Implied Measures of Relative Fund Performance*

Steve Hogan
Credit Suisse First Boston
New York, NY

Mitch Warachka†
Singapore Management University
Singapore

January 2005

Abstract

We provide a methodology which incorporates portfolio characteristics into the evaluation of fund performance and exploits their commonality across different fund managers. By measuring relative performance, our procedure also circumvents the need to specify benchmark returns or peer funds. Instead, conditional on portfolio weights in distinct investment classes, implied metrics for selection and market timing ability are inferred from a cross-section of fund returns. Consequently, our technique is robust to herding as well as window dressing, and partially mitigates survivorship bias over longer horizons. Empirically, the conditional information contained in classes defined by industry sectors, assets and geographical regions is found to be of critical importance when evaluating fund management. For each set of portfolio characteristics, we identify fund managers with moderate success at either selecting securities or timing-the-market. However, individual funds cannot overperform across all three criteria and evidence of investment skill dissipates at a higher performance threshold.

JEL Code: G11
Keywords: Mutual Funds, Performance Evaluation

*We thank Morningstar for providing us with the required data. In particular, the efforts of Stephen Murphy at Morningstar are greatly appreciated. The comments of Sheridan Titman significantly improved an earlier version of this paper. All remaining errors are our own. We also acknowledge the generous assistance provided by the Center for Academic Computing at SMU. Mitch Warachka is grateful for financial support from the SMU-Wharton Research Center, Singapore Management University.

†Please address correspondence to Mitch Warachka, Business Block #04-10, 469 Bukit Timah Road, Singapore Management University, Singapore, 259756. Phone: (65) 6822-0249 Fax: (65) 6822-0777 E-mail: mitchell@smu.edu.sg

Implied Measures of Relative Fund Performance

January 2005

Abstract

We provide a methodology which incorporates portfolio characteristics into the evaluation of fund performance and exploits their commonality across different fund managers. By measuring relative performance, our procedure also circumvents the need to specify benchmark returns or peer funds. Instead, conditional on portfolio weights in distinct investment classes, implied metrics for selection and market timing ability are inferred from a cross-section of fund returns. Consequently, our technique is robust to herding as well as window dressing, and partially mitigates survivorship bias over longer horizons. Empirically, the conditional information contained in classes defined by industry sectors, assets and geographical regions is found to be of critical importance when evaluating fund management. For each set of portfolio characteristics, we identify fund managers with moderate success at either selecting securities or timing-the-market. However, individual funds cannot overperform across all three criteria and evidence of investment skill dissipates at a higher performance threshold.

1 Introduction

This paper offers a methodology that circumvents the specification of exogenous benchmarks or peer funds when evaluating fund management. By focusing on relative fund performance, we construct fund-specific benchmarks conditioned on portfolio characteristics in different *classes*. Class definitions may pertain to fund investments in distinct industry sectors (healthcare vs. energy), assets (equity vs. bonds), geographical regions (United States vs. Japan) or any other conditional information capable of being expressed in terms of portfolio weights.[1] Four compelling advantages of using portfolio characteristics instead of traditional factor sensitivities when evaluating funds are reported in Daniel, Grinblatt, Titman and Wermers (1997). Furthermore, Chan, Chen and Lakonishok (2002) confirm the superiority of portfolio characteristics versus factor sensitivities when predicting fund performance. Therefore, through the class-specific portfolio weights, we continue to condition fund performance on portfolio characteristics.

However, in contrast to our cross-sectional procedure, Daniel, Grinblatt, Titman and Wermers (1997) construct "hypothetical" peer funds using book-to-market, size and past return criteria expressed in quintiles. Their procedure considers the portfolio weights of individual securities, and involves a sorting exercise based on their properties from the previous year. Furthermore, the benchmark return for each of the 125 passive portfolios underlying their characteristic-based procedure is a value-weighted average which does not necessarily reflect actual fund investments.[2]

[1]The proposed methodology could refine the performance measurement of funds adhering to a specific style. For example, including only value funds in our cross-sectional procedure while conditioning on industry portfolio weights allows industry criteria to complement an existing style classification. Another alternative would condition directly on portfolio weights defined jointly by value and industry characteristics.

[2]For example, the CS_t measure of Daniel, Grinblatt, Titman and Wermers (1997) has the portfolio weights of individual securities weighting their excess returns, which are defined via a benchmark return that is not conditioned on observed portfolio weights. In particular, this benchmark return could involve securities held by none of the funds. Conversely, Section 3 demonstrates that the benchmark returns for each class (and fund) in

Moreover, none of these passive portfolios are available in the market for investment. Instead, Daniel, Grinblatt, Titman and Wermers (1997) study the ease at which a fund's performance is replicated by the untraded passive portfolios.

Chan, Chen and Lakonishok (2002) compare a fund's performance with the S&P 500 in terms of its respective book-to-market and size properties expressed as deciles. While the S&P 500 is a popular benchmark, it is inappropriate for funds consisting of both equity and bonds, international funds, or US-focused funds concentrated in specific industries. Indeed, as documented in the existing literature, specifying an appropriate benchmark is a highly contentious issue when evaluating an individual fund's performance.

Thus, the existing literature determines whether fund managers have investment ability in comparison to absolute benchmarks, whose specification taints subsequent performance conclusions. In addition, absolute benchmarks imply the majority of fund managers (all of them in theory) could either underperform or overperform each period.[3] This criticism undermines the ability of existing performance metrics to advise investors regarding the selection of funds between available alternatives.

Consequently, we introduce a relative evaluation framework. Thus, our objective is not to investigate whether fund managers are capable of outperforming "mechanical trading rules" as in Daniel, Grinblatt, Titman and Wermers (1997) but whether they exhibit relative overperformance. Thus, when confronted with a decision between funds available for investment, our evaluation methodology is more appropriate. Indeed, the concept of relative evaluation underlies the recent research of Cohen, Coval and Pástor (2004) who compare a fund's portfolio weights in

our relative procedure are implied from a conditional analysis involving observed fund portfolio weights.

[3]Analogously, when assigning grades, a professor cannot have 80% of their students be above or below average. Thus, a relative normalization is often conducted to eliminate criticism that their absolute standards are inappropriate. Our relative procedure guarantees that exactly 50% of fund managers are above (below) their implied fund-specific benchmark return each period, with details provided in Section 3.

individual securities (as well as their changes) with those of "distinguished" funds. By exploiting the "overlap" in fund investments, their technique motivates the relative rather than absolute benchmarks constructed in our framework. However, as elaborated on below, we derive implied relative performance measures rather than forming a weighted sum of existing performance measures (such as Jensen's measure) as in Cohen, Coval and Pástor (2004).[4]

Thus, the fund evaluation procedure in this paper bridges the gap between the streams of literature which have focused on the conditional information in portfolio characteristics and their commonality across multiple fund managers. Moreover, the computational complexity involved in manipulating the portfolio weights of individual securities is eliminated by our approach which only conditions on class-specific portfolio weights. Therefore, our methodology is robust to fluctuations in the portfolio weights of individual securities within each class. A detailed comparison of our methodology with the existing literature is contained in Section 4 and Appendix A.

For emphasis, the *fund versus fund* analysis in this paper circumvents the need to exogenously specify peer funds or benchmarks which compromises conclusions from traditional *fund versus absolute benchmark* studies. Instead, our evaluation procedure solves for a benchmark return and variance which is customized to each fund. By conditioning these moments on the portfolio weights of an individual fund, peer funds are not required. Hence, we incorporate conditional information with greater precision than existing studies based on quintiles or deciles to describe portfolio characteristics. Given the relevant conditional information (class definitions), a fund

[4]To be more precise, the approach of Cohen, Coval and Pástor (2004) does not permit relative evaluation as it remains dependent on absolute benchmarks such as Jensen's measure (or those of Daniel, Grinblatt, Titman and Wermers (1997)). Nonetheless, these authors recognize the importance of comparing portfolio weights between funds to assess management ability. Indeed, overlap between the portfolio weights of different fund managers is crucial to both our procedures. As seen in Section 3, distinguished fund managers could determine the class-specific returns and variances underlying our proposed methodology, at the expense of having only half the funds be above (below) their benchmark returns.

manager cannot claim their implied benchmark fails to represent their portfolio. In addition, the implied benchmark return and variance form "t-statistics" which enables the extent of a fund's overperformance to be easily computed, and offers greater intuition when interpreting a fund manager's skill.

To briefly summarize our implied methodology, we first infer returns and variances for each class using a *cross-sectional* procedure that captures relative performance. Class-specific returns are obtained by "regressing" aggregate fund returns on their portfolio weights in each class. Furthermore, to calibrate variances for each class, the regression residuals are also functions of these portfolio weights. Combining these class-specific implied moments with the portfolio weights of individual funds yield fund-specific benchmark returns and variances for both selection and market timing ability. Separate t-statistics for these attributes are then created. Selectivity is related to a fund's choice of individual securities within a class, while market timing ability is concerned with a fund manager's portfolio allocation between the classes.[5]

In the subsequent and final step, a fund's performance is evaluated by examining its *time series* of implied t-statistics. Overperforming fund managers have selectivity and market timing statistics that consistently exceed a designated threshold. This threshold represents a critical value (or equivalent percentile). Intuitively, a fund manager with moderate selection ability has selectivity metrics which are frequently positive. Since each performance threshold has an equivalent expression in terms of percentile, our procedure is robust to outliers. For example, a threshold which requires a fund's performance to exceed its benchmark return corresponds to the 50^{th} percentile or median, which insensitive to distributional assumptions imposed on the benchmark's estimation.

Our relative evaluation approach is robust to window dressing and herding by fund managers.

[5]For convenience, we refer to funds and fund managers interchangeably although Baks (2003) reports that funds are more important than fund managers when explaining performance.

Indeed, transactions within a class leave the portfolio weights underlying our analysis unaltered. Furthermore, a fund which engages in window dressing by exchanging low for high return investments prior to a portfolio disclosure has its benchmark comprised of higher return characteristics. Consequently, the relatively low return earned by its actual portfolio over the prior period causes the fund to display underperformance. Similarly, investment decisions that arise from herding, either within a class or between classes, prevent fund managers from distinguishing themselves from their selectivity and market timing benchmarks.

By construction, our cross-sectional metrics are free from survivorship bias as they are computed at a single timepoint. Furthermore, provided survival is related to having invested (or not having invested) in certain classes, survivorship bias is partially mitigated over longer time horizons. Specifically, each cross-sectional t-statistic in the time series underlying performance measurement is conditioned on portfolio weights that reflect survival.[6]

An empirical illustration of our methodology utilizes a survivorship bias-free set of Morningstar data that consists of portfolio weights for different industry sectors, assets and geographical regions, of which there are twelve, four and five classes respectively. For all three classifications, moderate selection and market timing ability is concentrated in a small but statistically significant subset of fund managers. However, investment skill dissipates at a higher performance threshold which requires a fund's return to be one standard deviation above its benchmark.

Moreover, the "intersection" of moderately overperforming funds across the three portfolio weight classifications is nearly empty. Hence, conditioning on different portfolio weight characteristics yields distinct subsets of overperforming funds. Therefore, selection and market timing ability is not diffused across the industry, asset and regional classifications since an individual fund rarely overperforms across all three criteria. Furthermore, funds may possess the abil-

[6]However, factors that influence survival but are unrelated to portfolio weights, such as a fund's age, may continue to bias our performance measure.

ity to either select securities or time-the-market, but not both these attributes. Overall, fund management skill appears to be specialized.

We also document the critical importance of conditioning on portfolio weights when evaluating fund management. Specifically, our conditional performance measure has little in common with its unconditional counterpart that ignores the information in portfolio weights. In addition, fund performance appears unrelated to a fund's expense ratio, size and turnover. Instead, performance is explained by variability in a fund's class-specific portfolio weights over time and their dispersion across the different classes. For example, funds that focus their investments in a small number of classes are more prone to exhibit selectivity but less likely to time-the-market.

The remainder of this paper begins with the introduction of our proposed methodology in Section 2. Our estimation procedure is then described in Section 3 while properties of the implied performance metrics are discussed in Section 4. Section 5 explains the data underlying our empirical study whose results are presented in Section 6. The conclusion and suggestions for further research follow in Section 7.

2 Implied Metrics and Performance Measurement

This section describes the economic content underlying our implied statistics for selection and market timing ability. Classes are designated $c = 1, \ldots, C$ while the funds themselves are indexed by the subscript $p = 1, \ldots, P$. Note that our methodology only requires portfolio weights in each class. The holdings of individual securities, reported in Falkenstein (1996) as being influenced by their visibility, transactions costs and idiosyncratic volatility, are unnecessary.

