Graduate School ETD Form 9
(Revised 12/07)

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Nikita Tuzov

Entitled MUTUAL FUND PERFORMANCE EVALUATION METHODOLOGY AND LOCAL FALSE DISCOVERY RATE APPROACH

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Prof. Frederi Viens
Chair

Prof. Dabao Zhang

Prof. Michael Levine

Prof. Bruce Craig

To the best of my knowledge and as understood by the student in the *Research Integrity and Copyright Disclaimer (Graduate School Form 20)*, this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Approved by Major Professor(s): Frederi Viens

Approved by: Prof. Jun Xie 03/04/2009
Head of the Graduate Program Date

Graduate School Form 20
(Revised 10/07)

PURDUE UNIVERSITY
GRADUATE SCHOOL

Research Integrity and Copyright Disclaimer

Title of Thesis/Dissertation:

MUTUAL FUND PERFORMANCE EVALUATION METHODOLOGY AND LOCAL FALSE DISCOVERY RATE APPROACH

For the degree of Doctor of Philosophy

I certify that in the preparation of this thesis, I have observed the provisions of *Purdue University Executive Memorandum No. C-22,* September 6, 1991, *Policy on Integrity in Research.**

Further, I certify that this work is free of plagiarism and all materials appearing in this thesis/dissertation have been properly quoted and attributed.

I certify that all copyrighted material incorporated into this thesis/dissertation is in compliance with the United States' copyright law and that I have received written permission from the copyright owners for my use of their work, which is beyond the scope of the law. I agree to indemnify and save harmless Purdue University from any and all claims that may be asserted or that may arise from any copyright violation.

Nikita Tuzov
Signature of Candidate

03/04/2009
Date

*Located at http://www.purdue.edu/policies/pages/teach_res_outreach/c_22.html

MUTUAL FUND PERFORMANCE EVALUATION METHODOLOGY AND LOCAL FALSE DISCOVERY RATE APPROACH

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Nikita Tuzov

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

May 2009

Purdue University

West Lafayette, Indiana

To Dr. Yuri S. Kan who kindled my interest in something that, as I eventually discovered, is termed “Quantitative Finance”.

ACKNOWLEDGMENTS

I would like to thank Mr. Michael Cohn, CFA, and other members of Walter Raquet Capital Management for drawing my interest to the area of mutual fund performance evaluation during my internship in summer 2006.

During the initial stage of this research, I asked and received helpful answers from Profs. Bradley Efron and Brit Turnbull from Stanford University as well as Prof. Korbinian Strimmer from University of Leipzig.

I would like to thank the members of my PhD committee, Profs. Frederi Viens, Dabao Zhang, Michael Levine and Bruce Craig for reading the draft and making helpful suggestions.

Finally, I am eternally grateful to Prof. Olga Bezhanova of Cornell University for her unquenchable spiritual support as well as proofreading.

TABLE OF CONTENTS

ABSTRACT

Tuzov, Nikita V. Ph.D., Purdue University, May, 2009. Mutual Fund Performance Evaluation Methodology and Local False Discovery Rate Approach. Major Professor: Frederi Viens.

The history of applying statistical simultaneous inference methods to a financial problem of mutual fund performance evaluation is very short. A major problem in applying simultaneous inference methods is the non-trivial dependence among the utilized test statistics. When the number of tests is large, the explicit modeling of dependence structure becomes difficult. As a result, assumptions that are too restrictive are made, which can substantially bias the inference. In addition, the initial performance evaluation model itself can be misspecified and thus distort the results. For instance, the recent study of Barras, Scaillet and Wermers (2008) utilizes a multiple inference procedure with oversimplifying assumptions and, therefore, is prone to both sources of bias.

Another under-investigated issue is the statistical power in a typical mutual fund study. The study of Kothari and Warner (2001) makes some progress but their research is not based on real mutual fund data.

This paper catches up with the recent developments in Statistics by applying a state-of-the-art “empirical null hypothesis” concept combined with the” local false discovery rate” method, developed by Efron in 2001-2007. That offers a viable alternative to the explicit modeling of high-dimensional dependence structure. In addition, the findings of Efron suggest that the new procedure may account for the performance evaluation model misspecification. The new method also

provides informative power measures and an elegant way of comparing the performance of mutual fund subgroups.

A comprehensive investigation is performed for about 1900 actively managed US equity mutual funds observed monthly between 1993 and 2007. The results provide a significant extension to the findings of Barras et al. whose method can be seen as a restricted version of the method in this study. It is shown that the version of Barras et al. has both statistically and practically significant bias.

We conclude that, unfortunately, Barras et al. are too optimistic about the performance of US mutual funds. In addition, a detailed power analysis reveals that a typical mutual fund study with monthly dataset and multifactor performance evaluation model has a very low power. Even when outperformers are present in the sample, it usually requires too many years of data to single them out.

CHAPTER 1. INTRODUCTION

1.1. Objectives

The studies of portfolio manager performance evaluation and, in particular, performance of mutual funds (later referred to as MF), go back as much as 40 years. Over these years, a typical agenda for a MF performance study included the following steps:

1) Selecting performance measure(s);
2) Estimating performance for each MF individually;
3) Interpreting the results. In the MF context, this usually involves an attempt to find association between the performance and fund characteristics such as the fund's investment objective, its turnover, total net asset value (TNA), and so on. The persistence of performance (e.g., if past winners continue to win in the future) is also of interest.

The issues 1)-3) have been addressed thoroughly by a large number of financial researchers. The development of more adequate performance measures and utilization of higher quality datasets can be traced through the works of Jensen (1968), Ippolito (1989), Elton et al. (1993), Hendricks et al. (1993), Ferson and Schadt (1996), Carhart (1997), Daniel et al. (1997), Chen et al. (2000), Wermers (2000) and many others. A discussion of recent results and an extensive reference list can be found in Nitzsche et al. (2006).

The issue of simultaneous testing, on the other hand, has received significantly less attention. Its importance can be illustrated as follows: suppose that we want

to evaluate the performance of m MF managers, of whom m_0 do not perform well. The performance is measured by a certain test statistic obtained from a performance evaluation model, e.g. Carhart alpha. The corresponding p-value under the null hypothesis of "no outperformance" is also provided. Testing each manager separately at the significance level α one should expect to get αm_0 "false discoveries", i.e. the cases where the null hypothesis of "no outperformance" is rejected incorrectly. To distinguish between true and false discoveries, a multiple inference procedure has to be utilized.

However, the application of any multiple inference procedure is far from straightforward when a large number of test statistics have a non-trivial dependence structure and /or the model used to obtain those statistics is misspecified in the first place. The most recent MF study of Barras, Scaillet and Wermers (2008) does employ a multiple inference procedure but hardly addresses either of the abovementioned issues.

Yet another poorly explored but important question is the statistical power of the performance evaluation model. In a typical MF study, no power diagnostics are provided. The study of Kothari and Warner (2001) tries to shed some light on the issue but does not appear exhaustive, especially given that it is not based on the real MF data.

The overall objective of this research is to address the questions of multiplicity and power through a method that accounts for the high-dimensional dependence structure of test statistics and a possible misspecification of the performance evaluation model. These real data features have to be taken into account without imposing oversimplifying assumptions. In that sense, a new approach developed by Efron in 2001-2007 appears to be a viable option. The original purpose of Efron's method was to handle complex multiple testing problems of Statistical Biology. It has never been used for financial studies before, but, as shown below,

it allows us, at least to a certain extent, to address the issues of interest outlined above.

1.2. Organization

This dissertation has five chapters. This chapter (Chapter 1) provides a brief introduction and outlines the research objectives. Chapter 2 provides an overview of simultaneous inference techniques and their application to portfolio manager performance evaluation with the stress on assumption sensitivity and practical implementation issues. Chapter 3 describes the essence and advantages of Efron's method. Chapter 4 looks into the performance evaluation of a large sample of US mutual funds from this new angle. Chapter 5 summarizes the findings and suggests further financial applications of Efron's technique.

CHAPTER 2. SIMULTANEOUS INFERENCE AND ITS FINANCIAL APPLICATIONS

2.1. Possible approaches to simultaneous inference

Let us consider the following framework. Suppose we need to perform m hypothesis tests of the form:

$$H_i^0 \text{ vs. } H_i^a \ , i = \overline{1,m}; \qquad (2.1.1)$$

$P = (P^0, P^a)$ - a random vector of p-values corresponding to null and alternative hypotheses;

m_0 - unknown number of null cases;

V - (random & unobserved) number of rejected true null hypotheses or "false discoveries"

S - (random & unobserved) number of rejected non-true null hypotheses

V+S = R - (random & observed) number of all rejections

Q = V / max(R, 1) - proportion of rejected true nulls among all rejections

The following quantities may be of interest:

FDR = E[Q] - expected value of Q called "False Discovery Rate" (2.1.2)

PFER = E[V] - expected number of "false discoveries"

PCER = E[V]/ m - "Per Comparison Error Rate"

FWER = P{ V ≥ 1 } - "Family-Wise Error Rate"

k-FWER = P{ V ≥ k } - "k-Family-Wise Error Rate"

Any approach, FDR, FWER, PCER or PFER can be used to perform simultaneous inference (Dudoit et al. (2003)), but the choice depends on a

articular application. In the realm of financial performance evaluation, a FWER (based on Bonferroni method) is used by Ferson and Schadt (1996).
In the same context, Romano and Wolf (2005) and Romano et al. (2008) illustrate the control of FWER and k-FWER based on a number of methods, including their own StepM procedure.

A typical part of a MF study is to try to construct an outperforming portfolio of MF. The portfolio has to consist of presumably outperforming mutual funds, but it is admissible to have a relatively small proportion of non-performing funds as long as the overall performance is good. Let us consider the choice among different quantities in (2.1.2) in this context.

FWER usage is justified when a conclusion drawn from m tests is erroneous as soon as one (or more) out of m individual inferences is erroneous. Therefore, FWER is conservative and tends to have a low power, especially when *m* is large. In the abovementioned context, it is not crucial to require that every single one of the identified good performers is a genuine good performer. That rules out FWER as a tool of choice.

Likewise, we are not interested in controlling the absolute number of false discoveries V in terms of its average (PFER) or the probability that V exceeds a certain threshold (k-FWER). PCER is more relevant, but the false discovery proportion, Q, has a direct interpretation as the proportion of useless funds in the outperforming portfolio, so it makes sense to control its expected value, FDR. Another meaningful alternative to FDR is to control not $FDR = E[Q]$ but a certain quantile of Q itself (Romano et al. (2008)), but here we intend to focus on FDR.

The p-values in (2.1.1) can be derived from any particular MF performance evaluation model. Moreover, several models can be used simultaneously if, for instance, it is believed that different types of mutual funds should be evaluated

differently. For equity mutual funds, the four-factor Carhart (1997) performance evaluation model is as follows:

$$r_{i,t} = \alpha_i + b_i \cdot r_{m,t} + s_i \cdot r_{smb,t} + h_i \cdot r_{hml,t} + m_i \cdot r_{mom,t} + \varepsilon_{i,t} \quad (2.1.3)$$
$$t = 1,...T$$
$$i = 1,...m$$

where $r_{i,t}$ is the time period t excess return over the risk-free rate for the MF number i; $r_{m,t}$ is the excess return on the overall equity market portfolio; $r_{smb,t}$, $r_{hml,t}$, $r_{mom,t}$ are the returns on so-called factor portfolios for size, book-to-market, and momentum factors (all can be obtained from CRSP database, see Appendix); $\varepsilon_{i,t}$ is the residual error term. All returns are observed and the quantities α_i, b_i, s_i, h_i, m_i are estimated through multiple linear regression (see Section 4.1).

The parameter α_i is measured in % per time period t (usually one month) and its value shows by how much per one time period the fund outperforms ($\alpha_i > 0$) or underperforms ($\alpha_i < 0$) the benchmark model. Such funds will also be called "skilled" and "unskilled", respectively.

The m p-values in (2.1.1) may correspond to one-sided hypotheses

$$H_i^0 : \alpha_i = 0 \text{ vs. } H_i^a : \alpha_i > 0 \quad (2.1.4)$$

or two-sided hypotheses

$$H_i^0 : \alpha_i = 0 \text{ vs. } H_i^a : \alpha_i \neq 0 \quad (2.1.5)$$

One-sided testing corresponds to identifying significantly good performers and two-sided testing corresponds to identifying significantly "non-zero" (both good and bad) performers.

In a recent series of working papers made public between 2005 and 2008, Barras, Scaillet and Wermers (later referred to as BSW) utilize FDR approach and four-factor Carhart model to estimate the performance of 2076 US equity mutual funds over the period 1975-2006. BSW paper will be the main reference point for our study. In another working paper, Cuthbertson et al. (2008B) borrows the method developed in BSW study to perform a similar analysis of UK mutual funds. Likewise, the very same method is used for German mutual funds by Otamendi et al. (2008).

In order to extend BSW study, let us overview the assumptions underlying the FDR method and look into some issues pertaining to its practical implementation.

2.2. Practical restrictions of FDR-based methods

FDR method was properly introduced by Benjamini and Hochberg (1995) who produced the following result.

Assumption 1. The components of vector P^0 are independent and for any null p-value p_0

$$P\{p_0 \le u\} \le u \quad \forall\, u \in (0,1)$$

Theorem 1. Specify a fixed value $q \in (0,1)$. Under Assumption 1,

$$FDR = E[Q] \le \frac{m_0}{m} q \le q \tag{2.2.1}$$

if all the hypotheses with p-values less than γ are rejected. The cutoff γ is determined according to a certain data-driven stepwise procedure. It is also possible to solve an equivalent “inverse” problem: fix the test size γ and determine the minimal q such that (2.2.1) holds when all hypotheses with p-values less than γ are rejected.

An immediate extension of Theorem 1 is to try to estimate m_0, the unknown number of null cases, in order to make the procedure more powerful. Benjamini and Hochberg (2000) proposed a method (extended in Benjamini et al. (2006)) that essentially relies on one more assumption:

Assumption 2. The marginal distribution of each component of vector P^0 is U(0, 1).

Thus, under Assumptions 1 and 2, the components of vector P^0 are i.i.d. U(0,1) which is called "null distribution".

Then, consider the following subset of observed p-values, $\{p_i,\ i=\overline{1,m}\}$:

$$p_\lambda = \{\, p_i : p_i > \lambda \quad \},\ \lambda \in (0,1) \tag{2.2.2}$$

For λ large enough, p_λ will consist mostly of p-values corresponding to true nulls, i.e. the points in p_λ will approximately have $U(\lambda,\ 1)$ distribution. This fact can be used to estimate λ : e.g., in the histogram of p-values, the plot should "level off" to the right of a certain point on the horizontal axis, and that point is $\hat{\lambda}$. Then the estimate of m_0 is:

$$\hat{m}_0 = (\text{number of points in } \hat{p}_\lambda)/(1-\hat{\lambda}) \tag{2.2.3}$$

The spline estimator of Storey and Tibshirani (2003) and the bootstrap estimator of Storey, Taylor and Siegmund (2004) (the latter used in BSW) are based on the same two assumptions. Therefore, they may fail to work as soon as Assumption 1 or Assumption 2 does not hold.

If Assumption 2 does not hold (e.g., in the case of composite null hypothesis such as $H_i^0 : \alpha_i \le 0$ or a discrete distribution of p-values) the FDR control property (2.2.1) is still valid (Benjamini and Yekutieli (2001)). In that case, given that m_0 is

close to m, one could just take $m_0 = m$ without having to estimate m_0 and that is not going to result in much power loss. Besides, Pounds and Cheng (2006) propose a way of estimating m_0 that works for discrete p-values and the composite null.

If the independence requirement in Assumption 1 is violated, however, the FDR control property (2.2.1) is no longer valid. For that reason, much effort has been invested into adapting the FDR-based methods for the case of dependence among the components of P^0.

Here we would like make a clarification as to the terminology used. For a multidimensional vector with m dependent components, the joint distribution is defined by a c.d.f. that maps R^m into [0; 1]. It can be simplified when the distribution is assumed in a certain parametric form, e.g. for a multivariate normal we only need to know the mean and variance-covariance matrix. Alternatively, one may believe that the mean and variance somehow deliver a good approximation to the true joint distribution which, strictly speaking, is not normal. For the purpose of multiple inference, it is fairy common to assume that it is enough to know the variance-covariance matrix of test statistics. In the subsequent analysis, we are going to use the terms “dependence structure”, “dependence”, “correlation structure”, “variance-covariance matrix”, “joint distribution” interchangeably. For instance, in the case of Carhart model, the dependence structure of test statistics is determined by the $m \times m$ variance-covariance matrix of error terms. Below we are going to look into a few previously used approaches to working with dependent tests statistics.

The first and simplest way to do that is a straightforward modification of the original FDR procedure that works for any dependence structure (Benjamini and

Yekutieli (2001)). However, the corresponding power loss in a large-scale simultaneous testing situation is quite substantial.

In the same study, they show that FDR procedure is still adequate if the vector P^0 has so-called "positive dependency on each one from a subset" structure (PRDS). For instance, assume that the vector of test statistics is multivariate normal $N(\mu,\Sigma)$. Then, if each null statistic has a non-negative correlation with any other statistic, the joint distribution is PRDS. The verification of PRDS property is not a problem in some controlled experiments, where the design itself provides ways to simplify the dependence structure. For example, in clinical trials the researcher often has enough grounds to consider the subjects independent of each other. In fact, all examples of applied problems in Benjamini and Yekutieli (2001)are carefully designed experiments. On the other hand, MF study is an observational study where we have no luxury of simplifying the dependence through experimental design. Even if we are willing to assume that the joint distribution of test statistics is multivariate normal, the belief that each and every null statistic is non-negatively correlated with the rest (m - 1) statistics appears too restrictive. In addition, we cannot attain greater power by estimating m_0, since it is not clear how to do that when the statistics are PRDS-dependent.

Another approach to dependency is to try to estimate the joint distribution of components of P^0 non-parametrically. In particular, in Yekutieli and Benjamini (1999) a bootstrap procedure generates m-dimensional samples of p-values under "complete null" setting, i.e. when all m hypotheses are null. In MF performance evaluation context , Kosowski et al. (2006) introduce so-called "cross-sectional bootstrap". Essentially, they estimate the joint distribution of more than two thousand $\alpha_i's$ via resampling under "complete null" (with T being about 300). The study of Cuthbertson et al. (2008A) borrows this approach to apply it to about 900 UK mutual funds (with T about 340).

In the realm of Econometrics, a similar resampling scheme was proposed in the well-known paper of White (2000), whose approach is developed in Romano and Wolf (2005) and Romano et al. (2007, 2008). The latter develop a procedure called StepM, which can be used to control FWER, FWER-k, FDR and even quantiles of False Discovery Proportion (denoted Q in (2.1.1)). Also, Romano et al. (2008) mention that the StepM procedure is similar to the approach developed for biostatistical purposes by van der Laan et al. in a number of papers, e.g. van der Laan and Hubbard (2005).

At this point, non-parametric estimation of the dependence structure appears to be a fairly reasonable approach. The only flaw of bootstrap approach is that MF time series are usually of different length, and that can render the estimated variance-covariance matrix non-p.s.d. For instance, that can happen when we apply the bootstrap approach of White (2000). We shall say more about these methods in a few paragraphs.

The third approach is to model the dependence structure parametrically. In case of a multifactor performance evaluation model such as that of Carhart, it implies proposing a few “residual factors” that presumably account for all or almost all of the cross-sectional dependence of error terms. The residual factors can be assigned based on common economical sense, e.g., one may assume that error terms coming from MF with the same investment objective are correlated with the same correlation coefficient. It is also possible to derive the residual factors from the data using one of many available “dimension reduction” techniques. Let us describe one of these methods called Principal Component Analysis (PCA), which is also closely related to so-called Ridge Regression and LASSO Regression. Suppose we observe the data matrix *X* of size $m \times T$. For instance, under Carhart ‘s framework, *X* corresponds to the matrix of residual terms (assumed centered).

Then, the estimate of residual variance-covariance matrix is $\Sigma^1 = XX'$. The estimate, Σ^1, is not of full rank because $m > T$. The purpose of PCA is to identify a relatively small number, $p < T$, of linear combinations of columns of X and use these combinations to approximate Σ^1. It can be shown (Hastie et al. (2001)) that the most useful linear combinations correspond to the eigenvectors ("principal components") of Σ^1 that have the largest eigenvalues. The *p* most useful eigenvectors can be found from the eigen decomposition of Σ^1, and then they serve as an input to form the *p* "residual factors". The factors are used to create Σ^2, an approximation to Σ^1. A successful "dimension reduction" means that Σ^2 is a good approximation to Σ^1 with p being much less than T.

For example, Jones and Shanken (2005) utilize a combination of "economically sensible" residual factors (that correspond to MF investment objectives) and PCA-based residual factors.

However, one should be aware that the residual correlation matrix in (2.1.3) is not constant over time. For instance, the correlation between two otherwise weakly correlated equity MF goes up during the so-called "flight-to-quality" periods. The following example, taken from Avellaneda and Lee (2008), illustrates the "flight-to-quality" effect. They take a large number of US stocks observed daily between October 2002 and February 2008. The return correlation matrix is computed based on 1 year (252 business days) rolling window. They perform PCA and estimate the number of principal components that are necessary to explain 55% of the variance in the system.

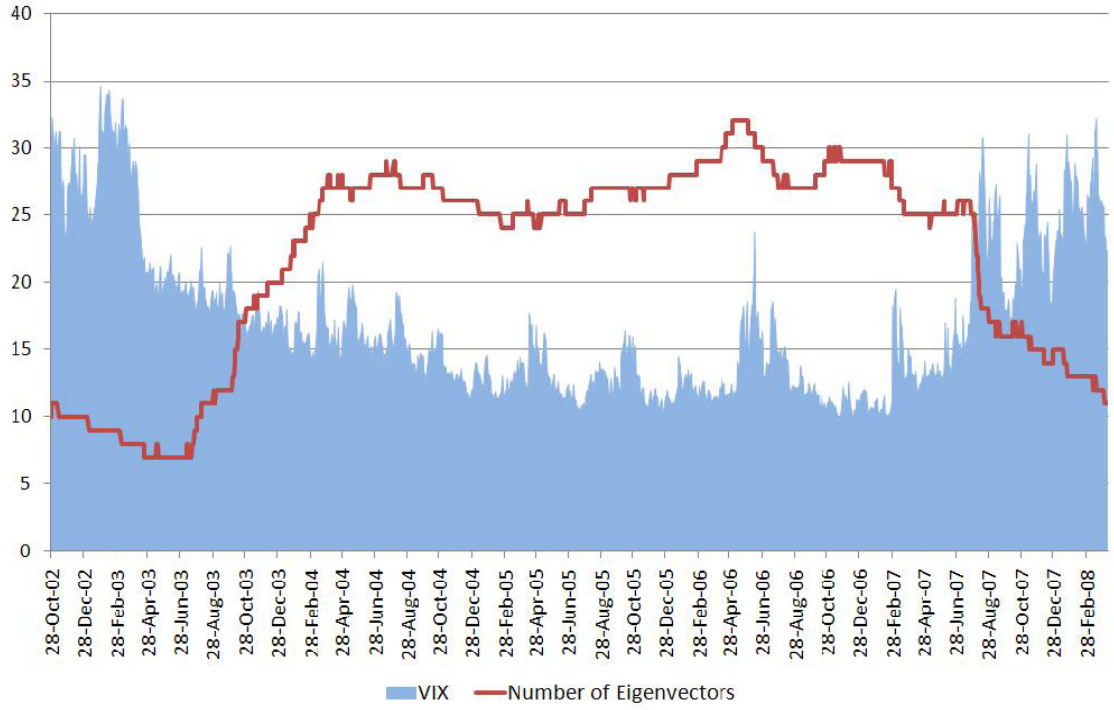

Figure 2..2.1 Market volatility index (blue) and the number of principal components (brown) required to capture 55% of total variance in US stock returns

The blue ragged outline on Figure 2.2.1 shows the market volatility index (VIX) with distinct peaks corresponding to the burst of Internet bubble around 2002 and the subprime crisis of 2007-2008. The red curve shows the number of principal components that is necessary to capture 55% of total variance in the system. Apparently, during the “good times” of 2004-2006, the number of components is much higher (over 25) than it is during the “bad times” (between 7 and 15).

