Performance Measurement of Mutual Funds, Hedge Funds, and Institutional Accounts

Russ Wermers

Robert H. Smith School of Business, University of Maryland at College Park, College Park, Maryland; email: wermers@umd.edu

Abstract

This review describes several important recent advances in the measurement of the performance of actively managed portfolios. For returns-based performance evaluation, we discuss several innovations, such as conditional performance evaluation, Bayesian approaches, and a new multiple-testing approach—the false-discovery rate. For portfolio holdings–based performance evaluation, our discussion ranges from extensions of the standard Daniel, Grinblatt, Titman, and Wermers (DGTW) stock return adjustment procedure to conditional holdings-based approaches. Applications of these approaches in the mutual fund, hedge fund, and institutional account universes are presented.

Keywords

portfolio performance, alpha, performance evaluation, performance attribution
1. INTRODUCTION

An estimated $71.3 trillion is invested in managed portfolios worldwide, as of 2009 (see TheCityUK 2010). This wealth is, by far, invested mostly in actively managed mandates. Expenses and trading costs of such actively managed mandates amount to at least 2%/year, making active asset management and its trades at least a $1.4 trillion/year industry.1 However, recent studies, such as by Barras, Scaillet, and Wermers (BSW), indicate that active manager skills, on average, have been declining over the past couple decades (Barras et al. 2010). Clearly, in this environment, precise, cutting-edge methods for evaluating active managers are extremely important to both a well-functioning economy and to society at large. Fortunately, the confluence of inexpensive computer processing power and storage with the increased disclosure of fund-level information has enabled the analysis of active managers to evolve substantially over the past few decades. [For instance, the Securities and Exchange Commission’s (SEC) Edgar site is a windfall for academic researchers who collect data that were previously available only in hardcopy form.]

Concurrently, a rich literature has developed since Jensen (1968) on methodologies to test for the skills of asset managers. These techniques can be classified into two major approaches: (a) returns-based performance evaluation and (b) portfolio holdings–based performance evaluation. In turn, researchers have applied both approaches in very simple ways, as well as more sophisticated and innovative ways.

Each approach has its advantages and disadvantages. Returns-based approaches rely on less information from fund managers, which is useful in situations where managers do not disclose much information, such as in hedge fund markets. In addition, even in situations where portfolio holdings are readily available, as in mutual fund markets (with quarterly holdings data), returns data are usually available on a much more frequent basis—daily, in the case of mutual funds.

All performance evaluation approaches start with the same premise. Managers should be rewarded for bets not easily replicated by uninformed investors; that is, managers should not be rewarded for easy bets that represent passive or known simple mechanical strategies (to borrow from a recent President, “it’s the alpha, stupid!”2). In effect, models of performance evaluation attempt to measure, with the best possible statistical accuracy, the precision of an asset manager’s private information about security, sector, or market returns.

Our objective is to survey the most useful recent academic research covering performance evaluation of asset managers. We begin with a discussion of recent performance evaluation advances, both returns-based and portfolio holdings–based, that are applicable across these different classes of conventional asset managers: mutual funds, hedge funds, and institutional accounts. (We note here that several of these approaches are applicable to other types of asset managers, such as those overseeing private equity or active exchange-traded funds.) Then, we progress to a discussion of empirical results from some recent, important studies that are specific to each class of managers. We conclude with some suggested directions for future research.

---

1See, for example, French (2008) for evidence on the size and expenses of U.S. active asset managers. For more recent evidence on U.S. mutual funds, see http://www.ici.org. For evidence on European funds, see http://www.efama.org. For evidence on world mutual fund markets, see http://www.iimf.org.

2James Carville popularized “It’s the economy, stupid” during Bill Clinton’s successful 1992 presidential campaign, and our expression is one of many different versions that have since been snowcloned.
2. GOALS, GUIDELINES, AND PERILS OF PERFORMANCE EVALUATION

Before proceeding with our discussion of benchmarks, performance measures, and models, let us set out some standards.

2.1. Benchmarks

To use a particular benchmark (or set of benchmarks) as a measuring tape, against which to compare a fund, a manager should be able to size up the competition (i.e., the benchmark). Accordingly, Bailey (1995) proposes that a valid benchmark should be

1. **Unambiguous**: The names and weights of component securities should be known (this rules out unknown derived benchmarks, such as those based on the arbitrage pricing theory factors);
2. **Tradeable**: It should be available as a passive investment alternative for the manager;
3. **Measurable**: It must be possible to compute a valid return on the benchmark periodically (might not be possible for benchmarks with illiquid assets);
4. **Appropriate**: The benchmark must reflect the manager’s style;
5. **Reflective of current investment opinions**: A manager should be able to form an opinion on the expected rate of return on the benchmark; and
6. **Specified in advance**: It should be able to give the manager a passive alternative ahead of time, to make clear the measuring tape.

2.2. Performance Measures

Chen & Knez (1996) propose that a performance measure should have, at minimum, four properties:

1. **Fit**: It should capture the strategies that could reasonably be used by an uninformed investor with control factors, and assign zero performance to portfolios that result from these simple strategies, whether they be passive or active;
2. **Be scalable**: Linear combinations of the manager measures should equal the measure for the same linear combination of their portfolios;
3. **Be continuous**: Two managers with arbitrarily close skills should have arbitrarily close performance measures; and
4. **Exhibit monotonicity**: It should assign higher measures for more skilled managers.

These properties ensure that performance measures are not easily gamed by unskilled asset managers, and that investors do not pay managers for strategies that they could easily replicate themselves.

2.3. Manipulation-Proof Performance Measures

Goetzmann, Ingersoll, Spiegel, and Welch (GISW) provide further detail on properties of performance measures that resist gaming (Goetzmann et al. 2007). An often-quoted way to game a simple returns-based regression model that assumes normally distributed returns, for example, is to sell short out-of-the-money call or put options on an index, then invest
the proceeds in the risk-free asset. This strategy generates left-skewed returns, with greater skewness present for more out-of-the-moneyness of the derivatives; resulting small-sample regression alphas (assuming normality) will be positive (and low volatility) most of the time, even though the strategy requires no real manager skill.

GISW ask whether a manipulation-proof performance measure (MPPM) is possible, if we define an MPPM as one that has four properties (Goetzmann et al. 2007):

1. The measure should produce a single valued score with which to rank each subject;
2. The score’s value should not depend upon the portfolio’s dollar value;
3. An uninformed investor cannot expect to enhance his estimated score by deviating from the benchmark portfolio. At the same time, informed investors should be able to produce higher scoring portfolios, and can always do so by taking advantage of arbitrage opportunities; and
4. The measure should be consistent with standard financial market equilibrium conditions.

GISW find that an MPPM is possible, and that its formula has a simple interpretation: It is the average per period welfare of a power utility investor in the managed fund. Unfortunately, most performance measures used in the literature are not perfectly manipulation proof, which makes it very important to understand the source of the performance of managers. Short of personal knowledge of the portfolio manager (which did not seem to work very well with Bernard Madoff’s hedge fund), there are two main ways to accomplish this goal. First, extract as much information as possible from the reported returns of the fund. Second, obtain detailed portfolio holdings—or, even better, a complete listing of trades including prices, sizes, and dates. This review aims to set out some of the best recent advances in each of these two areas.

2.4. Type I or Type II Error (Which Would You Prefer?)

Avoiding manipulation can be accomplished in a very easy way: dogmatically assign all active managers with zero performance, ex ante. Of course, this comes with a huge price: We miss out on the superior returns of truly skilled managers. So, all performance models and benchmarks must be chosen with an eye toward which is more important: Type I error, falsely identifying a skilled manager, or Type II error, falsely identifying an unskilled manager. Surely, only the most dogmatic investor would completely focus on minimizing Type I error, as shown by Baks et al. (2001). But, it would be a bigger mistake to focus completely on minimizing Type II error, as shown by BSW (Barras et al. 2010). Clearly, models and benchmarks that follow the above guidelines help to reduce both types of errors. Because no model is perfect, however, the researcher should attempt to apply as many models as is practical, reasonably adding and changing assumptions about benchmarks and model specifications to test for the sensitivity of measured performance to different benchmarks and models. [For example, Pastor and Stambaugh (PS) recommend adding nonmarket benchmarks to improve inferences about manager ability—such as a technology index when modeling technology funds (Pastor & Stambaugh 2002b).] As with the electrical engineering student trying to determine what is in the black box (capacitors, resistors, inductors, etc.), the aggregate evidence from several different tests (models and benchmarks) should lead to stronger conclusions about manager skills.
2.5. The Confounding Role of Risk Aversion

It has long been known that risk aversion may mitigate a skilled manager’s ability to produce alpha. In an extreme example, Verrecchia (1980) shows that a manager with quadratic utility will exhibit a reduction in alpha when the manager receives a signal of a large market return, due to risk aversion increasing in wealth for quadratic utility. (Also, Admati & Ross 1985 show, in a one-period model, that alphas are a function of both ability and risk aversion.) In response, Koijen (2010) applies a structural model (with fund managers having constant relative risk aversion preferences) that allows the separation of risk aversion and skill using time-varying alphas, betas, and residual risk. The variation in alphas, betas, and residual risk, together, are informative about the parameters describing preferences, technology (skill and benchmarks), and the incentive contract of the manager. Interestingly, Koijen finds a positive correlation between estimates of ability and risk aversion among U.S. domestic equity mutual fund managers, which indicates that skilled managers may be difficult to locate because they invest rather conservatively. In using Koijen’s approach, we must assume explicitly something about the \( a \) type of preferences (e.g., constant relative risk aversion), \( b \) benchmark, and \( c \) incentive contract of the portfolio manager. In many cases, we do not know these parameters, and Koijen (2010) demonstrates that we must use care in interpreting the results of regression approaches that do not account for the interplay of these three parameters over time, such as the returns-based measures discussed in the next section. Holdings-based performance evaluation, discussed in a later section, allows a more precise (but, still imperfect) inference about ability in the presence of risk aversion. However, if we can assume that the asset managers that we examine have roughly similar levels of risk aversion, then we believe that comparisons using the standard measures below are reasonable.

3. RETURNS-BASED ANALYSIS

All asset managers provide net returns to their clients, and many make these return data public. The widespread availability of returns data makes it imperative to extract the maximum information possible about fund performance, strategy, and risk-taking from returns. This goal involves the application of the best possible econometric models, based on \( a \) a knowledge of the types of risks taken by fund managers (i.e., the application of proper benchmarks), \( b \) the breakdown of systematic versus idiosyncratic risks taken by the managers (this involves both proper benchmarks and the detection of dynamic risk-taking strategies by managers), and \( c \) the statistical distribution of the returns to both systematic and idiosyncratic risks (i.e., both can be non-normally distributed).

Many biases can result from the improper application of returns-based models, all of which [as described by \( a \), \( b \), and \( c \) above] require assumptions about the set of strategies from which a manager generates returns. The most well-documented problem is that of choosing a benchmark, or set of benchmarks, that are mean-variance inefficient when measuring performance with a mean-variance model. Roll (1978) shows how such a choice of inefficient benchmarks can result in any conceivable ranking of investment managers, making performance evaluation an ambiguous undertaking.

In the next section, we describe the current “best practices” regarding the choice of benchmarks and models for common types of asset managers (e.g., mutual funds and
hedge funds). In subsequent sections, we discuss solutions to several other problems confronting returns-based analysis, including those described by (b) and (c) above.3

3.1. Baseline Models

The returns-based model most widely used among academics in analyzing equity managers is the four-factor model of Carhart (1997),

$$\tilde{r}_t = \alpha + \beta \cdot \tilde{R}MRF_t + s \cdot \tilde{SMB}_t + b \cdot \tilde{HML}_t + u \cdot \tilde{UMD}_t + \tilde{\epsilon}_t, \tag{1}$$

where $r_t$ is the month-$t$ excess return on the managed portfolio (net return minus T-bill return); $\tilde{R}MRF_t$ is the month-$t$ excess return on a value-weighted aggregate market proxy portfolio; and $\tilde{SMB}_t$, $\tilde{HML}_t$, and $\tilde{UMD}_t$ are the month-$t$ returns on value-weighted, zero-investment factor-mimicking portfolios for size, book-to-market (BTM) equity, and one-year momentum in stock returns, respectively. This model is based on empirical research by Fama & French (1992, 1993, 1996) and Jegadeesh & Titman (1993) that finds these factors closely capture the cross-sectional and time-series variation in stock returns. A regression of an individual stock return on the four-factor model gives an R-squared of only approximately 10%-20%, but a regression of a managed (long-only) portfolio usually gives an R-squared in excess of 90% due to the high amount of diversification usually present in such portfolios.