2.1 Selectivity Metric

Our first cross-sectional metric evaluates a fund manager's selection ability. We begin our analysis by introducing some additional notation:

1. W: a $C \times P$ matrix of observed portfolio weights in class c for fund p.

2. $\hat{R}$: a C dimensional implied vector of returns $\hat{r}_c$ for each class.

3. $\hat{\Theta}$: a $C \times C$ implied matrix of variances and covariances $\hat{\sigma}_{c,c'}$ for each class.

Conditional on portfolio weights in each class, a cross-sectional "regression" described in the next section infers the $\hat{R}$ and $\hat{\Theta}$ parameters. These implied class-specific quantities are the fundamental "building-blocks" of our technique. However, we first consider the economic structure of our evaluation approach assuming they have been calibrated.

Let $w_{p,c}$ denote the portfolio weight of fund p in class c while $r_{p,c}$ signifies its unobservable corresponding return. A fund-specific implied benchmark return

$$\hat{r}_p = \sum_{c=1}^{C} w_{p,c}\, \hat{r}_c \tag{1}$$

is formed by combining a fund's portfolio weights with the class-specific returns $\hat{r}_c$ comprising $\hat{R}$. Using this benchmark return, the observed aggregate return of a fund is decomposed as

$$\begin{aligned} r_p &= \sum_{c=1}^{C} w_{p,c}\, r_{p,c} \\ &= \sum_{c=1}^{C} w_{p,c}\, \hat{r}_c + \sum_{c=1}^{C} w_{p,c}\, (r_{p,c} - \hat{r}_c) \\ &= \hat{r}_p + (r_p - \hat{r}_p) \qquad (2) \\ &= \text{implied benchmark without selection ability} + \text{selection ability}\,. \end{aligned}$$

Each $r_{p,c} - \hat{r}_c$ deviation results from a fund manager's selection of securities within class c, and ultimately yields the weighted difference $r_p - \hat{r}_p$. In particular, a fund manager skilled at selecting individual securities within the various classes has positive deviations from their implied benchmark with r_p exceeding $\hat{r}_p$. For emphasis, the decomposition in equation (2) does not require the unobservable returns $r_{p,c}$ as these terms only serve an intermediary role in our analysis.

The implied benchmark return for selection ability in equation (1) is written more succinctly as

$$\hat{r}_p \;=\; \mathbf{w}_p^T \, \hat{R} \,, \tag{3}$$

where $\mathbf{w}_p$ is the vector of portfolio weights for fund p in each class. Specifically, $\mathbf{w}_p$ is a column of W containing the portfolio weights $w_{p,c}$ in each of the C classes. Using standard operations from the portfolio theory literature, the corresponding variance of a fund's benchmark return equals

$$\begin{aligned} \hat{\sigma}_p^2 &= \sum_{c=1}^{C} w_{p,c} \, \hat{\sigma}_c^2 + \sum_{c=1}^{C} \sum_{c' \neq 1}^{C} w_{p,c} \, w_{p,c'} \, \hat{\sigma}_{c,c'} \\ &= \mathbf{w}_p^T \, \hat{\Theta} \, \mathbf{w}_p \,. \end{aligned} \tag{4}$$

Observe that equations (3) and (4) condition $\hat{r}_p$ and $\hat{\sigma}_p^2$ on a fund's investments in each class since both quantities are functions of $\mathbf{w}_p$.

To clarify, $\hat{\sigma}_p$ does not *directly* reflect variability in a fund's return or the returns of individual securities. Instead, $\hat{\sigma}_p$ reflects variability in the cross-section of observed fund returns, after conditioning on their portfolio weights in each class. For example, if all fund managers hold identical positions in each individual security, $\hat{\sigma}_p$ equals zero even when the securities are highly volatile.

For each fund, the selectivity statistic

$$S_p \;=\; \frac{r_p - \hat{r}_p}{\hat{\sigma}_p} \tag{5}$$

is formed. Thus, the proposed selectivity metric evaluates the deviation between a fund's return and its benchmark, normalized by the benchmark's volatility. Under the null hypothesis of no selection ability, $S_p \overset{d}{\sim} \mathcal{N}(0,1)$ with positive (negative) values indicating fund overperformance (underperformance).[7]

Although similar in appearance, several important differences between the Sharpe ratio and the S_p metric in equation (5) are apparent. First, the implied benchmark return $\hat{r}_p$ replaces the riskfree rate. Second, $\hat{r}_p$ and $\hat{\sigma}_p$ involve implied parameters. Third, both these quantities are conditioned on a fund's class portfolio weights, hence the return and volatility of the benchmark are fund-specific. Fourth, as mentioned above, $\hat{\sigma}_p$ refers to variability in the cross-section of fund returns, not the volatility of individual fund or security returns.

2.2 Market Timing Metric

Besides selecting securities within a class, fund managers also allocate their portfolio across the various classes. In particular, funds that successfully time-the-market earn higher returns by deviating from *benchmark portfolio weights* denoted $w_{B,c}$ which are discussed later in this subsection.

Conditional on fund and benchmark portfolio weights, the previous benchmark return $\hat{r}_p$ is

[7] More formally, S_p has a t-distribution. However, we assume the number of available funds is sufficiently large to invoke normality throughout the paper.

decomposed as

$$
\begin{aligned}
\hat{r}_p &= \sum_{c=1}^{C} w_{p,c}\,\hat{r}_c \\
&= \sum_{c=1}^{C} w_{B,c}\,\hat{r}_c + \sum_{c=1}^{C} \left(w_{p,c} - w_{B,c}\right)\,\hat{r}_c \\
&= \hat{r}_B + \left(\hat{r}_p - \hat{r}_B\right)\,. \qquad (6)
\end{aligned}
$$

The term $\hat{r}_B$ represents the return of a fund manager without market timing or selection ability who simply earns the benchmark return in each class and holds the benchmark portfolio weights.

Extending the decomposition in equation (2), the observed return of a fund becomes

$$
\begin{aligned}
r_p &= \sum_{c=1}^{C} w_{B,c}\,\hat{r}_c + \sum_{c=1}^{C} \left(w_{p,c} - w_{B,c}\right)\,\hat{r}_c + \sum_{c=1}^{C} w_{p,c}\,\left(r_{p,c} - \hat{r}_c\right) \\
&= \hat{r}_B + \left(\hat{r}_p - \hat{r}_B\right) + \left(r_p - \hat{r}_p\right) \qquad (7) \\
&= \text{implied benchmark without market timing or selection ability} \\
&\quad + \text{market timing ability} + \text{selection ability}\,.
\end{aligned}
$$

For emphasis, the selectivity metric S_p evaluates the deviation $r_p - \hat{r}_p$ conditional on a fund's portfolio characteristics, while market timing considers the difference $\hat{r}_p - \hat{r}_B$ by also conditioning on a set of benchmark portfolio weights.

The benchmark return and variance of the market timing measure, denoted $\hat{r}_B$ and $\hat{\sigma}_B^2$ respectively, equal

$$\hat{r}_B = \sum_{c=1}^{C} w_{B,c}\, \hat{r}_c$$
$$= \mathbf{w}_B^T \hat{R}, \quad \text{and} \tag{8}$$

$$\hat{\sigma}_B^2 = \sum_{c=1}^{C} (w_{p,c} - w_{B,c})\; \hat{\sigma}_c^2 + \sum_{c=1}^{C}\sum_{c'\neq 1}^{C} (w_{p,c} - w_{B,c})\; (w_{p,c'} - w_{B,c'})\; \hat{\sigma}_{c,c'}$$
$$= (\mathbf{w}_p - \mathbf{w}_B)^T\; \hat{\Theta}\; (\mathbf{w}_p - \mathbf{w}_B)\ . \tag{9}$$

The above information combined with equation (6) yields the following T_p statistic for market timing ability

$$T_P = \frac{\hat{r}_p - \hat{r}_B}{\hat{\sigma}_B}\,. \tag{10}$$

Under the null hypothesis of no market timing ability, this metric has a $\mathcal{N}(0,1)$ distribution.

Combining portfolio weights with returns to measure market timing ability has also been investigated in Becker, Ferson, Myers and Schill (1999). However, the existing literature examines the correlation between changes in portfolio weights and future returns of individual securities rather than a return decomposition. As pointed out in Grinblatt and Titman (1993), the correlation approach is problematic when fund managers exploit serial correlation in returns, alter their portfolio's risk across time, or target securities whose expected return and risk have recently risen.

The flexibility to specify different benchmark portfolio weights is important as they may depend on an investor's objective. For example, investors concerned with obtaining a high overall return could utilize benchmark weights that reflect the implied class returns. Another alternative which parallels Grinblatt and Titman (1993) could invoke changes in the class return between consecutive periods.

However, as discussed in Chan, Chen and Lakonishok (2002), investors seeking diversification are not necessarily well-served by fund managers who attempt to time-the-market. Indeed, if a fund deviates from an investor's preferred allocation strategy, then evidence of market timing is not desirable.[8]

Without assuming specific investor preferences, we consider benchmark portfolio weights defined as

$$w_{B,c} = \frac{1}{P}\sum_{p=1}^{P} w_{p,c} \quad \text{for } c = 1, \ldots, C \tag{11}$$

which equal the average portfolio weights across funds at a point in time. A robustness test in Section 6 excludes funds from the computation of $w_{B,c}$ that seldom alter their portfolio weights. This modification recognizes that certain funds may have constraints on their allocation between the classes. However, as seen in equation (7), any enhancement preserves the relative nature of our procedure since deviations from the benchmark portfolio weights are multiplied by the implied class returns.

Although our selectivity and market timing metrics are both cross-sectional, their time series is the basis for evaluating fund managers over longer horizons as seen in the next subsection.

2.3 Performance Measurement for Individual Funds

Intuitively, evidence of investment skill requires a fund's implied statistics to consistently exceed a specified threshold. However, evaluating the percentage of funds in the tail probability at a single point in time is not informative. Indeed, such an analysis merely verifies or contradicts the underlying assumption of normality invoked when calibrating $\hat{R}$ and $\hat{\Theta}$ as discussed in the next section.

For brevity, our current exposition focuses on the selectivity attribute but the statistical technique is immediately applicable to market timing. Let n_p denote the number of observations

[8] Funds that release their "target" portfolio weights in each class enable market timing ability to be inferred relative to their stated portfolio objectives.

for fund p during the sample period. As seen from equation (5), each $S_{p,i}$ metric has an i.i.d. $\mathcal{N}(0,1)$ distribution under the null hypothesis of no selection ability.

Although a simple test of performance could involve the average of the implied metrics,

$$\bar{S}_p = \frac{1}{n_p} \sum_{i=1}^{n_p} S_{p,i} \overset{d}{\sim} \frac{1}{\sqrt{n_p}} \mathcal{N}(0,1) , \tag{12}$$

a more robust procedure is desirable when large $S_{p,i}$ values result from risky activities that cannot be successfully replicated over time. Therefore, we consider the frequency of $S_{p,i}$ statistics above a designated threshold denoted K. More precisely, the value one is assigned to every $S_{p,i}$ statistic greater than K, and zero to its counterpart below this threshold. Under this binary classification scheme, the number of occurences where $S_{p,i} > K$ is defined as

$$X_p = \sum_{i=1}^{n_p} 1_{\{S_{p,i} > K\}} , \tag{13}$$

implying X_p has a binomial distribution. As a result, our multiperiod performance measure is related to the binomial approach of Agarwal and Naik (2000) when $K = 0$.

Let α represent the probability that $S_{p,i} > K$. Given the distribution of $S_{p,i}$ under the null, the relationship between α and K is available from the standard normal cumulative distribution function as α equals the tail probability associated with $\mathcal{N}(0,1) \geq K$. Hence, K and α are often used interchangeably to signify the performance threshold.

The associated null hypothesis of no performance is $H_0 : p \leq \alpha$, where p denotes the sample probability that $S_{p,i} > K$, estimated as $\hat{p} = \frac{X_p}{n_p}$. Thus, $\hat{p}$ equals the proportion of the selectivity statistics that exceed the threshold. The resulting test statistic for selectivity is denoted $B_p(S_{p,i}|\alpha, n_p)$ with a binomial distribution, $Bin(\alpha, n_p)$, under the null. Thus, the corresponding p-value of the performance measure equals the probability that $Bin(\alpha, n_p) \geq X_p$, implying the test statistic is rejected whenever

$$p\text{-value of } B_p(S_{p,i}|\alpha, n_p) = \sum_{i=X_p}^{n_p} \begin{pmatrix} n_p \\ i \end{pmatrix} \alpha^i \, (1-\alpha)^{n_p - i} \tag{14}$$

is below its specified Type I error (denoted γ in Section 6).

