Review of Finance (2007) 11: 359–400
doi: 10.1093/rof/rfm018
Advance Access publication: 25 August 2007

Improved Forecasting of Mutual Fund Alphas and Betas*

HARRY MAMAYSKY[1], MATTHEW SPIEGEL[2] and HONG ZHANG[3]
[1] *Old Lane LP;* [2] *Yale School of Management;* [3] *INSEAD*

Abstract. This paper proposes a simple back testing procedure that is shown to dramatically improve a panel data model's ability to produce out of sample forecasts. Here the procedure is used to forecast mutual fund alphas. Using monthly data with an OLS model it has been difficult to consistently predict which portfolio managers will produce above market returns for their investors. This paper provides empirical evidence that sorting on the estimated alphas populates the top and bottom deciles not with the best and worst funds, but with those having the greatest estimation error. This problem can be attenuated by back testing the statistical model fund by fund. The back test used here requires a statistical model to exhibit some past predictive success for a particular fund before it is allowed to make predictions about that fund in the current period. Another estimation problem concerns the use of a single statistical model for all available mutual funds. Since no one statistical model is likely to fit every fund, the result is a great deal of misspecification error. This paper shows that the combined use of an OLS and Kalman filter model increases the number of funds with predictable out of sample alphas by about 60%. Overall, a strategy that uses very modest ex-ante filters to eliminate funds whose parameters likely derive primarily from estimation error produces an out of sample risk-adjusted return of over 4% per annum.

JEL Classification: G12, G13

Over the last 20 years the mutual fund industry has grown at an incredible rate, and this has naturally attracted a lot of attention from the academic and financial community. One area of particular interest has been whether or not it is possible to identify fund managers with skills that investors can capitalize

* We thank David Musto whose critique of an earlier paper lead to the creation of the eight-factor model used in this one. Additional thanks go to Jonathan Berk, Mark Carhart, Joshua Coval, Wayne Ferson, William Goetzmann, Peter Starr, Peter Bossaerts (the editor), and three referees for their comments. Finally, we thank seminar participants at INSEAD, Rutgers, the University of Michigan at Ann Arbor, the University of Calgary, and the University of Alberta and conference participants at the 2005 Winter Finance Summit, the 2005 Meetings of the Western Finance Association, and the 2005 Meetings of the European Finance Association.

on. The approach taken in this literature has been to apply a single statistical model to every available fund. But at any one time funds exhibit substantial differences in their strategies and holdings (Brown and Goetzmann (1997)). This makes it likely that any one statistical model will be incorrectly specified for at least some funds in the data pool. Also, as noted by Timmermann and Granger (2004), behavioral changes over time can cause a model that fits a subject in one time interval to fail in another. As a result many of the estimated parameter values used to forecast fund returns (especially those in the extreme deciles) may reflect misspecification error rather than reality. This paper proposes a very simple back testing procedure to help alleviate this problem. The results indicate that with back testing even simple ordinary least squares (OLS) models, which have previously been found to exhibit little predictive power, produce useful forecasts for large subsets of funds. Back testing is also shown to generate substantial improvements for the nonlinear models tested here. Overall, back testing, along with a combination of models, can produce portfolios encompassing over 15% of the mutual fund population that yield economically and statistically significant predictable above market performance in any one period.

In general, the problem with using a single model for every fund in a time series panel database is that sorting on the estimated alphas (or any other attribute) may simultaneously sort on misspecification errors. If this happens, then models may not select funds with predictable superior performance as "best" but rather those with the poorest parameter estimates, as these will tend to be the most extreme.[1] One possible way to help identify possible misspecification errors, and the one pursued here, is the following algorithm: prior to using a model to forecast a particular subject's performance in the current period it must first generate acceptable out of sample forecasts in the recent past. This helps to avoid using the model's in sample attributes to determine if it will do well out of sample. Intuitively, if one wishes to use a model to find funds that will produce future above (or below) market returns it seems natural to require at least some past success in this regard. If an extreme high alpha fund underperforms in the recent past, one might strongly suspect that there may be some problem with the model's forecasting ability, at least for the moment. The same should be true for an extreme low alpha fund that overperforms. Consequently, the back test proposed here requires a model to correctly predict the sign of a fund's excess return in the previous month before

[1] This argument is similar to the size anomaly critique found in Berk (1995). There he argues that misestimated betas will lead to the appearance that small firms outperform large ones on a risk-adjusted basis. Roughly, all else equal, higher discount rates lead to smaller market capitalizations. To the degree that a model misestimates beta, the firm's market capitalization will then proxy for the true cash flow risk.

allowing it to make predictions about the fund's future behavior. Applying even this simple criterion yields a dramatic improvement of the risk-adjusted return for top mutual fund deciles selected by either the one or Carhart's four–factor OLS model. Monthly returns jump from initial values of −8 basis points (bps) and 18 bps to the economically and statistically significant values of 21 bps and 37 bps, respectively.

Beyond that, this paper also shows the benefits of simultaneously using multiple models. Back testing implies that any one model will not be used to generate forecasts for every single subject. Thus, by using multiple models the set of funds for which one might produce useful forecasts potentially widens. To illustrate the point a dynamic Kalman filter model is tested along with the standard rolling OLS model.[2] These models provide a good pairing as they offer very different costs and benefits and (with the help of back testing to clean out misspecification errors) capture different types of managerial skills. Rolling OLS models are simple to estimate. But they require that a portfolio's parameters drift slowly over time, if at all. If a fund actively trades securities during the (typically) estimated 5 year window, the resulting parameter estimates may not accurately represent the current situation.[3] A Kalman filter model can potentially adapt itself to such changes and avoid this problem.[4]

Similar to the OLS case, with back testing the Kalman filter models successfully select funds with out of sample risk-adjusted abnormal returns in excess of 3.5% per annum. More importantly, each model selects a relatively unique set of funds with an overlap of only about a third. This implies that it is possible to find a remarkably wide variety of funds with positive (or negative) predictable risk-adjusted returns if one is willing to employ a variety of models. Hence, instead of running a horse race among different models and picking a "winner," this paper demonstrates the benefit of simultaneously using more than one model. By going from one model to two, the set of funds with predictable super normal returns increases by about 50%. Meanwhile,

[2] See Mamaysky et al. (2007) for a detailed derivation of the Kalman model based on dynamic selection ability.

[3] One solution can be found in Grinblatt and Titman (1994). The methodology they use avoids a direct comparison against a specific portfolio, and instead uses an "endogenous" benchmark. However, their technique requires knowledge of the fund's actual composition, which may not always be available. Ferson and Khang (2002) extend the technique to condition the portfolio betas on exogenous variables such as macro economic data.

[4] Grinblatt and Titman (1989) also propose a technique that can detect market timing abilities that arise from a fund's dynamic asset allocation strategy, and implement it in their 1994 paper. However, as Ferson and Schadt (1996) point out, correlations between factor loadings and market returns may also be due to predictable changes in time varying expected returns, and thus implement a technique for handling this case.

funds jointly selected by the OLS and Kalman models within the top decile, presumably the funds run by managers with more than one type of managerial talent, can have risk-adjusted returns as high as 6.0% per annum.

Whether or not statistical models can identify fund managers who will produce positive risk-adjusted returns for their investors has implications regarding the general functioning of markets. If such skills do not exist then this calls into question the value of fundamental analysis and active management, at least for mutual fund investors. On the other hand, if such skills do exist, but are somehow bid away when discovered, then this offers support for the Berk and Green (2004) hypothesis regarding fund returns in an efficient competitive environment.[5] Overall, this paper's findings offer support for at least part of their thesis: managerial skill exists but its benefit to mutual fund investors is short lived.

Carhart's (1997) paper documents how momentum returns can make it appear that some managers can, at least temporarily, appear to produce positive alphas going forward even when they cannot. In order to get around his findings, performance studies have since tried using more comprehensive data sets and/or improved methodologies. One approach has been to use the underlying holdings data, as in Chen et al. (2000), Cohen et al. (2005), and Kacperczyk et al. (2005a), and Baker et al. (2004). Another has been to use Bayesian models as in Avramov and Wermers (2005) and Busse and Irvine (2005), or daily data as in Bollen and Busse (2004). Recently, Kosowski et al. (2007) looked at whether it may be possible to detect the existence of funds that have outperformed the market by dropping the assumption of normality associated with classic t tests and instead using boot strapped standard errors. They find that boot strapped standard errors indicate that some managers are able to produce positive alphas. However, out of sample their results are similar to Carhart (1997). Top deciles funds yield alphas of 0.08% per month. With the boot strapped standard errors this is statistically significant while under the standard t test it is not. This paper can be seen as both complementing and extending this literature. The simple back test proposed here allows even the one factor OLS model to reliably identify funds that will produce above market returns of economic significance using classic t tests. This observation suggests that misspecification error is conceivably the main impediment preventing traditional models from identifying superior funds. It also suggests that it is perhaps too early to give up on traditional models and

[5] The published model excludes the cost of searching for funds with superior managers and thus states that investors literally earn zero excess returns from their mutual fund investments. However, as discussed later on in this paper, with positive search costs well designed strategies should be profitable.

data sets (e.g., Hendricks et al. (1993), Brown and Goetzmann (1995)) in the quest to find managerial talent ex-ante.

The back testing procedure suggested here can also potentially strengthen the findings and methodologies within the above and other related papers. Indeed, this has started happening. Kacperczyk et al. (2005b) use a variant of the back test proposed here to improve their model's ability to forecast fund returns based on the difference between observed returns and those calculated from the reported holdings data.[6] Future research will undoubtedly show that other (better) back testing procedures can be employed. Still, it should be emphasized that the goal here is not to produce an "optimal" estimator or back test. Rather, the goal of this paper is to develop an effective and simple procedure that is likely to be robust across a variety of possible situations and thus potentially a heuristic for future research. Having said that, it is not obvious that more complex filters alone will in fact yield better results. The difficulty is that misspecification errors can influence the values generated by any statistical model in a variety of ways and thus, to the degree, more complex filters are less robust they may yield inferior results. In fact, this paper finds just that. When the estimated betas are used as part of the back test the resulting portfolios exhibit poorer performance. Alternatively, one might think to use a model's diagnostics, like the estimated alpha's t statistic. However, this too yields inferior results relative to simply using the model's forecasting success in the previous period.

The paper's negative finding regarding more complex back testing methodologies may be due to the possibility that misspecification errors lead not only to large parameter estimates but also to erroneously good in sample diagnostics. The problem is that admitting that a model may be misspecified is tantamount to admitting "we do not know what we do not know." This can be seen in the negative results generated by Bossaerts and Hillion (1999) and Goyal and Welch (2006). Both studies try various model switching criteria and find no out of sample predictability regarding the equity premium (Goyal and Welch even find it degrades the resulting portfolio's performance). Both studies also suggest that "model instability," which is a form of misspecification error, may be to blame. Thus, as indicated here, simpler (and potentially more robust) tests may in the end produce the best results.

Since mutual funds use "closing prices" to calculate end of day net asset values (NAV), any model may appear to select funds with positive abnormal

[6] They cite this paper as the source for the procedure.

returns by taking advantage of the resulting stale pricing problem.[7] This is an important issue since such abnormal returns are not related to managerial ability but rather to the use of a poor (and exploitable) pricing algorithm for calculating end of day net asset values. To control for this possibility the paper introduces a "fifth through eighth factor," the four-factor returns from month $t-1$ to month t. This "eight factor model" should capture any stale pricing impacts, as returns due to past market returns should be absorbed by the historical return parameters. These stale pricing factors reduce the out of sample returns by about four to eight basis points per month, but typically do not eliminate them.