Hedge funds are notoriously difficult to model. However, in hedge fund studies, a most popular (current) model has evolved: the seven-factor model of Fung & Hsieh (2004),

$$\tilde{r}_t = \alpha + \beta \cdot \tilde{S}P\tilde{R}F_t + s \cdot \tilde{S}MB_t + g \cdot \tilde{T}R\tilde{E}A\tilde{S}10\text{YR}_t + c \cdot \tilde{C}R\tilde{E}D\it{I}t + b \cdot \tilde{B}O\it{N}D\it{P}T\it{F}S_t + d \cdot \tilde{C}U\it{R}R\it{P}T\it{F}S_t + o \cdot \tilde{C}O\it{M}M\it{P}T\it{F}S_t + \tilde{\epsilon}_t, \tag{2}$$

where the factors are, successively, the S&P 500 return minus risk-free rate, Wilshire small cap minus large cap return, change in the constant maturity yield of the 10-year U.S. Treasury, change in the spread of Moody’s Baa-rated corporate bonds over the 10-year U.S. Treasury, and three primitive trend-following strategies—derived from bond markets, currency markets, and commodities markets, respectively. Note that this model, to avoid the data biases in hedge fund databases, constructs benchmarks based on asset returns, rather than hedge fund returns. Notably, Fung & Hsieh (2004) find that significant loadings on these seven risk factors are present in 57% of the hedge funds in Lipper’s TASS database, and 37% of the funds in the Hedge Fund Research (HFR) database.

There are many assumptions behind the models represented by Equations 1 and 2. For example, both models assume normally distributed asset and factor returns, as well as constant risk loadings. In addition, they assume no temporal dependency in risk factor returns or residual (idiosyncratic) returns. Next, we explore the treatments of violations of some assumptions of these models.

However, there are issues that returns-based analysis will not be able to solve. For example, Goetzmann et al. (2000) show that measuring performance on a less frequent basis than the manager’s trading horizon can result in an interim trading bias in measured alphas. Such issues require holdings-based analysis, which is discussed in a later section.
3.2. Return Smoothing

Bollen & Pool (2009) find that return smoothing is common among hedge funds, apparently to fool investors into believing returns are less volatile. [For example, a fund having true returns of +12% and −8% may report (compounded equivalent) returns of +1.51% and +1.51% to lower both the observed fund volatility and the fund's measured exposure to risk factors.] Getmansky, Lo, and Makarov (GLM) suggest that Equation 2, or any chosen model, can be modified to account for return smoothing (Getmansky et al. 2004). GLM model the observed hedge fund return as a smoothing of lagged and contemporaneous returns,

$$\bar{R}_t = \theta_0 \bar{R}_t + \theta_1 \bar{R}_{t-1} + \theta_2 \bar{R}_{t-2},$$

where $\theta_0 + \theta_1 + \theta_2 = 1$. GLM show how a seven-factor model augmented with the smoothing equation can be estimated with maximum likelihood techniques.

3.3. Non-Normal Alphas (Model Residuals)

Mutual funds, and especially hedge funds, have return distributions that render suspect any analysis using standard assumptions about the shape of these distributions as well as the models and risk factors that are appropriate to measure performance. There are several irregularities in managed fund returns that deserve attention. First, hedge funds often have a large skewness and/or kurtosis in their alpha distributions. If we do not control for these non-normal distributions, and, instead, measure performance with a standard factor model (that assumes normality), both large Type I and Type II errors can easily result.

Kosowski, Timmermann, Wermers, and White (KTWW) show how to implement a bootstrap technique with any arbitrary model of returns in a complicated setting with large numbers of active managers having different (potentially non-normal) distributions (Kosowski et al. 2006). KTWW show that managers with high or low estimated alphas, relative to the group of all active managers, tend to have return distributions that have greater levels of skewness and kurtosis present—making a bootstrap especially important in the extreme regions of the performance spectrum.

A brief illustration of the bootstrap using the Carhart (1997) four-factor model follows; the application of the bootstrap procedure to other models is very similar, with the only modification of the following steps being the substitution of the appropriate benchmark model of performance. (The simplest bootstrap from KTWW, the residual-only bootstrap, is illustrated. More complicated bootstrapping procedures are also described in KTWW.)

First, the Carhart model is used to compute ordinary least-squares-estimated alphas, factor loadings, and residuals using the time series of monthly net returns (minus the T-bill rate) for fund $i$ ($r_{it}$):

$$r_{it} = \hat{\alpha}_i + \hat{\beta}_i RMRF_t + \hat{s}_i SMB_t + \hat{h}_i HML_t + \hat{u}_i UMD_t + \hat{e}_{i,t}. \tag{3}$$

For fund $i$, the coefficient estimates, $\{\hat{\alpha}_i, \hat{\beta}_i, \hat{s}_i, \hat{h}_i, \hat{u}_i\}$, as well as the time series of estimated residuals, $\{\hat{e}_{i,t}, t = T_{i0}, \ldots, T_{i1}\}$, and the $t$-statistic of estimated alpha, $t_{\hat{\alpha}_i}$, are saved, where $T_{i0}$ and $T_{i1}$ are the dates of the first and last monthly returns available for fund $i$, respectively. Next, for each fund $i$, we draw a sample with replacement from the fund residuals that are saved in the first step above, creating a pseudo time series of resampled residuals, $\{\hat{\epsilon}_{i,b,t}, t_s = T_{i0}^b, \ldots, T_{i1}^b\}$, where $b$ is an index for the bootstrap number (so, $b = 1$...
for bootstrap resample number one), and where each of the time indices, $s_t^b, \ldots, s_{T_t}^b$, are drawn randomly from $[T_{i0}, \ldots, T_{i1}]$ in such a way that reorders the original sample of $T_{i0} - T_{i0} + 1$ residuals for fund $i$ (with replacement). Conversely, the original chronological ordering of the factor returns is unaltered. 4

Next, we construct a time series of pseudo monthly excess returns for this fund, imposing the null hypothesis of zero true performance ($z_i = 0$, or, equivalently, $t_{z_i} = 0$):

$$\left\{ r_t^b = \tilde{b}_0 RMRF_t + \tilde{s}_i SMB_t + \tilde{b}_1 HML_t + \hat{u}_i UMD_t + \tilde{e}_{b,t} \right\},$$

(4)

for $t = T_{i0}, \ldots, T_{i1}$ and $t_{z_i} = s_t^b, \ldots, s_{T_t}^b$. As Equation 4 indicates, this sequence of artificial returns has a true alpha (and $t$-statistic of alpha) that is zero by construction. However, when we next regress the returns for a given bootstrap sample, $b$, on the Carhart factors, a positive estimated alpha (and $t$-statistic) may result, because that bootstrap may have drawn an abnormally high number of positive residuals, or, conversely, a negative alpha (and $t$-statistic) may result if an abnormally high number of negative residuals are drawn.

Repeating the above steps across all funds $i = 1, \ldots, N$, we arrive at a draw from the cross section of bootstrapped alphas. Repeating this for all bootstrap iterations, $b = 1, \ldots, 1000$, we then build the distribution of these cross-sectional draws of alphas, $\{\tilde{z}_i^b, i = 1, \ldots, N\}$, or their $t$-statistics, $\{t_{z_i}^b, i = 1, \ldots, N\}$, that result purely from sampling variation, while imposing the null of a true alpha that is equal to zero. For example, the distribution of alphas (or $t$-statistics) for the top fund is constructed as the distribution of the maximum alpha (or, maximum $t$-statistic) generated across all bootstraps. 5

(Importantly, this maximum alpha distribution can be non-normal, even if every individual fund alpha is normally distributed, as described in KTWW.) If we find that our bootstrap iterations generate far fewer extreme positive values of $\tilde{z}$ (or $t_z$) compared to those observed in the actual data, then we conclude that sampling variation (luck) is unlikely to be the sole source of high alphas—genuine stock-picking skills likely exist.

3.4. Nonstable Regression Parameters

The standard assumption of stable model parameters in a linear model is often violated by mutual funds, and, especially, by hedge funds. For instance, hedge funds often shift strategies or leverage, which can result in a flawed estimate of alpha by standard fixed-parameter regression techniques.

Bollen & Whaley (2009) demonstrate how a standard linear model can be modified using the changepoint regression technique of Andrews, Lee, and Ploberger (ALP), to pick up unknown structural shifts in risk loadings or alphas by funds (Andrews et al. 1996). (By structural shifts, we mean intended shifts in model parameters, not those that result from unexpected, temporary shifts in factor loadings of the underlying assets or strategies.) For instance, a changepoint regression (in a single-factor model) with a single break

---

4The authors also describe a version of the bootstrap that draws the residual from the same time period for all funds (that exist during that period)—a cross-sectional bootstrap. Fama & French (2010) recommend simultaneously drawing factor returns from these randomized time periods for improved inference.

5Of course, this maximum alpha can potentially be associated with a different fund during each bootstrap iteration, depending on the outcome of the draw from each fund’s residuals.
(changepoint) can be written as (for some fraction of the sample, \( \pi \), that lies between 0 and 1)

\[
\tilde{r}_t = \alpha_1 + \beta_1 \tilde{F}_t + \varepsilon_t \quad \text{for } t = 1, \ldots, T\pi
\]

\[
\tilde{r}_t = \alpha_1 + \alpha_2 + (\beta_1 + \beta_2) \tilde{F}_t + \varepsilon_t \quad \text{for } t = T\pi + 1, \ldots, T.
\]

Then, the econometrician tests the null hypothesis \( H_0: \alpha_2 = \beta_2 = 0 \). ALP show how to construct a test for an arbitrary number of changepoints, which is particularly useful for managed funds with long histories and different managers or strategies over time.

### 3.5. Unpriced Benchmarks

PS show how the addition of an unpriced benchmark to a regression model can improve inference about manager skills (Pastor & Stambaugh 2002b). The key is that the longer history of the unpriced benchmark, relative to an active portfolio that chooses securities in that benchmark, helps to reduce sampling error in the regression. For instance, a technology fund manager in the United States may be modeled with the four-factor model of Carhart, augmented with a passive technology index,

\[
\tilde{r}_t = \alpha^{PS} + \beta \cdot \tilde{RMRF}_t + s \cdot \tilde{SMB}_t + b \cdot \tilde{HML}_t + u \cdot \tilde{UMD}_t + \eta \cdot \tilde{TECH}_t + \tilde{\varepsilon}_t.
\]

The intercept from the above augmented regression, \( \alpha^{PS} \), has tighter standard errors than the intercept from the standard Carhart model,

\[
\tilde{r}_t = \alpha + \beta \cdot \tilde{RMRF}_t + s \cdot \tilde{SMB}_t + b \cdot \tilde{HML}_t + u \cdot \tilde{UMD}_t + \tilde{\varepsilon}_t,
\]

due to the ability of \( \tilde{TECH} \) to better capture passive, unpriced risk contained in the active technology portfolio. To apply the augmented model, PS advocate separately running a first-stage regression of \( \tilde{r}_t \) and \( \tilde{TECH}_t \) on the four factors using the entire time series for the index (which is usually much longer than that for a managed fund), saving, respectively, the regression intercepts \( \alpha \) and \( \alpha^{TECH} \). In the second stage, the resulting \( \alpha^{TECH} \) is used to adjust the first-stage \( \alpha \) of the fund. \( \alpha^{TECH} \) helps to control for sampling error that is contained in \( \alpha \).