The “flight-to-quality” suggests that, in a multifactor model with a fixed number of factors, the cross-sectional dependence structure of the residuals can change drastically over time. All other things being equal, the overall residual variance in (2.1.3) is a lot smaller during the “bad” times when all equities behave more or less alike.

For a MF dataset, our concern is that, even after dimension reduction, the number of factor in the model is still large compared to the number of observations. For instance, if we introduce the latent factors of Jones and Shanken (2005) into (2.1.3) we will end up having to estimate over 13 regression coefficients with an average number of observations equal to 129. Even if we assume that the dimension reduction is unbiased (Σ^2 is an unbiased estimator of Σ^1), the estimate, Σ^2, will still have a lot of variance. Consider also that the actual number of factors can be as large as 25 (Figure 2.2.1) and / or the loadings on residual factors are not constant over time (i.e., the residual correlation matrix is time-dependent). One may also take into account that, as mentioned in Section 2.1, a few different performance evaluation models can be used simultaneously, which will significantly complicate the explicit modeling of the dependence structure.

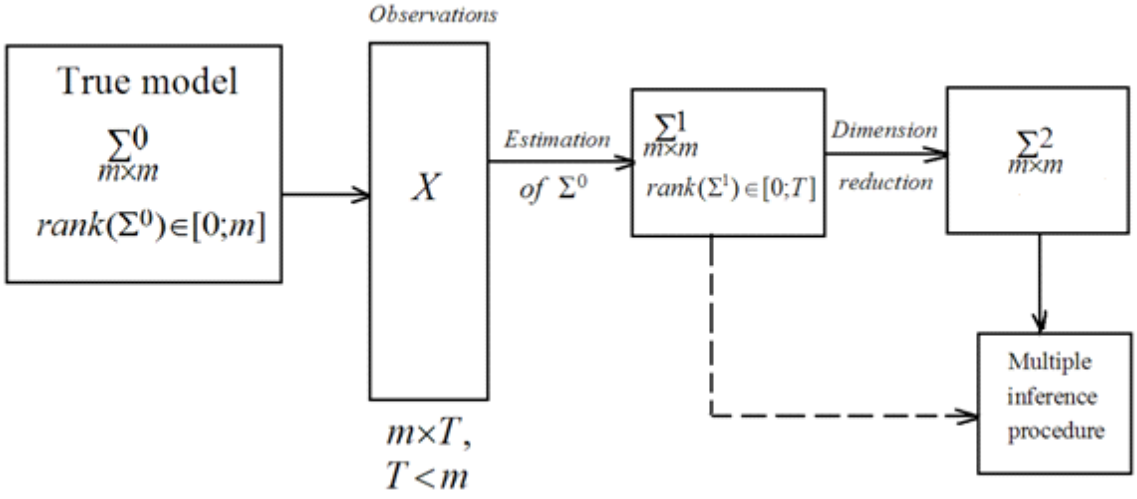

Figure 2.2.2 Estimation of the dependence structure for the purpose of multiple inference

Figure 2.2.2 dispalys a general scheme for the estimation of the dependence structure. For instance, the bootstrap approach of White (2000) corresponds to using the “crude” estimate, Σ^1. Jones and Shanken (2005) take one more step,

dimension reduction, and obtain PCA-based Σ^2 (even though they do not use it for the purpose of multiple testing). It is also possible to use Rigde Regression based $\Sigma^2 = (XX' + \omega I)$ where *I* is the identity matrix and ω is a positive smoothing parameter. For PCA, the smoothing parameter is the number of retained principal components, *p*.

Unfortunately, both parametric and non-parametric modeling of the dependence structure appear to have a fundamental problem: they only work when the utilized estimate, be it Σ^1 or Σ^2, is a "good" estimate of the true variance-covariance matrix, Σ^0. To put it in strict terms, the asymptotic results of Yekutieli and Benjamini (1999) and White (2000) state that the control of FDR is attained only asymptotically, for a fixed *m* and $T \to \infty$.

To look into this in more detail, let us consider White (2000) whose non-parametric approach is the foundation of so-called StepM procedure developed later by Romano and Wolf (2005) and Romano et al. (2008). White (2000) considers the following problem: suppose there are *m* forecasting strategies. For each strategy, its predictions are compared to those of a "naïve" strategy. The corresponding statistic is greater than zero when the "naïve" strategy is worse. For the best strategy (with the largest statistic), what is its p-value after multiplicity adjustment?

Suppose the statistics are multivariate normal with a $m \times m$ variance-covariance matrix Ω. Then we can get the desired p-value based on the distribution of the extreme value of m-dimensional $N(0, \Omega)$ vector. Note that even for a given Ω, the analytical expression for the distribution of extreme value is unknown. However, the proposed bootstrap procedure conveniently provides both $\hat{\Omega}$ and the cdf estimate for the extreme value.

This reasoning makes it perfectly clear that one assumes that there are enough data to obtain a good estimate of $m \times m$ variance-covariance matrix Ω. All theoretical results are derived for a fixed number of tests and large sample size (*m* is fixed and $T \to \infty$ in (2.1.3)), but in practice it is obvious that the "large enough" value of T depends on the value of m. In particular, it certainly makes little sense to rely on asymptotic results unless T is many times as large as m. Unfortunately, this crucial rule of thumb is obscured in practice because the bootstrap in StepM and similar procedures do not directly involve the estimation of variance-covariance matrix and, technically, can produce a result even when m is larger than T.

Ths "size problem" by itself has received lots of attention. Fan et al. (2008) provide simulation results that demonstrate the inadequacy of a variance-covariance matrix estimator when the data are insufficient. Romano et al. (2008) admit that the StepM procedure is similar to the approach developed in BioStatistics by Van der Laan et al. On the other hand, Efron (2006D, Section 6) refers to the work of van der Laan et al. to emphasize that the corresponding results are applicable only asymptotically and are of very limited use for a typical large-scale simultaneous inference problem.

When the data are insufficient, the researcher often has no choice but to hope that, somehow, his estimate of the dependence structure is still not far from the truth. For instance, Yekutieli and Benjamini (1999) give a weather analysis example where m = 1977 and T = 39. Remarkably, when they used another, simulated dataset to show that FDR is controlled they have to set m = 40 and T in between 200 and 1000.

In the context of MF studies, we have m about 2000 and T between 100 and 300, which amounts to a severe "size problem". Note that while the rank of Σ^0 may be anywhere between 300 and 2000, the rank of Σ^1 is always under 300. That is, we

know so little about Σ^0 that we cannot even provide a reasonable estimate of its rank, let alone more delicate statistical properties such as PRDS. The various dimension reduction techniques allow us to "reduce the dimension" of the available data (i.e., use the available data efficiently), but they do not solve the "size problem".

Yet another way to handle the dependence is the assumption of "weak dependence" outlined in Storey, Taylor, and Siegmund (2004), Storey and Tibshirani (2003), and Storey (2003). When the assumption is satisfied, the p-values are treated as if independent and the (asymptotic) FDR control still takes place.

Unfortunately, there is no statistical procedure to test for weak dependence, even though one could make a qualitative argument that it holds for particular datasets. For instance, it is likely to hold when the test statistics are dependent (if at all) within small groups with the groups being independent of each other. In particular, Storey and Tibshirani (2003) provide a qualitative argument for weak dependence assumption being true for some (but not all) microarray gene-expression datasets: genes behave dependently in "pathways" (small groups) with pathways being independent of each other. To demonstrate FDR control, Storey, Taylor, and Siegmund (2004) give a simulated example with m = 3000 and the group size of 10. They also show that under weak dependence FDR can be controlled for any fixed value of $\hat{\lambda}$ in (2.2.3). The choice of optimal $\hat{\lambda}$ is a bias-variance tradeoff problem which they solve via bootstrapping from the *m* p-values. Resampling from a set of (weakly) dependent p-values is a questionable technique and no analytical justification for that was ever developed; still, some numerical examples show that the bootstrap estimation of $\hat{\lambda}$ is robust under "small group" type of weak dependence (Storey and Tibshirani (2001)).

Correspondingly, the application of FDR in BSW study rests on the assumption of weak dependence for the purpose of both FDR control and the estimation of the optimal $\hat{\lambda}$ via bootstrap method. The same is true for the study of Otamendi et al. (2008), which is based on pFDR, a slight modification of FDR introduced in Storey (2002).

At first sight, it seems reasonable for BSW to assume that MF operate in small independent groups and the dependence between the estimated performance measures ($\hat{\alpha}_i's$ in this case), if any, can exist only within a group. However, there are certain reservations to utilizing this convenient assumption. As stated in BSW study itself, MF may exhibit correlated trading behaviors in large groups that can be caused, for instance, by being exposed to the same industrial sector or "herding" into particular stock(s). In the MF context, a natural candidate for a "small group" of funds is a fund family, with families being hopefully independent of each other. However, the findings of Wermers (1999) suggest that "herding" is not significantly less among different fund families than it is among funds within a family. While the absolute magnitude of "herding" is low, its qualitative nature shows that common sense-based qualitative assumptions w.r.t. the dependence structure may be not true at all.

Note that the mutual independence of $\hat{\alpha}_i's$ and their $p_i's$ is in no way implied by the model (2.1.3) itself. When a multifactor asset pricing model is perfectly specified, the asset returns are not forecastable, meaning that the residual terms $\varepsilon_{i,t}$ are not serially correlated. In that case, $\varepsilon_{i,t}$ can still be very well correlated cross-sectionally, i.e. across $i = 1, m$ for a fixed t (see Cochrane (2005)). It means that in case of the perfectly specified and estimated model the null p-values $p_i's$ can marginally follow the pre-specified null distribution (e.g., U(0,1)) and be cross-sectionally correlated at the same time.

BSW put a sizable effort into justifying the weak dependence assumption for their study. First of all, BSW argue that the funds' alphas are not very dependent because 15% of the fund histories in their sample do not overlap in time, and on average only 55% of return observations overlap. For funds i and j, non-overlapping of returns means that, given model (2.1.3), the estimates $\hat{\alpha}_i$ and $\hat{\alpha}_j$ and the corresponding p_i, p_j are not correlated (under another mute assumption that there is no serial correlation in error terms $\varepsilon_{i,t}$ and $\varepsilon_{j,t}$). How much independence does the "lack of overlap" introduce? Compare this to an example of a weakly dependent structure with *m*=3000 and the group size of 10 in Storey, Taylor, and Siegmund (2004). If we translate it into MF setting with m=2000, where the degree of independence is associated with the absence of overlap, we obtain the following: the entire time period should be divided into subintervals with only 10 funds observed on each subinterval. Hence, it requires 200 subintervals. Given that an average fund is observed for over 10 years, it implies the study's time span has to be over 2000 years. In reality, BSW data span only 32 years, which makes the "lack of overlap" argument doubtful. Besides, for a shorter time period (like in this study) the overlap has to be much greater than 55% while the number of funds is about the same. In fact, our data span 14 ½ years with an average of 10 ¾ return-years per fund.

BSW (05/2007 version) present two simulated examples to show that their multiple inference procedure works even when $\hat{\alpha}_i$'s have a non-trivial (but pre-specified) correlation structure. The first example is similar to the abovementioned weak / "small group" dependence simulation study of Storey and Tibshirani (2001). In particular, the simulated correlation matrix in BSW example has 30 non-zero blocks that comprise only about 5% of all elements in the correlation matrix. Therefore, it is not surprising that BSW multiple testing procedure (which ignores dependence) still produces reasonable multiple inference results.

The second example is based on the method of including the “residual factors” in the right-hand side of performance evaluation model (2.1.3) for the purpose of “whitening” the residual terms cross-sectionally. These factors in BSW are indicators of whether the fund has a zero, positive, or negative performance.

Since the latent residual structure in BSW was simulated, there is no proof that the real structure is in any way close to it. Note that if we were to try to prove that, for instance, the correlation coefficients are the same within the same investment objective, we would have deal with a much larger-dimensional problem. When we test that all $\alpha's$ of 100 MF of the same investment objective are equal to the same constant (such as zero), we have to know the dependence structure for the corresponding vector of estimates, $(\hat{\alpha}_i,\ i=1,100)$. It is usually approximated by 100×100 variance-covariance matrix. Now, suppose that we also want to test

$$H_0 : \rho_{ij} = \rho,\ \ i,j = 1,100\ \ i \neq j \qquad (2.2.4)$$

where ρ_{ij} is the correlation between $\varepsilon_{i,t}$ and $\varepsilon_{j,t}$ in (2.1.3) (it is assumed constant w.r.t. time). Similarly, (2.2.4) is a joint hypothesis test w.r.t. 4950 fixed parameters. In order to do it properly, one would have to be given a 4950×4950 variance-covariance matrix for the vector $(\hat{\rho}_{i,j},\ i \neq j)$.

In yet another example, BSW (05/2008 version) actually try to estimate the residual variance-covariance matrix of size 898*898 based on 898 funds observed for 60 months (2002-2006) in order to use it for dependence sensitivity analysis. The rank of such cross-product matrix is 60 at most and it cannot provide a more or less good estimate of the variance-covariance matrix. It would have taken at least 898 months of data (almost 75 years) just to make the 898*898 cross-product matrix non-singular. That can only be simplified via imposing some restrictions on the correlation structure which takes us back to the examples above.

For what it is worth, in that estimated matrix the pairwise correlation term has 25%, 50% and 75% quantiles equal to –0.09, 0.05, 0.19 with the mean of 0.08 (not too far from zero), which is another argument used in BSW to justify the weak dependence assumption. However, the seemingly close-to-zero range of pairwise correlation does not necessarily imply the weak dependence property. This particular issue will be considered in more detail in Section 3.2.

Therefore, a large-scale MF study being a high-dimensional observational study, the weak dependence property inevitably implies some rather questionable and/or hard-to-check assumptions about the data dependence structure. Explicit modeling of the high-dimensional correlation structure is not feasible either , unless, yet again, one is willing to tolerate a number of probably unrealistic assumptions. Moreover, even fairly restrictive assumptions may not reduce the number of estimated parameters to the point where the amount of available data appears enough for estimation.

There is one more source of error in a multiple inference procedure: even when independence or small-group dependence hold in theory, the multiple test procedure works with the estimated $\hat{\alpha}_i's$ and $p_i's$. The estimated $p_i's$ can correspond to null cases and at the same time they may deviate from the assumed null distribution. Efron (2006C) describes some “technical” causes of why that can happen in microarray studies. It is likely to take place when the model used to obtain the individual test statistics and p-values is misspecified and/or improperly estimated in some way. That can occur in MF studies just as well.

Possible sources of misspecification in a model like (2.1.3) are: using an inappropriate correlation structure for the error terms; failing to account for the temporal heteroskedasticity of the error terms; applying asymptotically valid results when the sample size (T in 2.1.3) is not large enough. For instance,

omitting an important (but unknown) factor in the right-hand side of the model can induce the serial correlation of error terms which may remain unaccounted for. It is also likely to cause dependence in estimated performance measures across all funds that have significant loadings on the omitted factor (BSW). The number of such funds can be quite large. Applying robust estimation methods (e.g., non-parametric bootstrap) can take care of some of these problems, but such methods are not bulletproof.

If any of these inconsistencies take place, they may result in the marginal distributions of null $p_i's$ being far away from U(0,1). Even if such $p_i's$ are independent, their ensemble is not going to behave like i.i.d. U(0, 1). In some cases, their behavior resembles that of dependent and marginally U(0, 1) $p_i's$ (see Section 3.2 for examples). Thus, as a result of model misspecification, even independent $p_i's$ can be seen as dependent “in effect”. In practice, both “genuine” dependence and the misspecification of marginal distribution are likely to be present. While one can try to ignore the former via justifying the independence / weak dependence assumption, the contribution of the latter is impossible to assess a priori, at least in a large-scale situation.

There is no argument that knowing the dependence structure of test statistics is sufficient to perform a multiple inference procedure. But what if it is not necessary? Given the “size problem” in MF studies, it would be very desirable to avoid the the modeling of high-dimensional dependence structure. The next section introduces a novel approach to large-scale simultaneous inference that can help us circumvent both the weak dependence assumption and the explicit modeling of high-dimensional correlation structure.

CHAPTER 3. LOCAL FALSE DISCOVERY RATE

3.1. Local false discovery rate: definition and properties

Suppose that for the model (2.1.3) we compute the individual one-sided p-values for the test:

$$H_i^0 : \alpha_i = 0 \text{ vs. } H_i^a : \alpha_i > 0 \tag{3.1.1}$$

The obtained p-values, $\{p_i\},\ i = 1, m$ are converted to normal z-scores:

$$z_i = \Phi^{-1}(1 - p_i) \tag{3.1.2}$$

where $\Phi^{-1}(.)$ is the inverse normal cdf. For instance, p_i = 0.025 corresponds to the fund that is likely to outperform and its z_i will be 1.96; if, on the other hand, p_i = 0.975 (obtained from a negative α_i) the fund is likely to underperform and its z_i will be -1.96.

Efron (2004) proposed the following structural model that ties together α and z values:

$$\begin{aligned} &\alpha \sim g(\alpha) \\ &z \mid \alpha \sim N(\alpha, \sigma_0^2) \\ &f(z) = g(\alpha) \ * \ N(0, \sigma_0^2) \end{aligned} \tag{3.1.3}$$

where $g(\alpha)$ is an arbitrary distribution and " * " denotes convolution. Our interest is in testing some hypothesis about α, and the support of $g(\alpha)$ can be arbitrarily split into two disjoint "null" and "non-null" sets. Then, $g(\alpha)$ itself will be a sum of two terms:

$$g(\alpha) = p_0 g_0(\alpha) + p_1 g_1(\alpha) \qquad (3.1.4)$$

where

$g_0(\alpha)$ – "null" component

$g_1(\alpha)$ – "non-null" component

$p_0 = P_g\{\alpha \text{ is null}\}$

$p_1 = P_g\{\alpha \text{ is non-null}\}$

$p_0 + p_1 = 1$

In terms of corresponding z-values this will result in:

$$f_0(z) = g_0 * N(0, \sigma_0^2) \text{ - null density of z's} \qquad (3.1.5)$$

$$f_1(z) = g_1 * N(0, \sigma_0^2) \text{ - non-null density of z's}$$

$$f(z) = p_0 f_0(z) + p_1 f_1(z) \text{ - mixture density of z's}$$

For instance, the "null" set can consist of one point $\{\alpha = 0\}$ ($g(\alpha)$ does not have to be absolutely continuous) and the "non-null" set is the corresponding complement $\{\alpha \neq 0\}$. Also, $p_0 = P_g\{\alpha = 0\}$ and under $\sigma_0^2 = 1$ the null subdensity $f_0(z)$ is N(0,1). This particular case of the structural model corresponds to the setting from Section 2.2: p_0 is the same as m_0 / m and if all null p-values are i.i.d. U(0,1), the corresponding null density $f_0(z)$ is $\Phi^{-1}(U(0,1))$ which is nothing but N(0,1).

Our inference will utilize the Bayesian concept of "local false discovery rate" (fdr) introduced in Efron (2001). It can be interpreted as a "local" version of Benjamini and Hochberg's FDR and it is defined as follows:

$$fdr(z) = P\{\text{case i is null} \mid z_i = z\} = \frac{p_0 f_0(z)}{f(z)} \qquad (3.1.6)$$

Local fdr, $fdr(z)$, is the posterior probability that the test with corresponding z-score came from the null distribution $f_0(z)$. One can also define

$$Fdr(z) = P\{\text{case i is null} \mid z_i \leq z\} = \frac{p_0 F_0(z)}{F(z)} \qquad (3.1.7)$$

where F_0 and F are cdf's corresponding to f_0 and f. $Fdr(z)$ is a closely related Bayesian version of Benjamini and Hochberg's FDR and the connection between the two is detailed in Efron(2002). Both Fdr and FDR are of tail-area type; in particular,

$$Fdr(z) = \frac{\int_{-\infty}^{z} fdr(t) f(t) dt}{\int_{-\infty}^{z} f(t) dt} = E_f[fdr(t) \mid t \le z] \tag{3.1.8}$$

Thus, FDR and Fdr characterize the average false discovery rate within a tail region. On the other hand, fdr has a local nature and provides more precision in interpreting $z_i's$ on an individual basis which is an obvious advantage of fdr.

The second advantage of this approach is that neither (3.1.6) nor (3.1.7) assume any particular dependence structure of z's such as PRDS of Benjamini and Yekutieli (2001) or the weak dependence assumption of Storey, Taylor, and Siegmund (2004) and BSW.

There are two ways that such flexibility is paid for: first, (3.1.6) is "one-at-a-time" statement: if we are given $fdr(z)$ and then observe two dependent values z_1 and z_2, $fdr(z_1)$ is not conditioned on z_2 and, if the probability structure of the entire vector $\vec{Z}$ were known, then $P\{\text{case 1 is null} \mid \vec{Z} = \vec{z}\}$ could be very different from $fdr(z_1)$ (Efron (2004)). Therefore, local fdr method is appropriate for the applications where the entire probability structure is not only unknown but also quite unknowable (Efron (2005), Section 2). In the context of a large-scale MF study, the estimation of a high-dimensional dependence structure is severely hindered by the lack of data (Section 2.2) and that is the reason why the word "unknowable" seems to apply to it quite well.

Secondly, the local fdr approach is that of "empirical Bayes" kind: in (3.1.5) we do not pre-specify the mixture density f(z) (which is an advantage) because, unlike in the "classical Bayes" setting, f(z) is estimated from the data. If the numerator in (3.1.6 – 3.1.7) is somehow pre-specified, all we need is a consistent estimator of f(z). This, however, adds a certain amount variability to our estimates of fdr(z) and Fdr(z). This is especially relevant for fdr(z) because in order to estimate f(z) properly one needs a large number of observations (at least a few hundred).

Nevertheless, in certain cases it makes sense to estimate p_0 and $f_0(.)$ from the data also, which is the subject of the next section.

3.2. Empirical null hypthesis

Under standard FDR approach from Section 2.2, the null density $f_0(.)$ is pre-specified as $N(0,\sigma_0^2=1)$) while the ratio m_0/m (equivalent of p_0) is estimated from the data. Efron (2003, 2004, 2006C, D) introduced the concept of "empirical null" where $f_0(z)$ is approximated by $N(\delta_0,\ \sigma_0^2)$ and the parameters $p_{0,}\ \delta_0,\ \sigma_0^2$ are estimated from the data also.

One may ask why not specify $f_0(z)$ a priori, e.g. $f_0(z) \sim N(0,1)$ ("theoretical null"). To understand that, note that such $f_0(z)$ is based on Assumptions 1 and 2 from Section 2.2, i.e. if the null p-values are i.i.d. U[0,1] then the corresponding z-scores are i.i.d. $N(0,1)$. As underlined in Section 2.2, either of these two assumptions can be violated. If the null p-values are not marginally U(0,1) because of model (2.1.3) misspecification, the corresponding z-scores will not behave like i.i.d. N(0,1) even if they are independent. In that case (under independence) we can see that by making σ_0 a free parameter in (3.1.5), the

model accounts for the case when the marginal distribution of null z-scores is $N(0,\sigma_0^2)$ instead of $N(0,1)$.

On the other hand, if the null p-values are marginally uniform (model (2.1.3) is well specified) but dependent, the corresponding z-scores will not behave like i.i.d. N(0,1) either. In practice, both of these forces are likely to be at work and, as a result, the histogram of null z-scores can be quite different from that of $N(0,1)$ distribution.

Efron (2006D) provides an explicit example of how the correlation structure can affect the inference. Suppose that z's are marginally N(0,1) , that is, all z's are null. Each pair (z_i, z_j) is bivariate normal with a distinct correlation coefficient ρ_{ij} drawn randomly from a certain normal distribution $N(0,\tau^2)$. Further, let A be a single independent realization (called "dispersion variate") from $N(0,\tau^2)$. It can be shown that the ensemble of all z-values will behave closely to an ensemble of i.i.d. $N(0,\sigma_0^2)$ where $\sigma_0^2 = 1+\sqrt{2}A$. The positive realizations of A produce $\sigma_0^2 > 1$ ("overdispersion") and the negative realizations of A produce $\sigma_0^2 < 1$ ("underdispersion").

On the other hand, the ensemble of i.i.d. $N(0,\sigma_0^2)$ can be seen as a family of independent z-values coming from a misspecified performance evaluation model that produces null z's that are marginally $N(0,\sigma_0^2)$ instead of $N(0,1)$. The result above implies that such z's can be treated as marginally N(0,1) and dependent with the correlation density $\rho \sim N(0,\ \tau^2)$.