For large n_p, the following approximation is applicable

$$B_p(S_{p,i}|\alpha, n_p) = \frac{\hat{p} - \alpha}{\sqrt{\frac{\alpha\,(1-\alpha)}{n_p}}} \overset{d}{\sim} \mathcal{N}(0,1)\,. \tag{15}$$

To examine an individual fund's market timing ability, a $B_p(T_{p,i}|\alpha, n_p)$ test statistic is formed from their $T_{p,i}$ time series.[9]

A fund manager's return is often compared to an index such as the S&P 500 without any regard to the benchmark's volatility. In these circumstances, overperformance (underperformance) is simply indicated by a fund's return being above (below) the benchmark. This situation is equivalent to equation (14) with $K = 0$ ($\alpha = 0.50$), providing a performance measure independent of the benchmark's variance. In particular, when $K = 0$, only symmetry of the implied distribution is assumed which results in a highly robust performance measure. Indeed, every threshold has an equivalent expression as a percentile (tail probability) which is robust to outliers. For completeness, we also implement a higher performance threshold equal to $K = 1$ in our empirical study. This threshold requires a fund's return to exceed its benchmark by one standard deviation.

Finally, observe that our performance measure analyzes the *frequency* at which an individual fund's $S_{p,i}$ and $T_{p,i}$ metrics exceed a threshold. This concept is related to persistence which examines the *consecutive* nature of fund returns over adjacent subperiods.

Having introduced the implied selection and market timing metrics as well as their performance measure, the next section details our statistical methodology for inferring the benchmark parameters.

[9] To ensure later statistical tests are of the stated significance level when employing the discrete binomial distribution, a randomization correction is incorporated with details in Casella and Berger (1990).

3 Implied Estimation Procedure

This section focuses on estimating the implied $\hat{R}$ and $\hat{\Theta}$ parameters. The vector of observed fund returns denoted R_P is assumed to be normally distributed as $\mathcal{N}\left(W^T\hat{R}, \operatorname{diag}\left\{W^T\hat{\Theta}W\right\}\right)$ where $\operatorname{diag}\left\{W^T\hat{\Theta}W\right\}$ denotes the diagonal of the $W^T\hat{\Theta}W$ matrix, written explicitly as

$$\begin{bmatrix} \mathbf{w}_1^T\,\hat{\Theta}\,\mathbf{w}_1 & 0 & \dots & 0 \\ 0 & \mathbf{w}_2^T\,\hat{\Theta}\,\mathbf{w}_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & \mathbf{w}_P^T\,\hat{\Theta}\,\mathbf{w}_P \end{bmatrix} .$$

Intuitively, our cross-sectional statistical technique may be expressed in terms of the following "regression" formulation

$$r_p \;=\; \mathbf{w}_p^T\,\hat{R} + \hat{\epsilon}_p \qquad \text{where} \qquad \hat{\epsilon}_p \overset{d}{\sim} \mathcal{N}\left(0, \mathbf{w}_p^T\,\hat{\Theta}\,\mathbf{w}_p\right) , \tag{16}$$

which may be summarized as

$$r_p \;=\; \hat{r}_p + \mathcal{N}\left(0, \hat{\sigma}_p^2\right) . \tag{17}$$

Therefore, after conditioning on portfolio weights, the error terms $\hat{\epsilon}_p$ are independent but not identically distributed due to their dependence on a fund's portfolio weights. The *conditional* independence of the regression residuals is consistent with considering the diagonal elements of $W^T\hat{\Theta}W$, and found to be appropriate in a later robustness test. For emphasis, the possibility of unconditional correlation between fund returns is not excluded, nor are any restrictions imposed on $\hat{\Theta}$. Furthermore, equation (17) highlights the relative nature of our procedure as only 50% of the funds (at a single point in time) are able to exceed their benchmark return. The conceptual problem with exogenously specifying class returns and variances, instead of implementing our implied procedure, is explained in Appendix A.

Overall, a maximum likelihood procedure solves for class-specific returns and variances that best explain observed fund returns, conditional on their portfolio weights. Maximizing the conditional log-likelihood of observed fund returns involves maximizing the function[10]

$$\ln L(\hat{R}, \hat{\Theta}) = -\frac{P}{2}\ln(2\,\pi) - \frac{1}{2}\sum_{p=1}^{P}\ln\left(\mathbf{w}_p^T\,\hat{\Theta}\,\mathbf{w}_p\right) - \frac{1}{2}\sum_{p=1}^{P}\frac{\left(r_p - \mathbf{w}_p^T\,\hat{R}\right)^2}{\mathbf{w}_p^T\,\hat{\Theta}\,\mathbf{w}_p}\,, \tag{18}$$

or simply minimizing

$$\sum_{p=1}^{P}\ln\left(\mathbf{w}_p^T\,\hat{\Theta}\,\mathbf{w}_p\right) + \sum_{p=1}^{P}\frac{\left(r_p - \mathbf{w}_p^T\,\hat{R}\right)^2}{\mathbf{w}_p^T\,\hat{\Theta}\,\mathbf{w}_p} \tag{19}$$

with respect to $\hat{R}$ and $\hat{\Theta}$. Since their solutions are intertwined, an iterative procedure is necessary. For notational simplicity, denote $\hat{\Lambda}$ as $\text{diag}\left\{W^T\hat{\Theta}W\right\}$. First, given $\hat{\Theta}$, the solution for $\hat{R}$ equals

$$\hat{R} = \left(W^T\,\hat{\Lambda}^{-1}\,W\right)^{-1}\,W^T\,\hat{\Lambda}^{-1}\,R_P\,, \tag{20}$$

which resembles a weighted least-squares estimator. Second, conditional on $\hat{R}$, elements of the variance-covariance matrix are obtained by minimizing equation (19) with respect to $\hat{\Theta}$. Thus, the following iterative scheme is available:

1. Given $\hat{\Theta}_j$, solve for $\hat{R}_j$ using equation (20).

2. Given $\hat{R}_j$, solve for $\hat{\Theta}_{j+1}$ by minimizing equation (19) using non-linear optimization.

3. Repeat steps 1 and 2 for $j = 0, 1, 2, \ldots$ until convergence is achieved.

It is important to emphasize that the class-specific benchmark parameters $\hat{r}_c$ and $\hat{\sigma}_c^2$ incorporate the returns of any fund with an investment in class c. In other words, every fund which

[10]For clarification, our estimation procedure is not related to the two-stage technique of Fama and MacBeth (1973) in which estimated coefficients are subsequently used as independent variables. Instead, our calibration exercise requires an iterative procedure to estimate both moments of the normal distribution but only utilizes observed returns and portfolio weights.

invests in this class contributes to their estimation. Indeed, by conditioning on class portfolio weights, there is no need to specify peer funds. Furthermore, unlike benchmark returns computed as the average return of individual securities in a passive portfolio (Daniel, Grinblatt, Titman and Wermers (1997)), which need not reflect the actual positions of fund managers, the estimation of $\hat{r}_c$ conditions on their portfolio weights.[11] Appendix A briefly explains the execution of our methodology when portfolio weights for individual securities are available.

4 Properties of Implied Performance Measures

In this section, the influence of window dressing and herding as well as survival on the performance measure is discussed. We also compare and contrast our evaluation framework with the existing literature.

To begin with, we return to the problem of self-misspecification highlighted in Brown and Goetzmann (1997). Along with Chan, Chen and Lakonishok (2002), they find that changes in a fund's reported style usually coincide with poor previous performance. In contrast, the time-varying portfolio weights underlying our assessment procedure provide greater objectivity into a fund's investment characteristics.

4.1 Window Dressing and Herding Behavior

Fund managers may engage in window dressing by selling "embarrassing" positions prior to a portfolio disclosure date. Lakonishok, Shleifer, Thaler and Vishny (1991) and Musto (1997, 1999) as well as Carhart, Kaniel, Musto and Reed (2002) provide empirical evidence of this behavior. Furthermore, Wermers (1999) reports that growth orientated funds often herd. Conflicting em-

[11]Passively managed investments such as index funds could augment the calibration of the implied benchmarks. The Bayesian studies of Pástor and Stambaugh (2002a, 2002b) utilize non-benchmark passive investments to better distinguish fund management skill from model inaccuracy.

pirical evidence on this phenomena is contained in Lakonishok, Shleifer and Vishny (1992) as well as Grinblatt, Titman and Wermers (1995).

However, the class portfolio weights in our analysis are unaltered by transactions within a class. Furthermore, selling securities in low return classes and purchasing their high return counterparts prior to a portfolio disclosure compromises a fund's relative selection ability. Specifically, the selectivity benchmark conditions on high return characteristics despite the relatively low realized return resulting from the fund's actual investments over the prior period. Thus, window dressing would undermine a fund's selection ability.[12]

In addition, the benchmark portfolio weights in equation (11) account for commonality in fund allocations between classes, while common investment strategies within the classes are reflected in the selectivity benchmark. Consequently, funds which herd are less able to distinguish themselves from their implied benchmarks that reflect relative investment skill. Therefore, herding behavior reduces a fund's ability to display either selection or market timing ability.

4.2 Survivorship Bias

Although survivorship bias is inherently a database issue, our approach may partially reduce its impact on performance measurement when survival is related to a fund's class portfolio weights. In discussing the potential problem of survivorship bias, we return to the existing literature on persistence which also operates over an extended time horizon.

[12]Funds cannot "reverse" window dress by investing in poorly performing classes prior to a portfolio disclosure. Indeed, such behavior modifies the relative benchmark returns and would constitute herding. Only a very small subset of funds willing to disclose positions in poorly performing classes could successfully engage in this deception without the implied benchmarks being adjusted. To further investigate this issue, a robustness test involving portfolio weights from the period prior to the reported return is conducted. However, consistent with the infrequent rebalancing of class portfolio weights reported in Subsection 6.3, no material impact on our empirical conclusions is detected.

As discussed in Brown, Goetzmann, Ibottson and Ross (1992) as well as Carpenter and Lynch (1999), persistence may be overstated as a result of heteroskedastic fund volatility.[13] In contrast, multiperiod persistence tests are biased towards reversals as funds with poor previous performance survive only by improving. Brown and Goetzmann (1995) and Carhart (1997) find empirical evidence consistent with multiperiod survival criteria. Furthermore, both Elton, Gruber and Blake (1996) as well as Carhart, Carpenter, Lynch and Musto (2002) report that relationships between performance and fund characteristics are distorted by survivorship bias.

Since the selection and market timing ability metrics in equations (5) and (10) respectively are inferred from a cross-section of returns and portfolio weights at a single point in time, survivorship bias does not influence their values. Furthermore, the attrition of poor funds actually induces a higher implied benchmark return.[14]

More importantly, provided survival is linked to investments in particular classes, it becomes incorporated into the subsequent performance measure of equation (14). For example, when survival depends on having invested (or not having invested) in specific classes, survival is manifested into the time series of S_p and T_p metrics through the portfolio weights. Therefore, the performance measure implicitly accounts for survival. However, factors such as a fund's age which are relevant to its survival but unrelated to its portfolio weights may continue to bias the proposed performance measure.

[13] Funds with greater volatility should offer higher returns, provided they survive. However, our methodology is independent of individual fund return volatility. Instead, we calibrate cross-sectional variation between fund returns conditional on portfolio weights.

[14] Brown and Goetzmann (1995) among others find empirical evidence that fund attrition is a consequence of poor returns.

4.3 Comparison with Existing Fund Performance Measures

There exists a vast literature on fund evaluation. The Jensen measure considers the regression intercept of excess fund returns on those of its benchmark. Treynor and Mazuy (1966) include an additional independent variable, the squared excess return, to measure market timing ability via the product of its coefficient with the benchmark return's variance. Both of these methods depend on a specified benchmark, while the Sharpe ratio and Treynor measures normalize a fund's excess return by its volatility and market beta respectively.

A fund's alpha is defined as its return minus the average return of all funds adopting a similar strategy. However, alpha is sensitive to leverage as it ignores return volatility. To alleviate this shortcoming, the appraisal ratio divides alpha by the residual standard deviation obtained from regressing a fund's return on the average of its peers.[15]

However, the above performance measures are not conditioned on a fund's portfolio characteristics. In contrast, our approach directly instills this conditional information into the assessment of fund management. Moreover, we are relieved of having to exogenously specify benchmarks or peer funds.