As another version of this approach, Hunter et al. (2011) suggest augmenting with the unpriced active peer-group risk,

\[
\tilde{r}_t = \alpha^{HKKW} + \beta \cdot \tilde{RMRF}_t + s \cdot \tilde{SMB}_t + b \cdot \tilde{HML}_t + u \cdot \tilde{UMD}_t + \gamma \cdot \tilde{TECHFUND}_t + \tilde{\varepsilon}_t,
\]

where \( \tilde{TECHFUND} \) is the residual return of the equal-weighted portfolio of peer funds, after controlling for the four factors in a first-stage regression. There are several advantages of using an active peer-group rather than a passive index. First, adding a peer group factor controls for strategies (even those that involve dynamic risk-factor loadings) used in common by many funds, which can easily be exploited by investing equal amounts in all active funds (and not requiring, therefore, to rank the funds). Second, estimation errors can be dramatically reduced. For instance, this approach reduces the well-known problem that highly ranked funds tend to have underestimated betas during the ranking period, and, therefore, do not have persistent performance once the higher beta is estimated during the out-of-sample period—the addition of \( \tilde{TECHFUND} \) helps to capture fund residuals that are, by chance, (negatively) correlated with risk factors.
3.6. Bayesian Methods

The large levels of idiosyncratic risk taken by active asset managers mean that large estimation errors can result when estimating models. In addition, models and benchmarks never perfectly control for priced risk. Although a careful choice of models, as described above, can help to tone down these issues, a Bayesian point of view can also be used to mitigate the influence of noise and potential model misspecification.

PS provide a clear demonstration of a Bayesian evaluation of mutual funds, and help to usher in several further contributions using a Bayesian approach (Pastor & Stambaugh 2002a). An appealing feature of their model is that it allows a prior belief to be specified, separately, for the skill of a particular manager and for the ability of a particular model to properly price all passive assets. PS show that optimal portfolios of mutual funds are influenced greatly by prior beliefs about both; this echoes the findings of Baks et al. (2001), who find that even a very skeptical investor can substantially improve her expected utility by having access to actively managed funds, rather than being forced to dogmatically accept a prior belief that no active funds can outperform passive strategies. An excellent in-depth discussion of Bayesian methods in asset pricing and investment fund performance evaluation can also be found in Avramov & Zhou (2010).

3.7. Conditional Returns-Based Performance Measurement

A new class of models examines whether manager skills and risk-taking are different at differing points in the business cycle. Ferson & Schadt (1996) show that asset managers change their risk exposure as the level of various business cycle indicators evolve over time, partly in response to changes in their inflows from investors. Christopherson et al. (1998) show that manager alphas may also change, as managers may have specialized skills that work well only in certain macroeconomic environments.

To the extent that a fund’s alpha varies systematically over time, this could be due to either (a) embedded macroeconomic sensitivities (e.g., sector-wide persistent mispricing), (b) time-varying skill, or (c) time-varying opportunities for managers to benefit from their skills. Although all three explanations may play a role, studies lend the most support to the third explanation—namely, that certain environments offer more mispricing opportunities where managers can take advantage of their superior insights. For example, many contrarian managers underperformed during the tech bubble of the late 1990s, when prices diverged significantly from fundamentals, but outperformed by a huge margin when the bubble burst. Did their skills suddenly change so dramatically, or did the market simply provide more opportunities for them in one period versus the other? It is likely the latter.

Studies by Moskowitz (2000) and Kosowski (2006) reveal that the average active manager is more likely to outperform the market during recessions. This is probably not the result of holding cash in down markets, given that Kosowski, in particular, adjusts returns for market risk. Instead, it seems likely that recessions are periods of above-average uncertainty, when superior information and analysis can be particularly valuable. Consistent with this explanation, Kosowski (and others) also find that the average active fund performs better in periods of higher return dispersion and volatility, which are also likely to be periods of heightened uncertainty—and opportunity.
Conditional returns-based analysis is simple, and involves a straightforward extension of linear unconditional models. Suppose that we do believe that a manager’s alpha and beta, in a capital asset pricing model (CAPM; Sharpe 1964) setting, dynamically evolve as the level of a particular macroeconomic factor, say, end-of-month t short-term interest rate \( z_t \), changes through time:

\[
\begin{align*}
\alpha_t &= \alpha(z_{t-1}) \\
\beta_t &= \beta(z_{t-1})
\end{align*}
\]

We can approximate these functional forms through a Taylor-series expansion, assuming they are continuous and differentiable:

\[
\begin{align*}
\alpha_t &= \alpha_0 + \alpha_0' \cdot z_{t-1} + \frac{1}{2} \alpha_0'' \cdot z_{t-1}^2 + \ldots \\
\beta_t &= \beta_0 + \beta_0' \cdot z_{t-1} + \frac{1}{2} \beta_0'' \cdot z_{t-1}^2 + \ldots
\end{align*}
\]

Dropping higher-order terms, which means that we assume the manager is not engaging in nonlinear strategies with respect to interest rates, then substituting into the Jensen (CAPM) model, gives

\[
\tilde{r}_t = \alpha_0 + \alpha_1 \cdot z_{t-1} + \beta_0 \cdot \overline{RMRF}_t + \beta_1 \cdot z_{t-1} \cdot \overline{RMRF}_t + \tilde{\epsilon}_{t, t}, \tag{5}
\]

where \( \alpha_1 \) and \( \beta_1 \) represent the (first-order) sensitivity of fund alpha and beta, respectively, to short interest rates. The above model suggests implementing a three-factor model, in place of the usual one-factor model of Jensen. Although this is a simple solution to time-varying skills and risk-taking, it is easy to see that adding more factors and more macroeconomic variables geometrically increases the number of explanatory variables. Thus, factors and macrovariables should be chosen judiciously; some degrees of freedom can be saved by limiting the interaction of macrovariables to a subset of the risk factors. Specifically, it is helpful to know which risk factors the manager is attempting to time using macrovariables, as well as the particular macrovariables being used by the manager for each risk factor. (For instance, a common approach is to interact the macrovariables only with the market factor in a Carhart four-factor regression.)

### 3.8. Stochastic Discount Factors

Chen & Knez (1996), Dahlquist and Soderlind (DS) (Dahlquist & Soderlind 1999), and Ferson et al. (2006) propose a stochastic discount factor (SDF) approach to measuring performance. [Grinblatt and Titman (GT) proposed a time-varying marginal utility-weighted return as a precursor to the SDF approach (Grinblatt & Titman 1989b).] The idea is quite elegant: If the law of one price holds, then there exists an SDF, \( m_t \), that prices any passive portfolio (with one-period return ending at \( t \) of \( \tilde{R}_t \)) either unconditionally,

\[
E[m_t \tilde{R}_t - 1] = 0,
\]

or conditional on any vector of public information available at the end of the prior period, \( z_{t-1} \),

\[
E[z_{t-1} m_t \tilde{R}_t - z_{t-1}] = 0.
\]
This set of restrictions implies moment conditions that can be used to estimate the SDF, $m_t$, with the generalized method of moments (GMM). Using the GMM allows the researcher to avoid choosing a particular assumed distribution for returns of passive or active portfolios, which, as we saw above, can be a great advantage with real portfolios of poorly behaved securities.6

Once the SDF is estimated that prices all passive portfolios and strategies based on public information, the SDF is used to estimate the alpha or performance of any active portfolio, $p_t$, during time periods $t = 1$ to $T$ as

$$
\alpha^p = \hat{E}_t\left[ z_{t-1} m_t R_t^p - z_t \right] = \frac{1}{T} \sum_{t=1}^{T} \left( z_{t-1} m_t R_t^p - z_t \right).
$$

Further, DS show that, if positivity is not imposed on the SDF, the above test of zero performance can be interpreted in a mean-variance world as whether a managed portfolio expands the mean-variance frontier. (In the absence of arbitrage, a positive SDF will exist.)

DS study several issues about SDFs with Monte Carlo simulations, including their small-sample properties—which are important in light of managed funds having relatively short histories. They find that a limitation of the SDF approach is that the managed portfolio must have a considerable actual alpha over a long period of time for the GMM to reject that the fund has zero alpha.

### 3.9. False-Discovery Rate Approach to Measuring Performance of a Group of Funds

Often, a researcher will wish to estimate the number, or proportion, of active fund managers having true skills in beating their benchmarks. Indeed, this is really the central question in the active/passive management debate: Does an economically significant number of skilled managers exist, and, if so, can we locate them ahead of time?

BSW address exactly this question with the false-discovery rate (FDR) approach (Barras et al 2010). The FDR is a statistical approach that has proven fruitful in studies in other academic areas, where an experiment deals with large numbers of subjects, such as studies of the commonality of substrands in the DNA of large groups of people having a common disease.

The idea behind the FDR approach is extremely simple, as is its application, which enhances its appeal. Consider the multiple hypothesis test of skills across $M$ active fund managers,

$$
H_{0,1} : \alpha_1 = 0, \quad H_{A,1} : \alpha_1 \neq 0, \\
\vdots \quad \cdots \\
H_{0,M} : \alpha_M = 0, \quad H_{A,M} : \alpha_M \neq 0.
$$

Traditional tests require the inversion of large covariance matrices, which are difficult to construct with hundreds or thousands of funds, many of which do not overlap in time.

---

6In addition to the non-normal distributions mentioned in a prior section, managed (and passive) portfolios can have problems with heteroskedastic and serially correlated returns. These issues pose no large problems when using the GMM.
The FDR approach, as developed by Storey (2002), instead relies only on the (two-sided) \( p \)-values associated with the (alpha) \( t \)-statistics of each of the \( M \) funds. By definition, zero-alpha funds satisfy the null hypothesis, \( H_{0,i} : \alpha_i = 0 \), and, therefore, have \( p \)-values that are uniformly distributed over the interval \([0, 1]\) as shown in the (nearly) uniform distribution of an empirical sample from BSW shown in Figure 1 above, over the interval \( 0.2 < p < 1.7 \). On the other hand, \( p \)-values of unskilled and skilled funds tend to be very small because their estimated \( t \)-statistics tend to be far from zero. We can exploit this information to estimate \( \pi_0 \), the proportion of funds with no skills, without knowing the exact distribution of the \( p \)-values of the unskilled and skilled funds.

The FDR approach merely estimates where the horizontal line lies by choosing a \( \lambda^* \), above which \( p \)-values are assumed to come from unskilled fund managers (thus, allowing a certain level of Type II error as acceptable). Once the hurdle (horizontal density line) is estimated, a simple integration of the density above the hurdle gives the estimated proportion of truly skilled funds.

\footnote{To see this we denote by \( T_i \) and \( P_i \), the \( t \)-statistic and \( p \)-value of the zero-alpha fund; by \( \hat{t}_i \) and \( \hat{p}_i \) their estimated values; and by \( T_i(P_i) \) the \( t \)-statistic associated with \( p \)-value, \( P_i \). We have \( \hat{p}_i = 1 - F(\hat{t}_i) \), where \( F(\hat{t}_i) = \text{prob}(|T_i| < \hat{t}_i | \alpha_i = 0) \). The \( p \)-value \( P_i \) is uniformly distributed over \([0,1]\) given that its CDF, \( \text{prob}(P_i < \hat{p}_i) = \text{prob}(1 - F(T_i(P_i)) < \hat{p}_i) = \text{prob}(|T_i(P_i)| > F^{-1}(1 - \hat{p}_i)) = 1 - F(F^{-1}(1 - \hat{p}_i)) = \hat{p}_i \).}
4. HOLDINGS-BASED ANALYSIS

Recent advances in methods that examine performance at the security level allow researchers (and investors) to reexamine the active/passive issue in much the same way as advances in DNA profiling have reopened many criminal cases for a more thorough analysis of the probable guilt of a defendant beyond more traditional fingerprinting methods. More detailed information about the actions of the “defendant” (the fund manager) can only serve to improve the precision of our inferences about the manager’s alleged talents. These security-level performance evaluation approaches have become known as portfolio holdings–based performance evaluation and attribution or simply holdings-based performance measurement.