Efron (2006C) showed that, in this example, not only the point estimate of fdr(z) but also its estimated standard error, $s.e.(\hat{fdr}(z))$, are conditioned on the ancillary

statistic A, and, in that sense, are conditioned on the dependence structure of z's. Likewise, the standard errors of $\hat{p}_0, \hat{\delta}_0, \hat{\sigma}_0$ are also conditioned on the dependence structure.

In this example, using the empirical null is essentially a way to adjust the inference for the dependence structure of z's without having to model it explicitly. In addition, the empirical null takes into account the possible misspecification of the marginal distribution of null p-values. If there is strong evidence against the theoretical null, the empirical null has to be considered. Note that the usage of empirical null increases the variability of the estimates of fdr(z) and Fdr(z), and whether or not it is worth using is a bias-variance tradeoff question.

Based on model (2.1.3), one could roughly estimate the density of ρ based on the empirical distribution of pairwise correlations in the residual variance-covariance matrix. BSW estimated the 898*898 cross-product matrix and found the estimated 25%, 50% and 75% quantiles for ρ are equal to –0.09; 0.05; 0.19, correspondingly. Since each pairwise correlation was based on only 60 observations, the sampling error must have added some variability (see Efron (2006D, Remark A)). For the sake of argument, suppose that the three quantiles of true ρ are -0.09; 0.0; 0.09 and ρ is normal, which implies $\rho \sim N(0,\ \hat{\tau}^2 = 0.133^2)$.

To see how this can affect the inference, we introduce another version of (3.1.7):

$$\tilde{F}dr\ (x \mid A) = P\{z_i \text{ null} \mid z_i \geq x,\ A\} \qquad (3.2.1)$$

Thus, $\tilde{F}dr(x \mid 0)$ corresponds to the inference made under z's being independent, i.e. under the theoretical null. Suppose we are interested in detecting the positive performers, so set x = 2.5. The following plot shows the ratio of $\tilde{F}dr(x = 2.5 \mid A)$ to $\tilde{F}dr(x = 2.5 \mid 0)$ as a function of A.

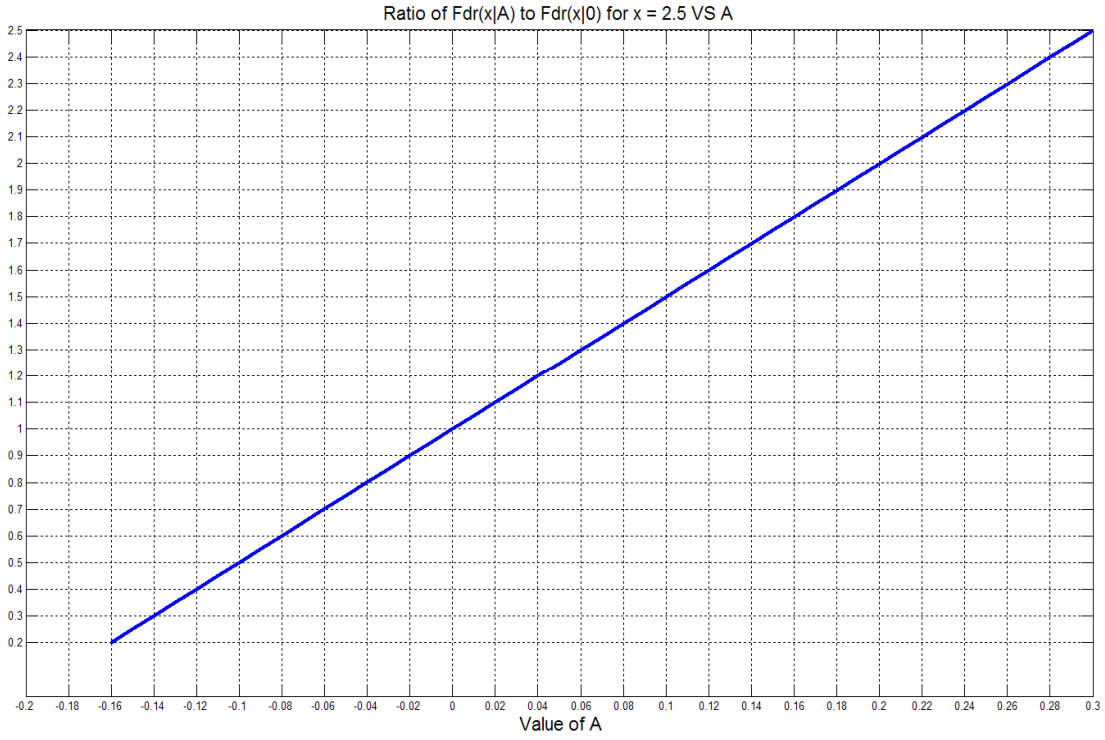

Figure 3.2.1 The ratio of Fdr(x | A) to Fdr(x | 0) as a function of A

For instance, if A took on the value of 0.16 (just 1.2 standard deviations from the mean of zero), the proportion of null z's in the tail region $\{z > 2.5\}$ is about 1.8 times as great as it is under A = 0 (theoretical null). Suppose $\tilde{F}dr(2.5 \mid 0)$ is 0.2, then $\tilde{F}dr(2.5 \mid 0.16)$ is 0.36. If 100 of z's fall above 2.5, 80 of them are "true discoveries" under the theoretical null, but under A = 0.16 the number of true discoveries is only 64.

If A = -0.16 then the proportion of null z's in the tail region $\{z > 2.5\}$ is five times less than that number under the theoretical null. In the example above, $\tilde{F}dr(2.5 \mid -0.16)$ = 0.04 and 96 out of 100 z's above 2.5 are true discoveries as opposed to only 80 under the theoretical null.

This illustrates that even a seemingly close-to-zero range of ρ can substantially bias the inference. If one chooses to use the theoretical null $f_0(.) = N(0,1)$ for overdispersed z's, too many null cases will be declared significant. On the other hand, using the theoretical null for underdispersed z's will ignore a lot of non-null cases. Apparently, the inference has to be adjusted for the estimated value of A. It is achieved through using the empirical null $N(0,\sigma_0^2)$ where $\sigma_0^2 = 1+\sqrt{2}A$. For this example, the empirical nulls are $N(0,\sigma_0^2 = 1.226)$ and $N(0,\sigma_0^2 = 0.774)$ for A = 0.16 and A = -0.16, correspondingly.

The advantage of the empirical approach can be summarized as follows: what we really need to know to be able to perform multiple inference is not the dependence structure per se, but the null component, $p_0 f_0(z)$. When we estimate the dependence structure explicitly, it is not immediately clear whether our method of modeling is adequate for the purpose of multiple testing. When the "size problem" (Section 2.2) is present, we know very little about the true dependence structure and it is hard to verify the weak dependence / independence assumption for test statistics. On the other hand, $p_0 f_0(z)$ is described by a small number of parameters that, by construction, are of direct relevance to our goal. Therefore, modeling $p_0 f_0(z)$ directly is a logical short-cut one may choose when the data allow for that. If the number of tests is large, we can obtain the information about $p_0 f_0(z)$ directly from the observed z-scores. In that sense, the parameters of empirical null do capture the main effect relevant to our ultimate goal, multiple inference (see Efron (2006C, D)).

Note that, as the number of tests, m, goes up, the performance of "explicit" approach deteriorates because the "size problem" (Section 2.2) becomes more severe. With the "empirical" approach, it is just the opposite: the larger m, the more precise is the estimation of $p_0 f_0(z)$.

While the theoretical null is always the first option to try, the abovementioned findings of Efron suggest that it is also worth checking whether there is strong evidence against the theoretical null. If that is the case, switching to the empirical null can be a justifiable option.

3.3. Parameter estimation

The numerical results in this study are obtained based on the R package *locfdr* which implements the fdr-based method of Efron.

Regardless of whether the empirical or theoretical null is used, the estimation of the parameters of null component, $p_0 f_0(z)$, is based the "zero assumption": it is assumed that only the null component is supported on a certain "zero interval" $(z_-; z_+)$. The parameters of interest are estimated with either MLE or so-called central matching (CME) (Efron (2006C)). The interval $U(\lambda, 1)$ from Section 2.2 corresponds to a symmetrical zero interval: e.g., U(0.05; 1) corresponds to the zero interval (-1.96; 1.96) . The following formula shows the relation between λ from (2.2.2), z_- and z_+:

$$\lambda = \Phi(z_-) + (1 - \Phi(z_+)) \qquad (3.3.1)$$

$\Phi(.)$ - standard normal cdf

For the theoretical null and a fixed zero interval, the point estimate of p_0 is the same in BSW method (formula (2.2.3)) and Efron's approach. If the empirical null is chosen, $f_0(z)$ can be approximated by a parametric distribution, such as symmetrical normal $N(\delta_0, \sigma_0^2)$ or skewed split-normal $SN(\delta_0, \sigma_1^2, \sigma_2^2)$. Fitting a heavy-tailed null distribution may be problematic in the sense that in order to fit the tail, one would have to expand the zero interval to the point where too many non-null z-values are included.

An additional restriction $p_0 \geq 0.9$ has to hold when we use the empirical null. Efron (2003) provides theoretical and numerical results that justify the restriction: if $p_0 \geq 0.9$ and the theoretical null is valid, then the MLE/CME estimates of δ_0 and σ_0 have to be very close to 0 and 1, respectively. If they are not, it implies that the theoretical null is inadequate. If $p_0 < 0.9$ then the estimates of (δ_0, σ_0) can be significantly different from (0, 1) even when the theoretical null is valid. Hence, if one wants to distinguish between the two types of nulls, first he has to make sure that $p_0 \geq 0.9$.

The choice of the zero interval itself is a bias-variance tradeoff problem: for a large interval, the estimate of p_0 (and, if applicable, the parameters of the empirical null) have low variance but a high bias since many non-null cases are likely to fall into the wide zero interval. For a narrow zero interval, the bias is small, but the estimates of p_0 and other parameters have large variance. The value of λ or the boundaries of $(z_-; z_+)$ are the corresponding smoothing parameters. BSW minimize $MSE(\hat{p}_0)$ using λ as a smoothing parameter. For a fixed λ, $MSE(\hat{p}_0)$ is calculated based on rather questionable bootstrap technique (see Section 2.2) and we are not going to use it for this study.

Instead, consider the error of $\hat{p}_0 \hat{f}_0(z)$ scaled by $1/f(z)$:

$$Error(z) = \frac{1}{f(z)}\left[p_0 f_0(z) - \hat{p}_0 \hat{f}_0(z) \right] \qquad (3.3.2)$$

The optimal zero interval is where the integrated $MSE(Error(z))$ is at the minimum, so we have to estimate the squared bias and variance. The locfdr package does not provide a direct estimate of $MSE(Error(z))$, and we are going to use some proxies to obtain the shape of bias-variance tradeoff curve.

First, we use the bias on the zero interval as a proxy for overall bias. On the zero interval we have

$$p_0 f_0(z) = f(z) \tag{3.3.3}$$

$$Error(z) = \frac{1}{f(z)}\left[p_0 f_0(z) - \hat{p}_0 \hat{f}_0(z) \right] = 1 - \frac{\hat{p}_0 \hat{f}_0(z)}{f(z)}$$

The mixture density $f(z)$ is unknown, but the expected error can be estimated by using an unbiased estimator of $f(z)$ which is obtained in *locfdr* via Poisson regression over the entire z axis. The estimator, $\hat{f}(z)$, is consistent even when z-scores are dependent (see Efron (2004, 2005)). The *locfdr package* also produces the estimates $\hat{fdr}(z)$ and $Var[\log(\hat{fdr}(z))]$.

As a result, the estimate of average squared bias is:

$$\hat{Bias}_\lambda^2 = \frac{1}{z_+ - z_-} \int_{z_-}^{z_+} (1 - \hat{fdr}(z))^2 \, dz \tag{3.3.4}$$

The error variance at point z will be

$$Var[Error(z)] = Var[\hat{fdr}(z)] \tag{3.3.5}$$

We are going to use the available $Var[\log(\hat{fdr}(z))]$ instead and then get the estimate of overall variance as:

$$\hat{Var}_\lambda = \int_{-\infty}^{\infty} Var[\log(\hat{fdr}(z))] dz \tag{3.3.6}$$

For the theoretical null, $f_0(z)$ is not estimated. $Var(\hat{f}(z))$ does not depend on λ and its magnitude is much larger than that of $Var(\hat{p}_0 \cdot N(0,1))$. For that reason, we are going to use $Var_\lambda(\hat{p}_0)$ instead of (3.3.6) for the theoretical null.

For the empirical null, we are using the full version (3.3.6). In that case, *locfdr* produces $Var[\log(\hat{fdr}(z))]$ where both numerator and denominator of $\frac{\hat{p}_0 \hat{f}_0(z)}{\hat{f}(z)}$

are considered random and $\hat{f}(z)$ can be seen as a random weight function. For the empirical null, the numerator strongly dominates the denominator and $\hat{Var}_\lambda$ is proportional to

$$\int_{-\infty}^{\infty} Var(\hat{p}_0 \cdot \hat{f}_0(z))dz \tag{3.3.7}$$

Because $\hat{Var}_\lambda$ and $\hat{Bias}_\lambda^2$ are not on the same scale, we divide each estimate by its median over the range of the smoothing parameter to get the value of bias-variance tradeoff, BVT_λ:

$$BVT_\lambda = \frac{Var_\lambda}{median_{\lambda'}(Var_{\lambda'})} + \frac{Bias^2{}_\lambda}{median_{\lambda'}(Bias^2{}_{\lambda'})} \tag{3.3.8}$$

BVT_λ is not equal to the equal to the integrated $MSE(Error(z))$, but it estimates the shape of MSE curve (see Storey and Tibshirani (2001)).The optimal value of λ is determined by minimizing BVT_λ over the range of λ. An alternative zero interval choice procedure based on $MSE(\hat{p}_0\hat{f}_0(z))$ is developed in Turnbull (2007) but the corresponding software is not publicly available.

Let us return to the example from Section 3.2 where marginally $N(0,1)$ z-values are correlated with the correlation density $\rho \sim N(0,\ \tau^2)$. In that case, if empirical null is used, the $\hat{fdr}(z)$ and $Var[\log(\hat{fdr}(z))]$ are adjusted for the dependence among z's in the sense that both estimates are conditioned on the value of dispersion variate A (Efron(2006C)). In that sense, these estimates are more adequate than the bootstrap estimate of variance used in BSW. However, if the theoretical null is used, the variance estimator, strictly speaking, works only under the independent z's which makes it akin to the bootstrap estimator of BSW.

Efron's method and *locfdr* package do not distinguish between significant z-values that are positive and significant z-values that are negative. For MF study, it is necessary to make that distinction because we need to separate outperformers from underperformers. Suppose all p-values are converted to corresponding z-scores via (3.1.2). The structural model (3.1.3 – 3.1.5) can be slightly modified as follows:

$$\alpha \sim g(.) \tag{3.3.9}$$
$$z \mid \alpha \sim N(\alpha, \sigma_0^2)$$
$$g(\alpha) = p_0 g_0(0) + p_1^+ g_1^+(\alpha) + p_1^- g_1^-(\alpha)$$
where
$$p_0 = P_g\{\alpha = 0\}, \quad p_1^+ = P_g\{\alpha > 0\}, \quad p_1^- = P_g\{\alpha < 0\}$$
$$g_0(\alpha) - \text{"zero" density equal to delta function}$$
$$g_1^+(\alpha) - \text{"positive" density with support on } \{\alpha > 0\}$$
$$g_1^-(\alpha) - \text{"negative" density with support on } \{\alpha < 0\}$$

In terms of z-values, we have

$$f(z) = p_0 f_0(z) + p_1 f_1(z) \quad \text{- mixture density of z's} \tag{3.3.10}$$
$$p_1 f_1(z) = p_1^+ f_1^+(z) + p_1^- f_1^-(z)$$
where
$$f_0(z) = N(0, \sigma_0^2) \quad \text{- "zero" density of z's}$$
$$f_1^+(z) = g_1^+ * N(0, \sigma_0^2) \quad \text{- "positive" density of z's}$$
$$f_1^-(z) = g_1^- * N(0, \sigma_0^2) \quad \text{- "negative" density of z's}$$
$$p_1^+ + p_1^- = p_1, \quad p_0 + p_1 = 1$$

The *locfdr* package produces the estimate of $p_1 f_1(z)$, but its decomposition into positive $p_1^+ f_1^+(z)$ and negative $p_1^- f_1^-(z)$ components is not identified. However, note that $f_1^-(z)$ is a (possibly continuous) mixture of normal densities

$$f_1^-(z) = g_1^-(\alpha) * N(0, \sigma_0^2), \quad \alpha < 0 \tag{3.3.11}$$

All normal densities in the mixture have negative means. Hence, $f_1^-(z)$ is non-increasing for z > 0. Typically, the estimation produces $\hat{fdr}(z) = \dfrac{\hat{p}_0 \hat{f}_0(z)}{\hat{f}(z)} = 1$ in

some interval $(-l;l)$ such as (-0.4; 0.4). It implies that $\hat{f}_1(z)$, $\hat{f}_1^-(z)$ and $\hat{f}_1^+(z)$ are equal to zero on $(-l;l)$. Hence, $\hat{f}_1^-(z)$ cannot have support for $z > l$ and $\hat{f}_1^-(z) = 0 \ \forall z > 0$. Similarly, $\hat{f}_1^+(z) = 0 \ \forall z < 0$.

Therefore, while in theory some $\alpha < 0$ can produce z > 0, the practical estimation procedure implies that z > 0 can be produced only by $\alpha \geq 0$ and z < 0 can only be produced by $\alpha \leq 0$. Then, for z > 0 we may formally define "positive" fdr as

$$fdr_+(z) = \frac{p_0 f_0(z) + p_1^- f_1^-(z)}{f(z)} \tag{3.3.12}$$

but since $f_1^-(z) = 0 \ \forall z > 0$, $fdr_+(z)$ will be the same as $\hat{fdr}(z)$ produced by *locfdr* package.

Then, the value of p_1^+ is estimated as follows:

$$\hat{p}_1^+ = \frac{\int_0^\infty [1 - \hat{fdr}(z)]dz}{\int_0^\infty \frac{\hat{f}_1^+(z)}{\hat{f}(z)} dz} \tag{3.3.13}$$

and similarly for $\hat{p}_1^-$, where integrals are computed as corresponding sums.

For two-component model, $s.e.(\hat{p}_0) = s.e.(\hat{p}_1)$, but as of now we don't have a way to get $s.e.(\hat{p}_1^+)$ and $s.e.(\hat{p}_1^-)$. Let us assume that

$$s.e.(\hat{p}_1^+) = s.e.(\hat{p}_1^-) = \kappa \quad \text{and} \quad corr(\hat{p}_1^+, \hat{p}_1^-) \leq 0 \tag{3.3.14}$$

then

$$Var[\hat{p}_0] \leq 2\kappa^2 \Rightarrow \kappa \geq s.e.(\hat{p}_0)/\sqrt{2} \tag{3.3.15}$$

Unless stated otherwise, the lower bound for κ will be reported instead of $s.e.(\hat{p}_1^+)$ and $s.e.(\hat{p}_1^-)$ whenever the three-component model (3.3.10) is used.

CHAPTER 4. US MUTUAL FUND PERFORMANCE EVALUATION

4.1. Data description and previous results

This study is focused on actively managed US equity MF. The first dataset consists of 1911 open-end, actively managed US equity MF selected from the CRSP mutual fund database. The monthly dataset covers 01/1993 –06/2007, inclusive. In this dataset, MF returns are net of management expenses, marketing fees, administration, and trading costs. The second dataset is obtained from the first one, with the original returns converted to “pre-expense” returns that are net of trading costs only. Because of the missing expense information, the second dataset includes 1876 funds. In every case, each MF has at least 50 monthly observations. The Appendix describes both samples in detail.

In reality, the managers are not going to work for free, but pre-expense analysis can still be useful. First, it is definitely an interesting theoretical question whether skilled stock pickers exist in principle, regardless of how much it costs to employ them. Second, if the good performers could be singled out, one could do some further analysis to see whether they earn more than their fees. This is especially relevant to institutional investors such as funds of funds because the MF fees for institutional investors are understandably lower than for individual investors. Besides, institutional investors can try to negotiate and lower the fees. Finally, another institution such as an equity hedge fund may be interested in obtaining a list of talented MF managers for the purpose of offering them a position. In that case, the fees charged by the corresponding MF are irrelevant.

The monthly factor returns (see Carhart (1997)) were obtained from the same CRSP database. The composition of sample (except for the time span) and the performance evaluation model correspond to BSW study that evaluates 2076 US equity mutual funds over the period 1975-2006; there has to be a significant overlap between the BSW sample after 1992 and the sample in this study.

The performance evaluation model is the four-factor Carhart model (2.1.3). BSW (10/2006) and Kosowski et al. (2006) consider a large number of possible extensions to the Carhart model that include time-dependent regression coefficients, serial correlation in error terms, and heteroskedasticity. They also apply time series bootstrap estimation. They report that none of these produce a significant change in results, and in the end focus on the Carhart model where regression coefficients are considered constant and error terms are considered serially uncorrelated. The model is estimated through a bootstrap procedure that, as shown in Kosowski et al. (2006), provides more adequate estimates. A similar approach is utilized in our study.

BSW (05/2008) estimated the proportions of funds with zero, positive and negative α, according to the model (2.1.3). For the entire period 1975-2006, they obtained the following results based on net returns of 2076 funds and pre-expense returns of 1836 funds:

Table 4.1.1 Summary of Barras et al. (05/2008) results

		Proportion, %	
	Zero	Positive	Negative
Pre-expense returns	85.9 (2.7)	9.6 (1.5)	4.5 (1.0)
95% CI	(80.61; 91.19)	(6.66; 12.54)	(2.54; 6.46)
Number of funds	1577	83	176
Net returns	75.4 (2.5)	0.6 (0.8)	24.0 (2.3)
95% CI	(70.5; 80.3)	(-0.97; 2.17)	(19.49; 28.51)
Number of funds	1565	12	499

BSW also found that for both net returns and pre-expense returns, the positive (skilled) proportion declined significantly and in a nearly monotone fashion between 1989 and 2006. Therefore, we expect the corresponding estimates for 1993-2007 period to be less than those in Table 4.1.1.

4.2. Pre-expense returns, Theoretical null

After the estimation of the Carhart model, the obtained p-values are converted to corresponding z-scores via (3.1.2). The next step is to estimate the structural model (3.3.9)-(3.3.10).

As mentioned in Section 2.2, BSW employ the theoretical null U(0,1) for two-sided null p-values (2.1.5). It is equivalent to $N(0,\sigma_0^2=1)$ for null z-scores (3.1.2). It is also possible to use the theoretical null in *locfdr* package which we will do first.

Let us start with pre-expense $\alpha's$ obtained from 1876 funds. Figure 4.2.1 shows the histogram of z-scores (y axis shows the counts of z-scores in each of 90 bins), the Poisson regression estimate of mixture density, $f(z)$, (green curve)

and the estimated null component, $\hat{p}_0 \cdot N(0,1)$, (blue dashed curve). The estimate $\hat{p}_0$ is equal to 0.8942 or 89.42%.