For an uninformed investor with power utility, the positive period weighting measure in Grinblatt and Titman (1989) utilizes a time-series of weights. These weights are the result of solving two equations; the first ensuring the benchmark's excess return is zero when summed against the weights, and the second equating the fund's performance (Jensen's measure) with the weighted sum of its excess returns. Consequently, these weights are not unique and require the specification of both a benchmark and the number of lagged excess returns. Grinblatt and Titman (1993) provide a benchmark-independent methodology for assessing fund performance by analyzing the covariance between the portfolio weights of individual assets and their subsequent returns for a zero-cost portfolio. However, this technique assumes no selection ability (uninformed

[15]Brown, Goetzmann, Ibbotson and Ross (1992) provide more details on alpha and the appraisal ratio.

investors) and constant expected returns.

The benchmarks constructed in Daniel, Grinblatt, Titman and Wermers (1997) facilitate a matching exercise involve 125 passive portfolios formed by sorting individual securities into quintiles based on their book-to-market, size and previous return characteristics. This procedure finds hypothetical peer funds for performance evaluation since these portfolios are not traded. Although they examine selection and market timing ability, as well as a separate style component, their technique is not an implied procedure designed to capture relative performance amongst available funds. Chan, Chen Lakonishok (2002) also incorporate portfolio characteristics (expressed as deciles) for comparison with the S&P 500, and find this method more adept at describing fund returns than its factor sensitivity counterpart.[16] Both of these papers consider absolute benchmarks to determine whether fund managers are capable of outperforming mechanical trading rules but not each other. Thus, the majority of funds, all in theory, could overperform (underperform) across time.

Cohen, Coval and Pástor (2004) also exploit the "overlap" in fund investments when comparing a fund manager with their distinguished counterparts. However, although we both condition performance on the portfolio weights of other fund managers, Cohen, Coval and Pástor (2004) continues to employ absolute benchmarks and considers individual securities. In addition, we derive original implied performance measures rather than constructing a "weighted" performance metric (such as Jensen's measure) over different fund managers. Nonetheless, our fund evaluation methodology may be viewed as merging the literature which conditions on portfolio character-

[16]Despite recent criticism of the factor approach, a time series of fund returns is often regressed on risk factors with performance measured by the resulting intercept coefficient. Ferson and Schadt (1996) and Christopherson, Ferson and Glassman (1998) introduce conditional information into the evaluation of mutual funds and pension funds respectively. However, multifactor regression models require the specification of proxies for the underlying risk factors. Furthermore, unknown sensitivities to these factors are estimated, rather than inferring relative benchmarks conditioned on known portfolio weights as in our approach.

istics with the more recent Cohen, Coval and Pástor (2004) approach designed to incorporate commonality in portfolio weights across multiple funds.

To summarize, portfolio weights in different classes have not been previously utilized to infer relative performance in selectivity and market timing ability. By incorporating this conditional information, our proposed methodology circumvents the need to specify benchmark returns or peer funds. Furthermore, our procedure is robust to fluctuations in the portfolio weights of individual securities between evaluation periods. In particular, only class-specific portfolio weights are required.

5 Data and Estimation

To illustrate our proposed test procedures, mutual fund data from Morningstar is utilized. Although Morningstar removes defunct funds (as if they never existed) from their published products, our sample is constructed from the original databases and includes all funds with available portfolio weight information.[17] Our sample period is from December 1992 to December 2001 which spans both bull and bear markets. The Morningstar data is ideal for our purposes as fund portfolios are classified according to their weights in specific industries, assets and geographical regions.[18]

Funds typically have quarterly disclosures causing some variability in the number of funds available each month. Nonetheless, returns and portfolio weights are simultaneously reported every month for a subset of funds. Consequently, selectivity and market timing statistics are computed monthly, with a visual illustration provided in Figure 1. These plots indicate no discernible differences within the quarterly horizon and confirm our assumption of normality.[19]

[17] We are indebted to Stephen Murphy at Morningstar for constructing the database in this manner.

[18] US companies with foreign operations have regional exposures that are unaccounted for by this analysis.

[19] The selectivity statistics have average skewness values of -0.05, 0.06 and 0.10 for the industry, asset and regional classes respectively with their average kurtosis being 0.83, 1.02 and 0.75. Thus, the implied distributions

As in Daniel, Grinblatt, Titman and Wermers (1997) and other studies focused on measuring selection and market timing ability, we consider returns before fees and expenses.

For industry sectors, Morningstar provides 12 different classes while the original asset categories are reduced to 4 by combining US and non-US equity positions as well as US and non-US bond holdings. Thus, our asset classifications are comprised of cash, equity, bonds and a separate class for preferred shares along with convertibles. Morningstar also reports portfolio weights for different geographical regions although very small entries are associated with Central and Latin America, Canada, Africa, Central and Eastern Europe as well as Australia. Therefore, three mergers are enacted to form 5 distinct classes. First, the United States, Canada as well as Central and Latin America are combined into an America class. Second, a Europe class is created by combining Western Europe with Eastern and Central Europe. This class also contains Africa but excludes the United Kingdom which remains a separate entity. Third, Asia is merged with Australia to yield an Asian class which is distinct from Japan.

Since we are interested in evaluating managerial ability, the usual convention of considering gross fund returns without adjustments for fees and expenses is adopted. For the cross-sectional inference of $\hat{R}$ and $\hat{\Theta}$, a total of 1,601, 1,754 and 1,551 unique funds are available for the respective industry, asset and regional analyses.

A summary of the portfolio weights underlying the three classifications is contained in Table 1. According to Table 1, portfolio weights are quite evenly distributed across the industry sectors. In contrast, the asset and region portfolio weights are dominated by equity and America respectively.[20] However, as demonstrated in the next section, the conditional information in these portfolio weights still exerts a significant influence on performance evaluation.

are symmetric and only slightly platykurtotic (thin tails).

[20]The industry and region portfolio weights only pertain to a fund's equity investments. Adjusting for the amount invested in remaining three classes (cash, bonds and the combination of preferred shares and convertibles) produces nearly identical results.

Each portfolio weight $w_{p,c}$ has an associated standard deviation $\sigma_{p,c}$ computed from its time series over the sample period with the quantity $\bar{\sigma}_c$ equaling

$$\bar{\sigma}_c = \frac{1}{P}\sum_{p=1}^{P}\sigma_{p,c}\,. \tag{21}$$

Larger values of $\bar{\sigma}_c$ imply funds are more inclined to adjust their portfolio weights in this class.

Additional information is also provided by Morningstar on the expense ratio, size and turnover of each fund. Therefore, relationships between these variables and our implied statistics are examined. In addition, a fund's *focus* is measured by the disparity between its portfolio weights. Formally, we define focus as

$$\text{Focus}_p = \max_{c=1,\ldots,C} w_{p,c} - \min_{c=1,\ldots,C} w_{p,c}\,, \tag{22}$$

which is an element of the $[0,1]$ interval. Consequently, funds which invest equal amounts in each class have zero focus, while those invested exclusively in a single class have a focus measure equaling one. Furthermore, we gauge the extent to which a fund alters their class portfolio weights by computing

$$\bar{\sigma}_p = \frac{1}{C}\sum_{c=1}^{C}\sigma_{p,c}\,. \tag{23}$$

In contrast to $\bar{\sigma}_c$ in equation (21), $\bar{\sigma}_p$ averages the standard deviations $\sigma_{p,c}$ over the number of classes rather than funds. Thus, $\bar{\sigma}_p$ supplements the fund's reported turnover by only accounting for transactions that alter their class portfolio weights. Indeed, transactions within a class have no influence on this variability measure.

To reduce the number of parameters requiring calibration, the $C(C+1)/2$ distinct off-diagonal entries of the symmetric matrix $\hat{\Theta}$ are set to zero. In economic terms, deviations from the benchmark return in one class, say software, are independent of deviations from the benchmark return in another class such as financial services.[21] A robustness test conducted in the next section justifies this simplification.

[21]To clarify, the diagonal assumption imposed on $\hat{\Theta}$ is unrelated to the diagonal nature of $\hat{\Lambda}$ in Section 3. In

The iterative procedure described in Section 3 is terminated when the difference in implied parameters between successive iterations is less than 10^{-6}. Starting values for the class-specific returns are the coefficients from a linear regression of the fund's return on its portfolio weights. For the return variances, the regression's standard error serves as the common starting value in all C classes.

6 Empirical Results for Fund Performance

In this section, we investigate the ability of mutual funds to select securities within a class and time-the-market by investing in different classes. We also examine the importance of conditioning a fund's performance on its class portfolio weights, and if investment skill is related to several fund characteristics.

However, the interpretation of our empirical results has to account for multiple Type I errors. For example, if one tests 1,000 funds for investment skill at the 5% level, then 50 false rejections of the null are expected. Therefore, we also test whether the subset of overperforming fund managers is statistically significant.

For a specified threshold, let *Percent* denote the percentage of funds (out of P) with significant test statistics at the γ level. Thus, *Percent* represents the subset of funds with p-values in equation (14) below γ, and is subjected to the following test procedure[22]

$$\frac{Percent - \gamma}{\sqrt{\frac{\gamma\,(1-\gamma)}{P}}} \overset{d}{\sim} \mathcal{N}(0,1)\,. \tag{24}$$

particular, $\hat{\Theta}$ pertains to the variance-covariance matrix of class-specific returns while $\hat{\Lambda}$ refers to individual fund returns.

[22]The number of funds in our sample is sufficiently large to justify the normal approximation to the binomial test procedure. Furthermore, we verify this approximation using the binomial cumulative distribution function and confirm the appropriateness of equation (24) in a later robustness test.

For clarification, the performance measure in equation (14) evaluates an *individual* fund. However, when examining *multiple* funds, equation (24) is invoked to determine whether the subset of overperforming funds is statistically significant.[23] To summarize, we are interested in answering the following canonical question, "At the 5% significance level ($\gamma = 0.05$), is the subset of funds ($Percent$) which consistently exceed their fund-specific implied benchmark ($K = 0$) significant?".

Statistically, it is possible for funds to overperform at $K = 1$ but underperform at $K = 0$. For example, a fund's return may be below its benchmark in 75% of the periods but above one standard deviation in the remaining 25%. In this instance, the fund's performance is highly variable with frequent underperformance interspersed with dramatic success. This effect may be produced by funds utilizing leverage or other risky trading strategies. Therefore, the results reported in this section require funds overperforming at the higher threshold to also overperform at its lower counterpart. In economic terms, this constraint ensures greater consistency in fund performance.

6.1 Fund Performance

Performance is computed for each fund with at least 20 observations. Recall that $K = 0$ tests a fund's ability to consistently exceed its benchmark return, while $K = 1$ examines performance one standard deviation above this threshold.

For $K = 0$, Table 2 reports that selection and market timing ability is detected after conditioning on all three portfolio weight classifications. Thus, a statistically significant subset of funds exhibit moderate investment skill in both attributes. However, at the higher $K = 1$ thresh-

[23] The threshold α does not serve as an upper bound on $Percent$. For example, let $\alpha = 0.10$ and suppose that in alternating (but not consecutive) months, an identical set of funds appear in the tail probability. In this instance, 20% of the funds appear in the 10% tail probability during half the periods in the sample with such nonrandom behavior being evidence of fund overperformance.

old, selection ability dissipates. Indeed, only market timing ability amongst industry classes is detected for a small subset of funds.

Correlation between securities within a class may be partially responsible for the decrease in selection ability. Specifically, higher correlation between securities implies greater difficulty in selecting investments that enable a fund to overperform. For example, if securities within the software industry are perfectly correlated, selectivity cannot be demonstrated with respect to this class.

One explanation for the large decline in market timing ability between the asset and region classes at the higher performance threshold is that fund managers are unwilling (or unable) to dramatically alter their exposure to non-American equities. Indeed, given the dominance of the equity and America class, Table 2 offers reassurance that our market timing methodology is performing appropriately.

An interesting feature of Table 2 is that as γ increases from 1% to 10%, the percentage of overperforming funds increases by less than a factor of ten, and the corresponding subsets are often less significant. This evidence indicates that investment skill is concentrated in a small number of fund managers who overperform in a convincing fashion by frequently exceeding their benchmark return. Less consistent funds require a larger value of γ before entering the overperforming subset. However, given the insensitivity of the subsets to γ, overperformance is often marginal as an increase in the Type I error does not induce a similar rise in the ranks of the overperformers.