There are several reasons why the use of portfolio-holdings data may provide new insights into managed performance, relative to returns-based methods. First, contrary to returns-based methods, recent approaches that use portfolio-holdings data allow a more precise construction of a benchmark to address the benchmark-error problem outlined by Roll (1978).

Second, the style orientation of a fund may shift nontrivially during short time periods—such as with a value manager who adds technology stocks during the heyday of growth stocks of the 1990s. Wermers (2010) finds a substantial level of so-called style drift among U.S. mutual funds during the 1975 to 2002 period. As shown by Ferson & Schadt (1996), inferences about manager ability may change substantially when we adjust for the shifting style or risk loadings of a professionally managed portfolio.

Third, holdings data allow an examination of fund manager abilities before expenses and trading costs, which can provide important insights about asset allocation or security selection talents. For example, a manager may hold talents in picking securities, but is handicapped by a fund that is too small, making the scale of the fund too expensive to support (through the expense ratio, which is usually beyond the direct control of the portfolio manager). Identifying such a manager may allow a restructuring of the fund (i.e., through a merger or by creating a new separate account) that overcomes such a friction while retaining the performance-generating aspects of the fund.

Fourth, an analysis of portfolio holdings adds to our ability to decompose the sources of value added by a manager. For example, a manager may have strengths in choosing technology stocks, but not telecommunication stocks. Or, the manager’s performance may be concentrated in her top 10 holdings, rather than being spread more evenly. Such insights may allow us to better predict whether specific talents in which we are interested are likely to persist.

Finally, benchmarking is more precise when applied on a security-by-security basis, given that each security holding constitutes a separate and (almost) independent observation of manager ability. Thus, portfolio holdings increase the speed of convergence of the estimated manager ability to her true ability by the law of large numbers. (Cohen et al.

---

8Much of this section is drawn from my prior survey of holdings-based analysis (Wermers 2006). More detailed analysis is described there.

9Although the adding-up constraint implies that the choice of holdings are not truly independent, we can treat them as approximately so given that the manager chooses weights from a very large set of securities. Note that this constraint only implies a correlation between weights on the order of \(1/N\). Thus, even for managers of relatively small portfolios, such as a manager of a biotechnology fund, portfolio weights provide a set of \(N\) approximately independent observations of fund manager ability.
2005 show that measuring the similarity in a fund’s portfolio holdings with prior successful funds increases the speed of this convergence.

4.1. The Self-Benchmarking Method of Performance Evaluation

A pioneering study that used portfolio-holdings data obtained from periodic SEC filings of mutual funds is Grinblatt & Titman (1989a). This paper broke new ground by examining performance at the portfolio-holdings level, that is, using holdings applied to changes in the closing price quotations of stocks (plus any cash dividends) to compute the return of an asset manager. GT labeled this as the “hypothetical performance” of a fund manager, given that the manager (or those who would mimic the manager) could not exactly replicate the performance computed from closing prices due to the reality of trading costs. (Another, less important complication is that closing prices may not reflect the actual value of securities if sufficient liquidity is not present.) GT proceeded to regress these computed hypothetical returns on that of one or more market-based benchmarks to determine the hypothetical, precost alpha of a fund manager (Grinblatt & Titman 1989a). Therefore, although this paper is the first to directly examine portfolio holdings, the authors do not overcome the benchmark-error problem outlined by Roll (1978).

4.1.1. Statistical foundations. In response to the Roll criticism, academic researchers have built on Grinblatt & Titman (1989a) by developing measures of portfolio performance that allow weights to play a more central role in the formation of the benchmark(s) against which performance is measured. These new, portfolio-holdings measures (PHM) are grounded in the concept of performance being measured as the covariance between lagged weights and current returns,

\[
PHM_t = \text{cov}(\tilde{w}_{t-1}, \tilde{R}_t).
\]

This concept is very simple: A skilled manager will exhibit portfolio weights that move in the same direction as future returns. Of course, given that covariances are unscaled (as opposed, for example, to correlation coefficients), such covariance-based measures assign a higher performance to a manager who is more aggressive; a good manager who aggressively turns over his portfolio will achieve a higher PHM than a great manager who is more cautious. (We discuss extensions to pure covariance-based measures that address this issue in a later section.) This issue aside, the PHM may be measured in a few alternative ways, simply by the various approaches to defining covariance, that is,

\[
PHM_t = \text{cov}(\tilde{w}_{t-1}, \tilde{R}_t) = E[(\tilde{w}_{t-1} - E[\tilde{w}_{t-1}]) (\tilde{R}_t - E[\tilde{R}_t])].
\]

These alternative approaches end up being much more than mathematical technicalities in practice. Specifically, these different expansions of covariance imply very different choices of benchmarks for a managed portfolio, and, therefore, have very different implications when applied to practical performance evaluation problems. For example, it may be easier to estimate portfolio weights, \(E[\tilde{w}_{t-1}]\), given available data, than expected security returns, \(E[\tilde{R}_t]\)—or, vice versa. Also, if precise estimates of both expected weights and expected returns are readily available, then we would expect that the expansion, \(PHM_t = E[(\tilde{w}_{t-1} - E[\tilde{w}_{t-1}]) (\tilde{R}_t - E[\tilde{R}_t])]\), might converge more rapidly to
the true performance of a manager than the alternative expansions, \( PHM_t = E[(\tilde{w}_{t-1} - E[\tilde{w}_{t-1}])\tilde{R}_t] \) or \( PHM_t = E[\tilde{w}_{t-1}(\tilde{R}_t - E[\tilde{R}_t])] \).

Copeland and Mayers (CM) use the first expansion of the covariance in Equation 9, \( PHM_t = \text{cov}(\tilde{w}_{t-1}, \tilde{R}_t) = E[\tilde{w}_{t-1}(\tilde{R}_t - E[\tilde{R}_t])] \) (Copeland & Mayers 1982), whereas GT apply the second, \( PHM_t = \text{cov}(\tilde{w}_{t-1}, \tilde{R}_t) = E[(\tilde{w}_{t-1} - E[\tilde{w}_{t-1}])\tilde{R}_t] \) (Grinblatt & Titman 1993). Either of these definitions of covariance should be interpreted as a time-series covariance for a single security holding; CM and GT then propose that a manager’s performance might reasonably be measured as the summed covariances across all security holdings, or

\[
CM = GT = \sum_{j=1}^{N} \text{cov}(\tilde{w}_{j,t-1}, \tilde{R}_{j,t}),
\]

which measures the aggregate correctness of the manager’s portfolio bets, across all securities. Of course, the true covariance must be estimated with time-series data, and here is where CM and GT differ. (It is easily shown that, in large samples, the two measures converge.) Specifically, CM recommend the estimation

\[
CM = \sum_{j=1}^{N} \frac{1}{T} \sum_{t=1}^{T} \tilde{w}_{j,t-1}(\tilde{R}_{j,t} - E[\tilde{R}_{j,t}]),
\]

whereas GT favor the estimation

\[
GT = \sum_{j=1}^{N} \frac{1}{T} \sum_{t=1}^{T} (\tilde{w}_{j,t-1} - E[\tilde{w}_{j,t-1}])\tilde{R}_{j,t},
\]

Equation 9 can be rewritten, for convenience, as

\[
GT = \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{N} (\tilde{w}_{j,t-1} - E[\tilde{w}_{j,t-1}])\tilde{R}_{j,t}.
\]

The next step is to decide on a proxy for the expected weight for security \( j \) at the end of period \( t - 1 \), \( E[\tilde{w}_{j,t-1}] \), or for the expected return of security \( j \), \( E[\tilde{R}_{j,t}] \). There are many approaches to this problem. GT propose that the past weight on security \( j \) is the best proxy for the security’s expected weight (Grinblatt & Titman 1993). Specifically, using a market- or peer-based benchmark portfolio may allow the manager to game the benchmark by overweighting (relative to the benchmark) securities with higher expected returns, and underweighting securities with lower expected returns.\(^{10}\) Further, using the manager’s future portfolio weights as a proxy for expected current weights biases the \( GT \) measure if the manager implements trading strategies that are conditional on past security returns. For instance, a manager who overweights securities with high past returns (a relative strength or momentum trader) will exhibit a future weight that is correlated with current returns—thus, spuriously reducing the estimated performance of the manager. For these reasons, GT recommend the past weight of the security in the manager’s portfolio as a proxy for the current expected, or bogey, weight.

\(^{10}\)Although a manager could also move toward higher expected return securities over time, thus gaming the past-weight benchmark, such a strategy would be much more difficult to sustain over time unless the manager overweights temporarily high-risk securities. One such strategy is investing in stocks with high 12-month momentum.
For the same reason, CM choose the future return of a security as a proxy for its expected return, $E[\hat{R}_{t,t}]$. With these proxy choices, note that the $GT$ expansion is advantageous, relative to the CM expansion, given that it does not require a security to survive past the current period, $t$, to be included in the current period’s performance calculation (recall that the CM expansion requires the future return of a security for the current period computation). Nevertheless, the $GT$ measure has its costs—it requires that the manager must exist for one full performance measurement period (e.g., one year) before we can measure performance, whereas the CM measure requires only a single observation of the manager’s weights—at the end of date $t - 1$. In most practical situations, we would expect that managers have existed for at least one year before we wish to evaluate them; thus, the $GT$ measure is preferred in most situations. As such, we proceed with the $GT$ measure, although the application of the CM measure follows similar logic. 11

4.1.2. Application of the method. At this point, we note another nice feature of the $GT$ measure: It is exactly equal to the difference between the next period return earned by the current portfolio and a historical portfolio held by a manager, or

$$GT_t = \sum_{j=1}^{N} (\hat{w}_{j,t-1} - \hat{w}_{j,t-k-1})\hat{R}_{j,t},$$

where $\hat{w}_{j,t-1}$ is the end of month $t - 1$ (beginning of month $t$) portfolio weight of stock $j$ held by the manager, $\hat{w}_{j,t-k-1}$ is the weight of the same stock lagged $k$ months, and $\hat{R}_{j,t}$ is the month $t$ return of stock $j$. Note that, with this measure, the benchmark used to adjust the return of a portfolio for its risk in a given month is the current month’s return earned by the portfolio (held by the same manager) $k$ months prior to the current month’s holdings. Taking this interpretation further, the time-series average $GT$ measure, $\bar{GT} = \frac{1}{T} \sum_{t=k+1}^{T} GT_t$ represents the mean return of a zero-investment portfolio—long the current portfolio and short the historical portfolio of a given manager.

Consider the risk implications of the $GT$ measure: If the systematic risk of the current and benchmark portfolios are the same (plus random noise) for a given manager having no selectivity or timing abilities (as defined by Grinblatt & Titman 1989b), then the average portfolio represented by $\bar{GT}$ will have diminishing systematic risk, and this average portfolio return will converge to zero as $T$ increases—in small samples, both risk and return will (under certain reasonable conditions) be insignificant for that manager.

In addition, although prior performance measures are susceptible to benchmark error, à la Roll (1978) and other forms of model misspecification, these errors are much less problematic with the $GT$ measure. This results from the form of the $GT$ measure—it differences portfolio returns, so any misspecification that remains is due only to differences in loadings (on such an omitted risk factor) between the current and historical portfolios. In effect, past holdings represent the normal or bogey risk taken by a particular manager.