This paper is also related to the general econometric literature on forecasting. For example, White (2000) shows how to produce exact p-values given a postulated distribution function for the hypothesis that some models are superior to the benchmark when several models are run on a data set. Pesaran and Timmermann (2005) examine model selection in real time and also seek tests to determine which among many models is the best at any moment in time. Here we are not interested in finding the best overall model, but rather seeing if a given model's performance can be improved by limiting its use to particular subjects within a data set. More closely related to the current article is Timmermann and Granger (2004). They discuss attempting to create forecasts under the assumption that markets, if not instantly efficient, will eventually work to invalidate any currently successful forecasting method. As they note, in this environment forecasters will frequently need to change the model they use. It is worth noting that the back testing procedure suggested here allows for potentially rapid switching across models to capture the changing behavior of particular subjects.

All of the empirical tests in this paper are conducted with data from the CRSP mutual fund files. Only equity funds (defined as funds with objective codes of AG, BL, GI, IN, LG, PM, SF, or UT) are used. These data are then augmented with the monthly CRSP value weighted return market index, and the Fama–French and Carhart factors as provided by the WRDS web site.

[7] The closing prices used to set mutual fund net asset values are in reality the last price a security traded at prior to 4:00 P.M. eastern time in the U.S. For infrequently traded securities this price may have been recorded hours if not days earlier. Investors can take advantage of this by purchasing funds with infrequently traded securities after the broad market has gone up, and by selling these funds following a market decline. This strategy sometimes goes by the name of market timing and the academic literature on this issue has grown extensively in recent years. Papers by Chalmers et al. (2001), Greene and Hodges (2002), Zitzewitz (2003), and Goetzmann et al. (2001) all show how investors can exploit stale net asset value prices in a variety of funds.

While funds with load fees are included in the analysis (to be consistent with past studies) the results are essentially unchanged when they are excluded.[8]

The remainder of the paper proceeds as follows. Section 1 provides evidence regarding the impact of estimation errors on sorts. Section 2 examines the out of sample returns for the various models. Section 3 looks at the degree to which the models pick similar funds. Section 4 discusses whether or not restricting the sample to funds with high levels of tracking error relative to various benchmarks helps to further refine the search for high alpha funds. Section 5 compares the results in this paper to those in Carhart (1997). Section 6 discusses the implications of this paper for the Berk and Green (2004) hypothesis. Section 7 looks at whether using a model's in sample betas or t-statistics improves or degrades its ability to find funds that will out perform in the future. Section 8 concludes. The Appendix (Section A) derives a dynamic model of mutual fund returns (used as an alternative to the typically employed static OLS specification) and explains how to estimate it via a modified Kalman filter.

1. Alpha–Beta Relationships in the OLS Model

As noted above, OLS models seeking to predict mutual fund returns typically begin by estimating a factor model within a 5 year rolling window. Next, funds are sorted on alpha and ranked by decile. Tests are then conducted on each decile's out of sample performance. However, sorting in this manner potentially ranks funds both by their alphas and by their estimation error. In a factor model this is especially problematic since misestimated betas will themselves induce misestimated alphas. To see why, consider a one–factor model and imagine that the estimated beta is too low. If the fund actually holds a portfolio of stocks then in a typical year it will have a higher return than the risk free asset. With an underestimated value of beta the model will then try to fit the returns by raising alpha. The opposite will also hold: overestimated betas will typically lead to underestimated alphas.

Table I examines how estimation errors feed back between alphas and betas. For each fund a one or four factor model of the form

$$r_{it} - r_t = \alpha_t + \beta_i{}'(r_{mt} - r_t) + \epsilon_t \tag{1}$$

[8] Readers can, to some degree, easily verify this for themselves. Out of sample statistics are provided for the top 5, 10, and 20 funds as well as the top decile. All four groups generate similar results. Thus removing the funds with loads from the top decile leaves more than enough funds to populate a top 5, 10, and 20 grouping.

is estimated. Here r_{it} is fund i's period t return, r_t the risk free rate, r_{mt} the vector of factor returns, and ϵ_t an error term. The estimated parameters are α and the vector β based on 5 year rolling windows. In an ideal world (without estimation error) sorting on alpha should not lead to systematic patterns in the beta parameters.

To generate Table I, first estimate Equation (1) fund-by-fund. Next, sort funds by alpha into ten portfolios and hold each portfolio for 12 months. The resulting 12 month return series are then regressed on the corresponding factor model to generate out of sample alpha and beta estimates. Let $\Delta\beta_{MKT}$ equal the difference between the predicted and realized beta ($\Delta\beta_{MKT} \equiv \beta_{MKT,predicted} - \beta_{MKT,realized}$). Table I reports by decile the mean forecasted alpha value, and the parameters from the regression $\Delta\beta_i = c_i + error$ for $i = MKT, SMB, HML, and\ MOM$ within each decile.

Most studies find that single factor OLS models do a poor job of detecting any out of sample performance. Yet, according to the sorted α column in Table I panel A on a risk–adjusted basis any number should. Based on the model's forecasts, the lowest decile should underperform by 8.9% and the top decile overperform by 7.3% per annum. These results are clearly at odds with anything one might realistically hope to generate. The next column (c_{MKT}) shows that the in sample alphas are perfectly negatively correlated with the out of sample beta forecast errors. Low alphas are associated with betas that out of sample are systematically lower than their in sample counterparts. For high alpha funds the results are reversed. These differences are both systematic and statistically highly significant. Looking at the final column in Panel A, one sees that the beta forecast errors are highest for the extreme deciles. This is particularly problematic since these are the deciles of greatest interest and this implies that their forecasts are the least reliable.

The results for the four factor OLS model in Table I Panel B once again indicate that the extreme deciles should produce out of sample returns that are far more extreme than anybody has ever detected. In terms of the beta forecast errors, while different from those in Panel A, they remain problematic. While the market beta forecast errors are no longer systematically associated with the alpha forecasts they are instead uniformly under estimated. Perhaps just as importantly, the worst beta forecasts are again associated with the highest and lowest deciles, which are the most important for studies concerned with detecting whether some managers can out perform the market. Thus, once again it appears that the estimates for the most important deciles are also the least reliable.[9]

[9] We thank Wayne Ferson for suggesting this test.

Table I. Beta errors in sorted alpha deciles

This table first calculates monthly OLS alphas and betas in month t from a rolling regression based on $t-60$ to $t-1$ returns. These estimated parameters are then used as forecasts in month t. This process is used over a 12 month period and the average forecasted alpha and beta is calculated. Next, the realized fund returns for the 12 month period is regressed on the market model to produce an in sample alpha and beta for the period in question. Let $\Delta\beta_{MKT}$ equal the difference between the predicted and realized beta ($\Delta\beta_{MKT} \equiv \beta_{MKT,predicted} - \beta_{MKT,realized}$). Similarly, for the four factor model the difference between the predicted and realized factor loadings will be referred to as $\Delta\beta_i$ for i equal to MKT, SMB, HML, and MOM. All of the mean alpha values and beta differences are then pooled across the different months for the testing period of January 1970 to December 2002. Ten deciles are then created based upon the estimated alphas. The table reports by decile the mean alpha value, the parameters from the regression $\Delta\beta_i = c_i + error$, and the standard deviation of $\Delta\beta_i$ ($i = MKT, SMB, HML, and MOM$) within each decile. $T_{NW,11}(c_{MKT})$ are the Newey-West t-statistics for c_{MKT} with eleven lags. $T_{OLS}(c_{MKT})/\sqrt{12}$ column reports the t-statistics as the OLS t-statistics divided by the square root of 12.

Deciles	Sorted α	c_{MKT}	$T_{NW,11}(c_{MKT})$	$T_{OLS}(c_{MKT})/\sqrt{12}$	Std($\Delta\beta_{MKT}$)	Std($\Delta\beta_{SMB}$)	Std($\Delta\beta_{HML}$)	Std($\Delta\beta_{MOM}$)
	A. CAPM							
1	−0.0077	0.0508	(16.37)	(6.02)	0.34			
2	−0.0033	0.0188	(10.78)	(3.15)	0.24			
3	−0.0021	0.0160	(9.81)	(2.87)	0.22			
4	−0.0012	0.0095	(6.24)	(1.77)	0.22			
5	−0.0005	0.0052	(3.77)	(1.08)	0.20			
6	0.0001	0.0016	(1.15)	(0.33)	0.19			
7	0.0008	−0.0050	(−3.53)	(−1.03)	0.20			
8	0.0015	−0.0119	(−7.85)	(−2.28)	0.21			
9	0.0027	−0.0247	(−14.97)	(−4.34)	0.23			
10	0.0059	−0.0388	(−19.47)	(−5.80)	0.27			
	B. Carhart 4-factor model							
1	−0.0067	−0.0631	(−16.93)	(−5.60)	0.48	0.58	0.68	0.58
2	−0.0033	−0.0274	(−13.63)	(−3.92)	0.30	0.34	0.48	0.36
3	−0.0022	−0.0262	(−14.17)	(−4.12)	0.27	0.31	0.43	0.34
4	−0.0014	−0.0195	(−10.90)	(−3.21)	0.26	0.29	0.40	0.31
5	−0.0007	−0.0226	(−13.94)	(−4.01)	0.24	0.27	0.37	0.29
6	−0.0001	−0.0224	(−13.84)	(−4.06)	0.23	0.27	0.37	0.29
7	0.0005	−0.0162	(−9.92)	(−2.87)	0.24	0.28	0.37	0.29
8	0.0013	−0.0263	(−15.26)	(−4.43)	0.25	0.31	0.40	0.32
9	0.0025	−0.0272	(−12.99)	(−3.76)	0.31	0.40	0.46	0.39
10	0.0060	−0.0631	(−20.59)	(−6.47)	0.41	0.49	0.62	0.55

Most forecasting papers have tended to focus on sampling error (as noted earlier, a recent example is Kosowski et al. (2007)). This is clearly an important issue. However, if the underlying model is not correctly specified then even the most sophisticated parameter and diagnostic test corrections will not yield useful forecasts. Thus, the next section begins the paper's examination of whether tests for specification errors can help improve our ability to detect managers who have the ability to out perform the market.

2. Out of Sample Returns with Filtering

In the tests that follow the text concentrates on whether a particular statistical technique can find mutual funds that generate positive alphas going forward. While one can also look at high-minus-low portfolio returns this is not a particularly productive exercise when examining mutual funds. From an investor's perspective, high-minus-low portfolios cannot be created since short positions in a mutual fund are not possible. From the perspective of the problem's economics many studies have found that it is easy to identify poor performing funds. Some funds charge high management fees or regularly incur high transactions costs. Not surprisingly, they regularly report returns that underperform on a risk-adjusted basis.[10]

When forecasting mutual fund alphas the standard approach uses a single model across the entire panel data set. This paper considers four models that have been used in the context of mutual fund research. The first two are the standard one and four factor OLS models. The second two are one and four factor Kalman filter models. This model was developed in Mamaysky et al. (2007) and assumes that managers receive a hidden signal that follows an AR(1) process, which they then use to rebalance their portfolios in order to earn excess returns. For interested readers, a brief derivation can be found in the Appendix. The important point is that the Kalman filter model explicitly allows for time variation in a fund's factor loadings, which offers a good contrast to the OLS models that do not.

Table II displays the results from using any one model across the entire universe of funds to forecast alphas. This exercise, with monthly rebalancing, serves as a benchmark from which any modifications to the statistical procedure can be judged. Panel A displays the out of sample four factor alphas based

[10] Like nearly every other study this one will also show that finding funds with predictably negative alphas is not difficult. A better question (and one not addressed here) is given how easy they are to identify, what allows them to remain in business? Berk and Xu (2004) attribute their survival to the fact that a substantial number of investors do not pull their money from funds even after they have done poorly. This allows these managers to continue operating despite having a poor past and predictably poor future set of returns.

Table II. Out of sample performance (no parameter forecast filters, no back testing)

For all domestic equity mutual funds having at least 5 years of monthly return data, both the Kalman and the OLS models are used to forecast a fund's alpha. Starting from January 1970, the 1-factor and 4-factor OLS models (OLS 1 and OLS 4) and 1-factor and 4-factor Kalman models (Kal 1 and Kal 4) independently sort funds into equally weighted portfolios based on forecasted alphas. Decile 10 contains funds with the highest forecasted alphas. The Top X portfolios contain an equally weighted portfolio of the X funds with the highest alphas. Each portfolio is held for 1 month and then rebalanced until December 2002. There are no restrictions on forecasted alphas and betas. Panel A reports the 4-factor (*MKT, SMB, HML, and MOM*) adjusted monthly returns realized during the entire period and the corresponding t-statistics. In Panel B, the portfolio returns are risk-adjusted by both the four factors and 1-month-lagged 4-factor returns (8 factors in total).