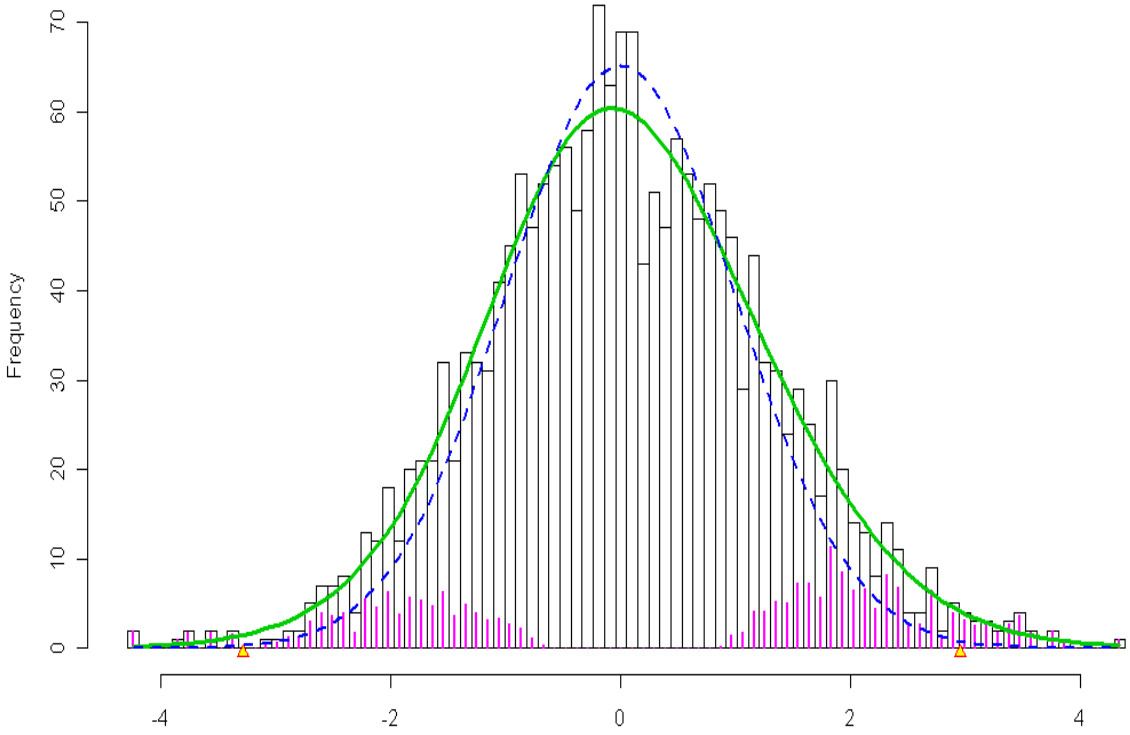

Figure 4.2.1 Estimated mixture density (green) and its null component (blue dashed) for pre-expense returns and theoretical null

The pink dashes in Figure 4.2.1 are so-called “thinned counts” that are equal to observed z counts times the estimated non-null component $\hat{p}_1 \hat{f}_1(z)$.

Table 4.2.1 summarizes the findings:

Table 4.2.1 Performance evaluation summary for pre-expense returns and theoretical null

	p0, %	p1+, %	p1-, %
	89.42 (0.75)	6.30 (0.53)	4.28 (0.53)
95% CI	(87.95; 90.89)	(5.26; 7.33)	(3.24; 5.31)
Number of funds	1678	118	80
Zero interval	(-1.5; 1.5)		
Lambda	0.1336		

The optimal zero interval (-1.5; 1.5) corresponds to using $\lambda = 0.1336$ in (2.2.3). It means that all z-values in (-1.5; 1.5) are considered to be i.i.d. from the theoretical null distribution N(0, 1). Equivalently, all two-sided p-values greater than 0.1336 are considered to be i.i.d. from U(0, 1).

We see that the confidence intervals for p_0, p_1^+, p_1^- in Table 4.2.1 have a lot of intersection with corresponding intervals in Table 4.1.1, even though the bootstrap procedure of BSW is dropped (Section 3.3). Secondly, the precision became considerably greater: in Table 4.1.1, the $s.e.(\hat{p}_0)$ is 2.7% whereas in Table 4.2.1 it is 0.75% (smaller by a factor of 3.6), which can make a practical difference because the point estimate of p_1 is not very large. The estimate of positive proportion drops from 9.6% in Table 4.1.1 to 6.3% in Table 4.2.1, possibly because of historical deterioration of MF performance mentioned in Section 4.1. Still, the proportion of positive performers is both practically and statistically significant.

The results of Table 4.2.1 suggest that some 118 money managers out of 1876 are outperforming on pre-expense basis. Unfortunately, knowing that some 118 funds are worth looking into is not the same as knowing those 118 skilled funds

by name. In order to single them out and, at the same time, avoid the useless false discoveries, one can try to select only the funds that fall in the bins where $fdr(z)$ is small, e.g. under 0.2 (see Efron (2006C)). The yellow triangles on Figure 4.2.1 mark these cutoffs. The funds to the right of the right triangle can be identified as skilled (outperforming) and the funds to the left of the left triangle can be identified as unskilled (underperforming). From the distribution of the thinned counts it becomes immediately clear that the majority of skilled and unskilled funds fall in between the cutoffs and therefore cannot be singled out. In other words, the study appears underpowered.

When the skilled/unskilled funds are identified based on the right/left z-value cutoffs like above, it is useful to know the tail false discovery rates:

$$FdrRight(z) = P\{\text{case i is null} \mid z_i \geq z\} = E_f[fdr(t) \mid t \geq z] \quad (4.2.1)$$
$$FdrLeft(z) = P\{\text{case i is null} \mid z_i \leq z\} = E_f[fdr(t) \mid t \leq z]$$

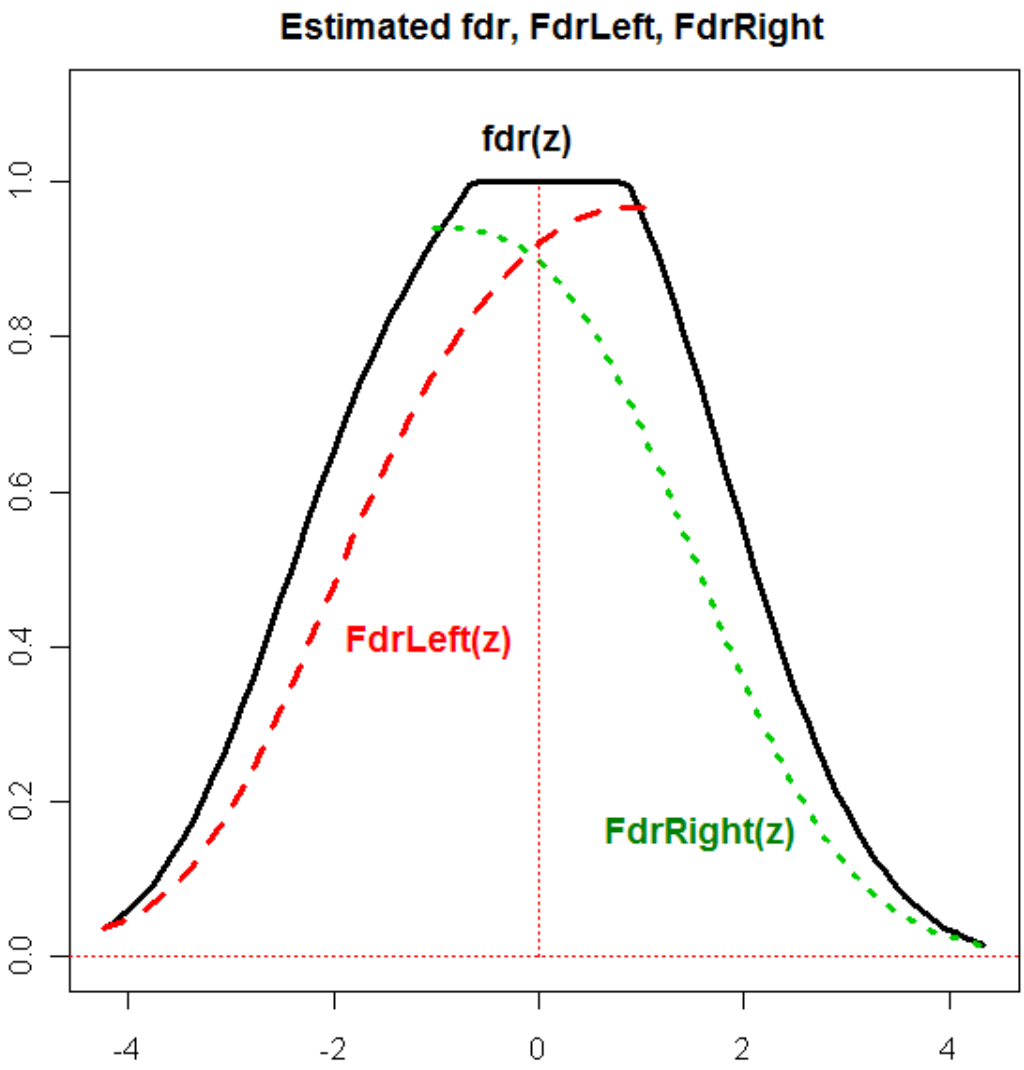

Figure 4.2.2 Estimated fdr (black), FdrLeft (red), FdrRight (green) for pre-expense returns and theoretical null

Figure 4.2.2 shows the estimates of fdr, $FdrRight$, and $FdrLeft$. For instance, Table 4.2.2 shows that fdr is under 0.2 to the right of z = 2.95. If we say that all funds with z-scores over 2.95 are outperforming, we will get FdrRight(2.95)=11.85% of false discoveries. Likewise, declaring all funds with z<-3.28 underperforming will produce FdrLeft(-3.28)=13.56% of false discoveries.

Table 4.2.2 Power statistics for pre-expense returns and theoretical null

Efdr	EfdrRight	EfdrLeft	
0.56	0.5	0.64	
fdr = 0.2 cutoffs	FdrLeft(-3.28)	FdrRight (2.95)	
(-3.28; 2.95)	0.1356	0.1185	
		Proportion of Identifiable performers based on fdr = 0.2 cutoff	
	Positive and negative	Positive only	Negative only
	11.04%	14.94%	5.62%
Number of funds	22 out of 198	18 out of 118	4 out of 80

A high power means that $fdr(z)$ is small on the support of $f_1(z)$, which can be described by an overall (post hoc) power measure:

$$Efdr = \frac{\int fdr(z) f_1(z) dz}{\int f_1(z) dz} = E_{f_1}[fdr(z)] \tag{4.2.2}$$

It can be adapted to measure the power in the left and right tails as follows:

$$EfdrRight = \frac{\int_0^{+\infty} fdr(z) f_1(z) dz}{\int_0^{+\infty} f_1(z) dz} = E_{f_1}[fdr(z) \mid z > 0] \tag{4.2.3}$$

$$EfdrLeft = \frac{\int_{-\infty}^{0} fdr(z) f_1(z) dz}{\int_{-\infty}^{0} f_1(z) dz} = E_{f_1}[fdr(z) \mid z < 0]$$

If a study has a good power, Efdr should be small, say, 0.2. Table 4.2.2 shows that, although there is more power in identifying the outperformers

(EfdrRight=0.5) than in identifying the underperformers (EfdrLeft = 0.64), the study is still very underpowered.

The lower part of Table 4.2.2 shows what such high Efdr values imply in practice. Suppose that we wish to identify the outperforming (underperforming) funds based on fdr = 0.2 right (left) cutoffs. Overall, we will be able to identity just 11.04% of “non-zero” (positive and negative combined) performers, which amounts to 22 funds out of 198. Focusing just on good performers, we can identify 14.94% of them, i.e. only 18 funds out of total 118 in the population. Given that we are willing to tolerate a sizable 11.85% of false discoveries, our ability to pick winners appears very limited. As for picking losers, it is even worse: we tolerate 13.56% of false discoveries and still are able to identify only 5.62% of negative performers, i.e. only 4 funds out of 80 underperformers in the population.

The only way to increase the proportion of identifiable performers for this sample is to try to tolerate a higher percentage of false discoveries, i.e. to move the left and right cutoffs closer to zero.

Table 4.2.3 Identified underperformers and false discoveries vs. FdrLeft for pre-expense returns and theoretical null

FdrLeft	Proportion of identified underperformers, %	Number of identified underperformers (rounded)	Number of false discoveries (rounded)
0.1356	5.62	4 out of 80	<1
0.2	12.5	10 out of 80	2
0.3	23	18 out of 80	8
0.4	36	30 out of 80	19
0.5	51	41 out of 80	41
0.6	68	54 out of 80	82
0.7	86	67 out of 80	161
0.8	100	80 out of 80	320

Table 4.2.4 Identified outperformers and false discoveries vs. FdrRight for pre-expense returns and theoretical null

FdrRight	Proportion of identified outperformers, %	Number of identified outperformers (rounded)	Number of false discoveries (rounded)
0.1185	14.94	18 out of 118	2
0.2	29	34 out of 118	9
0.3	47	55 out of 118	24
0.4	65	78 out of 118	51
0.5	83	98 out of 118	98
0.6	95	112 out of 118	168
0.7	100	118 out of 118	275

Tables 4.2.3 and 4.2.4 describe the corresponding tradeoff. For instance, to select about 50% (41 funds) out of all 80 underperformers one has to tolerate FdrLeft of 0.5. That means that getting this many underperformers is possible only in conjunction with just as many “zero” performers. For outperformers, the situation is better but not by much: to select 50% (59 funds) out of all 118 outperformers, one has to tolerate FdrRight of about 0.32 meaning that 28 useless funds (“zero” performers) have to be selected also: 28/(59 + 28)=0.32.

To obtain 90% (106 funds) of outperformers, one has to include about 135 “zero” performers that are not going to be distinguishable from outperformers.

Suppose that a fund of MF funds wants to construct an outperforming portfolio. After the expenses are deducted, all selected “zero” performers inevitably turn into underperformers and many outperformers turn into zero or even negative performers. Unless there remain some very strong performers who can make up for the rest, it is reasonable to require FdrRight be well under 50%, say, 20% at most. This corresponds to a portfolio of size under 41 (33 skilled and 8 unskilled funds for FdrRight = 0.2, with right z cutoff equal to 2.608). Further, some of these 41 funds may have to be dropped because of investor-specific restrictions (compliance, diversification, risk management, etc). This suggests that one’s ability to construct an outperforming portfolio of MF is fairly restricted.

The second goal of pre-expense analysis, identification of individual talents, is hard to achieve also: e.g. the list of “top 87” performers (87 = 59 + 28) will have 28 indistinguishable zero performers, which won’t make it very useful. The list of top 41 performers will have some 8 zero performers in it. The latter may be acceptable to someone who seeks to hire just one or two talented money managers, but still this warns one against the sizable amount of useless entries inevitably included in all kinds of “top performers” lists.

An interesting question is whether one could improve the situation by increasing the sample size and, thus, increasing the power. Here, increasing the sample size means increasing T in (2.1.3), e.g. the number of observations per fund. We are going to assume that the standard error of α_i in (2.1.3) is proportional to $1/\sqrt{T}$ and that the parameters such as p_0, p_1^+, p_1^- are fixed at their point estimates and only the number of observations per each fund is multiplied by a factor greater than one.

The current sample is 14 ½ years long with an average of 10 ¾ years of observations per fund; we can loosely think of this as having 10 ¾ years of data for each fund in the sample.

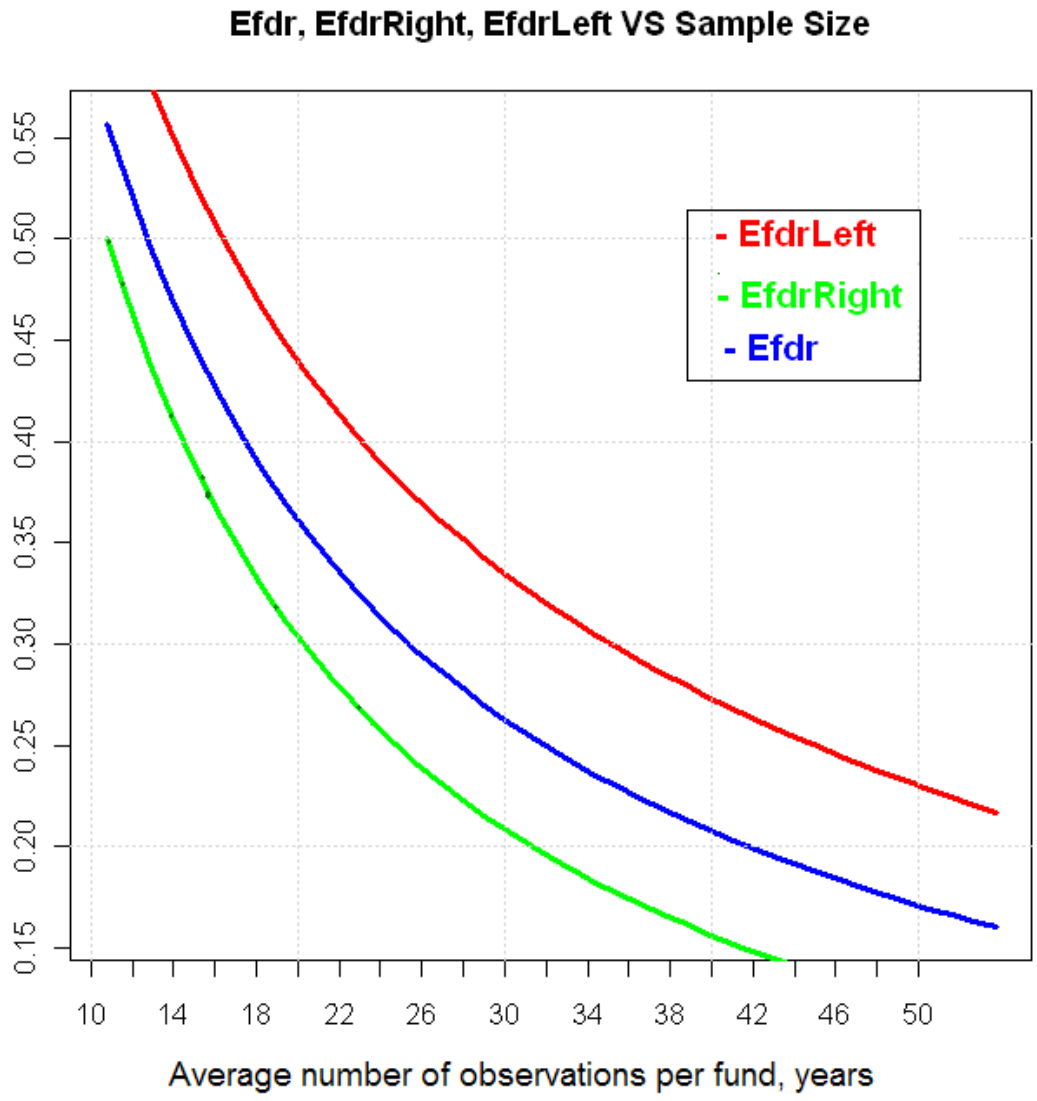

Figure 4.2.3 EfdrLeft (red), EfdrRight (green), Efdr (blue) vs. number of observations per fund (in years) for pre-expense returns and theoretical null

Figure 4.2.3 (obtained from *locfdr*) shows the increase in power vs. average sample size. For instance, if the current sample were doubled (to about 20 years

per fund on average) EfdrRight would decrease from 0.5 to 0.3. Having 32 years of data for each fund would decrease EfdrRight to a desirably low level of 0.2.

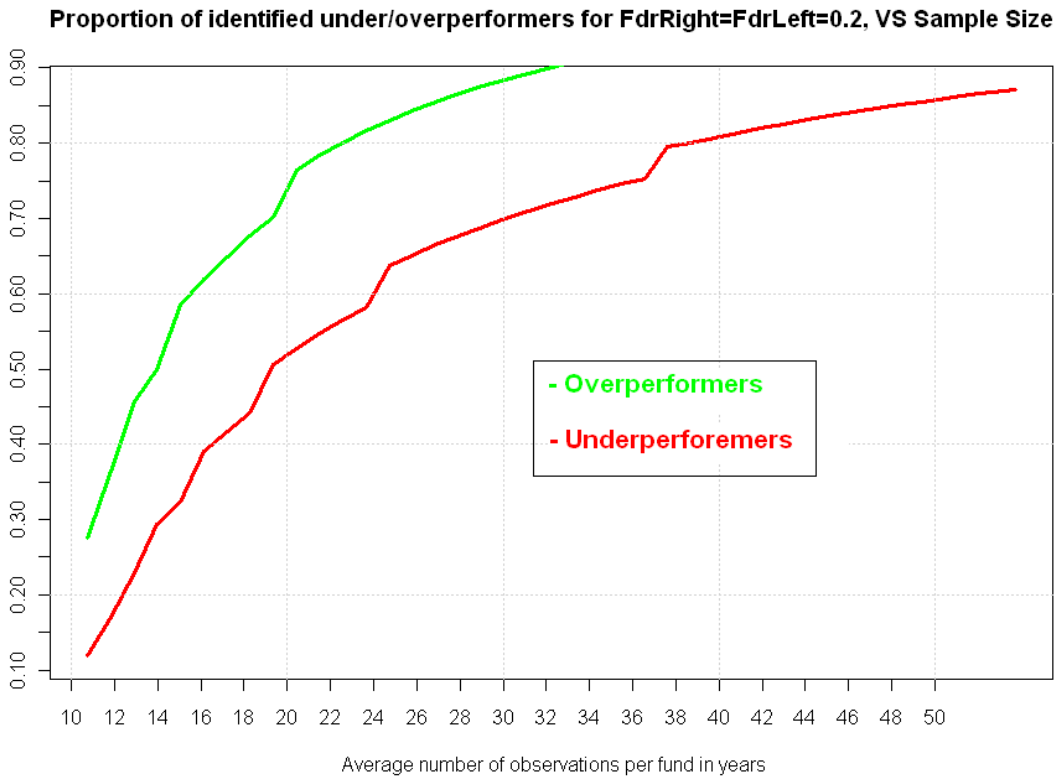

Figure 4.2.4 Proportion of identified outperformers (green), underperformers (red) vs. number of observations per fund (in years) for pre-expense returns and theoretical null. FdrRight and FdrLeft are fixed at 0.2

Figure 4.2.4 reflects our ability to identify more of the present over/underperformers thanks to a larger sample size given that FdrLeft and FdrRight are both fixed at 0.2. Roughly doubling the sample (from 10 ¾ to 20 years per each fund) will help us to identify about 76% (90 out of 118) of outperformers as opposed to 28% (33 out of 118) for the original sample. Having 32 years of observations for each fund could help identify 90% of outperformers. For underperformers the power is much worse: even with 40

years per each fund only about 81% (65 out of 80) of underperformers are identified.

Unfortunately, extending the sample back (e.g., BSW sample with 32-year span) can increase the number of funds but is not likely to produce many more observations per fund. For this study, the span is 14 ½ years with the mean of l0 ¾ years per fund and the standard error for the mean less than 1 month. Although 10% of the funds span the entire 14 ½ years, it is still unlikely to obtain a dataset with, say, more than 15 years of observations per fund on average, regardless of how far back it is extended. Therefore, power statistics obtained when there are 15 years of observations for each fund can be considered the upper bounds for the power. For the current dataset, having 15 years of data per each fund will not drive Efdr, EfdrRight and EfdrLeft much closer to 0.2 and only 58% (68 out of 118) outperformers will be identified with FdrRight = 0.2.

These findings suggest that unsatisfactory power is inherent to both the current and BSW study despite a much larger time span of the latter. It appears to be an issue to consider for any MF study that is based on monthly data and a similar multifactor performance evaluation model.

In addition, a long-living MF is likely to be managed by a few successive portfolio managers and, practically speaking, there are reservations about whether the 10-15 year-old data are relevant (unless the study is purely for historical purposes).

4.3. Pre-expense returns, Empirical null

All the inference in Section 4.2 was based on the theoretical null assumption. Therefore, there is no surprise that the obtained confidence intervals for p_0, p_1^+, p_1^- were consistent with those of BSW. Given that the 95% confidence interval for p_0 in Table 4.2.1 is (87.95; 90.89), it is possible to assume that $p_0 \geq 0.9$ in order to check whether the theoretical null N(0, 1) is adequate for the data.

We are going to use the same procedure as above but assume that the null distribution is $f_0(.) \sim N(\delta_0, \sigma_0^2)$. If the theoretical null is appropriate, the empirical parameters should not be significantly different from the corresponding theoretical ones.