Figures 2 and 3 reveal the realized distribution of p-values for selection and market timing ability respectively for every fund. As demonstrated in Appendix B, these density functions are uniform (not normal) under the null hypothesis. At the $K = 0$ threshold, the p-value distributions are often close to being uniform. However, for $K = 1$, probability mass in the density function shifts to the right which attests to the difficulty in overperforming at this higher threshold.

With respect to the recent literature, Daniel, Grinblatt, Titman and Wermers (1997) report that funds have moderate selection ability while Chan, Chen and Lakonishok (2002) document the tendency of fund managers to hold similar portfolios. Despite applying a distinct relative evaluation methodology to different conditional information, the inability of funds to overperform at the higher threshold is consistent with their findings.

6.2 Interaction between Investment Skills

We also study the "intersection" of significant overperformers across the three portfolio weight classifications to determine if the same funds are exhibiting overperformance. According to Table 2, the resulting subsets are negligible for both selectivity and market timing. Therefore, investment skill is not transmitted over the three sets of conditional information. For example, a fund manager able to successfully select securities (or portfolio weights) along industry characteristics cannot replicate this skill for the asset and regional criteria.

The intersection between selection and market timing ability for a single set of conditional information is also examined. This study attempts to discern whether funds are capable of exhibiting both attributes. Since investment skill is only prevalent for $K = 0$, we focus our attention on the lower performance threshold. Moreover, since the selectivity subsets are not sensitive to γ, the 10% significance level ($\gamma = 0.10$) is analyzed.

Table 3 finds that selection and market timing ability are unrelated, a finding which supports the notion that fund managers *specialize* in either selecting securities within a sector or allocating their portfolio between various sectors. Later in this section, a Logit analysis offers more insight into the relationship between these attributes in terms of fund characteristics.

6.3 Robustness Tests

We first determine the "power" of the test statistic in equation (24) via simulation. Furthermore, we analyze the consequences of correlation amongst classes and fund returns on our previous results. Second, the sensitivity of our market timing conclusions to the benchmark portfolio weights is explored.

The simulation experiment we perform has three objectives. First, correlation between the implied returns of different classes is studied. Second, we examine residual correlation between fund returns that is independent of the portfolio weights.[24] Thus, the "regression" procedure in equation (16) is generalized to permit correlation between the $\hat{\epsilon}_p$ errors. Third, the ability of equation (24) to appropriately account for the Type I errors associated with investigating multiple funds is analyzed.

Several scenarios are examined, including situations where the null hypothesis of no overperformance is true as well as false. More details regarding the simulation procedure are reported in Appendix C. As indicated in Table 4, our procedure is not sensitive to correlation between class returns or residual correlation among fund returns. Indeed, the use of diagonal $\hat{\Theta}$ and $\hat{\Lambda}$ matrices is justified. In addition, our methodology accepts and rejects the null hypothesis appropriately.[25]

One could also conjecture that evidence of market timing across various industry and asset classes in Table 2 arises from a subset of funds that strictly adhere to pre-determined portfolio weights. For example, suppose certain funds are constrained to invest in a specific set of industries or assets. Since these funds influence the benchmark portfolio weights defined in equation (11), their lack of investment flexibility could reduce the threshold required to exhibit market timing ability.

---

[24] Residual correlation refers to correlation between the deviations $r_p - \hat{r}_p$ of different funds after conditioning on the portfolio weights.

[25] As expected, the simulation results are even more supportive of our implementation when the $K = 1$ threshold is examined and we omit their presentation for brevity.

Therefore, a robustness test is conducted which only includes funds in the calculation of benchmark portfolio weights whose $\bar{\sigma}_p$ values in equation (23) are above a specified cutoff. However, this modification has no influence on the selectivity results which are independent of the benchmark portfolio weights. Note that this robustness test preserves the concept of relative evaluation but simply focuses on funds which are likely candidates for displaying this attribute.

Figure 4 contains histograms of $\bar{\sigma}_p$ for each classification and illustrates that many funds rarely, if ever, alter their class portfolio weights. Indeed, the minimum values are zero for all three criteria, indicating at least one fund never altered its portfolio weights once during the sample period. Median (mean) $\bar{\sigma}_p$ values of 0.0289 (0.0311), 0.0270 (0.0350) and 0.0070 (0.0115) are found for the industry, asset and regional classes respectively. Based on these results and the histograms in Figure 4, cutoff values of 0.03, 0.03 and 0.01 are chosen.

The market timing results in Table 5 for the robustness test parallel previous findings reported in Table 2. Indeed, for both performance thresholds, the percentage of funds with relative selection ability actually increases within industry classes.[26] For assets and regions, the subsets of overperforming funds become smaller (or are identical) but generally remain at similar significance levels. Thus, we conclude that evidence of market timing skill is not sensitive to the benchmark portfolio weights.

6.4 Importance of Conditional Information

The importance of conditioning on portfolio weights when evaluating funds is gauged by forming a naive selection statistic S_p^* defined as,

$$S_p^* = \frac{r_p - \bar{r}}{\sigma} \tag{25}$$

[26]This increase may be explained by a proportion of funds attempting to time-the-market but failing. These fund managers would benefit from simply maintaining a time-invariant position in each class, a strategy adopted by the funds removed from the benchmark portfolio weight computation in the robustness test.

where $\bar{r}$ and σ denote the *unconditional* mean and standard deviation of fund returns. Both $\bar{r}$ and σ are independent of the fund's portfolio composition. Thus, unlike the S_p metric in equation (5), S_p^* does not incorporate a fund's portfolio weights.

Investigating the importance of conditional information on market timing ability is difficult as this would involve an allocation between undefined classes. Consequently, we focus our attention on the importance of conditional information to the measurement of selection ability.

Formally, for a chosen significance level, we define the common performance ratio (abbreviated CPR) as

$$\text{CPR} = \frac{2 \times \text{Funds with both } B_p(S_{p,i}|\alpha, n_p) \text{ and } B_p(S_{p,i}^*|\alpha, n_p) \text{ being significant}}{\text{Funds with significant } B_p(S_{p,i}|\alpha, n_p) + \text{Funds with significant } B_p(S_{p,i}^*|\alpha, n_p)} . \quad (26)$$

Hence, the CPR is an element of the $[0, 1]$ interval and divides the number of funds in an intersection, which have common conditional and unconditional evaluations, by their sum. If the $S_{p,i}$ and $S_{p,i}^*$ metrics yield identical funds, then the intersection in the numerator has the same number of funds as each of the two components in the denominator and the CPR equals one. Conversely, if the conditional and unconditional performance measures have no funds in common, then the intersection is empty and the ratio equals zero. Overall, the lower the CPR, the more important is the conditional information contained in portfolio weights to fund evaluation.

Table 6 documents the CPR at the 10% significance level. Therefore, we only compare funds whose p-values in equation (14) are below $\gamma = 0.10$ when applied to either the time series of $S_{p,i}$ or $S_{p,i}^*$ metrics. As mentioned earlier in this section, the overperforming subsets are not sensitive to γ in the conditional analysis with the intersection inheriting a similar property.

At the lower performance threshold where selection ability is detected, less than 30% of the funds have common unconditional and conditional performance measures involving industry portfolio weights. The CPR increases slightly for assets and remains less than 50% for regions. Thus, incorporating class portfolio weights into performance measurement is critical since (1 - CPR)% of the funds have their evaluations misspecified when these characteristics are ignored.

Observe that the CPR increases with the performance threshold, although an insignificant number of funds overperform at the higher level. This property is intuitive since funds with extreme unconditional overperformance have a greater likelihood of remaining an overperformer after conditioning on portfolio weights in comparison to those with marginal unconditional overperformance.

Overall, the industry categories manifest the most conditional information although serious misspecifications in performance measurement arise if asset and regional portfolio weights are ignored.

6.5 Fund Characteristics and Performance

Grinblatt and Titman (1994) find that fund performance is positively related to portfolio turnover but unrelated to size and expenses. Carhart, Carpenter, Lynch and Musto (2002) examine the influence of similar variables on fund performance and report limited evidence of any relationships.[27] Our results parallel the existing literature with respect to size and expenses. However, our focus and σ_p variables constructed from the class portfolio weights partially explain variation in fund performance with the latter potentially supplanting turnover.

We examine a fund's average expense ratio, size, turnover, focus and fluctuations in its portfolio weights $\bar{\sigma}_p$ in relation to its implied statistics. Focus and $\bar{\sigma}_p$ are defined in equations

[27] Chevalier and Ellison (1999) examine the relationship between fund performance and the manager's age, SAT scores of their undergraduate university, MBA qualifications and tenure in present position. They find that attendance at institutions with higher SAT scores is correlated with higher risk-adjusted excess returns.

(22) and (23) respectively. A Logit model is estimated as

$$y_p = \frac{\exp\left\{\gamma_0 + \gamma_1 \text{ Expense}_p + \gamma_2 \text{ Size}_p + \gamma_3 \text{ Turnover}_p + \gamma_4 \text{ Focus}_p + \gamma_5 \ \bar{\sigma}_p\right\}}{1 + \exp\left\{\gamma_0 + \gamma_1 \text{ Expense}_p + \gamma_2 \text{ Size}_p + \gamma_3 \text{ Turnover}_p + \gamma_4 \text{ Focus}_p + \gamma_5 \ \bar{\sigma}_p\right\}} + \epsilon \,, \tag{27}$$

where y_p denotes the p-value of either the selectivity or market timing measure in equation (14) at the $K = 0$ performance threshold. Since the dependent variables are probabilities, and smaller entries coincide with greater overperformance, a positive t-statistic in Table 7 implies an increase in this variable corresponds to a reduction in performance.

Consistent with the existing literature, the expense ratio, turnover and size have little influence on fund performance. However, focused funds appear to have greater selectivity within an industry but less ability to time-the-market between different industries. These intuitive conclusions are supported by their negative and positive t-statistics for selectivity and market timing ability respectively. In addition, more focused funds that concentrate their investments in a small number of asset and regional classes are less able to time-the-market. This property confirms the reasonableness of our market timing formulation.

Besides focus, the other fund characteristic unique to the proposed evaluation framework is $\bar{\sigma}_p$, defined in equation (23), which captures variation between class portfolio weights. In agreement with our intuition, its negative t-statistic for market timing ability indicates that funds which alter their industry positions more aggressively have a higher likelihood of exhibiting this attribute.

The lack of significance for $\bar{\sigma}_p$ in explaining market timing ability is explained by its significant negative correlation with focus, -0.441 and -0.678 respectively for assets and regions. When focus is removed from equation (27), the t-statistics for $\bar{\sigma}_p$ become positive and significant (p-values below 10^{-3}). Therefore, as expected, funds with larger $\bar{\sigma}_p$ values experience a higher likelihood of timing-the-market. This relationship justifies our earlier use of $\bar{\sigma}_p$ in the robustness test of market timing ability.

7 Conclusion

This paper offers a methodology to evaluate the relative ability of fund managers to select securities and time-the-market. Selection and market timing metrics are inferred from a cross-section of fund returns, conditional on portfolio weights in different classes across all fund managers. A separate procedure assesses an individual fund's performance over longer horizons. Classes may represent any criteria capable of being expressed in terms of portfolio weights that investors deem relevant to the evaluation of fund performance. Thus, our approach merges two previously disparate literatures on performance evaluation by incorporating portfolio characteristics across multiple fund managers into an individual fund's assessment.

Our implied statistics gauge relative performance and circumvent the need to exogenously specify benchmarks or peer funds. Furthermore, the resulting performance measure is robust to herding and window dressing by fund managers. Over longer horizons, the effect of survivorship bias on performance measurement is also partially mitigated.

We investigate classes defined with respect to portfolio weights for different industry sectors, assets and geographical regions. A fund's ability to generate returns that exceed its implied benchmark and one standard deviation above this threshold are examined. For all three sets of conditional information, empirical evidence reveals that moderate selection and market timing ability at the lower threshold is concentrated in small number of fund managers. However, a negligible number of funds exhibit overperformance across all three portfolio classifications, and few funds possess both selection and market timing ability. Thus, investment skill appears to be specialized. Moreover, evidence of investment skill dissipates at a higher performance threshold.

The conditional information contained in portfolio weights is found to be critically important when evaluating funds. In particular, conditional performance measures and their unconditional counterparts that ignore portfolio characteristics are very different. Therefore, a failure to condition performance on class portfolio weights for industry, asset and regional criteria generates inaccurate fund evaluations.