Decile	OLS 1	OLS 4	Kal 1	Kal 4	OLS 1	OLS 4	Kal 1	Kal 4
	A1. Four-factor monthly alpha				A2. T-ratio			
1	−0.0009	−0.0018	−0.0015	−0.0019	(−1.15)	(−2.68)	(−1.97)	(−2.91)
2	−0.0000	−0.0011	−0.0009	−0.0009	(−0.08)	(−2.68)	(−1.73)	(−1.93)
3	−0.0002	−0.0006	−0.0005	−0.0008	(−0.37)	(−1.75)	(−1.13)	(−2.03)
4	−0.0004	−0.0007	−0.0001	−0.0007	(−0.97)	(−1.80)	(−0.14)	(−1.88)
5	−0.0003	−0.0005	−0.0003	−0.0003	(−0.70)	(−1.33)	(−0.89)	(−0.99)
6	−0.0006	−0.0003	−0.0002	−0.0004	(−1.79)	(−0.95)	(−0.63)	(−1.32)
7	−0.0004	−0.0005	−0.0005	−0.0002	(−1.03)	(−1.56)	(−1.39)	(−0.64)
8	−0.0001	−0.0004	−0.0008	0.0001	(−0.19)	(−0.89)	(−1.78)	(0.33)
9	−0.0003	0.0004	0.0000	0.0005	(−0.50)	(0.83)	(0.07)	(1.21)
10	−0.0008	0.0016	0.0005	0.0005	(−1.14)	(2.39)	(0.67)	(0.73)
Top 20	−0.0009	0.0030	0.0014	0.0001	(−0.96)	(2.64)	(1.11)	(0.15)
Top 10	−0.0005	0.0031	0.0013	−0.0004	(−0.42)	(2.18)	(0.80)	(−0.33)
Top 5	−0.0015	0.0037	0.0022	−0.0002	(−0.76)	(2.04)	(1.09)	(−0.14)
	B1. Eight-factor monthly alpha				B2. T-ratio			
1	−0.0009	−0.0017	−0.0010	−0.0016	(−1.13)	(−2.45)	(−1.27)	(−2.34)
2	−0.0000	−0.0009	−0.0010	−0.0008	(−0.08)	(−2.06)	(−1.84)	(−1.63)
3	−0.0000	−0.0004	−0.0004	−0.0006	(−0.08)	(−0.95)	(−0.90)	(−1.51)
4	−0.0003	−0.0006	−0.0000	−0.0006	(−0.59)	(−1.42)	(−0.08)	(−1.59)
5	−0.0004	−0.0003	−0.0003	−0.0003	(−0.87)	(−0.90)	(−0.89)	(−0.89)
6	−0.0005	−0.0002	−0.0002	−0.0004	(−1.35)	(−0.64)	(−0.61)	(−1.07)
7	−0.0002	−0.0005	−0.0005	−0.0002	(−0.48)	(−1.48)	(−1.28)	(−0.64)
8	0.0001	−0.0004	−0.0007	0.0002	(0.11)	(−1.04)	(−1.64)	(0.44)
9	−0.0001	0.0005	0.0002	0.0005	(−0.25)	(0.96)	(0.37)	(1.10)
10	−0.0007	0.0015	0.0006	0.0005	(−0.89)	(2.16)	(0.76)	(0.76)
Top 20	−0.0007	0.0026	0.0011	0.0000	(−0.70)	(2.19)	(0.83)	(0.00)
Top 10	−0.0007	0.0024	0.0011	−0.0005	(−0.47)	(1.57)	(0.63)	(−0.34)
Top 5	−0.0022	0.0029	0.0023	−0.0001	(−1.05)	(1.53)	(1.06)	(−0.08)

on the decile portfolios formed via each model. While the four factor OLS model does provide some predictive power (the top decile generates statistically significant returns) no other model does.

As noted in the introduction, many studies have documented the fact that mutual fund returns are impacted by stale prices. In part this is due to the tendency of some fund managers to invest in very illiquid securities for which a daily value must be assigned even if it has not traded in quite some time. From an empirical perspective, stale pricing causes a fund's reported current period return to depend on its return in both the current and past period. When conducting out of sample performance tests this problem opens up the possibility that a model will succeed not because it finds managers capable of producing above market risk-adjusted returns, but rather because it finds funds with currently stale prices whose net asset value can be expected to "catch up" next period. To see how a model can do this, imagine it finds funds that have betas of about one and has the portfolio invest in them when the market has risen in the prior month. Such funds will then generate reported returns with apparently positive alphas. Part of the reported return will come from the current month's return which will, on average, equal the market's. (If the market is down the fund will tend to report that it outperformed, since many its holdings will be priced on the basis of their pre market drop value. Conversely, for the same reason stale prices will lead the fund to report that it underperformed if the market has gone up.) In addition, however, these funds will also report, as part of the current month's return, an "increase" in the value of the illiquid securities that they hold as these securities trade and see their values updated accordingly. Since the goal of this paper is to see if statistical models, with the help of back testing, can find funds capable of generating truly positive alphas the stale pricing problem needs to be ameliorated in some manner. A simple solution is to run the out of sample portfolio returns on a model that includes each of the four-factor model's lagged and current values. If a model's ability to identify positive alpha funds is indeed due to its ability to find funds with particularly stale prices then the resulting alpha from this eight-factor model should equal zero.

Table II Panel B presents the results from extending the four factor model to include each factor's lagged value. This has a somewhat erratic impact on the performance of the four factor OLS model. Portfolios using the top 5 and 10 funds no longer produce statistically significant out of sample alphas. On the other hand, the broader top 20 and top decile portfolios do. It is perhaps possible that the very highest alphas are arising from the model's ability to exploit stale prices while the alphas for funds a bit further down the list are due to actual talent. In any event, even the best portfolios yield rather modest t-statistics making it difficult to confidently reject the null.

Based on the results in Table I the negative results in Table II are likely due to the fact that the extreme decile alphas and betas are of suspect quality. If so, then a simple solution may be to just drop any fund whose alpha or beta lies outside some boundary. Table III displays the impact of this filtering device on the portfolio returns. Comparing the results to those in Table II Panel A, one can see that the OLS models exhibit some improved out of sample performance when the beta range is substantially restricted, but none from restricting the alphas. For the Kalman filter model both the alpha and beta restrictions provide some modest improvements. Overall, though, simply dropping funds for which a model's coefficients are unreasonable does not appear to produce substantially better performing portfolios. However, many of the subsequent tests still drop funds with monthly alphas predicted to exceed 2% in absolute value. As will be seen, this primarily provides a formal way to define, in a manner repeatable by others, when the Kalman filter optimization algorithm does not converge. Other than that, this broad a range elimates very few funds.

From a Bayesian perspective, it may seem that parameter filters like those explored above or shrinkage estimators should successfully allow a model to identify funds with positive expected alphas, assuming such funds exist. However, this can only work if the model is properly specified to begin with. If it is not then large parameter estimates may arise not from sampling errors but from misspecification problems instead. In the latter case, shrinking what are effectively meaningless numbers towards zero does not help produce better forecasts. Even eliminating funds with extreme parameter estimates will not help since the meaningless numbers generated by funds incorrectly described by the model will be scattered throughout the real number range. Absent knowledge of the form of the misspecification error there is no way to know what numbers are likely to be generated.

The alpha and beta filters examined in Table III are an indirect way of checking to see if a particular statistical model fits a particular fund. A more direct approach (and one that can be used either in combination with the alpha and beta restrictions or on its own) is back testing. The general idea is that before one uses a statistical model to make predictions, some evidence should exist that it was successful in this regard some time in the past. Along these lines it seems reasonable to require predictive success in the period prior to the model's use as a forecasting device. The back testing procedure works as follows. The model is estimated with 60 months of data up to time $t-2$ and a predicted alpha and a set of betas for $t-1$ are calculated. If the fund's realized return minus the market's return $(r_{pt}-r_{mt})$ in period $t-1$ has the same sign as the predicted alpha, then the fund is added to the active pool. Otherwise the fund goes into the inactive pool. Note, that this procedure does

Table III. Out of sample performance with various α and β filters (no back testing)

For all domestic equity mutual funds having at least 5 years of monthly return data, both the Kalman and the OLS models are used to forecast a fund's alpha. Starting from January 1970 until December 2002, the 1-factor and 4-factor OLS models (OLS 1 and OLS 4), and the 1-factor and 4-factor Kalman models (Kal 1 and Kal 4) independently sort funds into 10 deciles based on their forecasted alphas. Decile 10 contains funds with the highest forecasted alphas. For each model, 10 equally weighted portfolios are constructed from stocks within the 10 deciles. Each portfolio is held for 1 month and rebalanced at the beginning of the next month. For any fund to be included in any decile the predicted alpha and beta values must lie within the specified boundary. In Panels A, the alpha boundaries are fixed at ±2% per month and the beta range changes. In Panels B, the alpha boundary changes but the beta range is fixed at 0 to 2. The table reports the 4-factor (*MKT, SMB, HML, and MOM*) adjusted monthly returns realized in the entire period for Decile 10 funds (those expected to generate the highest future return).

		OLS 1	OLS 4	Kal 1	Kal 4	OLS 1	OLS 4	Kal 1	Kal 4
A. Decile 10 performance with various beta ranges.									
β_{Min}	β_{Max}	Four-factor monthly alpha				T-ratio			
0.80	1.20	0.0004	0.0020	0.0005	0.0010	(0.70)	(2.79)	(0.93)	(1.84)
0.60	1.40	−0.0000	0.0019	0.0009	0.0008	(−0.03)	(2.86)	(1.49)	(1.42)
0.40	1.60	−0.0003	0.0017	0.0008	0.0007	(−0.42)	(2.69)	(1.27)	(1.33)
0.20	1.80	−0.0007	0.0016	0.0007	0.0007	(−1.03)	(2.34)	(1.05)	(1.25)
0.00	2.00	−0.0008	0.0016	0.0007	0.0007	(−1.22)	(2.35)	(1.04)	(1.28)
B. Decile 10 Performance with various α ranges.									
α_{Min}	α_{Max}	Four-factor monthly alpha				T-ratio			
−0.02	0.02	−0.0008	0.0016	0.0007	0.0007	(−1.22)	(2.35)	(1.04)	(1.28)
−0.04	0.04	−0.0010	0.0016	0.0005	0.0005	(−1.39)	(2.37)	(0.74)	(0.85)
−0.06	0.06	−0.0010	0.0016	0.0007	0.0006	(−1.39)	(2.37)	(0.91)	(0.95)
−0.08	0.08	−0.0010	0.0016	0.0006	0.0005	(−1.39)	(2.37)	(0.87)	(0.80)
−0.10	0.10	−0.0010	0.0016	0.0006	0.0004	(−1.39)	(2.37)	(0.77)	(0.74)
−1.00	1.00	−0.0010	0.0016	0.0004	0.0005	(−1.42)	(2.39)	(0.55)	(0.75)

not use risk-adjusted returns. This minimizes the use of the model's parameter estimates when looking to see if its forecasts are accurate, thus potentially preventing one error from masking another. Later on, additional tests will show that attempts to risk adjust returns at this stage only impede the search for nonmarket performers. Another way to look at the test used here is to view it as calculating the realized alpha by setting the market beta to one and all other betas, if any, to zero. Again, this guarantees that the results do not depend upon the model's own estimated factor loadings.[11]

[11] Initially there might be some concern that since estimated parameter values are serially correlated the back test may pick up these estimation errors and report them as alphas. However, since the in sample and out of sample parameters are estimated on disjoint data sets

At first it might seem like the above back test would bias the selections towards high-risk funds. These funds, after all, are the most likely to produce above market returns absent any special information. While that is true, the out of sample portfolio alphas are calculated from factor models that should catch and eliminate any such bias. Thus, even if in the end the back test tends to favor high beta funds it cannot bias the estimated out of sample alphas.