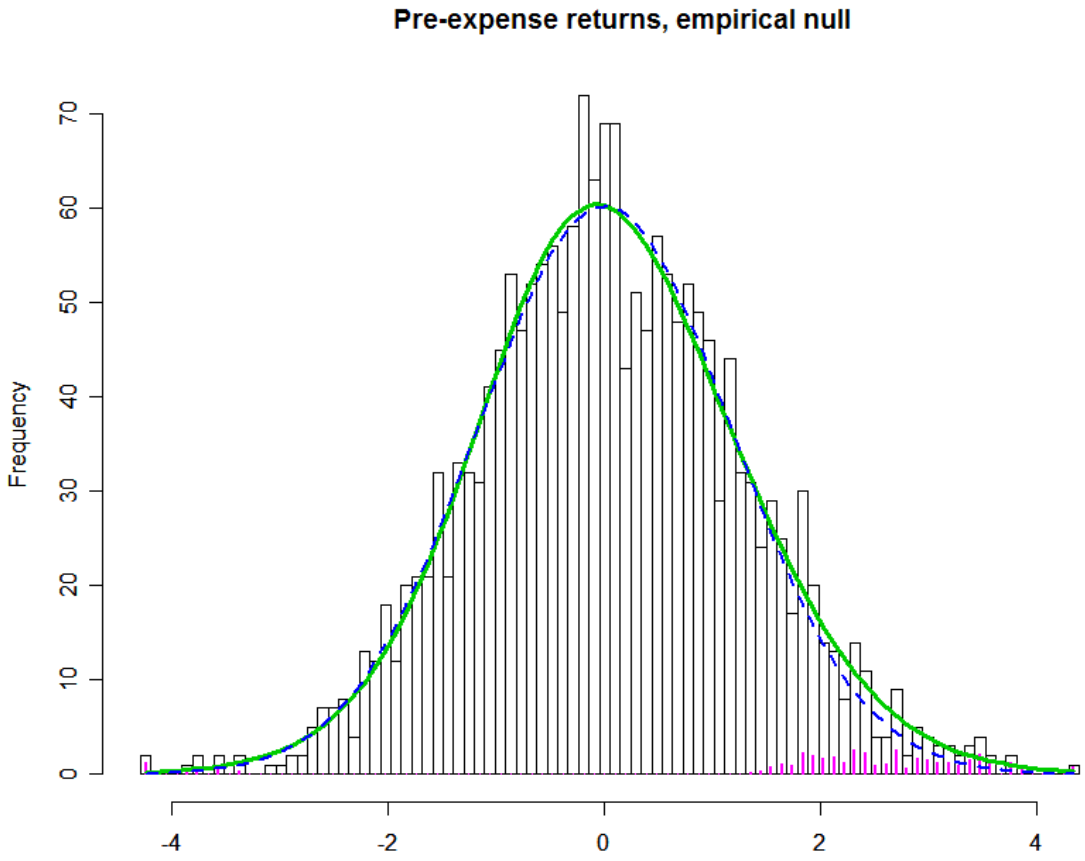

Figure 4.3.1 Estimated mixture density (green) and its null component (blue dashed) for pre-expense returns and empirical null

Figure 4.3.1 shows the fitted empirical null component $\hat{p}_0\hat{f}_0(z)$ (blue dashed curve) and the estimated mixture density (green curve, the same as on Figure 4.2.1).

Table 4.3.1 Performance evaluation summary and relevant statistics for pre-expense returns and empirical null

	p0, %	p1+, %	p1-, %
	98.11 (0.99)	1.85(0.99)	0.04
95% CI	(96.17; 100.05)	(-0.09; 3.79)	
Number of funds	1840	35	1
Zero interval	Lambda	EfdrRight	
(-1.7; 1.7)	0.0891	0.712	
delta0	sigma0	t-value for H0: sigma0 = 1	Dispersion variate A
0.0039 (0.0353)	1.179 (0.034)	5.29	0.276

As we see from Table 4.3.1, while δ_0 is indeed indistinguishable from zero, σ_0 is significantly greater than one with the corresponding t-value of 5.29. In other words, the z-values exhibit overdispersion which is significant, at least statistically. Since the estimate of p_1^- is very close to zero, the standard errors for $\hat{p}_0$ and $\hat{p}_1^+$ are given under the assumption that $p_1^- = 0$.

Speaking of practical significance, one may think of such z-values as being marginally N(0,1) and pairwise correlated with the correlation density $\rho \sim N(0,\ \tau^2)$ (see Efron's example in Section 3.2). Recall that in Section 3.2 the dispersion variate A is defined as a single independent realization from the correlation density. The estimate of the dispersion variate A in Table 4.3.1 is equal to 0.276. Therefore, the overdispersion appears to be even more significant than the preliminary guess of A = 0.16 discussed in Section 3.2. Returning to the example

based on $\tilde{F}dr(x \mid A)$ from (3.2.1), if we assume that $\tilde{F}dr(2.5 \mid 0) = 0.2$, then $\tilde{F}dr(2.5 \mid 0.276) = 2.37 * 0.2 = 0.474$.

It means that if 100 z's fall above 2.5, 80 of them are true discoveries if the theoretical null is used, but with the empirical null that number drops to about 53. Also, note that the value of λ in Table 4.3.1 and everywhere else is calculated under the theoretical null and, thus, underestimates the real cutoff p-value under overdispersion (i.e., the zero interval choice is less conservative than suggested by λ).

Comparing Figure 4.3.1 and 4.2.1, we see that the empirical null has a much better fit to $\hat{f}(z)$ in the central part of the histogram, i.e., the bias of the null distribution is reduced. In theory, the blue dashed curve, $\hat{p}_0 \hat{f}_0(z)$, must always be under the green curve, $\hat{f}(z)$. This is clearly violated on Figure 4.2.1, indicating high bias.

Naturally, the empirical null implies higher variance, but if we compare the measures of variance (3.3.6) and bias (3.3.4) of the theoretical and empirical nulls on the same zero interval (-1.7; 1.7) it turns out that the empirical null produces the variance that is 2.2 times as large and the bias that is 34.5 times as small. Therefore, there is both practically and statistically significant evidence against the theoretical null.

The usage of empirical null being justified, it implies that the theoretical null-based inference overestimated the number of both skilled and unskilled funds in the population. The 95% confidence interval for p_0 changes from (87.95; 90.89) under theoretical null to (96.17; 100.05) under empirical null. The latter means that it is possible that both underperformers and outperformers are not present

in the population at all. The estimated number of outperformers drops from 118 to 35 and the estimated number of underperformers drops from 80 to 1.

The estimated number of outperformers, 35, is unlikely to be significant practically. Besides, the power is extremely poor: the absence of yellow triangles on Figure 4.3.1 shows that in all bins fdr is above 0.2, and EfdrRight is 0.712.

Table 4.3.2 Identified outperformers and false discoveries vs. FdrRight for pre-expense returns and empirical null

FdrRight	Proportion of identified outperformers, %	Number of identified outperformers (rounded)	Number of false discoveries (rounded)
0.21	1	< 1 out of 35	< 1
0.3	7	2 out of 35	1
0.4	16	6 out of 35	4
0.5	31	11 out of 35	11
0.6	50	17 out of 35	26
0.7	72	25 out of 35	59
0.8	95	33 out of 35	133
0.9	100	35 out of 35	315

Table 4.3.2 shows that FdrRight is always greater than 0.2. In order to select 50% of outperformers (about 17 out of 35), one has to tolerate FdrRight of 0.6 by selecting about 26 “zero performers” as well.

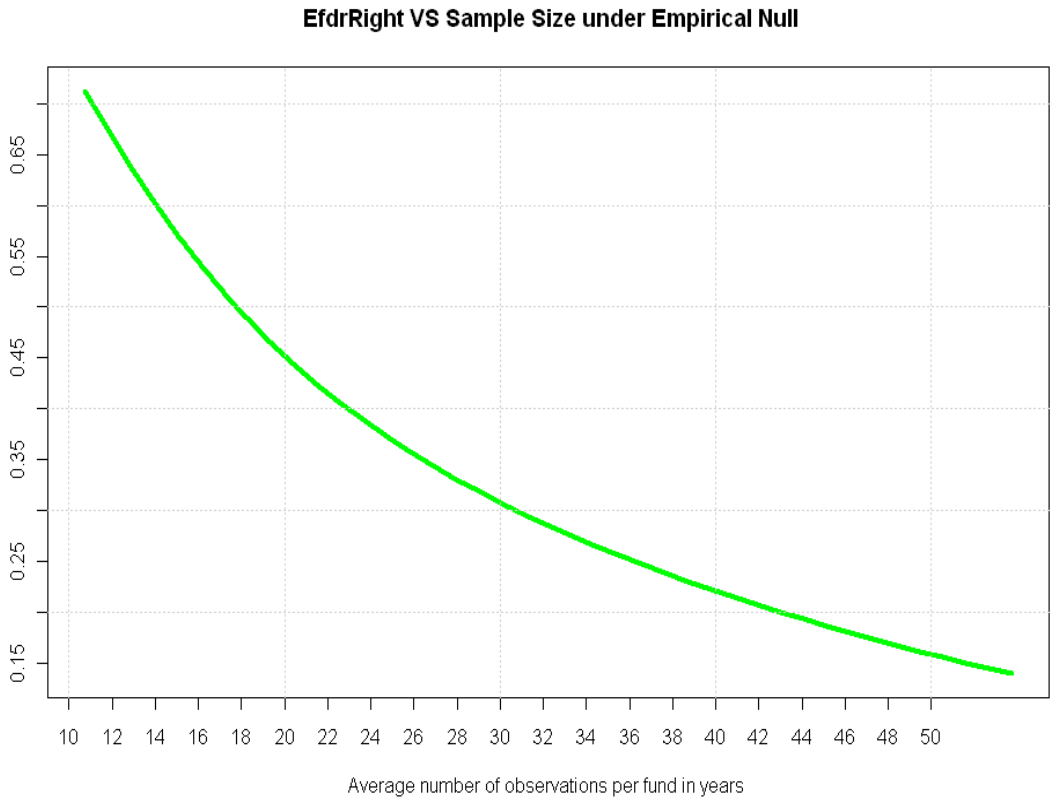

Figure 4.3.2 EfdrRight vs. number of observations per fund (in years) for pre-expense returns and empirical null

Figure 4.3.2 shows that it would take an unrealistic 43 years of observations per fund to obtain EfdrRight of 0.2. For 15 years of data per each fund, EfdrRight is still 0.57, far above 0.2.

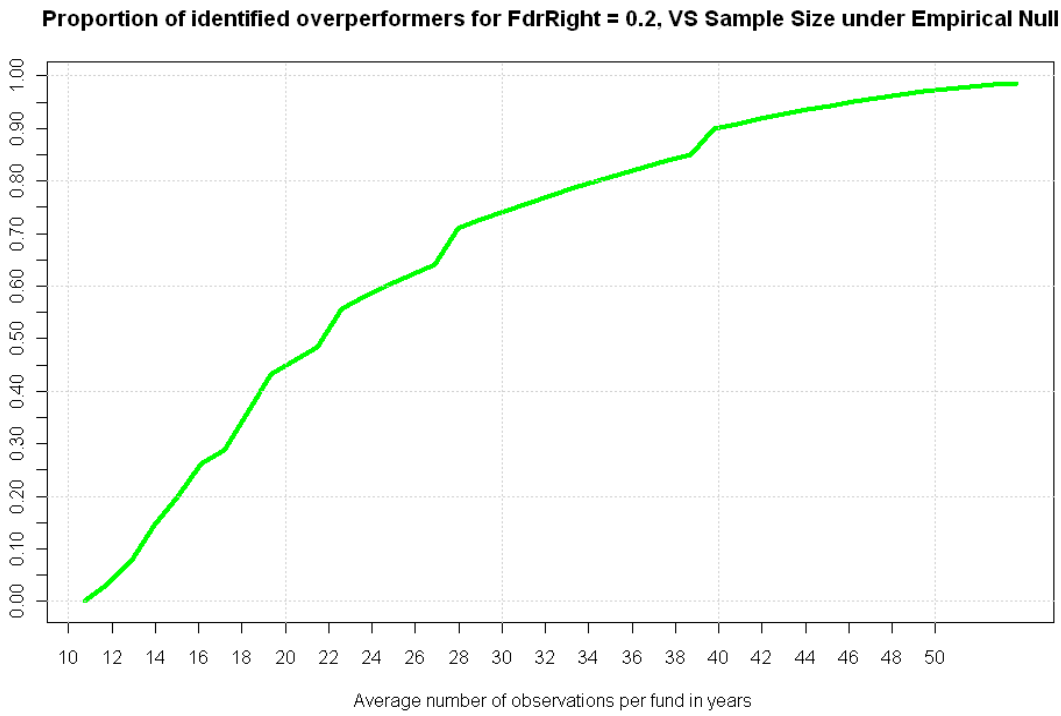

Figure 4.3.3 Proportion of identified outperformers vs. number of observations per fund (in years) for pre-expense returns and empirical null, FdrRight = 0.2

Figure 4.3.3 shows that at the level FdrRight = 0.2 outperformers are undetectable for the current sample. It would take roughly twice as much data (22 years per fund) to detect 50% (17 out of 35) outperformers. For 15 years per each fund, we are able to detect only 20% (7 out of 35) of outperformers.

We see that taking overdispersion into account leads us to conclusion that outperforming funds are both much fewer and much harder to single out than under the theoretical null. As for outperforming portfolio formation, it is impossible to construct one with FdrRight under 0.2. As for identifying individual talents, consider the “top 43” list of funds that will have 26 useless entries (43 = 17 + 26) and is of not much value. Therefore, while employing the theoretical null leaves a

little hope for obtaining a practical gain from performance evaluation, switching to the empirical none diminishes that hope to almost zero.

Since this study's sample has a significant overlap with that of BSW it is very likely that the overdispersion effect of similar magnitude was present in their sample also. It means that BSW study overestimated the percentage of skilled and unskilled funds in the population just as well. Under the empirical null, the percentage of outperformers in BSW sample will probably be greater than 1.85% in Table 4.3.1 but only because of better MF performance prior to 1993.

4.4. Net returns, Theoretical Null

The net returns dataset produces 1911 z-values.

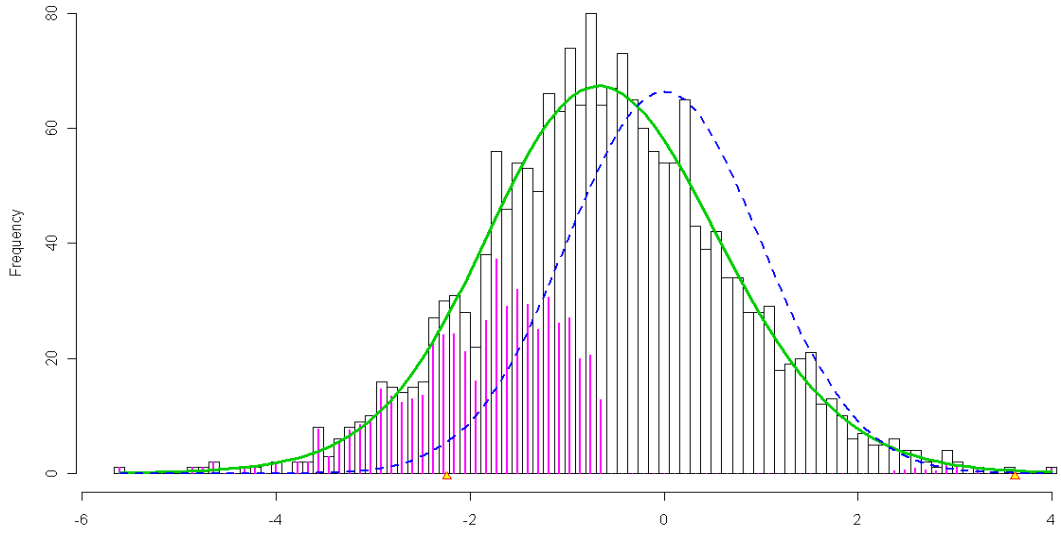

Figure 4.4.1 Estimated mixture density (green) and its null component (blue dashed) for net returns and theoretical null

Figure 4.4.1 shows the histogram of z's, the mixture density estimate (green curve fitted to 90 bins) and the fitted theoretical null component $\hat{p}_0 \cdot N(0,1)$ (blue dashed curve).

Table 4.4.1 Performance evaluation summary and relevant statistics for net returns and theoretical null

	p0, %	p1+, %	p1-, %
	70.91 (1.22)	0.45	28.64(1.22)
95% CI	(68.52; 73.30)		(26.25; 31.03)
Number of funds	1355	9	547
Zero interval	(-1.4; 1.4)		
Lambda	0.1615		

Table 4.4.1 and Table 4.1.1 show a good correspondence between the results for net returns. Since the estimate of p_1^+ is very close to zero, the standard errors for $\hat{p}_0$ and $\hat{p}_1^-$ in Table 4.4.1 are given under the assumption that $p_1^+ = 0$. Apparently, the estimated number of outperformers (9 funds out of 1911) is not significant neither statistically nor practically.

Table 4.4.2 Power statistics for net returns and theoretical null

Efdr	EfdrRight	EfdrLeft	
0.35	0.49	0.35	
fdr = 0.2 cutoffs	FdrLeft(-2.23)	FdrRight (3.61)	
(-2.23; 3.61)	0.11	0.17	
		Proportion of Identifiable performers based on fdr=0.2 cutoffs	
	Positive and negative	Positive only	Negative only
	29.17%	13.08%	29.42%
Number of funds	162 out of 556	1 out of 9	161 out of 547

Even though EfdrLeft = 0.35 is smaller than before, it is still well above 0.2 and the power is not too good.

Table 4.4.3 Identified underperformers and false discoveries vs. FdrLeft for net returns and theoretical null

FdrLeft	Proportion of identified underperformers, %	Number of identified underperformers (rounded)	Number of false discoveries (rounded)
0.11	29.42	161 out of 547	20
0.2	54	295 out of 547	75
0.3	80	438 out of 547	188
0.4	96	525 out of 547	350
0.5	100	547 out of 547	547

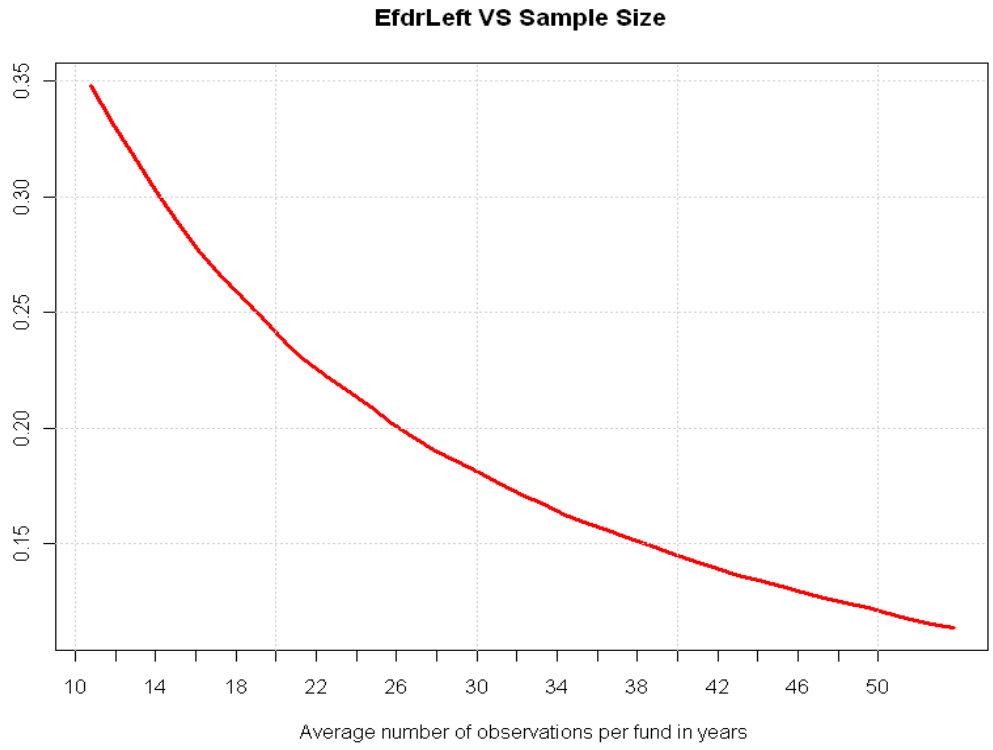

Figure 4.4.2 EfdrLeft vs. number of observations per fund (in years) for net returns and theoretical null

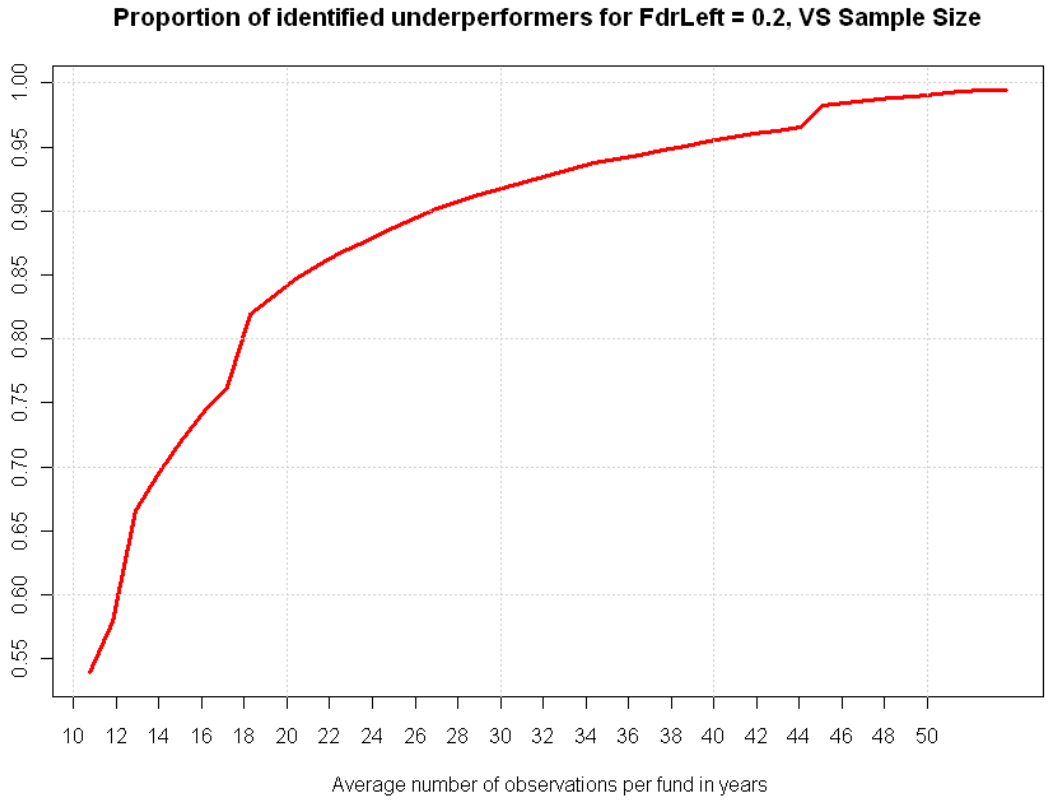

Figure 4.4.3 Proportion of identified underperformers vs. number of observations per fund (in years) for net returns and theoretical null, FdrLeft = 0.2

In particular, 54% of underperformers (295 out of 547) are identified with FdrLeft=0.2 (Table 4.4.3). Increasing the sample size to 15 years of data per each fund reduces EfdrLeft from 0.35 to 0.29, and only the unrealistic 26 years of data per fund brings EfdrLeft to 0.2 (Figure 4.4.2). Still, if it is possible to extend back the sample and obtain 15 years of data per fund, it pays off because the identifiable (under FdrLeft = 0.2) proportion of underperformers increases from 54% to 72% (394 funds out of 547). It is still far short of the 90% (492 funds out of 547) that could be obtained for 26-year sample (Figure 4.4.3).

Despite the low power, a high proportion of underperformers makes it much easier to create sizable “bottom lists”: e.g., the “bottom 156” list has FdrLeft of 0.1 which corresponds to about 16 useless funds with zero performance.

4.5. Net returns, Composite Empirical Null

For net returns data, it is not possible to fit the empirical null directly as in Section 3.3 because p_0 is way under 0.9. But the magnitude of overdispersion detected in Section 4.3 is not likely to change because of subtracting the expenses so it is safe to say that the theoretical null is inadequate for net returns just as well. When it is taken into account, the estimated number of outperformers (9 funds) will be reduced even more and the estimated number of underperformers will be reduced by a few percent.

Qualitatively, the results will remain about the same: the proportion of outperformers is both practically and statistically zero; proportion of underperformers is both practically and statistically positive (less than 28% but probably more than 18%); the majority of funds (well over 70%) have zero net performance.