Finally, the implied metrics for selectivity and market timing appear unrelated to a fund's expense ratio, size and turnover. Instead, variability in a fund's class portfolio weights is more adept at explaining performance. In addition, funds that restrict their investments to a small number of classes (and vary their class portfolio weights infrequently) are less able to time-the-market, but more likely to exhibit selectivity within industries.

Avenues for future research include modifying the proposed statistical methodology for applications to book-to-market, size and past return characteristics. In addition, passive investments such as index funds may also be utilized, exclusively or in conjunction with actively managed funds, when calibrating the implied return and variance parameters.

References

[1] Agarwal, V., and N. Y. Naik, 2000. Multi-period performance persistence analysis of hedge funds. Journal of Financial and Quantitative Analysis 35, 327-342.

[2] Bask, K. P., 2003. On the performance of mutual fund managers. Working Paper, Emory University.

[3] Becker, C., W. E. Ferson, D. H. Myers, and M. J. Schill, 1999. Conditional market timing with benchmark investors. Journal of Financial Economics 52, 119-148.

[4] Brown, S. J., and W. N. Goetzmann, 1997. Mutual fund styles. Journal of Financial Economics 43, 373-399.

[5] Brown, S. J., and W. N. Goetzmann, 1995. Performance persistence. Journal of Finance 50, 679-698.

[6] Brown, S. J., W. N. Goetzmann, R. G. Ibbotson, and S. A. Ross, 1992. Survivorship bias in performance studies. Review of Financial Studies 5, 553-580.

[7] Carhart, M. M., 1997. On persistence in mutual fund performance. Journal of Finance 52, 57-82.

[8] Carhart, M. M., R. Kaniel, D. K. Musto, and A. V. Reed, 2002. Learning for the tape: Evidence of gaming behavior in equity mutual funds. Journal of Finance 52, 1035-1058.

[9] Carhart, M. M., J. N. Carpenter, A. W. Lynch, and D. K. Musto, 2002. Mutual fund survivorship. Review of Financial Studies 15, 1439-1463.

[10] Carpenter, J. N., and A. W. Lynch, 1999. Survivorship bias and attrition effects in measures of performance persistence. Journal of Financial Economics 54, 337-374.

[11] Casella, G., and R. L. Berger, 1990. Statistical Inference. Duxbury Press.

[12] Chan, L. K. C., H. L. Chen, and J. Lakonishok, 2002. On mutual fund investment styles. Review of Financial Studies 15, 1407-1437.

[13] Chevalier, J., and G. Ellison, 1999. Are some mutual fund managers better than others? Cross-sectional patterns in behavior and performance. Journal of Finance 54, 875-899.

[14] Christopherson, J. A., W. E. Ferson, and D. A. Glassman, 1998. Conditioning manager alphas on economic information: Another look at the persistence of performance. Review of Financial Studies 11, 111-142.

[15] Cohen, R., J. Coval, and L. Pástor, 2004. Judging fund managers by the company they keep. Forthcoming, Journal of Finance.

[16] Daniel, K., M. Grinblatt, S. Titman, and R. Wermers, 1997. Measuring mutual fund performance with characteristic-based benchmarks. Journal of Finance 52, 1035-1058.

[17] Elton, E. J., M. J. Gruber, and C. R. Blake, 1996. Survivorship bias and mutual fund performance. Review of Financial Studies 9, 1097-1120.

[18] Falkenstein, E. G., 1996. Preferences for stock characteristics as revealed by mutual fund portfolio holdings. Journal of Finance 51, 111-136.

[19] Fama, E., and J. D. MacBeth, 1973. Risk, return, and equilibrium: Empirical tests. Journal of Political Economy 81, 607-636.

[20] Ferson, W. E., and R. W. Schadt, 1996. Measuring fund strategy and performance in changing economic conditions. Journal of Finance 51, 425-461.

[21] Grinblatt, M., and S. Titman, 1989. Portfolio performance evaluation: Old issues and new insights. Review of Financial Studies 2, 393-421.

[22] Grinblatt, M., and S. Titman, 1993. Performance measurement without benchmarks: An examination of mutual fund returns. Journal of Business 66, 47-68.

[23] Grinblatt, M., and S. Titman, 1994. A study of monthly mutual fund returns and performance evaluation techniques. Journal of Financial and Quantitative Analysis 29, 419-444.

[24] Grinblatt, M., S. Titman, and R. Wermers, 1995. Momentum investment strategies, portfolio performance, and herding: A study of mutual fund behavior. American Economic Review 85, 1088-1105.

[25] Lakonishok, J., A. Shleifer, R. Thaler, and R. W. Vishny, 1991. Window dressing by pension fund managers. American Economic Review 81, 227-231.

[26] Lakonishok, J., A. Shleifer, and R. W. Vishny, 1992. The impact of institutional trading on stock prices. Journal of Financial Economics 32, 23-44.

[27] Musto, D. K., 1997. Portfolio disclosures and year-end price shifts. Journal of Finance 52, 1563-1588.

[28] Musto, D. K., 1999. Investment decisions depend on portfolio disclosures. Journal of Finance 54, 935-952.

[29] Pástor, L., and R. F. Stambaugh, 2002a. Mutual fund performance and seemingly unrelated assets. Journal of Financial Economics 63, 315-349.

[30] Pástor, L., and R. F. Stambaugh, 2002b. Investing in equity mutual funds. Journal of Financial Economics 63, 351-380.

[31] Treynor, J., and F. Mazuy, 1966. Can mutual funds outguess the market? Harvard Business Review 45, 131-136.

[32] Wermers, R., 1999. Mutual fund herding and the impact on stock prices. Journal of Finance 54, 581-622.

Appendices

A Utilizing Information on Individual Securities

The portfolio weights of individual securities fluctuate between disclosure dates. Although these intermediate transactions influence the aggregate return r_p, they need not alter the class-specific portfolio weights. More importantly, our implied methodology extracts class-specific benchmark returns $\hat{r}_c$ that are consistent with a fund's aggregate reported return.

Suppose class c consists of N_c individual securities indexed by $k = 1, 2, \ldots, N_c$ whose returns are denoted r_k. This appendix highlights the inappropriateness of computing the class-specific expected returns and variances as

$$r_c = \frac{1}{N_c} \sum_{k=1}^{N_c} r_k \quad \text{and} \tag{28}$$

$$\sigma_c^2 = \frac{1}{N_c - 1} \sum_{k=1}^{N_c} (r_k - r_c)^2 \ , \tag{29}$$

using the realized returns of individual securities. In contrast to our implied statistical procedure, the estimates in equations (28) and (29) are not conditioned on fund portfolio weights in each class. Furthermore, the above estimates fail to measure relative performance as they have the following implications:

1. The values of r_c and σ_c^2 in equations (28) and (29) respectively are not conditioned on the portfolio weights of actual funds. For example, individual securities that are not held by any fund, but members of the class, contribute towards their estimation.

2. The variability measured by σ_c in equation (29) does not pertain to fund returns conditioned on class portfolio weights. Instead, the underlying variability of individual securities is calibrated. For example, equation (29) is positive even if all fund managers maintain identical portfolios over the previous period and therefore produce identical returns.

3. The majority of funds could overperform (underperform) any benchmark constructed from the estimates in equations (28) and (29). Thus, when deciding between available funds, a simple average and variance conveys less information.

Overall, our relative evaluation approach offers a *fund versus fund* comparison while equations (28) and (29) describe a *fund versus hypothetical benchmark* procedure. However, given market values for the individual securities, class portfolio weights may be constructed from the portfolio weights of individual securities after sorting them according to a specified criteria such as industry SIC codes. Once this initial step is completed, our implied estimation procedure is applicable.

B Uniform Distribution under Null

Let X represent a random test statistic with a continuous distribution described by a cumulative distribution function (cdf) $F_X(x)$ under the null hypothesis. The p-value of the test statistic X equals $1 - F_X(x)$ or the probability of X being greater than the observation x. In our context, $F_X(x)$ is the cdf of a standard normal.

According to Theorem 2.1.4 of Casella and Berger (1990), $F_X(X)$ has a uniform $U[0,1]$ distribution. This result implies the p-values of the test statistics are also $U[0,1]$ since $1-F_X(X)$ has a distribution identical to $F_X(X)$ in this circumstance.

C Simulation Study

The simulation analysis has $C = 12$ classes with a matrix W of portfolio weights sampled from the data for one thousand funds, $P = 1,000$. Returns in each class are drawn from a $\mathcal{N}(R_C, \Theta_C)$ distribution. Denote the randomly generated vector of class-specific returns as R, which creates the vector of fund returns $R_P = W^T R$.

Under the null hypothesis, scenarios in which Θ_C is both a diagonal and dense matrix are evaluated. In the latter case, correlations between class returns are first instilled into the analysis. Second, correlations amongst individual fund returns that exist even after conditioning on portfolio weights are introduced. This is accomplished by augmenting the R_p vector as follows

$$R_P^* \quad = \quad R_P + \epsilon^* \quad \text{where} \quad \epsilon^* \sim \mathcal{N}\left(\mathbf{0}, \Lambda_P\right) .$$

Thus, R_P^* represents fund returns generated by portfolio weights as well as a random component according to the variance-covariance matrix denoted Λ_P. Along with W and R_C, the diagonal elements of Θ_C as well as Λ_P originate from the industry data and are fixed throughout the simulation study. Three separate structures for Θ_C and Λ_P are investigated. In the first instance, both of these variance-covariance matrices are diagonal. In the second and third scenarios, Θ_C and Λ_P are randomly generated dense matrices respectively with a full complement of off-diagonal covariance terms. However, in all three cases, variances along the diagonal are identical.

A time series of $n_p = 24$ return vectors are simulated. The estimation procedure then produces $\hat{R}$ and $\hat{\Theta}$ estimates under the previous assumptions that $\hat{\Theta}$ and $\hat{\Lambda} = \text{diag}\left\{W^T \hat{\Theta} W\right\}$ are both diagonal as detailed in Section 3. These estimates are then converted into selectivity metrics before applying the performance measure in equation (14) to all 1,000 funds, yielding a *Percent* subset. The simulation process is repeated $N = 1,000$ times. Finally, each of the N subsets is then tested according to equation (24) for γ equal to 0.05 and 0.10.

Besides simulating under the null hypothesis of no selection ability, 15% of funds have their returns increased by a factor of 1.25. In this case, the null hypothesis of no selection ability is false at the $K = 0$ threshold and its rejection is anticipated.[28]

[28]The simulation exercise has identical funds exhibiting overperformance throughout the time series of 24 observations.