To summarize, the filtering and portfolio selection procedure proceeds as follows:

1. The back test is used to determine the active pool.
2. A second filter is then used to ensure that the parameter estimates are not unreasonable. This is done by first estimating the model parameters for funds in the active pool with the model in question using data through time t-1. If the estimated market beta lies between zero and two and the alpha between -0.02 and $+0.02$ the fund remains in the active pool. Otherwise it goes into the inactive pool.
3. Funds within the active pool are sorted by alpha. Then the decile and top 5, 10, and 20 fund portfolios are constructed.

The above procedure does *not* introduce any selection biases. Any investor can use the three steps in real time to select funds. To further ensure that there are no selection biases the out of sample alphas are calculated using the standard OLS procedure on the realized portfolio returns. That is, portfolios are formed each period and the return to the portfolio is calculated. The resulting return sequence is then regressed on the appropriate factor model (four or eight) and the estimated alpha is then reported. Note that the in sample estimated alphas and betas are not used at all when calculating the out of sample test statistics. Further, the standard OLS testing procedure is employed out of sample to further ensure that a particular model does not somehow feed itself risk estimates that will lead to a biased alpha value.

Table IV lists the results when both the back test and the parameter filters are used to select funds for the active pool prior to creating portfolios based upon the sorted alphas. Unlike the results in Carhart (1997) and Avramov and Wermers (2005), the top OLS portfolios produce statistically significant positive returns. This is not only true of the top decile but also of the *ninth* decile. This implies that with some back testing it is possible to find a substantial number of funds that will, on average, produce positive predictable returns. The same findings hold for the Kalman filter model. If one further restricts

that cannot happen. Furthermore, the back test discussed here does not even employ the in sample parameter estimates and thus cannot possibly focus on funds with poorly estimated factor loadings.

Table IV. Out of sample performance (with parameter forecast filters, with back testing)

For all domestic equity mutual funds having at least 5 years of monthly return data, both the Kalman and the OLS models are used to forecast a fund's alpha. Starting from January 1970 until December 2002, the 1-factor and 4-factor OLS models (OLS 1 and OLS 4), and 1-factor and 4-factor Kalman models (Kal 1 and Kal 4) independently sort funds into 10 deciles based on forecasted alphas. Decile 10 contains funds with the highest forecasted alphas. For each model, 10 equally weighted portfolios are constructed from stocks within the 10 deciles. Each portfolio is held for 1 month and rebalanced at the beginning of the next month. For any fund to be included in any decile the following must be true: (1) the absolute value of forecasted alpha must be less than 2% per month; (2) the beta must be greater than 0 but less than 2; and (3) in the previous month the out of sample forecasted alpha and the difference between the realized return and the market return must have the same sign. Finally, the table also constructs model by model equally weighted portfolios containing the 20, 10 and 5 funds with the highest alphas forecasts. Panels A reports the monthly excess return and Sharpe ratio for each portfolio after regressing the sequence of returns on the appropriate factor model. Panels B reports the 4-factor (*MKT, SMB, HML, and MOM*) adjusted monthly returns realized during the entire period and the corresponding *t*-statistics. Although not reported, the rank correlation between decile number and the average realized excess return for each model is highly significant (correlation >0.9, P-value <0.01).

Decile	OLS 1	OLS 4	Kal 1	Kal 4	OLS 1	OLS 4	Kal 1	Kal 4
	A1. Monthly return above the risk free rate				A2. Monthly Sharpe Ratio			
1	0.0011	0.0008	0.0013	0.0016	0.0215	0.0165	0.0266	0.0318
2	0.0020	0.0020	0.0026	0.0027	0.0402	0.0410	0.0539	0.0584
3	0.0021	0.0026	0.0027	0.0027	0.0444	0.0557	0.0565	0.0576
4	0.0032	0.0028	0.0036	0.0026	0.0702	0.0607	0.0803	0.0582
5	0.0042	0.0034	0.0033	0.0033	0.0987	0.0774	0.0757	0.0775
6	0.0045	0.0037	0.0043	0.0038	0.1062	0.0864	0.0996	0.0896
7	0.0052	0.0043	0.0052	0.0050	0.1245	0.1003	0.1225	0.1159
8	0.0052	0.0056	0.0050	0.0055	0.1228	0.1254	0.1179	0.1243
9	0.0059	0.0061	0.0064	0.0058	0.1318	0.1322	0.1427	0.1285
10	0.0070	0.0076	0.0078	0.0066	0.1474	0.1508	0.1706	0.1378
Top 20	0.0066	0.0074	0.0078	0.0057	0.1337	0.1343	0.1636	0.1159
Top 10	0.0078	0.0084	0.0084	0.0062	0.1494	0.1421	0.1769	0.1232
Top 5	0.0080	0.0078	0.0089	0.0063	0.1482	0.1245	0.1804	0.1186
	B1. Four-factor monthly alpha				B2. *T*-ratio			
1	−0.0026	−0.0037	−0.0023	−0.0031	(−2.88)	(−3.90)	(−2.57)	(−3.83)
2	−0.0020	−0.0025	−0.0016	−0.0016	(−2.72)	(−3.79)	(−2.51)	(−2.55)
3	−0.0021	−0.0018	−0.0015	−0.0018	(−3.38)	(−3.24)	(−2.45)	(−3.25)
4	−0.0010	−0.0016	−0.0005	−0.0019	(−1.73)	(−2.75)	(−1.00)	(−3.93)
5	0.0000	−0.0009	−0.0010	−0.0009	(0.04)	(−1.55)	(−2.04)	(−1.86)
6	−0.0000	−0.0009	−0.0003	−0.0005	(−0.00)	(−1.77)	(−0.57)	(−1.08)
7	0.0007	−0.0001	0.0007	0.0004	(1.09)	(−0.15)	(1.46)	(0.93)
8	0.0004	0.0008	0.0005	0.0010	(0.64)	(1.37)	(0.83)	(1.88)
9	0.0009	0.0016	0.0013	0.0013	(1.30)	(2.53)	(2.03)	(2.43)
10	0.0021	0.0037	0.0031	0.0023	(2.35)	(4.03)	(3.95)	(3.23)
Top 20	0.0019	0.0046	0.0031	0.0020	(2.01)	(3.93)	(3.57)	(2.45)
Top 10	0.0030	0.0057	0.0036	0.0027	(2.48)	(3.91)	(3.70)	(2.59)
Top 5	0.0032	0.0048	0.0040	0.0022	(2.20)	(2.73)	(3.34)	(1.81)

attention to the top twenty or better funds the excess returns are about 4.4% per annum. For every model the rank correlation between the deciles and their returns exceeds 0.95. This is statistically significant at any reasonable level, providing additional evidence that the alpha sorts are picking up relative out of sample fund performance. As a further test of the efficacy of the back testing procedure, the analysis in Table IV was repeated but this time with the funds from the inactive pool. To save space the table is not reported here. But the resulting decile portfolios have no predictive power further verifying that back testing helps remove from a model's forecast those funds it is unable to accurately track.[12]

The fact that back testing dramatically improves all four models tested here indicates that the results are not due simply to data snooping. By contrast, data snooping typically yields a small subset of "successful" models from the larger group under consideration. Furthermore, the improvement seen with back testing takes place not only in the high and low forecasted alpha deciles but overall as well. Without back testing the decile portfolios created with the in sample sorted alphas yield statistically insignificant rank correlations across deciles out of sample. However, with back testing the correlation coefficients rise dramatically and are statistically significant at any reasonable level.

Related to the issue of estimation is whether or not adding factors to the model improves its out of sample performance. For the OLS model the four factor model does appear to offer superior out of sample performance relative to the one factor model. The difference generally comes to about 2% per annum. For the Kalman filter model, however, the results are reversed. Under nearly every sample test comparison performed here, decisions based upon an in sample one factor model outperform decisions based upon an in sample four factor model. This indicates that if researchers try other nonlinear models they may find that the simpler in sample versions perform better out of sample.

As noted in the introduction, mutual fund NAVs and thus returns are impacted by stale pricing. This can potentially cause funds to generate spurious alphas even when their managers are incapable of generating nonmarket expected returns. However, this problem can be eliminated by including the lagged-factor returns when estimating the alphas. By doing so, any correlation between the four factor alphas and stale prices should be eliminated in these eight-factor regressions. Table V repeats the analysis in Table IV using the eight-factor model. Most of the models see their alphas decline by about five basis points per month. Nevertheless, the top deciles continue to produce statistically significant out of sample returns for all but the one factor OLS

[12] The authors thank Joshua Coval for suggesting that we run a test on funds from the inactive pool.

Table V. Eight–factor model out of sample performance (with parameter forecast filters, with back testing)

For all domestic equity mutual funds having at least 5 years of monthly return data, both the Kalman and the OLS models are used to forecast a fund's alpha. Starting from January 1970 until December 2002, the 1-factor and 4-factor OLS models (OLS 1 and OLS 4), and 1-factor and 4-factor Kalman models (Kal 1 and Kal 4) independently sort funds into 10 deciles based on forecasted alphas. Decile 10 contains funds with the highest forecasted alphas. For each model 10 equally weighted portfolios are constructed from stocks within the 10 deciles. At the beginning of month t, funds are sorted according to predicted alphas in month t. Each portfolio is held for 1 month and rebalanced at the beginning of the next month. For any fund to be included in any decile (1) the absolute value of forecasted alpha must be less than 2% a month; (2) the beta must be greater than 0 but less than 2; and (3) the forecasted alpha and the difference between the realized return and the market return in the previous month must have the same sign. Finally, the table also constructs equally weighted portfolios containing the 20, 10 and 5 funds with the highest alphas forecast by each model. The eight-factor model contains the four Carhart factors plus their 1 month lagged values.

Decile	OLS 1	OLS 4	Kal 1	Kal 4	OLS 1	OLS 4	Kal 1	Kal 4
	Eight-factor monthly alpha				T-ratio for lagged-factor adjusted return			
1	−0.0024	−0.0030	−0.0017	−0.0026	(−2.56)	(−3.04)	(−1.80)	(−3.06)
2	−0.0016	−0.0019	−0.0012	−0.0013	(−2.08)	(−2.80)	(−1.78)	(−1.97)
3	−0.0017	−0.0014	−0.0013	−0.0013	(−2.63)	(−2.31)	(−2.04)	(−2.26)
4	−0.0009	−0.0011	−0.0004	−0.0017	(−1.52)	(−1.82)	(−0.68)	(−3.37)
5	−0.0000	−0.0005	−0.0009	−0.0008	(−0.07)	(−0.84)	(−1.77)	(−1.61)
6	−0.0001	−0.0009	−0.0004	−0.0007	(−0.20)	(−1.63)	(−0.91)	(−1.38)
7	0.0005	−0.0002	0.0007	0.0003	(0.72)	(−0.36)	(1.30)	(0.65)
8	0.0002	0.0007	0.0002	0.0009	(0.34)	(1.15)	(0.29)	(1.51)
9	0.0004	0.0014	0.0011	0.0011	(0.62)	(2.15)	(1.59)	(1.93)
10	0.0016	0.0032	0.0027	0.0018	(1.75)	(3.29)	(3.17)	(2.46)
Top 20	0.0014	0.0038	0.0029	0.0016	(1.39)	(3.12)	(3.10)	(1.87)
Top 10	0.0025	0.0048	0.0033	0.0021	(1.92)	(3.14)	(3.15)	(1.91)
Top 5	0.0021	0.0036	0.0038	0.0016	(1.38)	(1.97)	(2.98)	(1.29)

model. In addition, the t-statistic for the ninth decile's alpha equals 2.15 for the funds selected by the four factor OLS model, and 1.93 for the funds selected by the four factor Kalman filter model. If one restricts attention to the top 20, 10, or 5 funds then the addition of the lagged factors reduces the estimated alphas by about eight basis points per month for the four factor OLS model, but only three basis points for the Kalman filter model. This provides some initial evidence that the two models are indeed picking up different funds and selecting on different skill sets. More importantly, however, the t-statistics for the portfolios containing the top 10 or 20 funds remain above three for both the four factor OLS and the one factor Kalman filter model. Thus, even when

one controls for stale pricing, a statistical model combined with back testing can still find a substantial number of funds that will outperform the four or eight-factor market benchmark over the next month.