Note that previously we tested simple nulls:

$$H_i^0 : \alpha_i = 0 \ \text{ VS } \ H_i^a : \alpha_i > 0 \qquad (4.5.1)$$

or

$$H_i^0 : \alpha_i = 0 \ \text{ VS } \ H_i^a : \alpha_i < 0$$

The test will become a lot more powerful if we could calculate the p-value under the composite null setting, i.e.

$$H_i^0 : \alpha_i \le 0 \ \text{ VS } \ H_i^a : \alpha_i > 0 \qquad (4.5.2)$$

Usually, the distribution of p-value under the composite null is unknown, so the simple null with $\alpha_i = 0$ is used instead. For this study, we can use the data itself to estimate the composite empirical null, just like we estimated the simple empirical null. In terms of the structural model, we have

$$\begin{aligned}
&\alpha \sim g(.) \\
&z \mid \alpha \sim N(\alpha, \sigma_0^2) \\
&g(\alpha) = p_0 g_0(\alpha) + p_1^+ g_1^+(\alpha) \\
&\textit{where} \\
&p_0 = P_g\{\alpha \le 0\}, \quad p_1^+ = P_g\{\alpha > 0\} \\
&g_0(\alpha) - \text{"null" density with support on } \{\alpha \le 0\} \\
&g_1^+(\alpha) - \text{"positive" density with support on } \{\alpha > 0\}
\end{aligned} \tag{4.5.3}$$

In terms of z-values, we have

$$\begin{aligned}
&f(z) = p_0 f_0(z) + p_1^+ f_1(z) \quad \text{- mixture density of z's} \\
&\textit{where} \\
&f_0(z) = g_0 * N(0, \sigma_0^2) \quad \text{- density of null z's} \\
&f_1^+(z) = g_1^+ * N(0, \sigma_0^2) \quad \text{- density of alternative z's} \\
&p_0 + p_1^+ = 1
\end{aligned} \tag{4.5.4}$$

The null density $f_0(z)$ is estimated on the zero interval is $(z_-; z_+)$ where z_- is some small value in the left tail, e.g. z_- = -4; for $z < z_-$, we assume that $p_0 f_0(z) = f(z)$; z_+ serves as a smoothing parameter.

From the results in the previous section, we would expect the optimal z_+ to be at least 1.4. Efron (2004) suggests a non-symmetrical parametric null, such as split-normal $f_0(.) \sim SN(\delta_0, \sigma_1^2, \sigma_2^2)$, in order to avoid the influence of the left-tail z's on the inference in the right tail. However, fitting a split-normal distribution along with normal $N(\delta_0, \eta_0^2)$ for $z_- = -4$ and $z_+ \in [1.4;\ 2.2]$ showed that the corresponding null components $\hat{p}_0 \hat{f}_0(z)$ are virtually identical and $N(\delta_0, \eta_0^2)$ is quite adequate for modeling the composite null.

The choice of zero interval was performed as described in Section (2.3) where the variance was calculated on the interval $(median(z);\ +\infty)$ and the bias on the interval $(median(z);\ z_+)$ because of our interest in the right-tail inference ($median(z)$ is very close to $\hat{\delta}_0$). The value of λ in this case is equal to $1-\Phi(z_+)$, where $\Phi(.)$ is the standard normal c.d.f.

Figure 4.5.1 shows the estimated null component $\hat{p}_0 \cdot N(\hat{\delta}_0, \hat{\eta}_0^2)$ (blue dashed curve).

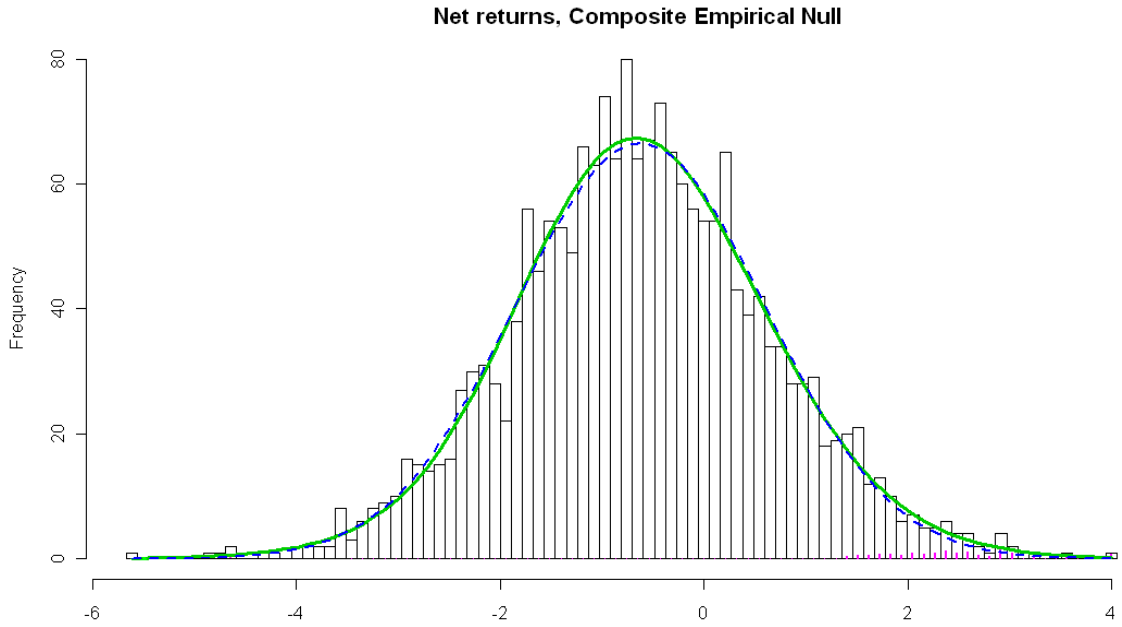

Figure 4.5.1 Estimated mixture density (green) and its null component (blue dashed) for net returns and composite empirical null

Table 4.5.1 Performance evaluation summary and relevant statistics for net returns and composite empirical null

	p0, %	p1+, %
	99.21 (0.7)	0.79 (0.7)
95% CI	(97.84; 100.58)	(-0.58; 2.16)
Number of funds	1896	15
Zero interval	Lambda	
(-4; 1.6)	0.055	
delta0	eta0	EfdrRight
-0.624 (0.033)	1.229 (0.028)	0.725

Table 4.5.1 shows that the bias-variance tradeoff is minimized on the zero interval (-4; 1.6). Here we were supposed to expect a much larger power to identify outperformers than for the test in Table 4.1.1. First, the mean of null density is shifted to the left by a sizable value of 0.624. Secondly, inclusion of z-values in [-4; -1.4] reduced the standard error of $\hat{p}_0$ by 0.38% without causing any increase in the bias in the right tail. Inclusion of z-values in [1.4; 1.6] reduced $s.e.(\hat{p}_0)$ by another 0.14% and overall it dropped from 1.22% in Table 4.1.1 to 0.7% in Table 4.5.1.

In spite of this, the estimated number of outperformers grows from 9 to only 15 (still practically insignificant) and is not statistically different from zero. The only explanation is that the estimated null distribution $\hat{f}_0(z) \sim N(\hat{\delta}_0, \hat{\eta}_0^2)$ reflects the fact that σ_0^2 in (4.5.4) is much greater than one. Taking that overdispersion into account drastically reduces the final estimated number of outperformers. It "negates" all the benefits we hoped to get from the composite empirical null.

Table 4.5.2 Identified outperformers and false discoveries vs. FdrRight for net returns and composite empirical null

FdrRight	Proportion of identified outperformers, %	Number of identified outperformers (rounded)	Number of false discoveries (rounded)
0.24	1	< 1 out of 15	< 1
0.3	5	1 out of 15	< 1
0.4	17	3 out of 15	2
0.5	30	5 out of 15	5
0.6	47	7 out of 15	11
0.7	68	10 out of 15	24
0.8	89	13 out of 15	53
0.9	100	15 out of 15	135

Besides, EfdrRight is over 0.725 and the power is abysmal. As Table 4.5.2 shows, FdrRight is always above 0.24. The list of "top 15" performers has FdrRight = 0.58 that amounts to about 9 useless funds in the list.

4.6. Net Performance vs. Mutual Fund Investment Objective

The 1911 funds in the sample are classified by the four investment objectives: "Small Company Growth" (SCG), "Other Aggressive Growth" (OAG), "Growth" (G) and "Growth and Income" (GI). We merge the first two groups as "Aggressive Growth" (AG) and consider only three groups. It would be interesting to look into the net performance versus investment objective. Statistically speaking, the findings of BSW suggest that one may be able to increase the power by using investment objective as a control factor.

BSW compare the fund categories by running their bootstrap-based procedure for each category separately. We can perform an fdr-based analysis which is not going to suffer from the misspecifications of null distribution since we use the

empirical null. First, let us compare the net outperformance across categories based on the composite empirical null from Section 4.5.

Efron (2007) proposes the following method. Suppose that all z-values are divided into two classes, A and B. Class A corresponds to the investment category of interest and class B corresponds to the rest of funds. Then the mixture density and fdr can be decomposed as follows:

$$f(z) = \pi_A \cdot f_A(z) + \pi_B \cdot f_B(z) \tag{4.6.1}$$

$\pi_A,\ \pi_B$ - a priori probabilities of class A and B

$f_A(z) = p_{A0} f_{A0}(z) + p_{A1} f_{A1}(z)$ - class A mixture density

$fdr_A(z) = p_{A0} f_{A0}(z) / f_A(z)$ - class A fdr

$f_B(z) = p_{B0} f_{B0}(z) + p_{B1} f_{B1}(z)$ - class B mixture density

$fdr_B(z) = p_{B0} f_{B0}(z) / f_B(z)$ - class B fdr

It can be shown that

$$fdr_A(z) = fdr(z)\frac{\pi_{A0}(z)}{\pi_A(z)} \tag{4.6.2}$$

where

$\pi_{A0}(z) = P\{\text{case from class A and null} \mid z\}$

$\pi_A(z) = P\{\text{case from class A} \mid z\}$

The difference between classes A and B is tested via the null hypothesis:

$$H_0 :\ fdr_A(z) = fdr(z) \tag{4.6.3}$$

If we assume that the null densities for A and B coincide for some z, then

$$f_{A0}(z) = f_{B0}(z) \Rightarrow \pi_{A0}(z) = \frac{\pi_A p_{A0}}{p_0} = const \tag{4.6.4}$$

and

$$fdr_A(z) = \frac{\pi_A p_{A0}}{p_0} \cdot fdr(z) \cdot \frac{1}{\pi_A(z)}$$

We don't have to run a separate fdr analysis for each group as long as the assumption $f_{A0}(z) = f_{B0}(z)$ holds in the area of interest, the right half of z-value histogram in this case.

In that case, (4.6.4) implies that the test (4.6.3) is equivalent to testing

$$H_0 : \pi_A(z) = const \tag{4.6.5}$$

To check the assumption $f_{A0}(z) = f_{B0}(z)$, we use another property:

$$fdr_A(z) = fdr_B(z) = 1 \Rightarrow \pi_A(z) = \pi_{A0}(z) \tag{4.6.6}$$

In particular, (4.6.6) is likely to hold for $z \in [-1; 0.5]$. If $\pi_A(z)$ (which can be estimated) is a constant in that interval, so is $\pi_{A0}(z)$. According to (4.6.4) this can be used as a diagnostic for the assumption $f_{A0}(z) = f_{B0}(z)$.

We use the same 90 bins as on Figure 4.5.1 and estimate $\pi_A(z)$ via binomial regression over the bins of interest with $z \geq -1$:

$$\begin{aligned} \text{logit}(\pi_A(z)) = \beta_0 + \beta_1 \cdot \max(0.5 - z, 0) + \beta_2 \cdot \max(z - 0.5, 0) + \\ + \beta_3 \cdot \max(z - 0.5, 0)^2 + \beta_4 \cdot \max(z - 0.5, 0)^3 \end{aligned} \tag{4.6.7}$$

The interval (-1; 0.5) corresponds to 14 non-empty bins and 858 z-values and the remaining (0.5; max(z)) corresponds to 27 non-empty bins and 357 z-values.

First, we keep the first covariate in the model and use a model selection procedure to include any of the other three covariates that are important. Then, if the p-value for $\hat{\beta}_1$ is small it suggests $f_{A0}(z) \neq f_{B0}(z)$. If the p-value is large, we can proceed under $f_{A0}(z) = f_{B0}(z)$. In that case, we drop the first covariate. Then, the p-value for H_0 :"no covariates are important in (4.6.7)" is used to test (4.6.5).

For instance, for AG funds the model selection step produces the model with two covariates corresponding to β_1 and β_3 (Figure 4.6.1). The stars indicate the observed proportions of AG funds in each bin and the blue curve is the fitted probability from (4.6.7).

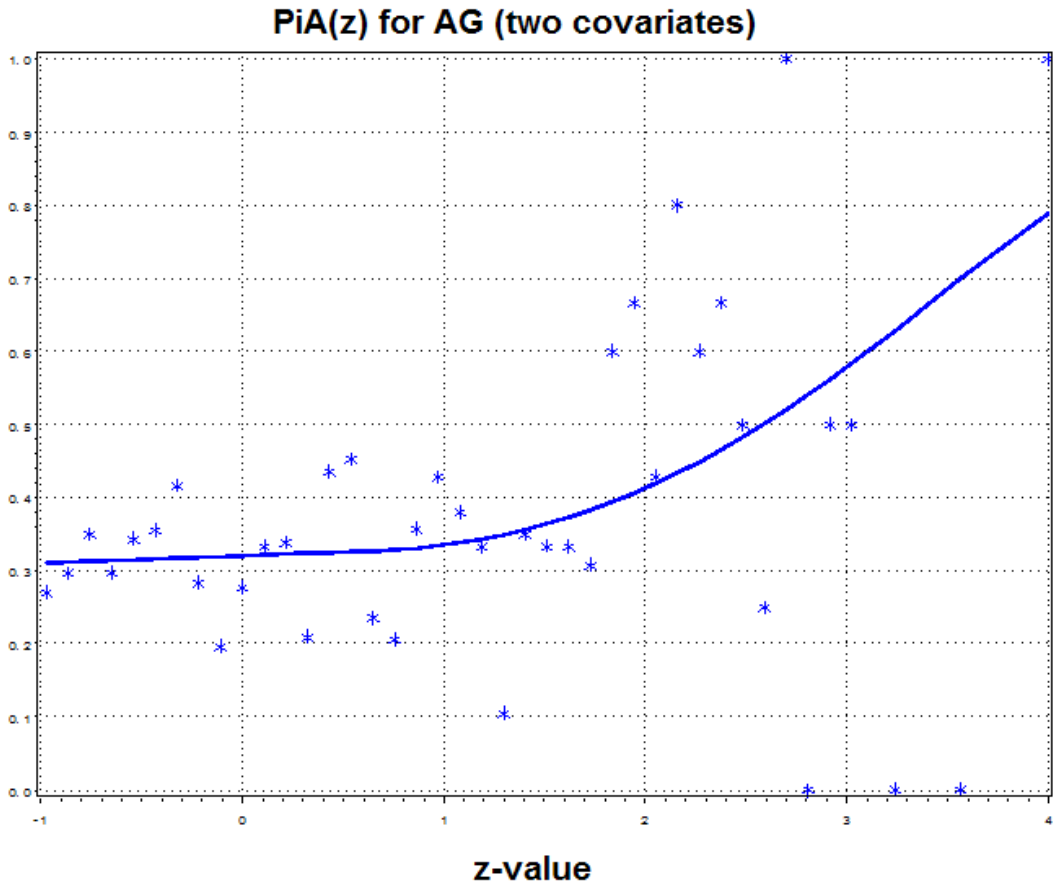

Figure 4.6.1 Probability Pi_A(z) for Aggressive Growth estimated with two covariates

The estimated probability does not change much in (-1; 0.5) and, indeed, the p-value for β_1 is 0.6997 (Table 4.6.1). After the first covariate is dropped, only the second order term remains (Figure 4.6.2) and its p-value is 0.0079 (Table 4.6.1).

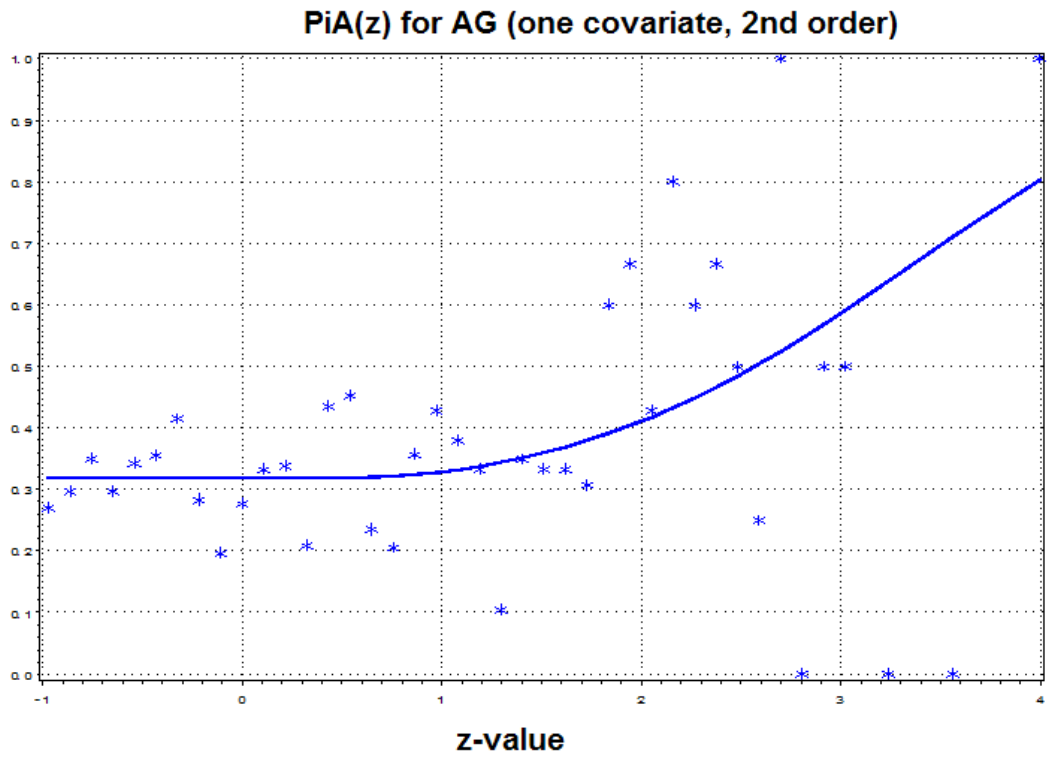

Figure 4.6.2 Final model fit for probability Pi_A(z) for Aggressive Growth

Table 4.6.1 Net outperformance vs. investment objective, composite empirical null

Category	Number of funds	Pvalue for H0: f_A0(z) = f_B0(z)	P-value for H0: fdr_A(z) = fdr(z)	Number of outperformers	Proportion
G	886	0.7083	0.5606	7	0.79%
GI	398	0.9698	0.0006	0	0%
AG	627	0.6997	0.0079	19	3%
Population	1911	n/a	n/a	15	0.79%

We therefore conclude that $fdr_{AG}(z) \neq fdr(z)$. Column 3 of Table 4.6.1 shows that the hypothesis $f_{A0}(z) = f_{B0}(z)$ is not rejected for any category. Column 4 suggests that $fdr_{GI}(z) \neq fdr(z)$ but we fail to reject $fdr_{G}(z) = fdr(z)$.

Figures 4.6.3 and 4.6.4 show the final models (with the first covariate dropped) for GI and G groups, correspondingly.

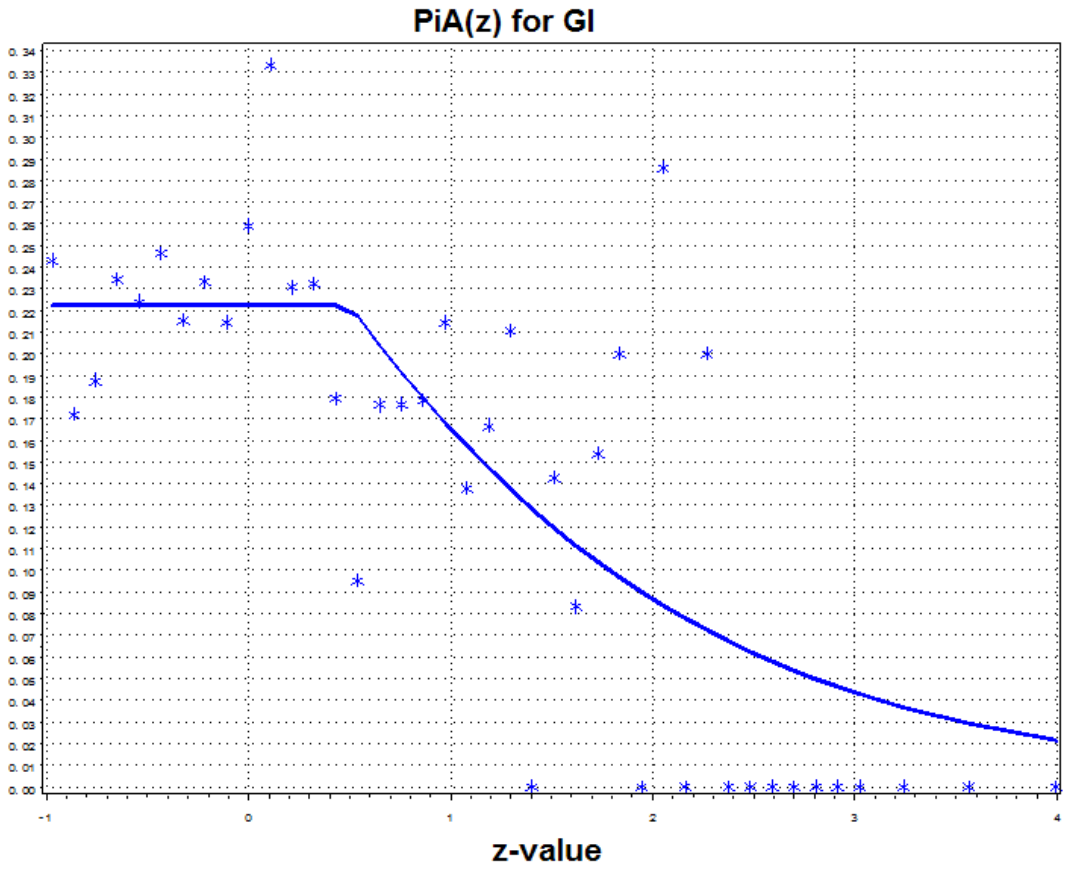

Figure 4.6.3 Final model fit for probability Pi_A(z) for Growth&Income

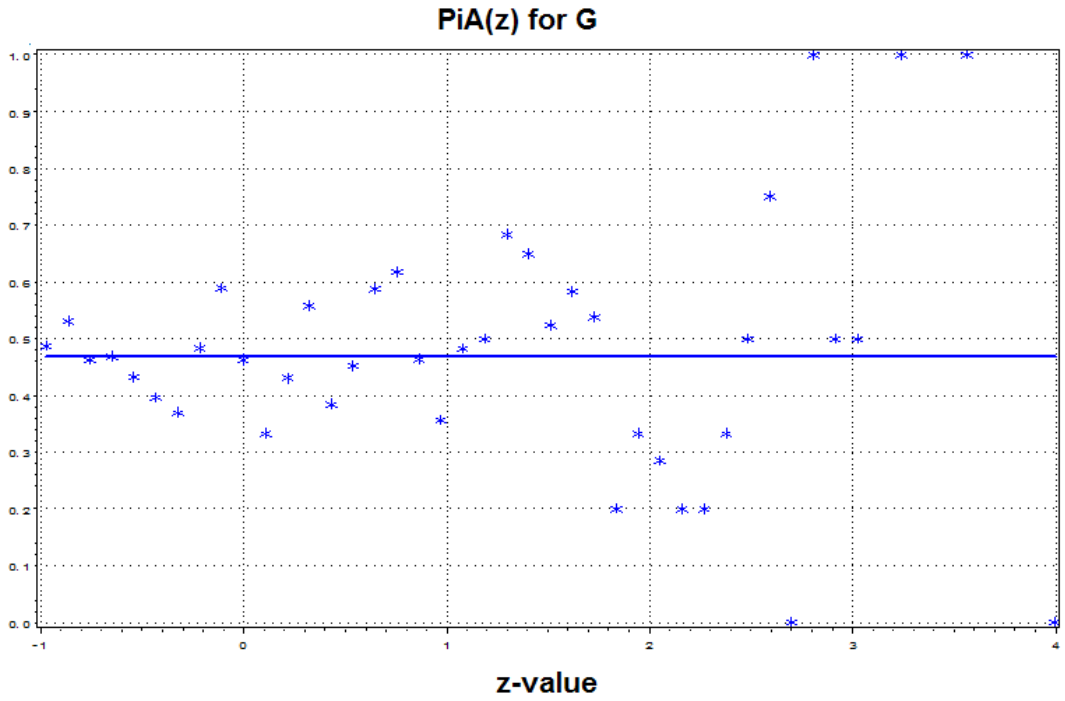

Figure 4.6.4 Final model fit: probability Pi_A(z) for Growth

A number of logistic regression diagnostics (not reported, for details see Pregibon (1981)) confirm the adequacy of all three final logistic models. It follows from (4.6.2), (4.6.4) and (4.6.6) that fdr for class A can be estimated as

$$\hat{fdr}_A(z) = \hat{fdr}(z)\frac{\hat{\pi}_A(0)}{\hat{\pi}_A(z)} \qquad (4.6.8)$$