Note, as mentioned above, that the $GT$ approach has a cost: In this case, the portfolio holdings for the first $k$ months must be set aside as benchmarks for future portfolios. Therefore, the performance of the manager is not available during these first $k$ months (which also eliminates short-lived funds), making this choice of $k$ important. (However, in

---

11In some cases, if historical portfolio weights and future returns are readily available for all securities held by a manager, the third expansion of Equation 9 might be used. If so, the application of this approach would, again, be similar to that shown below for the $GT$ measure.
of some cases, various alternative benchmarks might be applied to recover these first $k$ months, such as using peer-group weights.) The choice of $k$ has other implications for this measure as well. If $k$ is chosen to be small, we will eliminate performance that occurs beyond the first $k$ months that a manager holds a stock, given that, at that point, the difference in weights from Equation 10 will be zero if the manager holds the stock in a constant amount during these $k$ months; if $k$ is large, it is more likely that the measure may include some systematic risk differences in the manager’s portfolio over these longer time periods.\(^{12}\)

The $GT$ measure can also be decomposed into style bets, style timing bets, market timing bets, industry bets, etc., as well as the remaining residual stock-picking bets by simple decompositions of the sum in Equation 10. For example, if we wish to decompose into the overall performance from industry bets and the remaining residual stock selectivity bets, we would decompose as

$$GT_t = \sum_{j=1}^{N} (\bar{w}_{jt,k-1} - \bar{w}_{jt,k-2})(\bar{R}_t^{\text{IND}(j)}) + \sum_{j=1}^{N} (\bar{w}_{jt,k-1} - \bar{w}_{jt,k-2})(\bar{R}_t - \bar{R}_t^{\text{IND}(j)}),$$

where $\bar{R}_t^{\text{IND}(j)}$ is the month $t$ return of the industry portfolio to which stock $j$ belongs at the beginning of the month. Equation 11 can be decomposed into further partial sums, as desired, to explore the detailed sources of manager returns. Each partial sum is then averaged over all months. Or, if desired, each partial sum can be averaged over certain months only (such as Januaries) or certain time ranges (such as three-year periods) to analyze the time-series variation in manager performance.

In this section, we have argued that the $GT$ measure holds significant advantages over standard approaches that use market indexes as benchmarks given that the $GT$ approach uses lagged manager weights to form dynamic benchmarks. However, we may wish to use our knowledge of the factors or characteristics that, from past research, are known to drive security returns to obtain a more precise measure of performance. For instance, as mentioned above, a manager can potentially game the $GT$ measure by overweighting securities with temporarily high levels of risk. To overcome such concerns, we must define the sources of risk in our security universe—in effect, the cost of increased performance evaluation precision is that we must specify the main influences on security returns for all securities in the investable universe of a manager.\(^{13}\) For some asset managers, this may not be practical—thus, the $GT$ approach may be preferred. However, in the case of U.S. domestic equities, extensive research has documented the most important influences on returns. This research has outlined the drivers of cross-sectional differences in stock returns

\(^{12}\text{When applying this measure to U.S. domestic equity mutual funds, } GT \text{ find that fund performance appears to increase when varying } k \text{ from 1 through 12 months, beyond which further increases in fund performance are small (Grinblatt & Titman 1993). Thus, they recommend a 12-month portfolio lag for these U.S. funds. A similar approach may be used to choose a reasonable value of } k \text{ in other applications—increasing the chosen } k \text{ (within reasonable ranges) until further performance changes appear to be negligible or add more noise (increase in standard deviation) than signal (increase in point estimate).}\)

\(^{13}\text{This is fairly straightforward for U.S. domestic equities and (with recent research) for non-U.S. equities in many developed countries. However, the factors that drive bond returns as well as hedge fund returns are not as clearly understood as the factors that drive equities, as described in Section 3.1; thus, it is difficult to unambiguously create benchmark portfolios for these markets. In addition, a manager holding a mixed portfolio of both stocks and bonds presents a problem when using defined benchmarks developed for the stock market, although some approaches exist.}\)
as well as time-series variation in the returns of a given stock. I next describe an approach for evaluating equity portfolios in the United States—with a discussion on extensions in non-U.S. equity markets—that uses the results of this past research.

4.2. Characteristic-Based Portfolio Evaluation

DGTW’s approach applies the results of prior empirical research on the factors that drive stock returns (Daniel et al. 1997). This research, which includes Fama & French (1992, 1993, 1996) and Jegadeesh & Titman (1993), shows that the market index as well as indexes that proxy for the size, BTM, and momentum effects are sufficient to explain the vast majority of the cross-sectional and time-series variation in U.S. stock returns. International evidence, such as Rouwenhorst (1998), indicates that similar factors also explain non-U.S. stock returns. DGTW use the CM approach (Copeland & Mayers 1982) described in Section 4.1.1, but formulate a different proxy for expected return based on this past research.

A particular advantage of the holdings-based DGTW performance measures is documented by Kothari & Warner (2001), who use simulations patterned after actual mutual funds to show that returns-based performance measures such as Carhart’s four-factor model have little ability to detect economically large magnitudes (e.g., three percent per year) of abnormal fund performance, particularly if a fund’s style characteristics differ from those of the value-weighted market portfolio. The power to detect performance of holdings-based measures applied to trades of funds (as inferred from changes in portfolio holdings over time) vastly improves the power to evaluate performance.

4.2.1. The characteristic selectivity measure. The first component of performance measures the stock-picking ability of the fund manager, controlling for the particular style used by that manager. (For instance, Grinblatt et al. 1995 find that 77% of U.S. domestic equity funds use momentum in their stock-picking strategies.) This measure of stock-picking ability, which is called the characteristic selectivity (CS) measure, is computed during quarter t as

$$C_{St} = \sum_{j=1}^{N} \tilde{w}_{jt-1} \left( \tilde{R}_{jt} - \tilde{R}_{bjt-1} \right),$$  \hspace{1cm} (12)$$

where $\tilde{w}_{jt-1}$ is the portfolio weight on stock j at the end of quarter $t-1$, $\tilde{R}_{jt}$ is the quarter t buy-and-hold return of stock j, and $\tilde{R}_{bjt-1}$ is the quarter t buy-and-hold return of a value-weighted portfolio that is matched to stock j based on its characteristics at the end of quarter $t-1$.

To construct the characteristic-matched benchmark portfolio for a given stock at the beginning of a given quarter, we characterize that stock over three dimensions—the market capitalization of equity (size), the ratio of book value of equity to market value of equity (BTM), and the prior-year return. Forming these matching portfolios proceeds as follows (this procedure is based on DGTW, and is described in more detail in that paper): First, all stocks (listed on NYSE, AMEX, or Nasdaq) having book value of equity information in Compustat, and stock return and market capitalization of equity data in the Center for Research in Security Prices stock files, are ranked, at the end of each June, by their market capitalization. Quintile portfolios are formed (using NYSE size quintile breakpoints), and each quintile portfolio is further subdivided into BTM quintiles, based on their BTM data.
as of the end of the December immediately prior to the ranking year. Finally, each of the resulting 25 fractile portfolios are further subdivided into quintiles based on the 12-month past return of stocks through the end of May of the ranking year. This three-way ranking procedure results in 125 fractile portfolios, each having a distinct combination of size, BTM, and momentum characteristics. [Thus, a stock belonging to size portfolio one, BTM portfolio one, and prior return portfolio one is a small, low BTM (growth) stock having a low prior-year return.] The three-way ranking procedure is repeated at the end of June of each year, and the 125 portfolios are reconstituted at that date. Value-weighted returns are computed for each of the 125 fractile portfolios, and the benchmark for each stock during a given quarter is the buy-and-hold return of the fractile portfolio of which that stock is a member during that quarter. Therefore, the characteristic-adjusted return for a given stock is computed as the buy-and-hold stock return minus the buy-and-hold (value-weighted) matched benchmark return during the same quarter. Finally, the CS measure of the stock portfolio of a given mutual fund during quarter \( t \), \( CS_t \), is computed as the portfolio-weighted characteristic-adjusted return of the component stocks in the portfolio, where the stock portfolio is normalized so that the weights add to one (to focus on the performance of the equity portion of the portfolio).

A caveat is in order regarding the interpretation of the CS measure, as it controls for only three characteristic dimensions of stocks—size, BTM, and past returns. Recent research has shown that mutual funds show a distinct preference for other stock characteristics that are related to average returns—for example, stocks with greater liquidity (see Chen et al. 2000).\(^{14}\) For example, one might argue that our CS measure underestimates the stock-picking talents of funds given that we do not control for the lower average returns that accrue to stocks with greater liquidity. These potential missing factors probably do not impact the practical uses of the DGTW approach, but the literature has yet to fully explore such issues.

Further components of performance are also described by DGTW, including a characteristic timing measure, \( CT_t \), (which captures fund manager ability to time styles, such as value versus growth stock premia), and an average style return measure, \( AS_t \), (which captures the return premia of the average style of stocks held by a manager). We refer the interested reader to DGTW for details.

4.2.2. Comparison of DGTW measures with factor-based regression approaches. Table 1 shows the relation between the DGTW components and the Fama-French/Carhart (FF/C) components. One should note that the dynamic style control approach of DGTW allows for several advantages over the fixed style coefficient regression approach of FF/C. The most important of these advantages are the smaller standard errors for the selectivity component, CS, relative to \( z \), and the (correspondingly) more precise measure of returns attributable to style loadings, AS, relative to the sum of the fixed coefficients times average style return premia of FF/C.

4.2.3. Extensions. As with the GT performance measure of Equation 10, the CS performance measure of Equation 12 can easily be decomposed into partial sums. This decomposition is a useful technique to determine the attribution of performance in a manager’s

\(^{14}\)See Lee & Swaminathan (2000) and Datar et al. (1998) for evidence that more liquid stocks earn lower average returns.
portfolio. For instance, Kacperczyk et al. (2005) find that mutual funds with concentrated holdings in a few industries outperform those with more diverse holdings—suggesting that manager skills are industry specific. If so, we may wish to measure the performance of stocks held by the manager within a certain industry, ignoring the other holdings of that manager. Alternatively, we may wish to measure the performance of the manager’s top 10 holdings, regardless of their sector membership. Further, Wermers (2005) shows that stocks purchased by U.S. mutual fund managers in response to strong cash inflows from investors outperform the other holdings of the managers. In all of these cases, a modification of the performance measures may be used to capture the performance of the subportfolio of interest.

4.2.3.1. Attributing subportfolio performance. To illustrate, the CS measure is modified as follows. Suppose that we are interested in measuring the performance of the subportfolio of stocks belonging to set $S$, where $S$ refers, for example, to a certain sector of the market (e.g., technology stocks). The CS measure is decomposed to measure the partial sum corresponding to stocks within set $S$. That is, the selectivity performance of the manager that is attributable to holding stocks within set $S$ is measured as

$$CSS_{S,t} = \sum_{j \in S} \bar{w}_{j,t-1} \left( \bar{R}_{j,t} - \bar{R}_{S,t-1} \right).$$

A different question may also be addressed: How skilled (in theory) would this manager be if she were constrained to buy only stocks within set $S$? A further modification that normalizes set $S$ portfolio weights to add to unity allows an answer:

$$CSS_{S,t} = \frac{\sum_{j \in S} \bar{w}_{j,t-1} \left( \bar{R}_{j,t} - \bar{R}_{S,t-1} \right)}{\sum_{j \in S} \bar{w}_{j,t-1}}.$$

4.2.3.2. Normalizing for turnover. An ideal property of a performance measure is to rank managers by the precision of their private information, or simply put, skills that cannot be captured through simple strategies that may be implemented by an uninformed observer. Unfortunately, covariance-based measures assign a higher performance measure to managers implementing higher levels of portfolio turnover, holding constant their actual skills in forecasting security returns. Such differences in turnover can arise from the agency problems inherent in asset management, given that most managers own a very small fraction of their managed portfolios. Such agency problems include compensation that is

<table>
<thead>
<tr>
<th>Table 1. Comparison of attribution by DGTW and FF/C approaches$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Selectivity</strong></td>
</tr>
<tr>
<td>DGTW Component</td>
</tr>
<tr>
<td>FF/C Component</td>
</tr>
</tbody>
</table>

$^a$Abbreviations: DGTW, Daniel, Grinblatt, Titman, and Wermers; FF/C, Fama-French/Carhart
insufficiently sensitive to performance, or the labor-market concerns of a manager (i.e., the manager’s reputational concerns or concerns over being fired). These agency problems, in theory, should not affect an investor’s performance evaluation, given that an investor can scale investments (financed through borrowing or short selling other funds) to any level to increase or decrease the return desired. However, covariance-based measures, by their very nature, are biased.\(^{15}\)

In general, it is very difficult to properly adjust for the risk aversion of a manager, as the manager may not view all investment opportunities with the same level of risk aversion. For example, a durable-goods manager may be more willing to take a large position in an automobile stock relative to a technology stock with similar prospects. Thus, adjusting the covariance measure would require the knowledge of the manager’s risk aversion with respect to all potential investments.