Table VI examines the performance of each model with back testing but without the alpha and beta restrictions. Without them the Kalman filter model loses its predictive power. This is not too surprising. For nonlinear models, convergence of the parameter search algorithm is not guaranteed. Researchers frequently get around this problem by reporting only the results for which the search algorithm "converged." However, without a clearly defined rule for determining if convergence has occurred, one cannot guarantee that a particular result can be reproduced by other researchers. The alpha and beta restrictions provide the necessary definition in a way that can be clearly specified in advance. This prevents the creation of portfolios filled with funds for which the search algorithm "got lost" and yielded very high alpha values. However, even for the OLS models, the alpha and beta restrictions provide some additional value over and above the back testing requirement (in terms of an increased point estimate but not to the degree that there is a statistically significant difference).

The paper has argued that the proposed back testing procedure helps identify funds whose estimated alphas are likely to be generated from model specification errors. An alternative hypothesis is that the procedure finds funds whose alphas are unusually high or low because of random sampling error. If the former hypothesis is true then a fund in the active pool should be more likely than not to remain there over the next few months. If it does really have a positive alpha then (on average) it should outperform the market and remain in the active pool. Sampling error, however, implies a very short stay in the active pool. At its most extreme, if the model simply detects sampling errors then funds should be as likely to remain in the active pool as they are to remain out. This can be tested by simply looking at the autocorrelation between a fund's location in or out of the active pool in period t and in period $t + 1$. On average, the autocorrelation coefficient is 0.706 with a p-value of 0.004. If sampling error was the sole explanation then the autocorrelation statistic should be close to zero.

3. Fund Picks across Models

While it appears that appropriate back testing leads to better alpha forecasts it is also possible that both the Kalman filter and OLS models produce very similar rankings. If that is the case then clearly one should limit future work to the OLS model as it is far simpler to calculate. To address this possibility

Table VI. Out of sample performance (no parameter forecast filters, with back testing)

For all domestic equity mutual funds having at least 5 years of monthly return data, both the Kalman and the OLS models are used to forecast a fund's alpha. Starting in January 1970 until December 2002, the 1-factor and 4-factor OLS models (OLS 1 and OLS 4), and 1-factor and 4-factor Kalman models (Kal 1 and Kal 4) independently sort funds into 10 deciles based on forecasted alphas. Decile 10 contains funds with highest forecasted alphas. For each model, 10 equally weighted portfolios are constructed from the stocks within the 10 deciles. Each portfolio is then held for 1 month and rebalanced at the beginning of the next month. For any fund to be included in any decile in the previous month the out of sample forecasted alpha and the difference between the realized excess return and the market return must have the same sign. Also constructed are equally weighted portfolios containing the 20, 10, and 5 funds with the highest alpha forecasts by each model. Panel A reports the excess returns from the four-factor model, and Panel B from the eight-factor model.

Decile	OLS 1	OLS 4	Kal 1	Kal 4	OLS 1	OLS 4	Kal 1	Kal 4
	A1. Four-factor monthly alpha				A2. *R*-ratio			
1	−0.0017	−0.0024	−0.0012	−0.0017	(−1.20)	(−2.78)	(−1.49)	(−2.12)
2	−0.0017	−0.0014	−0.0012	−0.0013	(−1.67)	(−2.27)	(−2.05)	(−2.34)
3	−0.0010	−0.0010	−0.0006	−0.0012	(−2.90)	(−1.75)	(−1.13)	(−2.45)
4	−0.0013	−0.0010	−0.0003	−0.0013	(−1.83)	(−1.93)	(−0.59)	(−2.87)
5	−0.0003	−0.0007	−0.0012	−0.0009	(−1.06)	(−1.22)	(−2.72)	(−1.99)
6	−0.0004	−0.0011	−0.0006	−0.0006	(−0.54)	(−2.08)	(−1.33)	(−1.34)
7	−0.0000	−0.0009	0.0007	0.0002	(−0.29)	(−1.73)	(1.43)	(0.32)
8	0.0002	0.0004	−0.0003	0.0005	(−0.16)	(0.66)	(−0.57)	(0.95)
9	0.0010	0.0012	0.0009	0.0006	(1.09)	(1.84)	(1.32)	(1.01)
10	0.0021	0.0025	0.0016	0.0000	(0.85)	(2.67)	(1.89)	(0.06)
Top 20	0.0004	0.0030	0.0008	0.0002	(1.09)	(2.68)	(0.75)	(0.22)
Top 10	−0.0001	0.0044	−0.0004	−0.0002	(1.03)	(3.22)	(−0.31)	(−0.18)
Top 5	0.0003	0.0052	−0.0009	−0.0022	(0.66)	(2.90)	(−0.51)	(−1.34)
	B1. Eight−factor monthly alpha				B2. *T*-ratio			
1	−0.0016	−0.0023	−0.0014	−0.0017	(−1.21)	(−2.55)	(−1.59)	(−2.01)
2	−0.0017	−0.0013	−0.0011	−0.0012	(−1.72)	(−2.03)	(−1.74)	(−2.07)
3	−0.0009	−0.0010	−0.0006	−0.0012	(−2.40)	(−1.68)	(−1.03)	(−2.22)
4	−0.0013	−0.0009	−0.0002	−0.0013	(−1.39)	(−1.61)	(−0.44)	(−2.67)
5	−0.0003	−0.0003	−0.0011	−0.0009	(−0.63)	(−0.60)	(−2.37)	(−1.87)
6	−0.0004	−0.0008	−0.0008	−0.0006	(−0.29)	(−1.43)	(−1.55)	(−1.18)
7	0.0002	−0.0010	0.0009	0.0000	(0.08)	(−1.72)	(1.56)	(0.03)
8	0.0003	0.0003	−0.0002	0.0004	(0.05)	(0.48)	(−0.37)	(0.67)
9	0.0014	0.0014	0.0009	0.0005	(1.20)	(2.09)	(1.31)	(0.74)
10	0.0026	0.0026	0.0018	0.0003	(1.19)	(2.57)	(1.96)	(0.37)
Top 20	0.0017	0.0034	0.0014	0.0003	(1.45)	(2.89)	(1.28)	(0.38)
Top 10	0.0010	0.0046	0.0003	0.0002	(1.19)	(3.19)	(0.20)	(0.13)
Top 5	0.0013	0.0053	−0.0003	−0.0024	(0.76)	(2.77)	(−0.17)	(−1.35)

Table VII examines the degree to which the models select the same funds. Panel A reports the "Common Ratio." To construct this number, funds are first ranked by each model. Next the union and intersection of the funds picked by models i and j for a particular decile are collected. The reported figure is the intersection divided by the union. Thus, for example, consider the row for decile ten and the figure of 0.43. This number implies that of all the funds ranked in the top decile by either the one factor OLS or Kalman model 43% were selected by both models. Note that both models are likely to agree on which funds are particularly good or bad. Even so, by combining models one can expand the set of funds for which one can potentially predict performance by at least 50% to 60%.

Table VII Panels B and C look at the fraction of funds in the active pools (i.e., those funds that are accepted by the filters) based upon either one model (Panel B) or a pair of models (Panel C). Take, for example, the 0.051 figure in Panel A in the decile 10 row under Model 1. This implies that the one factor OLS model had 51% of all funds in its active pool. That, of course, means 5.1% of all funds were both in the active pool and ranked by the model in the top decile. Panel C shows what happens to the active ratio when more than one model is employed. In this case the decile numbers are no longer one-tenth of the funds in the active pool due to picks common across models. However, the important point is that by using multiple models the portion of all funds that are in the top decile increases from about 5% to 7.9% or more. Thus, by using two models simultaneously and back testing one obtains a top decile that contains nearly as many funds as one would have had by simply ranking all funds with a single model without back testing. More importantly, however, without the back test the resulting decile portfolios provide much less evidence that managerial talent can be detected. When combined, these results indicate that a wide array of funds can predictably outperform the market if one is willing to use a variety of models to find them.

If the Kalman and OLS models focus on different aspects of managerial talent then funds selected by multiple models might perform better than those selected by just one model. In addition, using two models can help reduce the level of estimation error as model specific noise may cancel. Table VIII examines this question by looking at the returns from a portfolio that selects only among the funds picked by a pair of models. Thus, when looking at models i and j, a fund is only selected if it is placed in the same decile by both models. The returns produced by these funds are in fact somewhat higher than those from the population as a whole. The top decile portfolio produces an excess return of nearly 6% for three out of the four possible model pairs. The one exception occurs for those funds picked by the two single factor models. Nevertheless, even here the point estimate increases somewhat from the returns

Table VII. Common fund selections across models

For all domestic equity mutual funds having at least 5 years of monthly return data, both the Kalman and the OLS models are used to forecast a fund's alpha. Starting in January 1970 until December 2002, the 1-factor and 4-factor OLS models (OLS 1 and OLS 4, when subscripted O1 and O4 respectively), and 1-factor and 4-factor Kalman models (Kal 1 and Kal 4, when subscripted K1 and K4 respectively) independently sort funds into 10 deciles based on forecasted alphas. Decile 10 contains funds with highest forecasted alphas. For each model, 10 equally weighted portfolios are constructed from the stocks within the 10 deciles. Each portfolio is then held for 1 month and rebalanced at the beginning of the next month. For any fund to be included in any decile the following must be true: (1) the absolute value of forecasted alpha must be less than 2% per month, (2) the beta must be greater than 0 but less than 2, and (3) in the previous month the out of sample forecasted alpha and the difference between the realized return and the market return must have the same sign. Also constructed are equally weighted portfolios containing the 10 and 5 funds with the highest alpha forecasts by each model. Panel A reports the average ratio of firms that are selected either by Models i or j (reported as $\eta_{i,j}$). The row "9 and 10" reports the common ratio for portfolios that combine the decile 9 and 10 model-portfolios. Panel B reports the active ratio (active funds divided by all available funds) selected by each model for the selected deciles. Panel C reports the active ratio for funds selected by either Model i or Model j (reported as $D_{i,j}$).

A.	Common ratio			
	$\eta_{O1,K1}$	$\eta_{O4,K4}$	$\eta_{O4,K1}$	$\eta_{O1,K4}$
1	0.46	0.39	0.34	0.31
2	0.25	0.20	0.17	0.16
3	0.20	0.16	0.12	0.12
4	0.18	0.13	0.10	0.09
5	0.17	0.12	0.09	0.08
6	0.17	0.11	0.09	0.07
7	0.18	0.13	0.10	0.08
8	0.20	0.17	0.13	0.12
9	0.26	0.23	0.18	0.16
10	0.43	0.38	0.34	0.29
Top 10	0.34	0.28	0.26	0.22
Top 5	0.27	0.20	0.21	0.17
9 and 10	0.35	0.31	0.26	0.23
B.	Active ratio for different models			
	OLS 1	OLS 4	Kal 1	Kal 4
10	0.051	0.051	0.050	0.049
9 and 10	0.101	0.102	0.100	0.098
C.	Active ratio across models			
	$D_{O1,K4}$	$D_{O4,K4}$	$D_{O4,K1}$	$D_{O1,K4}$
9	0.087	0.089	0.092	0.092
10	0.079	0.081	0.084	0.085
9 and 10	0.166	0.170	0.175	0.177

found in Table IV. Comparing the top performing model in Table IV with the top performing model pair in Table VIII yields an increased point estimate of 12 basis points per month. For the combination of the one factor OLS with the four factor Kalman the increased return is so large that it is statistically different from the average return produced by the two models alone (t-statistic of 2.36). Overall, it appears that top decile funds selected by more than one model tend to produce higher returns than funds selected by just one model. This is at least consistent with the hypothesis that various models pick up on different aspects of managerial ability.