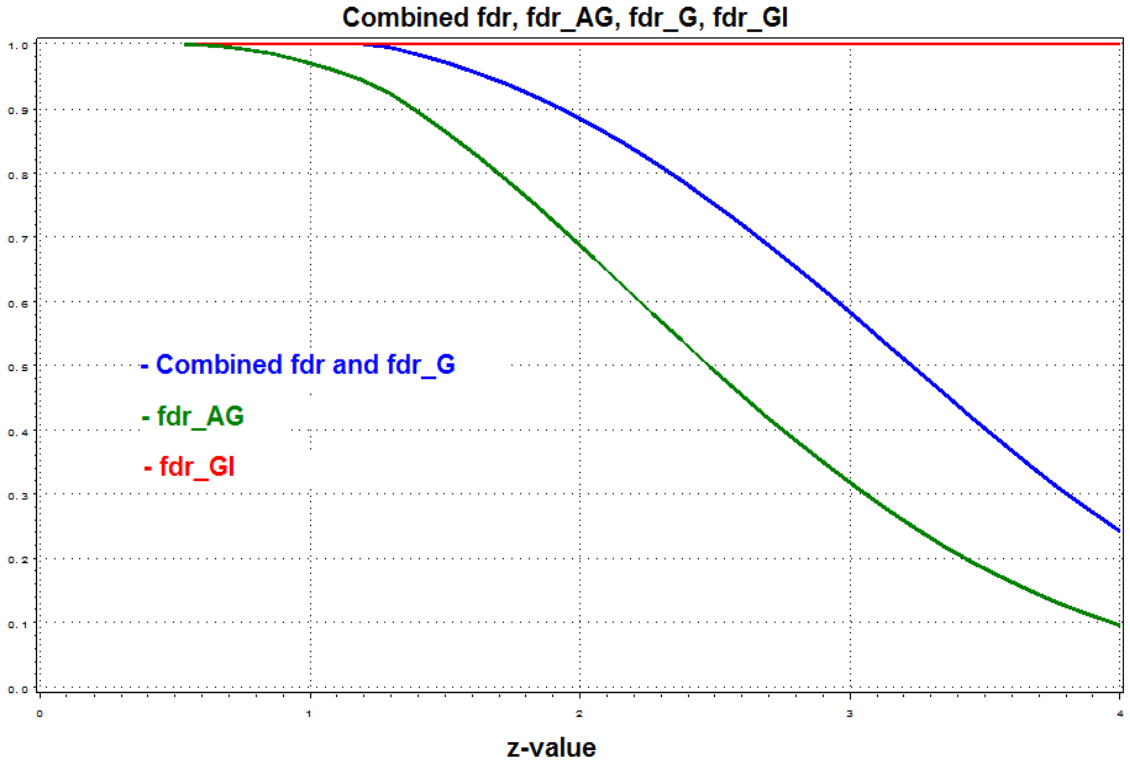

Figure 4.6.5 Combined fdr and Growth fdr (blue), Aggressive Growth fdr (green), Growth&Income fdr (red)

Figure 4.6.5 shows the curves corresponding to $\hat{fdr}(z)$ (which coincides with $\hat{fdr}_G(z)$), $\hat{fdr}_{GI}(z)$, and $\hat{fdr}_{AG}(z)$. The first and obvious conclusion is that there are no skilled managers in GI group.

Using the estimate $\hat{fdr}_{AG}(z)$ and $\hat{fdr}_G(z)$, we conclude that there are 19 outperformers among 627 AG funds and 7 outperformers among 886 G funds. Therefore, while the percentage of outperformers is 0.79% in the population (15 out of 1911), it is about 3% in AG group, 0.79% in G group and 0% in GI group (Table 4.6.1).

While $\hat{fdr}(z)$ is always above 0.24, $\hat{fdr}_{AG}(z)$ is under 0.2 for $z \geq 3.56$. Unfortunately, only one AG fund has $z \geq 3.56$ and can be identified as outperformer. Even if we raise the fdr cutoff from 0.2 to a quite aggressive level of 0.4 ($z \geq 2.807$), only 4 out of 19 AG outperformers are identified. Even a relatively superior AG group is unable to produce a practically significant number of identifiable outperformers.

The results of BSW for the same three groups (G, GI, AG) are not very consistent. In their 05/2007 version (based on 1464 funds, 1975-2002) they claim that GI funds have the lowest proportion of skilled managers (0%) and the AG funds are the best (8.0%). In BSW of 05/2008 (2076 funds, 1975-2006) they claim that “results for the three investment-objective subgroups… are similar” but do not provide the numbers. Instead, they look into the “short-term performance” (see Section 4.7) to find that AG is the best (4% of outperformers) and GI is the worst (0%).

BSW used the theoretical null, while the results in this section are based on the composite empirical null to provide extra power and adjust for apparent overdispersion. Our findings are consistent with the preliminary results of BSW and, at the same time, provide more realistic and statistically grounded picture of the relative investment category performance.

Comparison of investment categories based on pre-expense returns is of interest for the reasons outlined in Section 4.2 and because of additional theoretical implications which are discussed in Section 4.8. Using the simple empirical null from Section 4.3, we look into the distribution of both out- and underperformers across investment objectives.

Table 4.6.2 Pre-expense performance vs. investment objective, simple empirical null

Category	Number of funds	Number of under - performers	Proportion	Number of out- performers	Proportion
G	871	0	0.00%	16	1.84%
GI	387	1	0.26%	0	0.00%
AG	618	35	5.66%	29	4.69%
Population	1876	1	0.04%	35	1.85%

While there are statistical differences between the categories, it appears that the only practically significant result is that AG group has a higher proportion of outperformers and a higher proportion of underperformers than G and GI groups. However, the power is still low: for instance, only 2 out of 29 AG outperformers are identified with fdr = 0.2 cutoff and 10 out of 29 are identified with fdr = 0.4 cutoff. Out of 35 AG underperformers, zero are identified with 0.2 cutoff, and only 4 are identified with 0.4 cutoff. See Section 3.8 for further discussion.

4.7. Short-term net performance

The long-term results of net MF performance are quite disappointing because the number of outperformers is never practically significant: 12 in BSW study and the best result for this study is 26 (7 G and 19 AG funds discovered in Section 4.6).

However, the short-term performance may be better, as suggested by BSW.
To look into short-term performance, BSW partition the data into six non-overlapping subperiods of 5 years each, starting with 1977-1981 and ending with 2002-2006. If a fund has 60 observations on a subperiod, it is treated as a separate “fund” with 5-year history. They thus increase the number of estimated

alphas from 2076 to 3311 and the positive proportion goes up from 0.6 (0.8) % (Table 4.1.1) to a statistically significant 2.4 (0.7)%, correspondingly. In BSW this is interpreted as the evidence for superior “short-term” performance that exists for a while and gradually disappears because the “long-run equilibrium” has to settle. Berk and Green (2004) describe the equilibrium model, but BSW point out that if the model holds, the negative performance has to disappear just as well, which is not observed in reality.

All this seems to imply that investors are more capable of recognizing the good performance (and that is the reason why it is only short-term) than the bad performance (it is not spotted and, therefore, continues for a long time). That is not very convincing and we will try to make a case that “superior short-term performance” is merely a result of inadequate multiple inference technique employed by BSW.

Note that the extended dataset of 3311 “short-term funds” is a lot more likely to deviate from the weak dependence assumption. Many funds are included more than once, even though on different subperiods. But the major concern is that drastically reducing the number of observations per fund is very likely to increase the overdispersion of z-values. In the end, the “short-term” z-values will probably be more overdispersed than the original z-values. That alone could explain a higher estimated percentage of outperformers and, therefore, the utilization of empirical null is even more justified here.

Similarly, we partition our dataset into three non-overlapping 58-month subperiods. If a fund has 50 or more observations on a subperiod, it is treated as a separate “short-term fund”. In the end, there are 3636 of such “funds”. Applying the theoretical null (just like in Section 4.4) gives the following results:

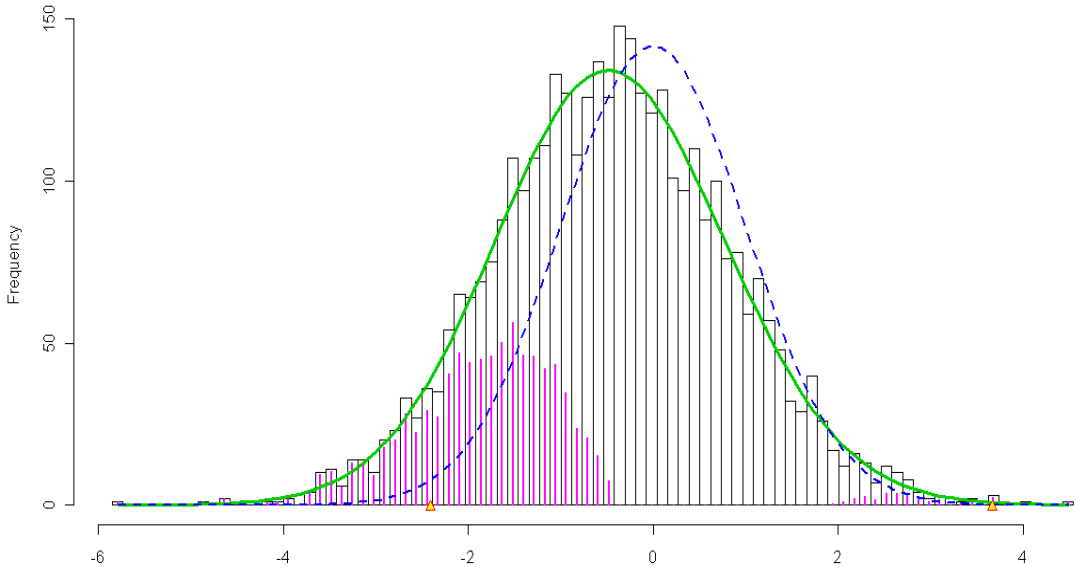

Figure 4.7.1 Estimated mixture density (green) and its null component (blue dashed) for 3636 “short-term” funds (net returns, theoretical null)

Table 4.7.1 Net performance summary and relevant statistics for 3636 “short-term” funds under theoretical null

	p0, %	p1+, %	p1-, %
	76.45 (0.74)	0.81	22.74 (0.74)
95% CI	(75.0; 77.9)		(21.29; 24.19)
Number of funds	2780	29	827
Zero interval	Lambda		
(-1.5; 1.5)	0.1336		
Efdr	EfdrLeft	EfdrRight	
0.411	0.405	0.562	

Comparing this to the results of Section 4.4, we see that the number of outperformers is larger (29 instead of 9) but is still practically insignificant. The standard error of $\hat{p}_0$ is reduced from 1.22% to 0.74% but $\hat{p}_1^+ = 0.81\%$, which is hardly statistically significant (standard errors in Table 4.7.1 are given under assumption $p_1^+ = 0$). That is, even when overdispersion is not taken into account, there is no evidence of short-term outperformance in 1993-2007, which is consistent with the overall deterioration of MF performance mentioned in Section 4.1.

Following the procedure of Section 4.5, we can try to empower the test via composite empirical null. Just like in Section 4.5, it turns out that $N(\delta_0, \eta_0^2)$ is enough and split normal is unnecessary. The results are as follows:

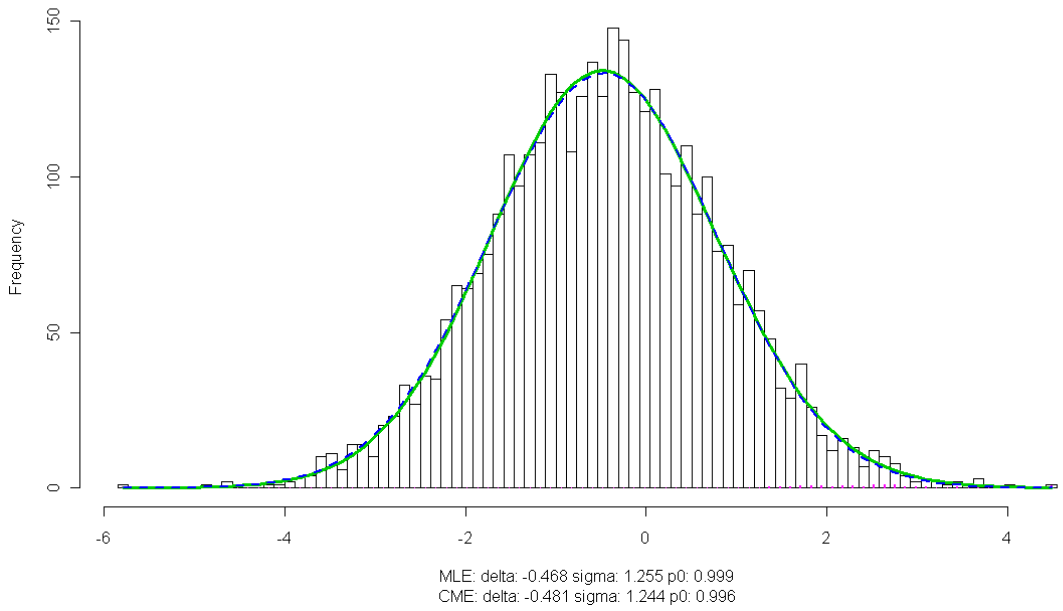

Figure 4.7.2 Estimated mixture density (green) and its null component (blue dashed) for 3636 “short-term” funds (net returns, composite empirical null)

Table 4.7.2 Net performance summary and relevant statistics for 3636 "short-term" funds under composite empirical null

	p0, %	p1+, %
	99.63 (0.69)	0.37 (0.69)
95% CI	(98.28; 100.98)	(-0.98; 1.72)
Number of funds	3623	13
Zero interval	Lambda	
(-3.5; 1.6)	0.055	
delta0	eta0	EfdrRight
-0.467 (0.026)	1.254 (0.024)	0.877

The composite empirical null is shifted to the left by 0.467, and because of inclusion of additional (mostly negative) z-values the standard error of $\hat{p}_0$ dropped from 0.74 to 0.69. Like in Section 3.5, this allows us to hope that more positive cases will be identified. However, as predicted above, the overdispersion is so severe that the estimated number of outperformers not only fails go up but actually drops from 29 to 13 funds and is statistically insignificant, as well. Therefore, we conclude that there is no compelling evidence of short-term outperformance in 1993-2007.

On the other hand, BSW manage to construct an outperforming portfolio based on minimizing its FDR. The portfolio is observed for 27 years (1980 – 2006) with yearly recalculation of FDR for all funds and corresponding rebalancing. In the end, the portfolio produced statistically significant annual alpha of 1.45% with a p-value of 0.04, even though its average FDR was 0.415. However, the BSW study provides compelling evidence that in 1980 – 1993 the proportion of outperformers in the population was much higher than in 1994 – 2006, with a

sharp and monotone decline from 1993 on. This pattern is so pronounced that it will likely to remain valid even when overdispersion is taken into account. Therefore, even though the decent performance of FDR-based portfolio is not spurious, it is more of historical interest. Our study does not find any evidence to state that the construction of outperforming MF portfolio would have been possible in 1993-2007.

The traditional approach to form an outperforming portfolio is to include the top (based on z-value ranking) k% of funds at each rebalancing. Without the multiplicity adjustment, the tail FDR or Fdr are not taken into account, and the proportion of useless funds in the portfolio is out of contol. Therefore, a multiple-comparison-based cutoff (e.g., include all funds with Fdr < 0.2), applied at each rebalancing, should work better.

However, any multiple inference procedure works with “input list” of z-values and the “quality” of this list is at least as important as an appropriate multiple inference method. In particular, in Section 3.2 it is suggested that the empirical null-based fdr procedure “takes into account” the asset pricing model misspecification. Suppose, for simplicity, that all z’s are independent and the only source of overdispersion ($\sigma_0^2 > 1$) is the model (2.1.3) misspecification, e.g. caused by a too small sample size T. Essentially, $\sigma_0^2 > 1$ tells us that there is some extra noise in $\hat{\alpha}_i$’s which we have to take into account by using $f_0 = N(0, \sigma_0^2)$ instead of $f_0 = N(0,1)$. Taking that into account will prevent us from making false discoveries, but it will not make the extra noise disappear. If the level of noise is very high, the procedure will simply declare that all or almost all cases are null.

As a result, one cannot just rely on a multiple inference method to substantially improve the portfolio performance. In particular, Mamaysky et al. (2007) argue

that it is unlikely for a single performance evaluation model to be equally good for each fund in the sample. They show that using a few competing models, combined with backtesting, can significantly improve the performance of portfolio of MF. Using such approach coupled with a multiple inference procedure can be an interesting topic for future research.

4.8. Size, Power and Asset pricing model misspecification

An important issue in the asset pricing theory is that of asset pricing model misspecification. There are a few ways to detect misspecification. For instance, the theoretical requirement that discounted returns are unforecastable implies that in (2.1.3) the residuals are not supposed to be serially correlated. In this study, we say that the model is misspecified when the marginal distribution of null z's is different from N(0,1). This can have many causes, including the abovementioned serial correlation.

A practical way to check for such misspecification is to see whether "naïve" stock portfolio formation strategies that presumably have a zero alpha show any abnormal performance. For instance, introducing their "conditional" (i.e., with time-dependent regression coefficients) multifactor model, Ferson and Schadt (1996) show that three "naïve" portfolio formation strategies produce abnormal performance under some "unconditional" (with time-independent regression coefficients) multifactor models. When "conditioning" is introduced, the abnormal performance disappears which is interpreted as evidence that the unconditional models are misspecified and the conditional models are not.

Kothari and Warner (2001) follow a similar path to investigate asset pricing model misspecification and power. First, they make a point that, in practice, the investor is interested in performance evaluation on a rather short time frame, from 3 to 5 years. They construct 348 "naïve" stock portfolios that mimic an average MF's general features, such as size, number of securities, book-to-market ratio, and

turnover. Each portfolio spans 3 years with a 1-month shift: the first portfolio is on [01/1966; 12/1968], the second is on [02/1966; 01/1969], and so on, until 1994 (it is similar to Fama-MacBeth procedure). As a result, the alphas are ordered sequentially and can be analyzed as a univariate and possibly autocorrelated stationary time series.

They find that for Carhart model the true nulls tend to get rejected too often, even though they do not investigate whether the over-rejection is statistically significant. Because little to no serial correlation is found in the sequential alphas, the over-rejection (which is the same as overdispersion) can be interpreted as evidence in favor of Carhart model misspecification. In our study, it would be incorrect to say that the overdispersion found in Section 4.3 and elsewhere is caused solely by the misspecification of Carhart model because we cannot offer any evidence that our z-values are independent.

Model misspecification makes it problematic to compare the output of different asset pricing models because the test size (type I error) gets out of control. Kothari and Warner find that at the nominal test size of 5% the rejection rate for Carhart model is 13% when all nulls are true and 80% when all nulls are false (outperformance is introduced artificially). For characteristic-based (CS), model the corresponding numbers are 3.4% and 59%. It means that Carhart model has a greater power but its actual test size is also larger than the nominal 5%.

In such a case, Kothari and Warner conclude that the comparison between the models is “clouded”. We would like to note that comparison can still be made if asset pricing models are considered binary classifiers (i.e., zero- VS outperformance) and then their overall discriminative ability can be compared via Receiver Operating Characteristic (ROC) curve, e.g. area under ROC curve can be a good criterion. However, that approach does not work for dependent z’s

since, as explained in Section 3.2, the “dependence effect” and model misspecification effect can be absolutely indistinguishable.

This study suggests an alternative approach to compare the power of different models that can work for dependent z’s and does not require the rather artificial “stretching out” of the portfolios in time like in Kothari and Warner. Construct a large number of “naïve” portfolios where the proportion of artificially introduced outperformers is under 10%. Run a few competing performance evaluation models and, like in Section 4.3, use the empirical nulls if necessary. Using the empirical nulls adjusts for both sources of null distribution misspecification: dependence among z’s and asset pricing model misspecification. It is not possible to tell these two effects apart. However, we believe that utilizing the empirical null puts the test sizes of different performance evaluation models on the same level. For instance, if we do this with Kothari and Warner portfolios under the Carhart model, we will get that σ_0>1, which is supposed to bring the inflated rejection rate of 13% closer to the target of 5%. In addition, a positive and significant estimate $\hat{p}_1 > 0$ is likely to become indistinguishable from zero. After that, the estimated proportions of non-null cases and the power measures (Efdr, EfdrLeft, EfdrRight, and such like) become comparable among the models.

Moreover, it may be unnecessary to introduce another layer of approximation by investigating artificial MF instead of real MF. For instance, if a preliminary analysis of a MF dataset shows that the proportion of interesting cases is a lot less than 10% (e.g., 1.85% in Section 4.3), one can artificially add some economically significant alphas to the existing MF in the sample and re-estimate to see how powerful the model is. The only thing we have to control is that the percentage of non-null cases be under 10%, which is doable, since there are enough bins with fdr = 1 that do not contain any “non-zero” performers. This approach looks especially attractive when the performance evaluation model is holdings-based, i.e. the exact composition of the portfolio is very relevant.

There is one more way to test the validity of an asset pricing model, which is probably the most traditional. A multifactor asset pricing model states that

$$E[R^{ei}] = \beta_i' E[f], \quad i = 1, m \tag{4.8.1}$$

where $E[R^{ei}]$ is the average return for the asset i in excess of risk-free rate, $E[f]$ is a p-dimensional vector of average excess factor returns.

The p-dimensional vector β_i is defined as the regression coefficient in

$$R_t^{ei} = \alpha_i + \beta_i' f_t + \varepsilon_t^i, \quad t = 1, T \tag{4.8.2}$$

where R_t^{ei} and f_t are random and observed excess returns for asset i and the factors at time t (see Cochrane (2005)). The Carhart model (2.1.3) is an example of (4.8.2) with p = 4 factors.

Taking expectations of both sides of (4.8.2) w.r.t time and comparing the result to (4.8.1), we get that (4.8.1) implies that all the intercepts in (4.8.2) should be zero. In practice, the attention is paid not to the statistical significance of this test but to how practically significant the values of $\hat{\alpha}_i$ are.

Further, consider so-called cross-sectional regression:

$$E[R^{ei}] = \beta_i' E[f] + a_i, \quad i = 1, m \tag{4.8.3}$$

where β_i are obtained from (4.8.2) and are considered fixed covariates, a_i are interpreted as pricing errors for model (4.8.1) and $E[f]$ is a p-dimensional vector of estimated regression coefficients. While it depends on the intricacies of the joint estimation of (4.8.2) and (4.8.3), one may roughly assume that the pricing errors a_i in (4.8.3) are equal to the corresponding intercepts α_i in (4.8.2).

In MF performance context, we may say that the pricing errors are negligible when the number of out- and underperformers (on pre-expense basis) is insignificant. Correspondingly, our only possible concern in this study is that in

AG group (Section 4.6) we find the total number of non-zero performers is 64 out of 618, or 10.35%. These portfolios are not “naïve” (they are actively managed MF), but, as we will see below, that is not the point. The 64 cases consist of 35 underperformers and 29 outperformers, which means that on average, the performance is not practically different from zero, i.e. in (4.8.3) $E[a_i] \approx 0$ even for AG group.

However, given that β_i in (4.8.3) are covariates, the distribution of error terms, a_i, is not supposed to depend on the covariates. There is not supposed to be any sort of pattern when we plot the residuals a_i (or α_i) against any of p components of β_i. This is a common test for a multifactor asset pricing model.