However, we can easily normalize portfolio-based performance measures for risk aversion with a simplifying assumption—that a manager views all stocks similarly. This assumption can be relaxed when measuring the performance of subportfolios, as described in the previous section—as such subportfolios likely have much more homogeneous risk characteristics, from the viewpoint of the manager, than securities within the overall portfolio.

To adjust for turnover, the investor can implement a turnover-adjusted version of the portfolio holdings–based measure. To illustrate, the turnover-adjusted \( GT_t \) measure during period \( t \) is computed as

\[
TAGT_t = \frac{GT_t}{\sum_{j=1}^{N} (\bar{w}_{j,t-1} - \bar{w}_{j,t-k-1})}.
\]

Similarly, the \( CS_t \) measure, adjusted for turnover, is computed as

\[
TACS_t = \frac{CS_t}{\sum_{j=1}^{N} (\bar{w}_{j,t-1} - \bar{w}_{j,t-k-1})}.
\]

### 4.2.4. The correlation between returns-based and holdings-based performance measures.

An interesting issue is the correlation, across funds, of performance at the stockholdings level (DGTW) and performance at the net returns level (FF/C). Although, at first blush, it would seem that the correlation would be near unity, the issue becomes more interesting if the expenses and transactions costs of funds are also positively correlated with their pre-expense performance. The tendency of funds with superior stock-picking skills to incur higher costs would be consistent with the equilibrium model of Grossman & Stiglitz (1980), where the (risk-adjusted expected) returns to information gathering and processing skills are equal to the costs. Analyzing both the portfolio holdings and the net returns of funds allows deeper insight into an accounting of the sources and uses of returns earned by a portfolio manager.

Table 2 investigates this issue by presenting cross-sectional correlations (across funds) between various measures of performance. These performance measures are computed over the entire life of each U.S. domestic equity mutual fund during the 1975 to 1994 period. The only restriction we place on a fund to be included in these correlations is that

\(^{15}\)It is important to note that returns-based performance measures are also plagued with the risk-aversion problem. One approach to adjusting such measures is to normalize (divide) returns-based regression alphas by the turnover of the manager.
the fund must have at least 24 valid monthly return observations (both for stockholdings and net returns) to provide a reasonable number of degrees of freedom in the regression-based measures.

Panel a presents Pearson correlations between three measures of performance at the stock portfolio level: the CS measure, the Carhart measure using the time series of excess monthly returns on the stock portfolio as the explained variable ($\alpha_\text{Gross Carhart}$), and the Jensen measure using the same explained variable ($\alpha_\text{Gross Jensen}$). In addition, two measures of performance at the net return level are included: the Carhart measure using the time series of excess monthly net returns as the explained variable ($\alpha_\text{Net Carhart}$) and the Jensen measure using the same explained variable ($\alpha_\text{Net Jensen}$). Panel b presents Spearman rank correlations between all of these performance measures.

Several observations may be drawn from the two correlation matrices. First, the crosssectional Pearson correlation (at the gross return level) between the Jensen and Carhart measures is 0.74, which indicates that adding the size, BTM, and momentum factors provides a modest increase in the precision of the performance estimate. This result indicates that mutual fund loadings on these omitted variables in the Jensen regression are correlated with the intercept of the Carhart regression. The Spearman rank correlation between these two measures is similar, 0.67.

Table 2. Mutual fund performance measure correlations, a,b

<table>
<thead>
<tr>
<th>Panel a. Pearson correlations</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_{\text{Pearson}}$</td>
</tr>
<tr>
<td>CS</td>
</tr>
<tr>
<td>$\alpha_\text{Gross Carhart}$</td>
</tr>
<tr>
<td>$\alpha_\text{Net Carhart}$</td>
</tr>
<tr>
<td>$\alpha_\text{Gross Jensen}$</td>
</tr>
<tr>
<td>$\alpha_\text{Net Jensen}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel b. Spearman rank correlations</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_{\text{Spearman}}$</td>
</tr>
<tr>
<td>CS</td>
</tr>
<tr>
<td>$\alpha_\text{Gross Carhart}$</td>
</tr>
<tr>
<td>$\alpha_\text{Net Carhart}$</td>
</tr>
<tr>
<td>$\alpha_\text{Gross Jensen}$</td>
</tr>
<tr>
<td>$\alpha_\text{Net Jensen}$</td>
</tr>
</tbody>
</table>

aThis table is replicated from Wermers (2000).
bThis table measures, for each U.S.-domiciled domestic equity mutual fund existing for at least 24 months during January 1, 1975 to December 31, 1994, correlations between different performance measures. The measures included in this table are the characteristic selectivity measure (CS), the Carhart four-factor regression alpha, and the Jensen regression alpha for the stock portfolio of each mutual fund—labeled “gross” alphas—and the Carhart four-factor regression alpha and the Jensen regression alpha for the realized net returns of the funds—labeled “net” alphas. All two-tailed $p$-values in both panels are less than 0.0001.
The correlations between the CS measure and the Carhart and Jensen gross performance measures are 0.57 and 0.58, respectively. This lower correlation supports the idea that the CS measure provides more precise adjustments for characteristic-based returns than the regression-based methods. Again, Spearman rank correlations are similar.

At the net return level, the Carhart and Jensen performance measures are again highly correlated, both with the Pearson correlation and with the nonparametric Spearman rank correlation. A comparison of these measures at the gross stockholdings return level and the net return level provides further insight. For example, the Carhart measure of the stockholdings of funds is highly correlated with the Carhart net return performance measure. The Pearson correlation is 0.62, whereas the Spearman correlation is 0.63. These high correlations between gross and net performance indicate that the level of mutual fund expenses and transactions costs, although possibly correlated with fund performance at the stockholdings level, do not eliminate the higher benchmark-adjusted net returns provided by funds with stock-picking talents.

However, as Kacperczyk et al. (2008) show, the difference between portfolio-level and net return–level performance also indicates that some managers game the infrequent disclosure of portfolios to take performance-destroying hidden actions. For instance, managers can hide badly performing trades by making a round-trip transaction between quarterly disclosures of portfolios. Further, Huang et al. (2011) find evidence that funds that game their performance measures through risk-shifting underperform other funds.

4.3. Conditional Holdings-Based Performance Measurement

The above unconditional performance measures, although straightforward to compute, may be subject to some nontrivial biases. These biases will occur if the manager of the measured portfolio uses macroeconomic information to (dynamically) sharpen her estimates of expected security returns over time. In such a case, the manager can game the unconditional performance measure by overweighting securities that, under current macroeconomic conditions, have higher-than-normal expected returns, and by underweighting securities that have lower-than-normal expected returns. Such a strategy will exhibit a positive unconditional performance measure, given that the manager’s weights will be correlated with returns over time even in the absence of any private information (i.e., not based purely on widely available macroeconomic information) about expected stock returns. Prior research by Keim & Stambaugh (1986) and Fama & French (1989) indicates that instruments that are useful in predicting future U.S. market-wide equity returns include the (a) aggregate dividend yield, (b) default spread, (c) term spread, and (d) yield on the three-month T-bill. Further, Ferson & Harvey (1999) find that these variables are relevant in explaining the cross section of U.S. stock returns.

For instance, suppose that a particular lagged macroeconomic factor (e.g., the T-bill yield) is positively correlated with the next-period return in one sector of the stock market (durable goods stocks), while inversely correlated with the return in another sector (utility stocks). Specifically, assume that the unconditional expected return on each sector is 12%

---

16At the end of each month, the dividend yield is measured as the aggregate cash dividends on the value-weighted Center for Research in Security Prices index over the previous 12 months divided by the current level of the index, the default spread is the yield differential between bonds rated BAA by Moodys and bonds rated AAA, and the term spread is the yield differential between Treasury bonds with more than 10 years to maturity and T-bills that mature in three months.
per year, while the conditional expected return for durables is 10% and 14% following low and high T-bill yield realizations (recessions and expansions), respectively, and the reverse is true for nondurables. A manager who uses the T-bill yield level to vary portfolio weights between 30% and 70% for each sector (averaging a 50% allocation for each) will appear to be informed, according to the unconditional $GT$ measure described in the last section. Specifically, during a transition from a year (quarters 1 through 4) following a low T-bill yield (a recession) to a year (quarters 5 through 8) following a high bill yield (an expansion), the manager’s one-year (industry component) $GT$ measure will be measured as

$$GT_t = \sum_{j=1}^{2} (\tilde{\mu}_{t, t-1} - \tilde{\mu}_{t, t-k-1}) R_{t}^{IND(j)}$$

$$= (\bar{\mu}_{durable, 4} - \bar{\mu}_{durable, 0}) \tilde{R}_{5-8} + (\bar{\mu}_{nondurable, 4} - \bar{\mu}_{nondurable, 0}) \tilde{R}_{5-8}$$

$$= (0.7 - 0.3)(0.14) + (0.3 - 0.7)(0.10) = 0.056 - 0.040 = 0.016 = 1.6\%,$$

where $k = 4$ quarters lag between current and historical (benchmark) portfolios. Thus, the manager will exhibit an outperformance of 160 basis points during the year, based simply on using publicly available information on macroeconomic indicators to shift portfolio weights. Over a large number of periods, this performance measure, in the absence of any private information–based skills possessed by the manager, will converge to the covariance between weights and returns, which is positive by assumption,

$$GT = \sum_{j=1}^{2} cov(\tilde{\mu}_{t, t-1}, \tilde{R}_{t,j}).$$

To further illustrate, suppose that, in the above example, the probability of expansion or recession are each 50% from the point of view of an uninformed observer, but that macroeconomic information gives the manager perfect foresight about whether an expansion or recession will occur during the following year. In this case, the unconditional expected weight, $E[\tilde{\mu}_{t, t-1}]$, equals 0.5 for each industry, and the $GT$ measure will converge to

$$GT = \sum_{j=1}^{2} cov(\tilde{\mu}_{t, t-1}, \tilde{R}_{t,j}) = \sum_{j=1}^{2} E[(\tilde{\mu}_{t, t-1} - E[\tilde{\mu}_{t, t-1}]) \tilde{R}_{t,j}]$$

$$= (2)(0.5)(0.7 - 0.5)(0.14) + 0.5(0.3 - 0.5)(0.10]$$

$$= 0.008 = 0.8\%$$

per year.

Note that this is half the level of the performance measure in the prior example above, given that a recession year is equally likely (by assumption) to be followed by another recession year as it is to be followed by an expansion year. If a recession is followed by a recession, the $GT$ measure will assign zero performance given that portfolio weights would not shift from one year to the next for the perfect foresight manager.

If we wish to control for abnormal returns derived from manager weight changes that are correlated with shifts in expected returns (predictable using macroeconomic

---

17Note that this is half the level of the performance measure in the prior example above, given that a recession year is equally likely (by assumption) to be followed by another recession year as it is to be followed by an expansion year. If a recession is followed by a recession, the $GT$ measure will assign zero performance given that portfolio weights would not shift from one year to the next for the perfect foresight manager.
variable information), then we need to adjust our unconditional measures to take account of such predictability-based returns. We can then make the decision of whether we should reward the manager for such public information–based returns; the key is to separate abnormal returns based on public information from those derived from private information. Such private information–based returns are unambiguously interpreted as manager skill.