Even though the number of funds selected by any pair of models is a rather modest fraction of each model's selections it is possible that all of the excess returns come from that group. If this is so, it would imply that only a very small number of fund managers can produce above market returns and that one can find them by seeing if they are selected by multiple models. To test this hypothesis Table IX examines the complement of the funds in Table VIII. For a fund to be included in a Table IX portfolio for each model pair it must be selected by one for inclusion in the top decile but not by the other. Given the previous results, it is not surprising that the portfolio returns produced by this group are somewhat lower than those found in Table IV. Nevertheless, every model's top decile continues to produce statistically significant positive alpha returns. Also, every model other than the one factor OLS model does so for the ninth decile portfolio as well. This reinforces the earlier conclusion that no single model appropriately describes the statistical properties of every single fund. It also buttresses the paper's thesis that a significant number of fund managers can produce predictable above market returns, but only if one uses a variety of statistical models to locate them, and also back tests each model for its use with each fund.

4. Tracking Error Filters

For a mutual fund to generate a nonzero alpha it must, by definition, avoid tracking an index. Nevertheless, funds that closely track an index may still generate nonzero estimated alphas. Various indices rise and fall relative to the overall market over, occasionally, long periods of time. Thus, a fund that tracks (for example) the S&P 500 will generate a positive alpha relative to the CAPM whenever large capitalization stocks outperform small capitalization stocks.

Since most active fund managers do not publicly admit that they purchase stocks that closely mirror those within an index it is necessary to estimate the degree to which this occurs. The solution followed here compares a

Table VIII. Out of sample performance: funds selected by multiple models

For all domestic equity mutual funds having at least 5 years of monthly return data, both the Kalman and the OLS models are used to forecast a fund's alpha. Starting from January 1970 until December 2002, the 1-factor and 4-factor OLS models (O1 and O4), and 1-factor and 4-factor Kalman models (K1 and K4) independently sort funds into 10 deciles based on their forecasted alphas. Decile 10 contains funds with the highest forecasted alphas. For each model, 10 equally weighted portfolios are constructed from the stocks within the 10 deciles. For any fund to be included in any decile the following must be true: (1) the absolute value of forecasted alpha must be less than 2% per month, (2) the beta must be greater than 0 but less than 2, and 3) in the previous month the out of sample forecasted alpha and the difference between the realized return and the market return must have the same sign. Next, new decile portfolios $C_{i,j}$ are constructed by equally investing into funds that are commonly selected by both model i and j for each decile. Each portfolio is then held for 1 month and rebalanced at the beginning of the next month. If no common funds are available for one specific decile during month t, then this decile invests in t-bills for that month. Also constructed are equally weighted portfolios containing the intersection of the 20, 10 and 5 funds with the highest alpha forecasts by the two models. Finally, Panel A reports monthly excess returns and the Sharpe Ratios for these portfolios. Panel B reports the 4-factor (MKT, SMB, HML, and MOM) adjusted monthly returns.

A.	Monthly excess return				Monthly Sharpe Ratio			
Decile	$C_{O1,K1}$	$C_{O4,K4}$	$C_{O4,K1}$	$C_{O1,K4}$	$C_{O1,K1}$	$C_{O4,K4}$	$C_{O4,K1}$	$C_{O1,K4}$
1	0.0005	−0.0002	0.0000	−0.0003	0.0104	−0.0039	0.0000	−0.0054
2	0.0025	0.0020	0.0026	0.0024	0.0479	0.0408	0.0514	0.0484
3	0.0029	0.0019	0.0026	0.0028	0.0602	0.0417	0.0557	0.0552
4	0.0039	0.0021	0.0030	0.0028	0.0851	0.0459	0.0648	0.0596
5	0.0041	0.0039	0.0040	0.0038	0.0917	0.0855	0.0915	0.0887
6	0.0044	0.0036	0.0042	0.0062	0.1061	0.0791	0.0951	0.1445
7	0.0045	0.0048	0.0043	0.0048	0.1067	0.1088	0.0996	0.1075
8	0.0037	0.0053	0.0052	0.0055	0.0855	0.1176	0.1177	0.1209
9	0.0065	0.0054	0.0056	0.0052	0.1384	0.1134	0.1189	0.1113
10	0.0078	0.0085	0.0079	0.0089	0.1669	0.1654	0.1597	0.1768
Top 20	0.0068	0.0079	0.0082	0.0077	0.1366	0.1372	0.1519	0.1428
Top 10	0.0076	0.0085	0.0079	0.0101	0.1434	0.1405	0.1399	0.1773
Top 5	0.0071	0.0079	0.0060	0.0079	0.1274	0.1325	0.1023	0.1385
B.	Four-factor monthly alpha				T-ratio			
1	−0.0024	−0.0044	−0.0031	−0.0040	(−2.53)	(−3.78)	(−2.61)	(−3.71)
2	−0.0017	−0.0023	−0.0016	−0.0019	(−2.16)	(−2.73)	(−2.15)	(−2.38)
3	−0.0015	−0.0018	−0.0014	−0.0019	(−2.06)	(−2.75)	(−1.91)	(−2.32)
4	0.0002	−0.0018	−0.0010	−0.0013	(0.28)	(−2.72)	(−1.60)	(−1.78)
5	0.0000	−0.0002	0.0001	−0.0002	(0.07)	(−0.32)	(0.12)	(−0.28)
6	−0.0001	−0.0009	−0.0002	0.0024	(−0.22)	(−1.39)	(−0.29)	(2.74)
7	0.0003	0.0007	0.0005	0.0007	(0.45)	(0.99)	(0.71)	(0.92)
8	−0.0008	0.0007	0.0006	0.0011	(−1.17)	(1.10)	(0.85)	(1.50)
9	0.0015	0.0008	0.0012	0.0003	(1.74)	(1.11)	(1.29)	(0.31)
10	0.0038	0.0049	0.0045	0.0048	(4.02)	(5.08)	(4.38)	(4.35)
Top 20	0.0024	0.0048	0.0050	0.0036	(2.14)	(3.63)	(3.67)	(2.88)
Top 10	0.0036	0.0062	0.0041	0.0075	(2.41)	(3.92)	(2.47)	(4.53)
Top 5	0.0028	0.0054	0.0024	0.0055	(1.67)	(3.31)	(1.21)	(3.26)

Table IX. Out of sample performance: funds selected by only one model

For all domestic equity mutual funds having at least 5 years of monthly return data, both the Kalman and the OLS models are used to forecast a fund's alpha. Starting from January 1970 until December 2002, the 1-factor and 4-factor OLS models (O1 and O4), and 1-factor and 4-factor Kalman models (K1 and K4) independently sort funds into 10 deciles based on their forecasted alphas. Decile 10 contains funds with the highest forecasted alphas. For each model, 10 equally weighted portfolios are constructed from the stocks within the 10 deciles. For any fund to be included in any decile the following must be true: (1) the absolute value of forecasted alpha must be less than 2% per month, (2) the beta must be greater than 0 but less than 2, and (3) in the previous month the out of sample forecasted alpha and the difference between the realized return and the market return must have the same sign. Next, new decile portfolios $C_{i,j}$ are constructed by equally investing into funds that are selected by model i but not model j and that are selected by model j but not model i for each decile. Each portfolio is then held for 1 month and rebalanced at the beginning of the next month. If no disjoint funds are available for one specific decile during month t, then this decile invests in t-bills for that month. Also constructed are equally weighted portfolios containing the intersection of the 20, 10 and 5 funds with the highest alpha forecasts by the two models. Finally, Panel A reports monthly excess returns and the Sharpe Ratios for these portfolios. Panel B reports the 4-factor (MKT, SMB, HML, and MOM) adjusted monthly returns.

A.	Monthly excess return				Monthly Sharpe Ratio			
Decile	$C_{O1,K1}$	$C_{O4,K4}$	$C_{O4,K1}$	$C_{O1,K4}$	$C_{O1,K1}$	$C_{O4,K4}$	$C_{O4,K1}$	$C_{O1,K4}$
1	0.0014	0.0020	0.0013	0.0016	0.0266	0.0410	0.0265	0.0323
2	0.0020	0.0022	0.0021	0.0022	0.0401	0.0455	0.0434	0.0451
3	0.0022	0.0025	0.0025	0.0022	0.0464	0.0538	0.0523	0.0460
4	0.0031	0.0027	0.0031	0.0027	0.0677	0.0593	0.0678	0.0606
5	0.0035	0.0032	0.0032	0.0035	0.0809	0.0743	0.0725	0.0815
6	0.0042	0.0035	0.0037	0.0038	0.0976	0.0828	0.0865	0.0891
7	0.0051	0.0044	0.0047	0.0049	0.1192	0.1020	0.1089	0.1151
8	0.0052	0.0053	0.0050	0.0051	0.1203	0.1210	0.1148	0.1169
9	0.0060	0.0057	0.0061	0.0057	0.1325	0.1252	0.1345	0.1277
10	0.0067	0.0061	0.0074	0.0057	0.1416	0.1279	0.1550	0.1234
Top 20	0.0070	0.0061	0.0075	0.0058	0.1437	0.1199	0.1485	0.1208
Top 10	0.0075	0.0067	0.0082	0.0062	0.1501	0.1260	0.1579	0.1241
Top 5	0.0075	0.0064	0.0081	0.0066	0.1480	0.1143	0.1494	0.1290
B.	Four-factor monthly alpha				T-ratio			
1	−0.0023	−0.0023	−0.0028	−0.0024	(−2.74)	(−2.90)	(−3.72)	(−3.19)
2	−0.0018	−0.0020	−0.0020	−0.0017	(−2.72)	(−3.30)	(−3.25)	(−2.64)
3	−0.0017	−0.0018	−0.0016	−0.0019	(−2.86)	(−3.37)	(−3.15)	(−3.60)
4	−0.0009	−0.0016	−0.0010	−0.0015	(−1.84)	(−3.10)	(−1.97)	(−3.19)
5	−0.0006	−0.0008	−0.0009	−0.0006	(−1.24)	(−1.77)	(−1.96)	(−1.35)
6	−0.0002	−0.0008	−0.0007	−0.0005	(−0.47)	(−1.72)	(−1.52)	(−1.02)
7	0.0006	0.0000	0.0003	0.0004	(1.16)	(0.05)	(0.73)	(0.88)
8	0.0006	0.0009	0.0005	0.0005	(1.03)	(1.69)	(0.95)	(0.92)
9	0.0011	0.0014	0.0015	0.0012	(1.69)	(2.57)	(2.48)	(2.09)
10	0.0018	0.0020	0.0031	0.0011	(2.23)	(2.64)	(3.86)	(1.57)
Top 20	0.0023	0.0027	0.0036	0.0015	(2.54)	(2.94)	(3.97)	(1.90)
Top 10	0.0026	0.0036	0.0046	0.0018	(2.64)	(3.23)	(4.35)	(1.90)
Top 5	0.0025	0.0027	0.0040	0.0020	(2.09)	(2.10)	(3.21)	(1.74)

fund's returns to various indices one at a time. For each benchmark the standard deviation of the difference between its return and that of the fund's is calculated. The minimum standard deviation across all benchmarks is then defined as the fund's tracking error. Next funds are sequentially sorted. First, funds are sorted by their tracking error into terciles. Within each tercile funds are then sorted into groups based upon the forecasted alpha. Table X reports the results from the procedure with and without back testing.

There are two primary conclusions to be drawn from Table X. First, Panel A shows that tracking error statistics can help identify funds with true alphas. For the low tracking error funds (indexers) Panel A1 shows that there is little evidence to support the idea that they can produce above market returns out of sample. In contrast, consider the funds in the top two tracking error terciles that also have high in sample four factor OLS alphas. These funds do produce statistically significant out of sample alphas. The same holds true for the high tracking error funds with high Kalman filter alphas. This is perhaps not surprising since high tracking error funds are likely to be the most dynamic in terms of their factor loadings, which is what the Kalman filter model is adapted to handle.

