performance of estimated portfolios for lower values of *delta* is due to diversification and to the low number of funds selected during $S2$. Results become in line with previous conclusions for *delta* 20% where the constructed FoF out-performs other constructed portfolios of the same kind. Nevertheless, the use of FF model induces a decrease in *alpha* and in the number of selected funds. On the other hand, focusing on the same level of accepted "luck" proportion ($\delta = 20\%$), using FF model the $EW - FoF$ is of higher performance in terms of total return reaching 9.10% and Sharpe ratio being 6% compared to $CAPM$ where the total risk return is valued to be 6.75% and the Sharpe ratio being 4%. Moreover, there is a decrease in volatility and in maximum drawdown. These results show that the out-of-sample portfolios deliver a lower performance than the in-sample one because they are not based on *ex-post* analysis as in-sample estimations are. The positive performance of the out-of-sample strategies compared to the lower benchmark performance for the same sample and for similar levels of risk let us conclude that FDR could be useful method in the multi-management investment industry.

5 Conclusion

We conducted FDR selection procedure to determine the set of fund managers having "true" positive *alpha* returns or being categorized as "skilled". The FDR selection method is based on bootstrap procedures therefore we conducted different tests to determine the optimal number of bootstraps needed to reach selection stability. The main results of this study include the stability of FDR sets as the number of funds increases with parameter *delta*, decreases with the number of factors composing the pricing model and impacts positively the portfolio performance of the constructed FoF. The out-of-sample constructed portfolios out-perform the benchmark but there is a lower performance compared to in-sample portfolios. However, the minimum-variance portfolio is the investment strategy that out-performs in absolute all other constructed FoF. The FDR technique applied to European mutual funds is useful to *ex-ante* select funds with "true" *alphas*. In this empirical study, the accepted level of "luck" is considered exogenous such as to give us the possibility to construct enough diversified portfolios but as soon as the FDR selection starts to be used in real life, official institutional rules could decide to bound the accepted level "luck" .

References

Barras, L., O. Scaillet, and R. Wermers (2010). False Discoveries in Mutual Fund Performance: Measuring Luck in Estimated Alphas. *Journal of Finance 65*(1), 179–216.

Benjamini, Y. and Y. Hochberg (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society B 57*(1), 289–300.

Carhart, M. (1997). On Persistence in Mutual Fund Performance. *Journal of Finance 52*, 57–82.

Cuthbertson, K. and D. Nitzsche (2010). Performance, Stock Selection and Market Timing of the German Equity Mutual Fund Industry. *Working Paper, Cass Business School*.

Cuthbertson, K., D. Nitzsche, and N. Sullivan (2008). UK Mutual Fund Performance: Skill or Luck? *Journal of Empirical Finance 15*, 613–634.

Hochberg, Y. (1988). A Sharper Bonferroni Procedure for Multiple Tests of Significance. *Biometrika 4*(75), 800–802.

Wermers, R., R. Kosowski, A. Timmermann, and H. White (2006). Can Mutual Fund Stars Really pick Stocks? New Evidence from a Bootstrap Analysis. *Journal of Finance 61*, 2551–2595.

Wolf, M. and D. Wunderli (2011). Fund-of-Funds Construction by Statistical Multiple Testing Methods. *The Oxford Handbook of Quantitative Asset Management, Oxford University Press, Oxford.*, 116–135.