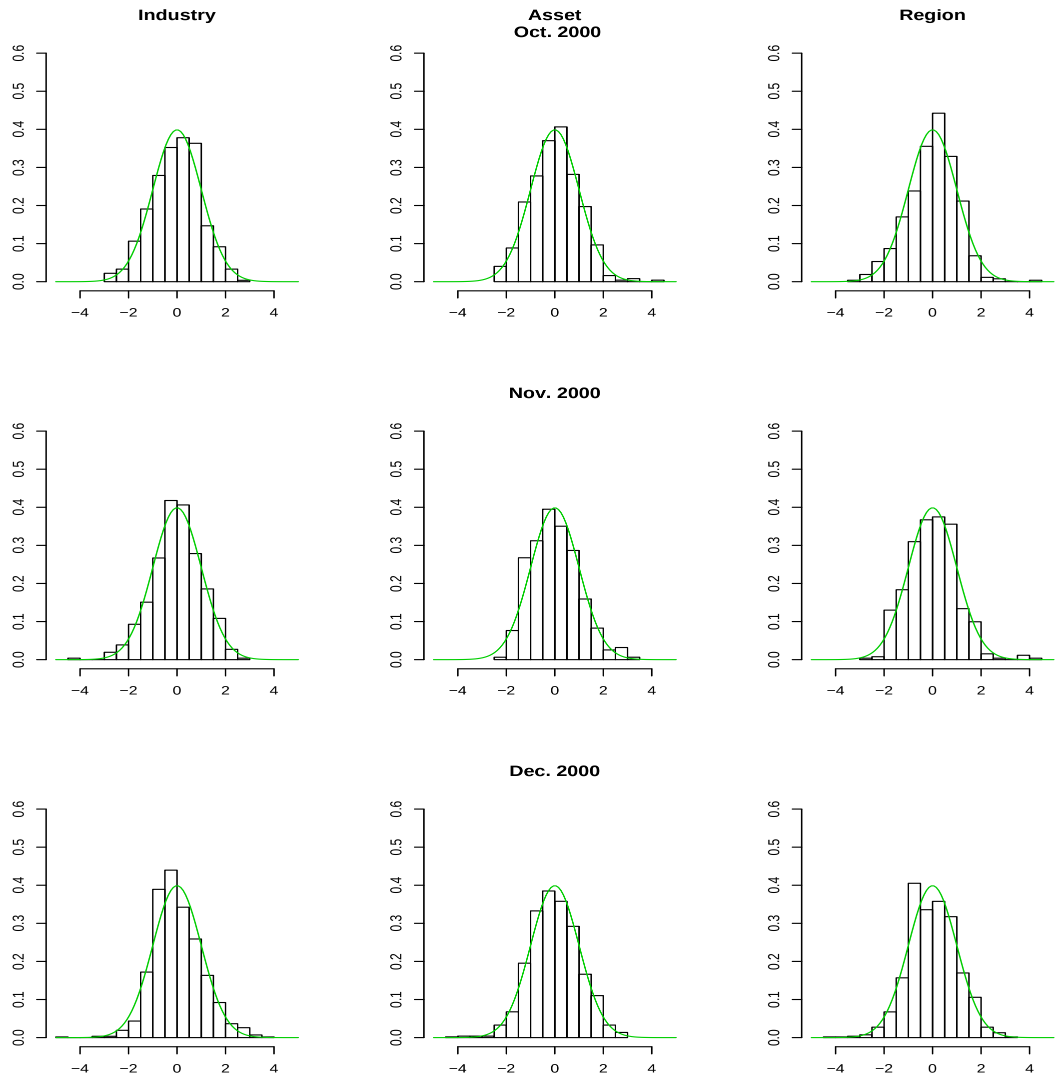

Figure 1: Implied distributions of selectivity statistics for three consecutive months during the sample period. The exact dates correspond to October, November and December of 2000. A normal distribution (with the same mean and variance) is superimposed on each histogram for ease of comparison. Observe the near normality of the distributions, with little evidence of skewness or kurtosis. Furthermore, the implied distributions within the quarterly period appear very similar.

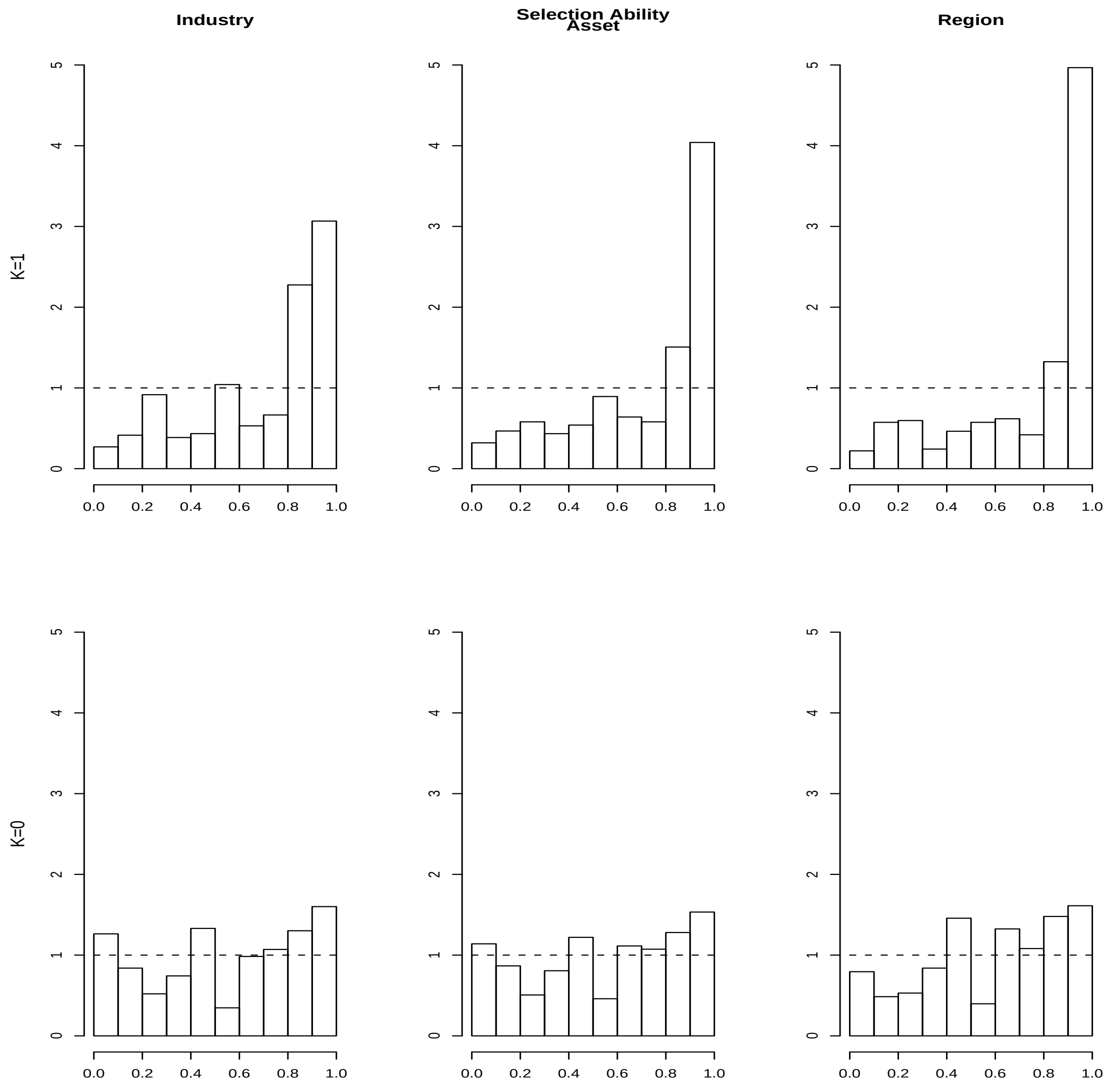

Figure 2: Discrete density functions (area under histogram sums to one) for the p-values of individual fund performance in selectivity. Although related to Table 2, the y-axis above represents a probability distribution, not a percentage. Superior relative performance is indicated by low p-values towards the left side of the plots. The top row designated $K = 1$ corresponds with performance one standard deviation above the implied benchmark, while the bottom row labeled $K = 0$ gauges a fund's ability to consistently exceed this fund-specific threshold. Note that normality is not required. Instead, under the null hypothesis of no selectivity, the p-values are uniformly distributed. For comparison, a uniform distribution is superimposed on each density function as a dashed line.

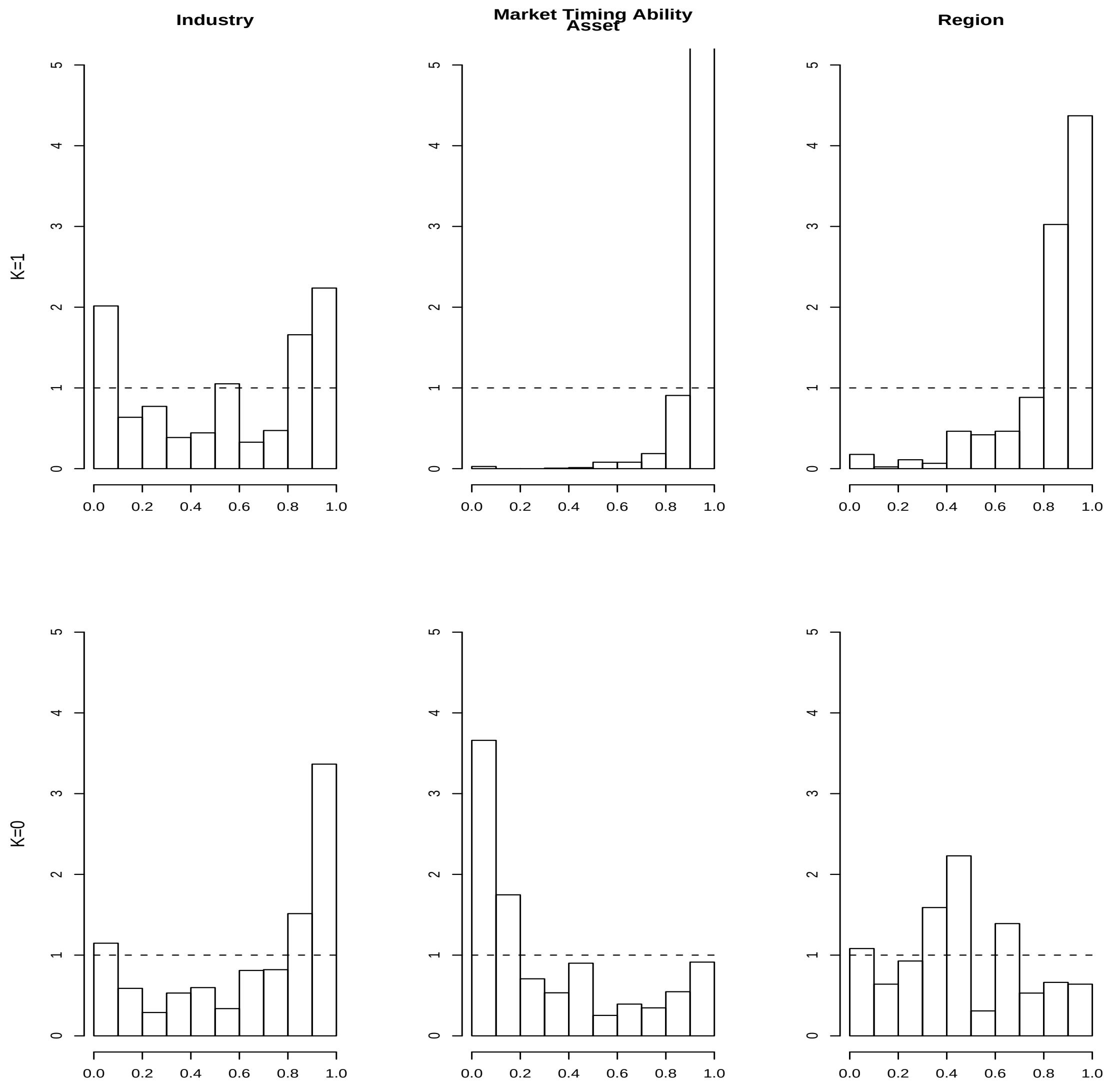

Figure 3: Discrete density functions (area under histogram sums to one) for the p-values of individual fund performance in market timing. Although related to Table 2, the y-axis above represents a probability distribution, not a percentage. The top row designated $K = 1$ corresponds with performance one standard deviation above the implied benchmark, while the bottom row labeled $K = 0$ gauges a fund's ability to consistently exceed this fund-specific threshold. Under the null hypothesis of no market timing skill, the p-values are uniformly (not normally) distributed with superior relative performance indicated by low p-values near the left side of the plots. For comparison, a uniform distribution is superimposed on each density function as a dashed line.