The second conclusion from Table X is that back testing is essential. Filtering funds on tracking error alone does not yield reliably positive out of sample alphas. This can be seen in Panel B. For the most part the t-statistics in this panel are nowhere near the standard significance levels. In addition, the point estimates in Panel B are almost universally lower than the corresponding point estimates in Panel A, and by economically significant amounts as well.

Overall, it appears that tracking error statistics do help identify funds with the potential to outperform the market. But, they only help if one also filters funds for potential model specification errors with a procedure like the back testing algorithm used here.

5. Comparison with Carhart's Results

One goal of this paper is to examine the degree to which back testing can be used to find mutual funds with superior out of sample performance. Perhaps, though, the efficacy of the back test proposed here really derives from differences in the estimation procedure used here versus Carhart (1995, 1997). In both of his papers a fund is ranked so long as it has at least two and a half years of available data at the time a decision has to be made. Thus, the portfolios based upon 3 and 5 year training periods also include funds with

Table X. Four-factor model out of sample performance (with tracking error, 1985--2002)

For all domestic equity mutual funds having at least 5 years of monthly return data, both the Kalman and the OLS models are used to forecast a fund's alpha. Starting from January 1985 until December 2002, in each month t all available funds are independently sorted into 3 tracking error groups. Within each group top 20%, top 20, 10, and 5 funds are then selected by each model according to forecasted alphas. Tracking error is calculated as the standard deviation of the difference between fund return and its benchmark return from $t-60$ to $t-1$. The benchmark is selected from S&P500, Barra Growth, Barra Value, and Russell 2000, as the one with the lowest tracking error in the previous 60 months. The models are the mean premium of the fund return in excess of its benchmark in the previous 60 months (BM), the 1-factor and 4-factor OLS models, and 1-factor and 4-factor Kalman models, respectively. Each portfolio is held for 1 month and rebalanced at the beginning of the next month. Panel A reports 4-factor adjusted return and corresponding t-statistics for funds satisfying: (1) the absolute value of forecasted alpha must be less than 2% a month, (2) the beta must be greater than 0 but less than 2, and (3) the forecasted alpha and the difference between the realized return and the market return in the previous month must have the same sign. In Panel B the last condition (back testing) is dropped.

	BM	OLS 1	OLS 4	Kal 1	Kal 4	BM	OLS 1	OLS 4	Kal 1	Kal 4
	Four-factor monthly alpha					T-ratio for four-factor adjusted return				
	A1. Low tracking error funds, with back testing									
Top 20%	−0.0002	−0.0000	0.0004	0.0001	0.0002	(−0.61)	(−0.06)	(0.76)	(0.12)	(0.35)
Top 20	−0.0002	−0.0005	0.0006	−0.0003	−0.0001	(−0.29)	(−0.81)	(0.83)	(−0.47)	(−0.24)
Top 10	−0.0000	−0.0007	0.0008	−0.0000	0.0002	(−0.02)	(−0.87)	(0.95)	(−0.01)	(0.30)
Top 5	−0.0013	−0.0003	0.0011	−0.0004	0.0001	(−1.25)	(−0.38)	(1.16)	(−0.44)	(0.15)
	A2. Media tracking error funds, with back testing									
Top 20%	−0.0005	0.0009	0.0018	0.0010	0.0009	(−0.74)	(1.22)	(2.39)	(1.31)	(1.31)
Top 20	0.0001	0.0009	0.0020	0.0013	0.0011	(0.18)	(1.06)	(2.15)	(1.46)	(1.44)
Top 10	0.0011	0.0016	0.0019	0.0016	0.0014	(1.21)	(1.62)	(1.97)	(1.49)	(1.49)
Top 5	0.0011	0.0025	0.0026	0.0025	0.0016	(0.99)	(2.20)	(2.19)	(1.91)	(1.48)
	A3. High tracking error funds, with back testing									
Top 20%	−0.0007	0.0013	0.0032	0.0027	0.0018	(−0.59)	(0.84)	(2.07)	(2.07)	(1.48)
Top 20	0.0003	0.0024	0.0038	0.0030	0.0015	(0.22)	(1.43)	(1.94)	(1.93)	(1.06)
Top 10	0.0003	0.0024	0.0054	0.0033	0.0023	(0.16)	(1.19)	(2.47)	(1.98)	(1.40)
Top 5	0.0015	0.0011	0.0048	0.0020	0.0020	(0.75)	(0.50)	(1.96)	(1.04)	(0.96)
	B1. Low tracking error funds, without back testing									
Top 20%	−0.0002	−0.0005	0.0002	−0.0003	−0.0004	(−0.61)	(−1.10)	(0.37)	(−0.58)	(−0.87)
Top 20	−0.0002	−0.0004	0.0007	−0.0004	−0.0007	(−0.29)	(−0.78)	(0.98)	(−0.63)	(−1.14)
Top 10	−0.0000	−0.0005	0.0013	−0.0005	−0.0009	(−0.02)	(−0.78)	(1.50)	(−0.70)	(−1.42)
Top 5	−0.0013	−0.0005	0.0007	−0.0001	−0.0010	(−1.25)	(−0.58)	(0.71)	(−0.13)	(−1.33)
	B2. Media tracking error funds, without back testing									
Top 20%	−0.0005	−0.0004	0.0003	0.0001	−0.0000	(−0.74)	(−0.62)	(0.42)	(0.20)	(−0.00)
Top 20	0.0001	0.0002	0.0010	0.0003	0.0005	(0.18)	(0.27)	(1.25)	(0.41)	(0.70)
Top 10	0.0011	0.0013	0.0020	0.0004	0.0008	(1.21)	(1.37)	(2.25)	(0.41)	(0.92)
Top 5	0.0011	0.0016	0.0019	0.0018	0.0008	(0.99)	(1.37)	(1.65)	(1.48)	(0.79)
	B3. High tracking error funds, without back testing									
Top 20%	−0.0007	−0.0015	0.0004	0.0005	0.0004	(−0.59)	(−1.34)	(0.39)	(0.47)	(0.41)
Top 20	0.0003	−0.0017	0.0020	0.0011	−0.0008	(0.22)	(−1.04)	(1.12)	(0.72)	(−0.56)
Top 10	0.0003	−0.0025	0.0020	0.0007	−0.0004	(0.16)	(−1.30)	(0.98)	(0.39)	(−0.20)
Top 5	0.0015	−0.0036	0.0020	0.0010	−0.0012	(0.75)	(−1.55)	(0.95)	(0.50)	(−0.50)

shorter training periods.[13] Once the funds are ranked they are then held for 1 year. This contrasts with the 1 month used up to now in this study. While reducing the data requirements for a fund's inclusion to only two and a half years increases the pool of available funds, it also potentially increases the impact of misspecification error.[14]

As previously discussed, many of the estimated fund parameters are of questionable reliability even when a 5 year training period is used. A shorter training period can potentially produce even less reliable estimates, especially in a multifactor model where several parameters must be accurately forecast. This can be seen in Table XI which compares the out of sample returns for sorted alpha portfolios using 3 and 5 year training periods and 1 year holding periods. Panel A shows the standard four factor risk-adjusted returns (on the left) along with their *t*-statistics (on the right). As one can see, when funds are sorted based upon the alphas from a 3 year training period, the top decile alpha returns are quite modest and one cannot reject the null hypothesis that they are zero at any reasonable level. However, when a 5 year training period is used the results change dramatically. Now, the top decile earns an abnormal return of 0.12% a month which is statistically significant at the 5% level. If, instead, the portfolio concentrates on the top funds (20, 10, or 5) the results improve even further.

As with earlier tests it is possible that the results in Table XI Panel A arise from the model's ability to exploit the stale pricing phenomenon inherent in reported fund returns. To test this Panel B regresses the alpha sorted portfolio returns on the eight-factor model (the four current and four past factors). The realized returns do in fact drop somewhat. Under the eight-factor model none of the decile portfolios produce statistically significant out of sample returns. Among the top fund portfolios only the one using the best 20 funds retains its statistically significant alpha, although the point estimate drops by three basis points per month. Thus, for the 1 year holding period used by Carhart the evidence also indicates that at least some of the improved performance is due to stale pricing rather than managerial ability. Overall, though, there is strong evidence that using a 5 year training period to estimate the model produces

[13] While Carhart (1997) analysis reports the results from portfolios formed on the basis of a 3 year training period, his unpublished (1995) dissertation includes those from a 5 year period as well. Since his results are similar across the two training periods the discussion that follows should be taken to encompass both techniques.

[14] However, reduced data requirements do help ameliorate the omitted data problem documented by Elton et al. (2001). To see if omitted data may be responsible for the results in this paper a number of tests were conducted. Overall, the evidence indicates that omitted data can account for *at most* half a basis point per month. Post 1985 its impact is close to zero, while the point estimates for the out of sample returns are similar to that of the whole sample.

Table XI. Out of sample performance: 12-month holding period

For all domestic equity mutual funds having at least 3 years or 5 years of monthly return data (depending on the length of the training period), the OLS models are used to forecast a fund's alpha. Starting from January 1970 until December 2002, the 1-factor and 4-factor OLS models, OLS 1 and OLS 4 respectively, independently sort funds into 10 deciles based on forecasted alphas, which are estimated over the prior 3 or 5 years. Decile 10 contains funds with the highest forecasted alphas. For each model, 10 equally weighted portfolios are constructed from stocks within the 10 deciles. Each portfolio is held for 12 months and rebalanced at the beginning of the next year. Finally, the table also constructs equally weighted portfolios containing the 20, 10, and 5 funds with the highest alpha forecast by each model. Panel A reports the 4-factor (MKT, SMB, HML, and MOM) adjusted monthly returns realized during the entire period and the corresponding t-statistics. In Panel B, the portfolio returns are risk-adjusted by both four factors and 1-month-lagged 4-factor returns (the 8-factor model).

Decile	OLS 1	OLS 2	OLS 1	OLS 4	OLS 1	OLS 4	OLS 1	OLS 4
	3 years		5 years		3 years		5 years	
	A1. Four-factor monthly alpha				A2. T-ratio			
1	−0.0011	−0.0013	−0.0006	−0.0019	(−1.26)	(−1.72)	(−0.83)	(−2.64)
2	−0.0004	−0.0011	−0.0001	−0.0007	(−0.70)	(−2.49)	(−0.23)	(−1.63)
3	−0.0002	−0.0007	−0.0001	−0.0003	(−0.29)	(−1.78)	(−0.25)	(−0.84)
4	−0.0004	−0.0007	0.0003	−0.0005	(−1.06)	(−2.03)	(0.65)	(−1.18)
5	−0.0004	−0.0006	−0.0002	−0.0004	(−1.12)	(−1.64)	(−0.45)	(−1.17)
6	−0.0005	−0.0004	−0.0004	−0.0002	(−1.57)	(−1.05)	(−1.04)	(−0.45)
7	−0.0003	−0.0002	−0.0004	−0.0001	(−1.00)	(−0.52)	(−0.92)	(−0.16)
8	−0.0002	−0.0001	−0.0005	0.0002	(−0.49)	(−0.31)	(−0.94)	(0.43)
9	−0.0004	0.0003	0.0002	0.0007	(−0.73)	(0.69)	(0.39)	(1.37)
10	−0.0003	0.0004	−0.0002	0.0012	(−0.33)	(0.52)	(−0.29)	(2.05)
Top 20	−0.0003	0.0008	−0.0004	0.0019	(−0.20)	(0.61)	(−0.40)	(2.48)
Top 10	−0.0001	0.0006	−0.0002	0.0025	(−0.04)	(0.38)	(−0.16)	(2.14)
Top 5	−0.0006	0.0009	0.0000	0.0028	(−0.29)	(0.44)	(0.02)	(1.60)
	B1. Eight-factor monthly alpha				B2. T-ratio			
1	−0.0011	−0.0012	−0.0009	−0.0019	(−1.14)	(−1.54)	(−1.07)	(−2.50)
2	−0.0002	−0.0009	−0.0001	−0.0008	(−0.24)	(−1.93)	(−0.13)	(−1.59)
3	−0.0002	−0.0006	−0.0001	−0.0002	(−0.29)	(−1.33)	(−0.18)	(−0.60)
4	−0.0003	−0.0007	0.0004	−0.0005	(−0.78)	(−1.75)	(0.98)	(−1.08)
5	−0.0004	−0.0005	−0.0003	−0.0004	(−1.01)	(−1.29)	(−0.61)	(−0.99)
6	−0.0005	−0.0004	−0.0003	−0.0001	(−1.51)	(−1.05)	(−0.84)	(−0.29)
7	−0.0003	−0.0001	−0.0005	−0.0002	(−0.89)	(−0.30)	(−1.00)	(−0.49)
8	−0.0001	−0.0000	−0.0004	0.0003	(−0.12)	(−0.04)	(−0.75)	(0.62)
9	−0.0002	0.0005	0.0002	0.0007	(−0.30)	(0.91)	(0.37)	(1.33)
10	0.0000	0.0006	−0.0003	0.0010	(0.02)	(0.83)	(−0.36)	(1.57)
Top 20	0.0001	0.0012	−0.0005	0.0016	(0.06)	(0.84)	(−0.45)	(1.95)
Top 10	0.0005	0.0011	−0.0005	0.0019	(0.28)	(0.61)	(−0.36)	(1.55)
Top 5	0.0001	0.0013	−0.0002	0.0019	(0.03)	(0.60)	(−0.08)	(1.04)

substantially better out of sample predictions than one obtains by using a 3 year training period.