4.3.1. The Ferson-Khang conditional portfolio-holdings approach. In response to the need to separate public- and private-based abnormal returns, Ferson and Khang (FK) develop a conditional portfolio holdings–based performance measurement methodology (Ferson & Khang 2002). FK show that the relation between unconditional and conditional weight-based covariance measures of performance is

\[
\sum_{j=1}^{N} \text{cov}(\tilde{w}_{j,t-1} - E[\tilde{w}_{j,t-1}], \tilde{R}_{j,t}) = \sum_{j=1}^{N} E\left[\text{cov}(\tilde{w}_{j,t-1} - E[\tilde{w}_{j,t-1}], \tilde{R}_{j,t}|Z_t)\right] \quad \text{UWM}
\]

\[
+ \sum_{j=1}^{N} E\left[\text{cov}(E[\tilde{w}_{j,t-1} - E[\tilde{w}_{j,t-1}]|Z], E[\tilde{R}_{j,t}|Z]|Z_t)\right] \quad \text{CWM}
\]

\[
+ \sum_{j=1}^{N} \text{cov}(E[\tilde{w}_{j,t-1} - E[\tilde{w}_{j,t-1}]|Z], E[\tilde{R}_{j,t}|Z]|Z_t) \quad \text{RESID}
\]

Here, UWM is the (summed) covariance-based performance measure of Grinblatt & Titman (1993). The right side of Equation 15 decomposes the GT measure of performance into two components. First, CWM is the performance of the manager, in excess of that attributable to using public information (Z) to choose holdings. The second, RESID, is performance that is due solely to using public information to predict future returns of securities, and varying portfolio weights accordingly. The quality of the manager’s private information (skills) is reflected in CWM, whereas purely mechanical trading strategies are reflected in RESID. Although the active portfolio manager should clearly be compensated for producing a positive CWM, similar rewards for producing RESID depend on one’s view toward returns that are predictable based on public information variables. An investor should judge the cost of producing such returns himself, and compare these costs to the fees and other costs charged by the manager who produces returns derived from macroeconomic information that is publicly available.

The CWM is the time-series average of its component at time t, which, in turn, is written as

\[
\text{CWM}_t = E\left[\sum_{j=1}^{N} (\tilde{w}_{n,t-1} - E[\tilde{w}_{j,t-1}]) (\tilde{R}_{j,t} - E[\tilde{R}_{j,t}|Z_{t-1}]) \right]|Z_{t-1}
\]

An analogous expression captures the time t UWM component,

\[
\text{UWM}_t = E\left[\sum_{j=1}^{N} (\tilde{w}_{j,t-1} - E[\tilde{w}_{j,t-1}]) (\tilde{R}_{j,t} - E[\tilde{R}_{j,t}]) \right].
\]

18RESID might be considered as returns on simple strategies, as referred to in Section 2.2.
Note that Equation 16 involves working with the conditional expected portfolio weight for each security $j$: $E[\tilde{\omega}_{j,t-1}|Z_{t-1}]$. In the spirit of Grinblatt & Titman (1993), FK suggest that a good choice of a proxy for these conditional expected weights are the buy-and-hold portfolio weights using the portfolio held $k$ periods ago, and held until the end of period $t-1$, that is,

$$\text{proxy for } E[\tilde{\omega}_{j,t-1}|Z_{t-1}] \equiv \tilde{\omega}_{j,t-k} = \tilde{\omega}_{j,t-k-1} \prod_{t=k}^{t-1} \left( \frac{1 + \tilde{R}_{j,t}}{1 + R_{p,t}} \right).$$

However, any reasonable proxy for the expected weights in the presence of conditioning variable information may be used—such as the security weights of a conditionally efficient benchmark portfolio according to some asset pricing model (e.g., the optimal portfolio of stocks and bonds under the conditional CAPM). The key to choosing a good proxy is that it should reflect the advantage of using current macroeconomic information to improve the portfolio allocation, but not any private information contained in the manager’s current (or very recent) weights.\(^{19}\)

5. RECENT MUTUAL FUND RESULTS

We now turn to recent empirical findings for mutual funds. First, we cover returns-based results, followed by holdings-based results.

5.1. Returns-Based Analysis

Several recent advances have been made in measuring the performance of active mutual fund managers using only their reported returns. Busse (1999) shows that analyzing daily mutual fund returns can uncover manager ability not evident in monthly returns. Specifically, Busse reveals that fund managers have the ability to time market volatility to achieve superior returns.

Avramov & Wermers (2006) show how the Bayesian approach discussed in Section 3.6 can be combined with the conditional models discussed in Section 3.7 to select individual mutual funds. They show that this approach can be used to incorporate predictability in (a) manager skills, (b) fund risk loadings, and (c) benchmark returns. When they apply their model to U.S. domestic equity funds from 1975 to 2002, they find that predictability in manager skills (alpha) is the dominant source of investment profitability. For instance, long-only fund selection strategies that incorporate such predictability produce a Carhart alpha of 2% to 4% per annum through timing industries over the business cycle (through investments in funds), and by an additional 3% to 6% per annum by choosing funds that outperform their industry benchmarks.

Some recent and innovative papers have applied the SDF approach to study the performance of active mutual funds. DS study 24 Swedish equity mutual funds from July 1991 to December 1995 using the SDF approach of Equation 7, and find some weak evidence of positive performance for some funds (Dahlquist & Soderlind 1999). Ferson et al. (2006) use a novel approach of deriving SDFs from continuous-time fixed-income pricing models

---

\(^{19}\)FK show that a simple procedure may be used to estimate CWM and UWM. We refer the interested reader to Ferson & Khang (2002) for further details.
to test for active government bond fund manager skills. They find little evidence of the active funds to beat the passive benchmarks implied by the SDF pricing factor.

Banegas et al. (2010) extend the conditional returns-based model described in Section 3.7 to study the performance of European active mutual fund managers from 1988 to 2008 using a framework similar to Avramov & Wermers (2006). This analysis is interesting, in light of the partial segmentation in European country stock markets, which decreases over the time period of their study. The contribution of their paper is that it demonstrates that macroeconomic information is useful in identifying business conditions under which local country active managers are more likely to outperform. For example, during the recent financial crisis, we might expect that active UK asset managers would be valuable because of their ties to London financial institutions, in the face of large asymmetric information on the value of banking stocks. Interestingly, Banegas et al. find that a similar set of macroeconomic variables is useful for predicting European active manager skills, relative to the U.S. study of Avramov & Wermers—short government rates, term spreads, credit spreads, and dividend yields—but, they also find that consumer confidence and producer inventories are also useful. They find that a number of investment strategies (that use macroeconomic variables to predict fund returns) generate alphas from 4%–8%/year (after fund-level trading costs and fees, except for load fees), when measured with a single-factor model, and from 7%–12%/year with a four-factor model that controls for fund exposures to size, BTM, and momentum (Carhart 1997).

5.2. Holdings-Based Analysis

Holdings-based performance models have also been applied outside the United States. Specifically, researchers have applied the DGTW portfolio-benchmarking technique to equity portfolios outside the United States by forming the characteristic-based benchmarks in the appropriate country stock market. For example, Wylie (2005) follows DGTW (Daniel et al. 1997) closely in forming characteristic-based benchmarks for the UK market.

The procedure is as follows. At the end of June of each year, all London Stock Exchange listed stocks are partitioned into 125 benchmark portfolios by repeated division into quintiles on the basis of market capitalization at the end of June, then BTM ratio during the most recent fiscal year-end, then prior 12-month return (ending in May rather than June to avoid the one-month return reversal described in Jegadeesh 1990). Then, the market capitalization–weighted average return of each benchmark portfolio is calculated for each of the 12 months following June to compute benchmark returns for each month of that year. Wylie (2005) finds that stocks that UK fund managers herd into (buy as a group) underperform those they herd out of by 0.8% during the following year.

Iihara et al. (2004) form characteristic-based portfolios from stocks listed on the Tokyo Stock Exchange (TSE) to study momentum in Japanese stocks. At the end of each June from 1975 to 1997, all TSE stocks are sorted into five equal groups from small to large based upon their market capitalization. They independently sort TSE stocks into five equal BTM groups, where BTM is equal to the ratio of book-value to market-value of equity at the end of June for each year. Finally, 25 benchmark portfolios are created from the

---

20Book value is defined as ordinary share capital plus reserves plus deferred and future taxation. Stocks are excluded from the portfolio formation unless they have market capitalization data for June of the current year and December of the preceding year, book value data for the previous two years, and at least six monthly returns in the year preceding May 31, and unless they are ordinary shares.
intersection of the five size and five BTM groups. Monthly equal-weighted returns for each of these 25 portfolios are calculated from July of year $t$ to June of year $t + 1$.

Other papers that form characteristic-based benchmarks, in the spirit of the DGTW approach, include Pinnock (2003), who forms benchmarks in Australia. In addition, Jern (2002) forms benchmarks for Finnish fund managers, while Nitibhon et al. (2005) form benchmarks for Thai equity fund managers. Given the popularity of the DGTW approach, it is likely that further developments in other markets will make these benchmarks more widely available to both academics and practitioners. In addition, once the risk factors in fixed-income markets become widely accepted, it is likely that a similar approach will be feasible for measuring the performance of fixed-income asset managers.

### 5.3. Covariates of Skilled Managers

Given the prior-noted difficulty in extracting fund manager skill from returns data, it is important to note that we can add other variables that have been found to be correlated with fund outperformance. Here, we briefly cover some of the more interesting recent findings, although there are many more studies that are also promising.

Cremers & Petajisto (2009) show that fund managers taking larger bets away from their benchmarks outperform benchmark huggers, on average. Active Share is defined as the absolute difference between a stock’s weight in the portfolio and its weight in the best fit benchmark, cumulated across all stocks in the portfolio and benchmark. They find that funds with the highest aggregate active share outperform those with the lowest active share by roughly 250 basis points per year. They attribute this to greater conviction on the part of the manager and conclude that the most active stock pickers create value for investors, whereas funds that hug the benchmark tend to destroy value.

Other manager characteristics that can help predict outperformance include social connections, academic background, and co-investment:

1. Cohen et al. (2008) find that managers take larger positions and perform better in equities of companies where they have social connections (i.e., where the officers or board members attended the same college as the manager). Cohen et al. conjecture that this is because connected managers have better access to private information, or are better able to assess the quality of the company’s management team.

2. Chevalier & Ellison (1999) find that managers who graduate from colleges where students have higher average SAT scores also tend to outperform, presumably because they are better qualified and, thus, better able to analyze information.

3. Gottesman & Morey (2006) find that the quality of a manager’s MBA program is positively correlated with future performance. They measure the quality of an MBA program using both the average GMAT score of students in the program and the annual *Businessweek* rankings. Interestingly, they find no relation between performance and other graduate degrees (including a PhD degree) or the CFA designation.

4. Ding & Wermers (2009) find that more experienced managers outperform when they manage large funds, but underperform when they manage small funds, implying that skilled managers are promoted, but unskilled managers are not always replaced.

5. Grinblatt et al. (2011) find that high-IQ investors exhibit superior investment performance.
Other studies have focused on the characteristics of the fund management company. Not surprisingly, funds sponsored by larger fund management companies tend to perform better, probably due to (a) economies of scale and scope (lower costs/fees), (b) greater resources for gathering and analyzing information, and (c) better technologies for executing trades with less price impact. In addition, fund companies with a greater number of independent directors also tend to perform better (Ding & Wermers 2009), possibly because they are more demanding of their managers.

Massa & Zhang (2009) find that funds managed by companies with a flatter organizational structure outperform funds managed by companies with a more hierarchical structure. Such “flat-structure” funds also tend to be more concentrated and exhibit less herding behavior. This may be because a hierarchical structure discourages managers from innovating, taking risks, and collecting private information—that is, in vertical organizations, managers may not feel as much direct responsibility (ownership) for their funds.

6. RECENT HEDGE FUND RESULTS

The size of the hedge fund industry has doubled almost every two years from 1994 to 2008, until it peaked at more than 11,000 active funds managing more than $2.5 trillion in July 2008 (Titman & Tiu 2011). With leverage, the assets controlled by hedge funds are many times larger. No longer can we ignore this industry as the playground of the ultra-wealthy, and academics have taken this advice to heart; many studies of hedge fund performance have been conducted over the past 15 years.

Unfortunately, hedge fund data present a number of problems for researchers. The chief problem is that hedge funds are generally not required to disclose anything to noninvestors, and can report only very basic information to investors. Fortunately, hedge funds want to market themselves to gain more clients, so they voluntarily provide data to vendors, such as HFR or Lipper/Tass. However, one can easily predict the problem with relying on such self-reported data; we will get data from funds that are likely more successful, or from funds wishing to grow (which may be less successful). Further, as shown in Figure 2 for December 2005 (from Titman & Tiu 2011), database vendors generally obtain data from differing sets of hedge funds, making the interpretation of studies using different datasets very difficult to compare. Nevertheless, several interesting and careful studies stand out.