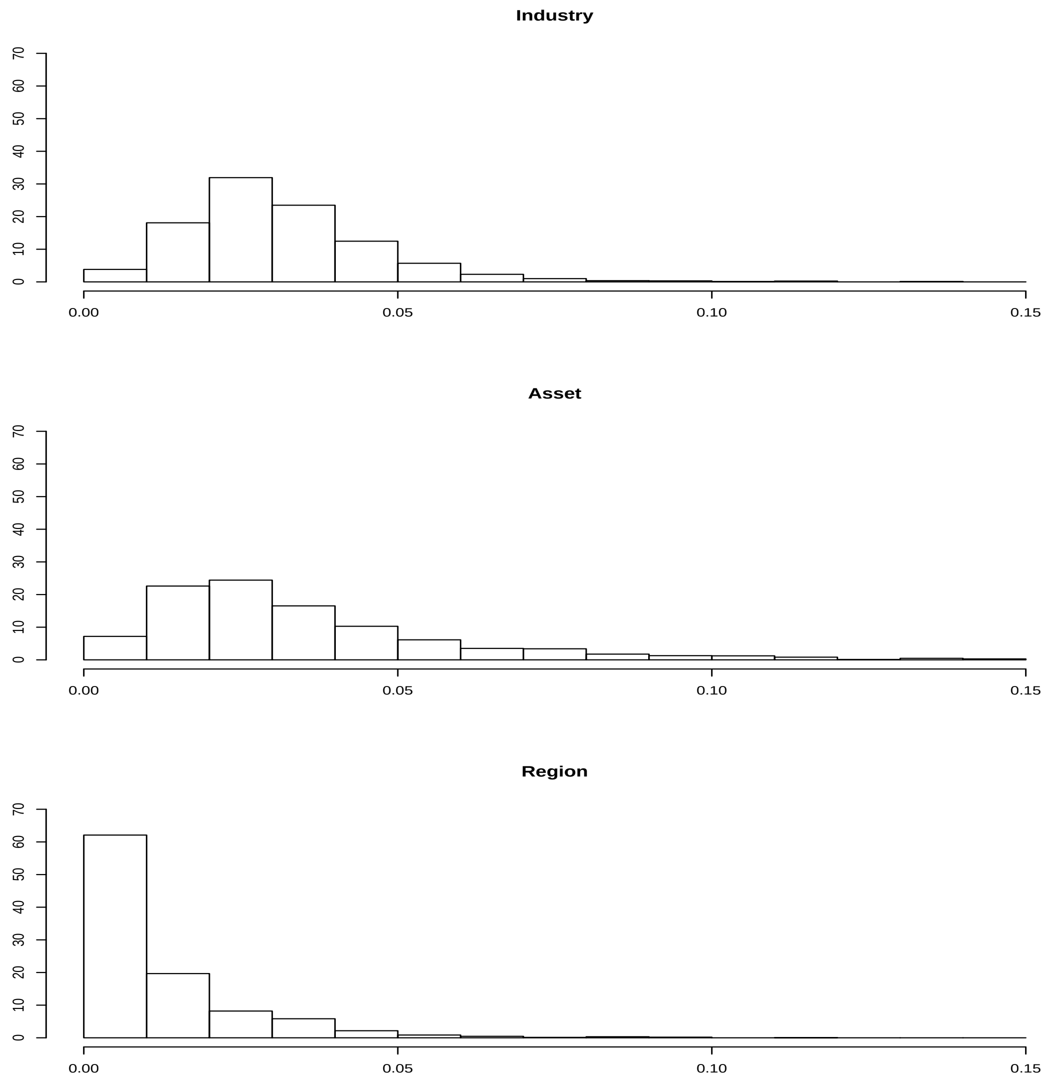

Figure 4: Histograms (discrete density functions) pertaining to the volatility of class portfolio weights defined in equation (23) for all three classifications. These plots illustrate that many funds rarely, if ever, adjust their portfolio weights across the various classes. Furthermore, they indicate that 0.03, 0.03 and 0.01 are reasonable cutoff values for the robustness test of market timing ability.

Table 1: Summary statistics for Morningstar data on industries, assets and regions. The mean and median of the class portfolio weights across all funds are reported. Entries in the final column are defined in equation (21) and measure variability in the portfolio weights. Note that the average portfolio weights sum to one. In addition, the asset and regional classifications are dominated by the equity and America class respectively, while industry portfolio weights are more uniformly distributed over the different classes.

Industries	12 Classes	Mean	Median	$\bar{\sigma}_c$
	1. Software	4.1	2.9	1.7
	2. Hardware	13.5	11.3	4.9
	3. Media	4.0	3.0	1.6
	4. Telecommunications	6.2	4.7	2.3
	5. Healthcare	10.7	10.0	2.8
	6. Consumer Services	8.0	7.1	2.5
	7. Business Services	6.4	4.8	2.1
	8. Financial Services	17.2	15.5	3.5
	9. Consumer Goods	9.1	8.4	2.7
	10. Industrial Materials	11.7	10.1	2.8
	11. Energy	5.6	5.0	2.0
	12. Utilities	3.5	1.5	1.4
Assets	4 Classes			
	1. Cash	5.4	3.6	4.2
	2. Equity	87.8	94.2	5.0
	3. Bonds	4.7	6.0	1.7
	4. Preferreds and Convertibles	2.1	1.8	1.7
Regions	5 Classes			
	1. America	82.3	97.3	1.4
	2. United Kingdom	3.1	3.4	0.7
	3. Europe	7.7	1.4	1.2
	4. Japan	3.3	3.5	0.6
	5. Asia	3.6	4.2	0.6

Table 2: Summary of selection and market timing performance with classes defined in terms of industries, assets and regions. For $K = 0$, a fund's ability to exceed its fund-specific implied benchmark return is examined while $K = 1$ ascertains performance one standard deviation above this threshold. The percentages reported below document the proportion of funds that exhibit investment skill by having p-values in equation (14) below the stated significance levels. The asterices *, ** and *** indicate statistical significance of these subsets at the 10%, 5% and 1% level respectively according to equation (24). In general, statistically significant selection and market timing ability is only detected at the lower performance threshold. The row entitled "All Criteria" corresponds to an intersection across all three sets of conditional information. This analysis indicates that investment skill is specialized since conditioning on different portfolio characteristics yields distinct subsets of overperforming fund managers.

Fund Evaluation: Percentage of Overperforming Funds

$K = 0$ Performance Threshold

Conditional Information	Selectivity			Market Timing		
	Significance Level			Significance Level		
	10%	5%	1%	10%	5%	1%
Industry	13.0***	11.2***	8.6***	11.5*	9.4***	5.5***
Asset	11.5*	10.3***	7.2***	36.7***	29.0***	14.9***
Region	7.3	6.2	4.6***	9.1	7.7***	7.3***
All Criteria	0.5	0.5	0.0	0.7	0.5	0.0

$K = 1$ Performance Threshold

Conditional Information	Selectivity			Market Timing		
	Significance Level			Significance Level		
	10%	5%	1%	10%	5%	1%
Industry	2.7	1.9	1.3	4.1	3.2	2.0***
Asset	3.2	2.2	1.5*	0.0	0.0	0.0
Region	2.0	1.3	0.4	0.9	0.9	0.9
All Criteria	0.2	0.2	0.0	0.0	0.0	0.0

Table 3: Success of funds in selecting securities and timing-the-market conditional on industry, asset and regional portfolio weights. The results below examine the intersection of individual funds with significant performance measures for selection and market timing ability at the 10% level. The reported percentages correspond to the $K = 0$ performance threshold which evaluates a fund manager's ability to exceed their fund-specific benchmark return. The results indicate that a fund's success in selecting securities within the classes is not duplicated by their allocation decisions between classes since each of the subsets are insignificant. Thus, a fund manager may possess selection or market timing ability but seldom both attributes.

Intersection of Selection and Market Timing Ability			
$K = 0$ Performance Threshold		Selectivity	
	Industry	Asset	Region
Market Timing			
Industry	2.1	3.9	1.6
Asset	5.9	7.7	2.9
Region	2.2	1.6	0.7

Table 4: Simulated "power" of statistical test in equation (24) based on performance measure in equation (14) under different scenarios. Under the null hypothesis, none of the 1,000 funds have expected returns above the benchmark. In contrast, scenarios in which the null is false instill 15% of the funds with expected returns that are 25% higher than their benchmark. The simulations are conducted under three different structures for Θ_C and Λ_P, the variance-covariance matrices of classes and individual fund returns respectively. The first set of studies are conducted under a diagonal structure, as in our estimation procedure, while the second and third incorporate dense matrices with a full complement of covariance elements. Details of the simulation procedure are found in Appendix C. Overall, the simulation analysis attests to the robustness of our statistical procedure. Specifically, the percentages between the columns for "Both Diagonal", "Θ_C Dense" and "Λ_P Dense" are nearly identical. Moreover, the test statistics appropriately accept and reject the null hypothesis.

"Power" of Performance Measure under Different Scenarios

Null True: No Overperformance	Both Diagonal	Θ_C Dense	Λ_P Dense
5% Significance Level			
Acceptances	93.71	95.12	93.72
Rejections	6.29	4.88	6.28
10% Significance Level			
Acceptances	92.00	92.57	91.86
Rejections	8.00	7.43	8.14

Null False: 15% of Funds Overperform	Both Diagonal	Θ_C Dense	Λ_P Dense
5% Significance Level			
Acceptances	2.13	2.23	2.15
Rejections	97.87	97.77	97.85
10% Significance Level			
Acceptances	0.00	0.00	0.00
Rejections	100.00	100.00	100.00

Table 5: Robustness test of market timing ability to benchmark portfolio weights. Funds whose $\bar{\sigma}_p$ values in equation (23) are below 0.03, 0.03 and 0.01 for the industry, asset and region classifications respectively are removed from the computation of benchmark portfolio weights in equation (11). This modification excludes funds from the portfolio weight benchmark calculation which are not (likely) attempting to time-the-market. At $K = 0$, a fund's ability to exceed its fund-specific implied benchmark return is examined while $K = 1$ ascertains performance one standard deviation above this threshold. The percentages reported below document the proportion of funds that individually exhibit investment skill, with p-values in equation (14) below the three significance levels stated as column headings. Equation (24) is then applied to these subsets, with asterices *, ** and *** indicating their statistical significance at the 10%, 5% and 1% level respectively. The results below are similar to those in Table 2, with an increase in overperformance reported for market timing between industries while the subsets become smaller or are identical for assets and regions.

Robustness Test of Market Timing Ability Involving Benchmark Portfolio Weights

	$K = 0$ Threshold			$K = 1$ Threshold		
	Significance Level			Significance Level		
Conditional Information	10%	5%	1%	10%	5%	1%
Industry	18.1***	13.7***	8.4***	10.4	9.1***	6.2***
Asset	29.1***	25.0***	14.9***	0.0	0.0	0.0
Region	7.9	5.8	5.0***	0.7	0.7	0.7

Table 6: Importance of conditional information to selection ability according to the common performance ratio (abbreviated CPR) defined in equation (26) for the 10% significance level. For $K = 0$, a fund's ability to exceed its fund-specific implied benchmark return is examined while $K = 1$ ascertains performance one standard deviation above this threshold. The ratios indicate that the conditional information in portfolio weights is critical to fund evaluation. In particular, if one considers a fund's ability to exceed its benchmark return and conditions on industry portfolio weights, there is only a 27.4% chance that a fund's performance evaluation coincides with its unconditional counterpart. Similar results apply for the asset and region classes. By not conditioning on the designated portfolio weights, (1 - CPR)% of the funds have their performance misspecified. Therefore, at the lower performance threshold where overperformance is detected, ignoring the information in portfolio weights causes the majority of funds to be erroneously evaluated.

Common Performance Ratio for Selectivity

	Performance Threshold	
Conditional Information	$K = 0$	$K = 1$
Industry	0.274	0.397
Asset	0.362	0.729
Region	0.472	0.613

Table 7: Relationships between selection and market timing ability versus fund characteristics. The results displayed below are reported for the Logit model in equation (27). Focus and $\bar{\sigma}_p$ are defined in equations (22) and (23) respectively. The p-values are reported in parentheses below the t-statistics with *, ** and *** denoting significance at the 10%, 5% and 1% level respectively. Since the dependent variables are the p-values of the performance metrics, smaller p-values imply more significant overperformance. Therefore, a positive (negative) t-statistic implies that larger variables coincide with a decreased (increased) likelihood of overperformance. Observe that the expense ratio, turnover and size have little influence on both selection or market timing ability. Across different industries, selection ability is more likely to be demonstrated by focused funds, although this concentration compromises their ability to time-the-market. In addition, highly focused funds that invest in a small number of asset and regional classes are less able to time-the-market. Another intuitive result is that funds with larger $\bar{\sigma}_p$ values, which are more willing to alter their industry portfolio weights, have greater market timing ability.

Selection Ability

	Intercept	Expense	Size	Turnover	Focus	$\bar{\sigma}_p$
Industry	2.14**	-1.91*	1.39	1.27	-3.47***	0.35
(p-value)	(0.032)	(0.056)	(0.166)	(0.205)	(0.000)	(0.725)
Asset	-1.40	-1.19	0.83	0.39	1.16	3.96***
(p-value)	(0.163)	(0.235)	(0.401)	(0.695)	(0.245)	(0.003)
Region	-0.48	-0.39	1.76*	0.90	0.24	0.21
(p-value)	(0.632)	(0.695)	(0.080)	(0.371)	(0.811)	(0.834)

Market Timing Ability

	Intercept	Expense	Size	Turnover	Focus	$\bar{\sigma}_p$
Industry	2.39***	-0.98	0.38	-0.15	3.79***	-10.08***
(p-value)	(0.017)	(0.328)	(0.700)	(0.883)	(0.000)	(0.000)
Asset	116.15***	0.31	0.74	1.06	-40.20***	-1.57
(p-value)	(0.000)	(0.755)	(0.460)	(0.289)	(0.000)	(0.116)
Region	47.40***	-1.76	0.23	0.58	-6.37***	-0.34
(p-value)	(0.000)	(0.173)	(0.818)	(0.395)	(0.000)	(0.735)