In particular, Huij and Verbeek (2008) employ this test and suggest that the four standard Carhart factors are inadequate for pricing MF. In particular, they suggest that the pricing errors depend on the value of h_i in (2.1.3), which is the same as the component of β_i corresponding to HML or “growth vs. value” factor. Huij and Verbeek find that “growth” funds tend to have positive pricing errors (outperform) and “value” funds have negative pricing errors (underperform) after the “growth vs. value” factor has been already included in the model.

In our case, we have three investment objectives (G, GI, AG) and although we do not explicitly compute the regression coefficients h_i of these groups w.r.t. HML factor, it is reasonable to assume that AG consists mostly of “growth” funds (small h_i), GI mostly of “value” funds (large h_i) and G is something in between. Correspondingly, results from Section 4.6 shows that AG group has unusually large (both positive and negative) pricing errors. That can be interpreted as follows: the variance of a_i in (4.8.3) depends on the level of h_i, i.e. for growth funds $Var(a_i)$ is much larger than for value funds.

While these results are not consistent with those of Huij and Verbeek (who used a very different time span and sample of MF), they resemble the mispricing anomaly reported in a well known paper of Fama and French (1993). They found that their 3-factor model could not properly price the stocks with the smallest values of h_i, i.e. growth stocks. Those stocks had significantly positive and negative pricing errors. The 4-factor Carhart model (tested in a manner similar to that of Fama and French) managed to correct that, but in this study we see that a similar anomaly reappeared.

One possible explanation, suggested by Huij and Verbeek, is that Carhart model uses factors constructed based on a very large subset of US stocks, which may not reflect the stock-picking restrictions that apply to MF. Therefore, all these findings suggest that creating more adequate benchmarks specifically for MF may be justified.

CHAPTER 5. SUMMARY AND CONCLUSIONS

5.1. Summary of Mutual Fund performance results

In this study, we look into the performance of about 1900 US equity mutual funds over the period 1993-2007. MF performance evaluation problem is handled with a state-of-the-art local false discovery rate approach combined with the utilization of empirical null hypothesis. While trying to extend the prior BSW study, we still see it as a reference point because it employs the theoretical null, which is a particular case of the empirical null.

It is reassuring that despite the difference in the employed datasets, whenever we use the theoretical null (Sections 4.2 and 4.4), our findings are consistent with theBSW results. As predicted in Sections 2.2 and 3.2, the introduction of empirical null is well grounded. First, we obtain compelling statistical evidence (Section 4.3) that the theoretical null is misspecified (overdispersion) and has to be replaced with the empirical null. The inference changes dramatically: over 10% of funds are either skilled or unskilled on pre-expense basis under the theoretical null, but under the empirical null that proportion is not distinguishable from zero.

The empirical Bayes method also allows us to test the net performance under the more powerful composite null that includes both “zero” and underperformance as opposed to the simple null of “zero” performance used in BSW and probably all other MF studies. Since even under that powerful setting the number of outperformers proves neither statistically nor practically significant (Section 4.5), the evidence for the absence of outperformance in MF industry in 1993-2007 is

substantially reinforced. We therefore believe that the outperformance of low FDR-based portfolio in BSW study is mostly due to a better performance of MF industry prior to 1993.

We use the local false discovery rate method to look into the net performance vs. MF investment objective (Section 4.6). We obtain compelling statistical evidence that “Aggressive Growth” funds have the largest number of outperformers and “Growth and Income” have no outperformers, which is consistent with the empirical findings of BSW study. Unfortunately, even the strongest “Aggressive Growth” category fails to produce a practically significant number of identifiable winners.

We provide evidence that BSW’s finding of “short-term superior performance” is likely to have been an effect of overdispersion, as opposed to the presence of true short-term winners. In any event, there is no evidence of “short-term outperformance” in our sample (Section 4.7).

If we are interested in practical applications of MF performance evaluation, the study has to have a high power. The detailed power analysis showed that regardless of whether the utilized null is theoretical or empirical and whether we are interested in picking winners or losers, our ability to do so is very limited. In particular, the “top N performers” lists (for both pre-expense and net returns) are likely to have a very small proportion of true outperformers. Essentially, in this study we can only be good at composing meaningful “worst net performers” lists thanks to a high proportion of net underperformers.

Power analysis calculations show that to obtain decent power, each fund in the sample has to have an unrealistically long history of returns, well over 15 years. It appears that any MF study that is based on monthly data and a similar multifactor performance evaluation model is bound to be very underpowered.

In Section 4.8 we suggest how we can leverage Efron's approach to investigate the comparative power of different performance evaluation models, which can be an interesting subject for future research. In addition, we discover some evidence of the misspecification of the volatility of error terms in Carhart model.

Returning to the question of performance, analysis in Section 4.5 shows that well over 70% of funds in the sample have net return alphas that are not distinguishable from zero. That proportion will probably remain large even after some unconsidered fees (such as loads) are taken into account. Zero alpha funds are of value because they essentially provide a free (on average) access to the US equity market. For a risk-neutral investor, zero-alpha funds are superior to index funds whose net alphas are negative, although close to zero. To estimate the total gain, one may use the study of Elton et al. (2004) who look into fifty-two S&P500 index funds over 1996-2001 and find that their average alpha is minus 0.41% p.a.

One may try to take a broader view and speculate that even the sizable proportion of underperformers (from 18% to 28%, Section 4.5) somehow adds value, even though that value does not go to the shareholders directly. Providing liquidity to the stock market is the most obvious contribution, but there may be others. It is easy to dismiss the equity research performed by MF on the grounds that one can do just about as well by indexing. But what if thousands of MF equity researchers do a good job of preventing overtly fraudulent companies from entering the US stock universe? If that check were not in place, it would be quite possible that the US stock market would be far less efficient than it is.

5.2. Possible applications outside Mutual Fund industry

The method of Efron can be also applied to model selection. In particular, consider such model selection criterion as AIC:

$$AIC = 2k - 2\ln(L) \qquad (5.2.1)$$

where *k* is the number of parameters in the model and *L* is the maximized value of likelihood function. AIC and similar criteria are routinely assumed to be deterministic, whereas in fact they are not. It may be the case that, after examining a large number of models, the "best AIC" model is just "lucky". This idea was originally proposed by White(2000), but, as discussed in Section 2.2, his direct approach of estimating the dependence structure via bootstrap is bound to fail when the number of tests is large.

The models of interest can be fairly similar to each other (e.g., based on almost the same set of covariates), and the assumption of mutual independence (or weak dependence) of AIC's is unlikely to hold. Similarly to MF case, explicit modeling of high-dimensional dependence structure is far from straightforward. Correspondingly, we can apply Efron's results and gain the same benefits as we enjoyed in this study.

Another possible area of application is Statistical Arbitrage. In Avellaneda and Lee (2008), the residuals from a multifactor model are integrated (from asset returns to asset levels) and then fed into a simple mean-reverting model. The goal is to select the stocks whose residuals have good mean-reverting properties, that is, the true parameters of the estimated mean-reverting model have to belong to a certain range. Given that the number of stocks can be quite large, it is nothing but a large-scale multiple inference problem. Yet again, the test statistics are certainly dependent but the dependence structure not transparent at all. Efron's approach can help gain an edge here, which may be an interesting subject for future research.

BIBLIOGRAPHY

BIBLIOGRAPHY

Avellaneda, M., Lee, J., 2008. Statistical Arbitrage in the U.S. Equities Market. Available at SSRN: HTTP://SSRN.COM/ABSTRACT=1153505

Barras, L., Scaillet, O., Wermers, R., 2008. False Discoveries in Mutual Fund Performance: Measuring Luck in Estimated Alphas. Available at SSRN: HTTP://PAPERS.SSRN.COM/SOL3/PAPERS.CFM?ABSTRACT_ID=869748

Benjamini, Y., Hochberg, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society, Vol. 57, No.1 (1995), pp. 289-300.

Benjamini, Y., Hochberg, Y., 2000. On the Adaptive Control of the False Discovery Rate in Multiple Testing with Independent Statistics. Journal of Educational and Behavioral Statistics, Vol. 25, No. 1, 60-83 (2000).

Benjamini, Y., Yekutieli, D., 2001. The Control of the False Discovery Rate in Multiple Testing under Dependency. The Annals of Statistics, Vol. 29, Number 4 (2001), 1165-1188.

Benjamini, Y., Krieger, A., Yekutieli, D., 2006. Adaptive linear step-up procedures that control the false discovery rate. Biometrika 93(3), 491–507.

Berk, J., Green, R., 2004. Mutual Fund Flows and Performance in Rational Markets. Journal of Political Economy 112, 1269-1295.

Carhart, M., 1997. On persistence of mutual fund performance. The Journal of Finance, Vol. 52, No. 1, (Mar., 1997).

Chen, H., Jegadeesh, N., Wermers, R., 2000. The value of active mutual fund management: An examination of the stockholdings and trades of fund managers. Journal of Financial and Quantitative Analysis 35, 343-368.

Cochrane, J., 2005. Asset Pricing. Princeton University Press; Revised edition.
Cuthbertson, K., Nitzsche, D., O'Sullivan, N., 2008, "A". UK mutual fund performance: Skill or luck? Journal of Empirical Finance 15 (2008) 613–634.

Cuthbertson, K., Nitzsche, D., O'Sullivan, N., 2008,"B". False discoveries: winners and losers in mutual fund performance. Available at SSRN: HTTP://PAPERS.SSRN.COM/SOL3/PAPERS.CFM?ABSTRACT_ID=1093624

Daniel, K., Grinblatt, M., Titman, S., Wermers, R., 1997. Measuring mutual fund performance with characteristic-based benchmarks. Journal of Finance, Volume 52, Issue 3, (Jul., 1997).

Dudoit, S., Shaffer, J., Boldrick, J., 2003. Multiple Hypothesis Testing in Microarray Experiments. Statistical Science, Vol. 18, No. 1 (2003), pp. 71-103.

Efron, B., 2001. Empirical Bayes Analysis of a Microarray Experiment. Journal of the American Statistical Association, Vol. 96, No. 456, (2001), pp. 1151-1160.

Efron, B., Tibshirani, R., 2002. Empirical Bayes Methods and False Discovery Rates for Microarrays. Genetic Epidemiology 23:70-86 (2002).

Efron, B., 2003. Large-Scale Simultaneous Hypothesis Testing: the choice of a null hypothesis. Journal of the American Statistical Association, Mar 2004.

Efron, B., 2004. Selection and Estimation for Large-Scale Simultaneous Inference. Available at: HTTP://WWW-STAT.STANFORD.EDU/~CKIRBY/BRAD/PAPERS/

Efron, B., 2005. Local False Discovery Rates. Available at: HTTP://WWW-STAT.STANFORD.EDU/~CKIRBY/BRAD/PAPERS/

Efron, B., 2006, "A". Microarrays, Empirical Bayes, and the Two-Groups Model. Statistical Science, Vol. 23, Number 1 (2008), 1-22.

Efron, B., 2006, "B". Testing the significance of sets of genes. The Annals of Applied Statistics, Vol. 1, Number 1 (2007), 107-129.

Efron, B., 2006, "C". Size, Power, and False Discovery Rates. The Annals of Statistics, 2007, Vol. 35, No. 4, 1351-1377.

Efron, B., 2006, "D". Correlation and Large-Scale Simultaneous Significance Testing. Available at: HTTP://WWW-STAT.STANFORD.EDU/~CKIRBY/BRAD/PAPERS/

Efron,B., 2007. Simultaneous inference: when should hypothesis testing problems be combined? The Annals of Applied Statistics, Vol. 2, Number 1 (2008), 197-223.

Elton, E., Gruber, M., Das, S., Hlavka, M., 1993. Efficiency with costly information: A reinterpretation of evidence from managed portfolios. The Review of Financial Studies, Vol. 6, No. 1, (1993).

Elton, E., Gruber, M., Busse, J., 2004. Are Investors Rational? Choices Among Index Funds. Journal of Finance, 59, 261-288.

Fama, E., French, K., 1993. Common risk factors in the returns on stocks and bonds. Journal of Financial Economics, 33, (1993).

Fan, J., Fan, Y., Lv, J., 2008. High dimensional covariance matrix estimation using a factor model. Journal of Econometrics, 147(2008) 186-197.

Ferson, W., Schadt, R., 1996. Measuring Fund Strategy and Performance in Changing Economic Conditions. Journal of Finance 51, 425-461.

Freedman, D., 1981. Bootstrapping Regression Models. The Annals of Statistics, vol. 9, No 6.

Hastie, T., Tibshirani, R., Friedman, J., 2001.The elements of statistical learning. Springer, ISBN 0-387-95284-5.

Hendricks, D., Patel, J., Zeckhauser, R., 1993. Hot Hands in mutual funds: short-term persistent of relative performance, 1974-1988. The Journal of Finance, Vol. 48, No. 1 (Mar., 1993).

Huij, J., Verbeek, M., 2008. On the use of multifactor models to evaluate mutual fund performance. Available at SSRN: HTTP://SSRN.COM/ABSTRACT=906723

Ippolito, R., 1989. Efficiency With Costly Information: A Study of Mutual Fund Performance,1965-1984. The Quarterly Journal of Economics, Vol. 104, No. 1. (Feb., 1989).

Jensen, M., 1968. The Performance of Mutual Funds in the Period 1945–1964. Journal of Finance, Vol. 23, No. 2 (1968).

Jones, C., Shanken, J., 2005. Mutual fund performance with learning across funds. Journal of financial economics, 78(2005), 507-552.

Kosowski, R., Timmermann, A., Wermers, R., White, H., 2006. Can Mutual Fund “Stars” Really Pick Stocks? New Evidence from a Bootstrap Analysis. Journal of Finance, Vol. 61, December 2006.

Kothari, S., Warner, J., 2001. Evaluating Mutual Fund Performance. The Journal of Finance, Vol. 56, No. 5 (Oct., 2001), pp. 1985-2010.

Mamaysky, H., Spiegel, M., Zhang, H., 2007. Improved Forecasting of Mutual Fund Alphas and Betas. Review of Finance, (2007) : 11.

Nitzsche, D., Cuthbertson, K., O'Sullivan, N., 2006. Mutual Fund Performance. Available at SSRN: HTTP://SSRN.COM/ABSTRACT=955807

Otamendi, J., Doncel, L., Grau, P., Sainz, J., 2008. An evaluation on the true statistical relevance of Jensen's alpha trough simulation: An application for Germany. Economics Bulletin, Vol. 7, No. 10 pp. 1-9.

Pastor, L., Stambaugh, R., 2002. Mutual Fund performance and seemingly unrelated assets. Journal of Financial Economics 63 (2002) 315-349.

Pounds, S., Cheng, C., 2006. Robust estimation of the false discovery rate. Bioinformatics, Aug 2006.

Pregibon, D., 1981. Logistic Regression Diagnostics. The Annals of Statistics, Vol. 9, No. 4.

Romano, J., Wolf, M., 2005. Stepwise Multiple Testing as Formalized Data Snooping. Econometrica 73, 1237-1282.

Romano, J., Shaikh, A., Wolf, M., 2007. Control of the False Discovery Rate Under Dependence Using the Bootstrap and Subsampling. University of Zurich Working Paper No. 337. Available at SSRN: HTTP://SSRN.COM/ABSTRACT=1025410

Romano, J., Shaikh, A., Wolf, M., 2008. Formalized Data Snooping Based on Generalized Error Rates. Econometric Theory, 24, 2008.

Scherer, B., 2004. Portfolio Construction and Risk Budgeting, 2nd Edition. Incisive Financial Publishing Ltd. See pp. 68, 142, 158.

Storey, D., Tibshirani, R., 2001. Estimating False Discovery Rates Under Dependence, with Applications to DNA Microarrays. Available at: HTTP://STAT.STANFORD.EDU/REPORTS/PAPERS2001.HTML

Storey, J., 2002. A Direct Approach to False Discovery Rates. Journal of the Royal Statistical Society B 64, 479-498.

Storey, D., Tibshirani, R., 2003. Statistical Significance for Genomewide Studies. Proceedings of the National Academy of Sciences of the United States of America, 2003, vol. 100, p. 9440.

Storey, D., 2003. The positive false discovery rate: a Bayesian interpretation and the q-value. The Annals of Statistics, Vol. 31, Number 6 (2003).

Storey, J., Taylor, J., Siegmund, D., 2004. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. Journal of the Royal Statistical Society, 2004, vol. 66, p. 187.

Strimmer, K., 2008. A Unified Approach to False Discovery Rate Estimation. BMC Bioinformatics, 2008, vol. 9, p. 303.

Turnbull, B., 2007. Optimal Estimation of False Discovery Rates. Available at HTTP://WWW.STANFORD.EDU/~BKATZEN/

van der Laan, M., Hubbard, A., 2005. Quantile function based null distribution in resampling based multiple testing. Statistical Applications in Genetics and Molecular Biology 5, article 14.

Wermers, R., 1999. Mutual Fund Herding and the Impact on Stock Prices. The Journal of Finance, 1999, vol. 54, p. 581.

Wermers, R., 2000. Mutual Fund performance: an empirical decomposition into stock-picking talent, style, transaction costs and expenses. Journal of Finance, Volume LV, No. 4., (Aug., 2000).

White, H., 2000. A Reality Check for Data Snooping. Econometrica, 68, 1097-1126.

Williams, E., 2003. Essays in Multiple Comparison Testing. Ph.D. Dissertation, Department of Economics, UCSD.

Yekutieli, D., Benjamini, Y., 1999. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. Journal of Statistical planning and inference, 1999, vol. 82, p. 171.

APPENDIX

APPENDIX

The dataset consists of monthly net return data obtained from the Center for Research in Security Prices (CRSP) MF database between January 1993 and June 2007 (i.e. 174 monthly observations for a fund that was open throughout that period). The sample was drawn before the CRSP MF database was re-engineered on April 21, 2008.

The term “net returns” means that these returns are adjusted for management expenses, marketing fees (a.k.a. 12b-1), administration costs and trading costs. Management expenses, marketing fees and administration costs comprise the fund’s expense ratio (ER).

Besides, there exist other expenses such as load fees. Because they are not taken into account in net returns, the performance estimate based on net returns is actually an upper bound on what the individual investor can expect.

CRSP MF database consists of all open-ended US mutual funds, but in extracting the target sample of actively managed US domestic equity funds the following two problems had to be solved: 1) identifying the fund’s investment objective; 2) if a fund consists of a few shareclasses, aggregating the returns across shareclasses to produce a single time series of returns. In the CRSP MFLINKS database, which is essentially a merger of abovementioned CRSP MF and so-called Thomson/CDA database, both problems can be solved easily and CRSP MFLINKS was the one used in BSW study. Unfortunately, it is very expensive and therefore this research relies on CRSP MF database.
To solve the investment objective problem, the following algorithm (similar to that of Pastor(2002)) was implemented: according to two investment objective codes,

Strategic Insight Objective (“sp_obj_cd”) and ICDI Objective (“icdi_obj_cd”) the funds of interest were selected and placed into subcategories as follows:

Table A.1 Net returns data for 1911 mutual funds

Assigned subcategory	sp_obj_cd	icdi_obj_cd	Number of funds
Small company growth (SCG)	SCG	n/a	463
Other aggressive growth (OAG)	AGG	AG, AGG	164
Growth (G)	GRO, GMC	LG	886
Growth and Income (GI)	GRI	GI	398
Total			1911

The assignment is performed in top-to-bottom priority, e.g. if a fund has sp_obj_cd = AGG and icdi_obj_cd = LG then it is assigned to OAG category. The categories were assessed yearly and the overall fund objective was determined by the majority.

The following funds were implicitly (via Pastor’s method) or explicitly (using some other indicators from CRSP MF) excluded from the sample:

- International funds
- Money market funds
- Bond funds
- Balanced funds
- Flexible funds
- Funds of funds
- Income funds
- Index funds
- Sector stocks (oil, precious metals, etc) funds
- Preferred stock funds

- Funds with no available objective
- Funds with no available name
- Funds with zero or not available expense ratio
- Funds with zero or not available turnover
- Finds with average yearly TNA less than $5M

To solve the multiple shareclass problem, a separate algorithm was developed based on the available shareclass code (“icdi”), date (to account for possible renaming) and shareclass name. The goal was to obtain a portfolio code to identify the shareclasses belonging to the same MF. Because for the period of 2003-2007 the true portfolio code (“port_code”) was available, it was possible to test the algorithm on a large sample of 29471 shareclass-years and only 51 of them (0.17%) were assigned an incorrect portfolio code. Since for the entire 1993-2007 sample the portfolio code had to be calculated only for 1993-2002, the overall error rate is probably less than 0.17%.

If a MF return is missing, the next non-missing return is discarded since it corresponds to the cumulative return over the entire missed period (CRSP convention). After that, the fund monthly net return was computed by weighting the net return of each shareclass by its monthly total net asset value (“mtna”). Each fund was required to have at least 50 (not necessarily consecutive) monthly returns.

The pre-expense MF data were obtained based on the sample of 1911 funds above. For each MF, its annual expense ratio was computed as a TNA-weighted average of expense ratios of its shareclasses. Then, for each month, 1/12 of the annual expense ratio was added to the MF monthly net return resulting in the return that would be obtained after trading costs but before all costs included in the expense ratio. Funds with less than 50 monthly observations were dropped.

Table A.2 Pre-expense data returns data for 1876 mutual funds

Assigned subcategory	sp_obj_cd	icdi_obj_cd	Number of funds
Small company growth (SCG)	SCG	n/a	457
Other aggressive growth (OAG)	AGG	AG, AGG	161
Growth (G)	GRO, GMC	LG	871
Growth and Income (GI)	GRI	GI	387
Total			1876

Because of some missing expense ratio information, the pre-expense sample includes 1876 funds (Table A.2). For both pre-expense and net returns data, the average number of observations per mutual fund is about 129 (10 3/4 years).

VITA

VITA

NIK TUZOV

Tel: (718) 877-6352
Email: NTUZOV@PURDUE.EDU
LinkedIn: HTTP://WWW.LINKEDIN.COM/IN/NTUZOV

EDUCATION

PURDUE UNIVERSITY, W. Lafayette, IN, USA

Ph.D. in Statistics with emphasis on Quantitative Finance, 05/2009

PhD Thesis: Performance Evaluation & Attribution of Equity Portfolio Managers

M. S. in Statistics with Computational Finance Specialization, 2006

MOSCOW AVIATION INSTITUTE, Moscow, Russia

M.S. in Applied Mathematics, 2000
Thesis: Equity portfolio optimization with credit risk constraints
Participated in project: Portfolio selection for Russian fixed income securities (GKO)

SUBJECTS OF STUDY

- Portfolio Theory
- Equity Derivatives
- Extreme Value in Finance
- Interest Rate Models
- Risk Management
- Market Microstructure
- Monte-Carlo Simulation
- Asset Pricing
- Time Series in Econometrics
- Bootstrap Methods
- Machine Learning
- Model Building & Validation

PROGRAMMING SKILLS

Main tools:

- C++
- Matlab

Have experience with:

- Java
- SAS
- Excel VBA
- S-PLUS / R
- CRSP Database
- Ox

Software samples can be downloaded from:

HTTP://WWW.STAT.PURDUE.EDU/~NTUZOV/

EXPERIENCE

PURDUE DEPARTMENT OF STATISTICS, Lafayette, IN, 2003 – present

Statistical Consultant

- Provide professional consulting in Applied Statistics/ Design of experiment/ Quality control area
- Analyze initial problem statements to proceed with *model building, diagnostics and validation*
- Maintain feedback with clients, superiors and co-workers to ensure high quality of consulting
- Document the client-consultant interaction process for the record

Teaching Assistant

- Assist students with SPSS during labs
- Continuously interact with students, fellow TAs (peer mentoring) and course instructors to improve teaching efficiency

WOOD CREEK CAPITAL, New Haven, CT, Summer 2007

Niche and alternative strategies hedge fund

Quantitative Analyst

- Performed advanced time series model building
- Created financial and general purpose applications in Matlab and VBA

WALTER RAQUET CAPITAL, Stamford, CT, Summer 2006

Equity hedge fund

Quantitative Analyst

- Developed models for performance assessment and hedging of money managers
- Contributed to WR asset allocation strategy
- Trained staff to implement the developed methodology

PUBLICATIONS

Kan, Y., Tuzov, N., 1998. Equity portfolio optimization with default risk constraints. Automation and Remote Control. Moscow, Nov. 1998.

PROFESSIONAL ASSOCIATIONS

- American Statistical Association
- SIAM
- IAFE
- American Finance Association