6.1. Returns-Based Analysis

Fung et al. (2008) merge the HFR, CISDM (the University of Massachusetts database), and Lipper TASS databases to conduct an investigation into the performance, risk, and capital formation of fund of funds from 1995 to 2004 using the seven-factor model of
Fung & Hsieh (2004). Although the average fund of funds delivers alpha only in the period between October 1998 and March 2000, a subset of funds of funds consistently delivers alpha. The alpha-producing funds are not as likely to liquidate as those that do not deliver alpha, and experience far greater and steadier capital inflows than their less fortunate counterparts. These capital inflows attenuate the ability of the alpha producers to continue to deliver alpha in the future. (In a related paper, Getmansky et al. 2011 find that flows outperform only when they are not impeded by flow restrictions, such as subscription periods.)

Jagannathan et al. (2010) examine the HFR database, correcting for funds that stop reporting to HFR with a statistical model. (Liang 2000 finds that poor performance is the main reason for fund disappearance from the TASS and HFR databases.) Hedge fund outperformance, relative to other same-category hedge funds, exhibits performance persistence that lasts for three years after a three-year ranking period. Jagannathan et al.
speculate that relative performance measures are better predictors than absolute measures due to relative measures neutralizing unknown common risk exposures of the managers within a given category.

Kosowski et al. (2007) extend the bootstrap methodology of KTWW (Kosowski et al. 2006) into the hedge fund universe, using the seven-factor Fung & Hsieh (2004) model in a Bayesian setting. It is particularly important to account for non-normal returns with hedge funds, given that their dynamic strategies and security choices make their portfolios distinctly non-normal. Kosowski et al. evaluate a sample of hedge funds from a merger of the TASS, HFR, and CISDM hedge fund databases: a total of 4,300 live and 1,233 dead hedge funds from 1990 to 2002. Top decile hedge funds, ranked by a two-year lagged Bayesian posterior alpha, outperform bottom decile funds by 5.8% during the following year, whereas the deciles do not differ with a simple ordinary least-squares ranking.

Avramov et al. (2009) extend the predictability models of Avramov & Wermers (2006) into the hedge fund universe, and find a substantial ability to predict hedge fund outperformance. Specifically, Avramov et al. find that several macroeconomic variables, including VIX and credit spreads, allow long-only strategies to outperform their Fung & Hsieh (2004) benchmarks by more than 17% per year.

6.2. Holdings-Based Analysis

Although most hedge funds need not disclose anything about their portfolio positions, those hedge funds holding at least $100 million in U.S. equities must file quarterly equity position reports with the SEC, with a maximum 45-day delay after the quarter-end. Griffin and Xu (GX) use these SEC 13f filings to study the performance of hedge funds (Griffin & Xu 2009). This is a promising line of research, although we have to remember that 13f filings are aggregated to the level of the management company. Although this may be fine for a stand-alone hedge fund managed by Jeff Vinik, it biases against examining the hedge funds of conglomerate asset managers, such as Goldman Sachs Asset Management. Thus, if large conglomerates have differing performance, this limited portfolio analysis can be problematic.

Nevertheless, GX have some interesting findings. Following the approach used with mutual funds by Chen et al. (2000), GX find that the aggregate equity holdings and trades of hedge funds do not forecast stock returns. They do, however, find that large positions taken by hedge funds, in aggregate, appear to forecast (weakly) stock returns for the next quarter, and negatively predict returns at longer horizons. Further, GX use the DGTW approach outlined in Section 4.2 to study the performance of individual hedge funds. Over the entire 1986–2004 period, they find that hedge funds held stocks that outperformed by 2.15%/year, compared to only 0.82% for mutual funds. However, most of the performance was generated during 1999 and 2000, and the differences between hedge funds and mutual funds disappear during other periods.

6.3. Covariates of Skilled Managers

With the limited transparency of hedge funds, the literature on characteristics of skilled hedge funds is much sparser than its counterpart in the mutual fund space. However, there are a few noteworthy studies. For instance, De Souza & Gokcan (2003) find that managers who invest their own capital in their funds are more likely to outperform, possibly because
such managers have greater conviction, or because they are more likely to avoid uncompensated risks.

TT study the breakdown of hedge fund risks into systematic versus idiosyncratic risks (Titman & Tiu 2011). Simply put, how much of the hedge fund return comes from a beta strategy and how much from a skill-based alpha strategy? TT, using the seven-factor Fung & Hsieh (2004) model augmented with leads and lags to account for potential return smoothing documented by GLM (Getmansky et al. 2004) and Bollen & Pool (2009), measure the model R-squared of individual hedge funds. TT combine six different databases to minimize selection bias over the 1994–2005 period, across all categories of hedge funds. They find that funds with lower R-squareds have better future performance; specifically, a 10% decline in the R-squared is associated with a 0.0787 increase in the annual Sharpe ratio during the following year, an increase of 2.37% in the annual alpha, and a 0.0945 increase in the information ratio. Interestingly, funds with longer lockups appear to outperform.

7. INSTITUTIONALLY SPONSORED PORTFOLIOS

The least studied of the three conventional asset management universes are those pools of money that are privately managed for pension funds, endowments, trusts, insurance companies, etc. Although the assets managed in this sector are roughly equal to those in the entire mutual fund universe, it is difficult to study funds that have limited public disclosure requirements. However, a few studies have managed to obtain the private data to investigate the performance of institutional funds.

Blake & Timmermann (2005) study the international market-timing ability of UK pension funds using a conditional model, where the conditioning variables include the default premium on U.S. bonds computed as the differential yield on Baa and Aaa-rated corporate bonds, the one-month U.S. T-bill rate, the U.S. to UK T-bill spread, and the local dividend yield in each region. They find that, controlling for dynamic strategies that use such predictive variables, there is no evidence of what Graham & Harvey (1996) refer to as “extra market-timing ability,” that is, anticipating return movements beyond those which could have been predicted using public information.


Finally, Blake et al. (2011) study a dataset of UK pension funds and fund managers, obtained from BNY Mellon Asset Servicing (formerly Russell-Mellon-CAPS—commonly known as CAPS), consisting of quarterly returns on the investment portfolios of 2,385 UK pension funds that had their performance monitored by CAPS at some stage between March 1984 and March 2004. The focus of the Blake et al. study is on the secular trend toward decentralization that has occurred among pension mandates over the past 20 years: Sponsors have moved from a single balanced manager across all asset classes to multiple specialist managers within each asset class. Such decentralization is found to be useful in terms of improved performance, due to the superior skills of the specialists, but some of the improved performance is lost through higher fees; also, poorer diversification results, due to the problem in coordinating the portfolio choices of separate active managers.
8. FUTURE RESEARCH DIRECTIONS

A few select research directions seem most promising at this point. First, the macroeconomic predictive model of Avramov & Wermers (2006) has shown great promise in the hedge fund universe, and it would be interesting to see future research on other asset classes, such as bond funds, private equity funds, and emerging market funds. Second, one glaring omission from the literature is an estimate of the trading costs of individual funds. Although Wermers (2000) provides a rough estimate of such costs, new data sources, such as the Abel Noser trade data of Ancerno, allows researchers to go far further in analyzing the daily trades of funds—their performance as well as their trading costs. Third, securities lending is a sideline business that is profitable for many managed funds. It would be of great interest to study the interface between the holdings of managed funds, securities lending activity, and the effects of short selling on markets.

Fourth, further research advances should be made for non-U.S. asset managers; some has been completed for non-U.S. mutual funds, but much more is necessary. Finally, the recently passed Dodd-Frank Act might provide further data for researchers, which could benefit the public in many ways with likely few harmful effects if disclosed properly. For instance, if hedge funds are required to disclose portfolio holdings periodically, with a reasonable time lag, researchers could better study the performance-generating abilities and operational risks of these funds and provide information providers, such as Lipper or Morningstar, with fresh research results. One of the externalities of such free academic research could be the quicker demise of unskilled managers, not to mention fund managers engaging in fraudulent activities, such as Ponzi schemes.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

I am grateful to Mark Grinblatt, Sheridan Titman, and the editor Rob Stambaugh for invaluable comments on an earlier draft of this review.

LITERATURE CITED

Avramov D, Wermers R. 2006. Investing in mutual funds when returns are predictable. J. Financ. Econ. 8:339–77


Roll R. 1978. Ambiguity when performance is measured by the securities line. *J. Finance* 33:1051–69


www.annualreviews.org • Performance Measurement of Active Funds 573


NEW FROM ANNUAL REVIEWS:

EconScholar App
Available for iOS and Android

Economics scholars can now access highly cited, mobile-optimized review articles from a variety of mobile devices. The EconScholar app, from Annual Reviews, allows immediate access to full-text review articles for users with personal or institutional subscriptions to the Annual Review of Economics, the Annual Review of Financial Economics, and the Annual Review of Resource Economics.

Also, non-subscribers and new users can access selected complimentary articles and all abstracts, and discover firsthand the breadth and quality of these review articles.

The app allows users to:

- read and cache full-text articles on a mobile device
- view high-resolution images and video
- bookmark articles (saving full-text indefinitely)
- search journal content
- read and share content through social media tools

Subscribers can either enter their personal login information or connect via institutional access to view full-text content.

To download the free EconScholar app, please visit the Apple AppStore or GooglePlay store.

For more information visit: http://www.annualreviews.org/r/econscholarpdf
## Contents

My Life in Finance  
*Eugene F. Fama*  .............................................. 1

Banking Crises: A Review  
*Luc Laeven*  ................................................ 17

The Consequences of Financial Innovation: A Counterfactual Research Agenda  
*Josh Lerner and Peter Tufano*  .................................. 41

Rediscovering the Macroeconomic Roots of Financial Stability Policy: Journey, Challenges, and a Way Forward  
*Claudio Borio*  .............................................. 87

Money Markets  
*Marvin Goodfriend*  ......................................... 119

Inflation-Indexed Bonds and the Expectations Hypothesis  
*Carolin E. Pflueger and Luis M. Viceira*  ......................... 139

The Economics of Mutual Funds  
*David K. Musto*  ........................................... 159

The Origins and Evolution of the Market for Mortgage-Backed Securities  
*John J. McConnell and Stephen A. Buser*  ......................... 173

Valuation and Risk Management of Collateralized Debt Obligations and Related Securities  
*Christian Bluhm and Christoph Wagner*  ........................ 193

Government Policy and the Fixed-Rate Mortgage  
*Michael Lea and Anthony B. Sanders*  ........................... 223

The Economics of Credit Default Swaps  
*Robert A. Jarrow*  .......................................... 235

Payment Systems  
*James McAndrews, Ed Nosal, and Guillaume Rocheteau*  ............ 259
Financial Intermediary Balance Sheet Management

Tobias Adrian and Hyun Song Shin ........................... 289

A Review of Empirical Capital Structure Research and Directions for the Future

John R. Graham and Mark T. Leary ........................... 309

Equilibrium in the Initial Public Offerings Market

Jay R. Ritter .................................................. 347

Finance and Governance in Developing Economies

Randall Morck .................................................. 375

Microfinance and Social Investment

Jonathan Conning and Jonathan Morduch ........................... 407

Global Asset Pricing

Karen K. Lewis .................................................. 435

Predictability of Returns and Cash Flows

Ralph S.J. Koijen and Stijn Van Nieuwerburgh ........................... 467

Momentum

Narasimhan Jegadeesh and Sheridan Titman ........................... 493

Carry Trade and Momentum in Currency Markets

Craig Burnside, Martin Eichenbaum, and Sergio Rebelo ........................... 511

Performance Measurement of Mutual Funds, Hedge Funds, and Institutional Accounts

Russ Wermers .................................................. 537

Errata

An online log of corrections to Annual Review of Financial Economics articles may be found at http://financial.annualreviews.org