6. Berk and Green Hypothesis

To what degree can one ever expect to find funds that can reliably outperform the market for over a year? Berk and Green (2004) proposed a model in which some managers have the ability to generate positive alphas but can only do so with a finite level of investable funds. If they receive inflows that exceed their capacity to generate positive alphas they simply index the excess funds. Thus, as a positive alpha fund receives inflows (beyond its capacity to generate a positive alpha) its overall alpha declines. Since investors have no way of knowing which funds are managed by high or zero alpha managers they end up return chasing, as good past returns provide a positive signal that the manager may be talented. However, these inflows then degrade the fund's future performance. Because managed funds are costly to run the net result is that in equilibrium investors expect to earn a zero alpha on their investment. That is, the expected alpha conditional on the funds under management equals zero net of costs for each and every fund. Investors do not withdraw their money as that would leave the fund with a positive expected alpha and do not add any further funds as that would leave the fund with a negative expected alpha.

This paper finds that it is possible to find funds with positive expected alphas. Initially, these results may appear to contradict the Berk and Green hypothesis. However, this is only true if one looks solely at the published model. In that model investors can discern one fund's alpha from another's at no cost. But, this is impossible. Indeed, the algorithms proposed here require a great deal of sophistication and access to costly databases. Modifying the Berk and Green model to accommodate this reality then implies that it is possible to find funds that produce positive out of sample alphas. However, such out performance should short lived. Presumably as a fund generates additional above market returns the cost of discovering its manager's ability will fall. This then attracts additional funds until at some point the fund's expected alpha does fall to zero.

If Berk and Green (2004) are right, then a 1 year holding period is likely to include long stretches past the point where a fund manager's ability can manifest itself. In such an environment a potentially superior technique for finding managerial ability is to reduce the holding period to 1 month. Compare Table XI with annual rebalancing to Panels A and B in Table II with monthly rebalancing. With monthly rebalancing the four factor OLS model produces

Table XII. Dissipation of top decile alphas (with parameter forecast filters, with back testing)

For all domestic equity mutual funds having at least 5 years of monthly return data, both the Kalman and the OLS models are used to forecast a fund's alpha. Starting from 1970 until 2002, the 1-factor and 4-factor OLS models (OLS 1 and OLS 4), and 1-factor and 4-factor Kalman models (Kal 1 and Kal 4) independently sort funds into 10 deciles based on forecasted alphas in each month. Decile 10 contains funds with the highest alpha forecasts. The table also constructs model by model portfolios containing the 20, 10 and 5 funds with the highest alphas forecasts. Next, for each month t, once top deciles are created, Panel A calculates the average number of months for funds within a decile to maintain positive forecasted alpha in the future according to corresponding model (thus if on average at month $t + 2$ funds begin to have the first negative alpha forecast then the length will be 1). Panel A then reports the time-series mean and standard deviation of these lengths. Panel B reports the similar statistics for the number of months for an average fund within these deciles to have positive return in excess of market. For any fund to be included in any decile the following must be true: (1) the absolute value of forecasted alpha must be less than 2% per month; (2) the beta must be greater than 0 but less than 2; and (3) in the previous month the out of sample forecasted alpha and the difference between the realized return and the market return must have the same sign.

Decile	OLS 1	OLS 4	Kal 1	Kal 4	OLS 1	OLS 4	Kal 1	Kal 4
	Number of months				Standard deviation			
	A. Number of months with positive forecasted alpha.							
Decile 9	3.03	2.73	4.33	4.76	(2.47)	(1.90)	(1.72)	(1.57)
Decile 10	2.39	2.00	4.21	4.26	(2.23)	(1.70)	(1.59)	(1.62)
Top 20	2.42	1.95	4.22	4.31	(2.38)	(1.66)	(1.45)	(1.37)
Top 10	2.16	1.66	4.17	4.02	(2.61)	(2.14)	(1.85)	(1.76)
Top 5	1.87	1.62	3.96	3.96	(3.01)	(3.03)	(2.39)	(2.48)
	B. Number of months with positive above market return.							
Decile 9	2.15	2.15	2.15	2.14	(0.71)	(0.67)	(0.72)	(0.67)
Decile 10	2.26	2.31	2.23	2.18	(0.76)	(0.76)	(0.76)	(0.64)
Top 20	2.26	2.35	2.22	2.17	(0.77)	(0.88)	(0.74)	(0.64)
Top 10	2.30	2.42	2.25	2.20	(0.88)	(1.09)	(0.82)	(0.74)
Top 5	2.32	2.41	2.24	2.22	(1.03)	(1.21)	(0.93)	(0.93)

higher risk-adjusted returns under both the four and eight-factor models. Furthermore, this time the eight-factor returns are statistically significant for both the top decile and the Top 20 portfolios. This provides at least some evidence for Berk and Green's hypothesis and for the use of shorter rebalancing periods if one wishes to detect managerial ability.

One can also directly test whether the Berk and Green model with search costs holds. This can be done by examining the number of months over which

a fund with a currently positive alpha tends to have a positive alpha going forward. Under their model one expects alphas to decline reasonably rapidly and, as Table XII shows, this is what appears to happen. Depending on the model one uses a fund with an alpha lying in the top two deciles will tend to see its forecasted alpha degrade to zero within 2 to 4 months. The standard deviations on this figure are in the range of 1.5 and thus few funds retain their positive alphas for much longer than 5 months.

7. Other Back Tests

The back testing procedure used so far has deliberately eschewed using a model's in sample beta estimates and diagnostic statistics. While this helps ensure the procedure's robustness, one might postulate that these in sample results can be employed to produce even better results than those displayed so far. This section tests that hypothesis by looking at one back test that employs the estimated model betas and another that uses the estimated alpha's t-statistic. In general, the results are inferior to those using the simpler one period back test analyzed so far.

Table XIII displays the results from using a back test that employs a model's estimated factor loadings. Unlike the back testing procedure in the previous sections, this variant adjusts a fund's previous period return using the model's beta estimates. Thus, the model is estimated with data up to month $t-2$. The resulting betas are then used to risk adjust the fund's realized return in period $t-1$. If the realized risk-adjusted return then matches the model's alpha prediction in sign the fund then goes into the active pool for period t. Using this procedure, across every model, the top decile fund portfolios produce point estimates lower than those in Table IV (in which the active pools do not depend on a model's beta estimates). Furthermore, for all but the single factor Kalman model the ninth decile fund returns are no longer statistically different from zero. Looking at the top 20, 10, or 5 funds the models in general see a significant degradation in predictive ability and for many of the portfolios the returns are no longer different from zero at the standard significance levels.

The above results reinforce those from Section 1. For high and low alpha funds the beta estimates are simply not very accurate. Using them to "help" pick funds therefore seems unlikely to generate consistent above or below market returns. Not necessarily because such funds do not exist, but because the in sample alpha sorts lead to estimation error sorts on the predicted betas which then feed back into poor alpha forecasts. To some degree this is expected given that Bossaerts and Hillion (1999) and Goyal and Welch (2006) both find that model selection criteria based on in sample statistics not only

Table XIII. Out of sample performance (parameter forecast filters, with back testing using model beta forecasts)

For all domestic equity mutual funds having at least 5 years of monthly return data, both the Kalman and the OLS models are used to forecast a fund's alpha. Starting from January 1970 until December 2002, Models 1 to 4 (the 1-factor and 4-factor OLS models, and 1-factor and 4-factor Kalman models, respectively) independently sort funds into 10 deciles based on forecasted alphas. Decile 10 contains funds with the highest forecasted alphas. For each model 10 equally weighted portfolio are constructed from stocks within the 10 deciles. At the beginning of month t, funds are sorted according to predicted alphas in month t. Each portfolio is held for 1 month and rebalanced at the beginning of the next month. For any fund to be included in any decile: (1) the absolute value of forecasted alpha must be less than 2% a month; (2) the beta must be greater than 0 but less than 2; and (3) the forecasted alpha and the risk adjusted return when calculated with the model's forecasted betas in the previous month must have the same sign. Finally, the table also constructs equally weighted portfolios containing the 20, 10 and 5 funds with the highest alphas forecast by each model.

Decile	OLS 1	OLS 4	Kal 1	Kal 4	OLS 1	OLS 4	Kal 1	Kal 4
	Four-factor monthly alpha				T-ratio			
1	−0.0021	−0.0028	−0.0023	−0.0035	(−2.41)	(−3.29)	(−2.69)	(−4.63)
2	−0.0017	−0.0022	−0.0016	−0.0011	(−2.45)	(−4.02)	(−2.52)	(−2.19)
3	−0.0015	−0.0011	−0.0013	−0.0005	(−2.42)	(−2.20)	(−2.23)	(−1.08)
4	−0.0007	−0.0012	−0.0006	−0.0006	(−1.36)	(−2.53)	(−1.28)	(−1.39)
5	−0.0003	−0.0007	−0.0008	−0.0008	(−0.46)	(−1.32)	(−1.70)	(−2.14)
6	0.0000	−0.0008	−0.0008	−0.0004	(0.08)	(−1.78)	(−1.69)	(−1.03)
7	0.0002	−0.0006	0.0003	0.0000	(0.35)	(−1.30)	(0.68)	(0.02)
8	0.0004	0.0009	−0.0000	−0.0001	(0.72)	(1.75)	(−0.09)	(−0.26)
9	0.0007	0.0009	0.0016	0.0005	(1.03)	(1.75)	(2.52)	(0.98)
10	0.0012	0.0030	0.0026	0.0014	(1.36)	(3.90)	(2.74)	(1.98)
Top 20	0.0015	0.0037	0.0024	0.0011	(1.51)	(3.34)	(2.20)	(1.33)
Top 10	0.0023	0.0041	0.0025	0.0021	(1.80)	(3.10)	(1.80)	(1.93)
Top 5	0.0024	0.0039	0.0028	0.0010	(1.42)	(2.36)	(1.53)	(0.66)

fail to improve the out of sample portfolio returns but actually reduce them. While both studies concern time variation in the equity premium, the general problem of forecasting returns with a battery of possible models is similar to the issues explored here.

Another possible back test is to use the t-statistic on the estimated alphas. Intuitively, funds with high alphas but low t-statistics might be of limited use. This alternative procedure is tested in Table XIV. The results reported there use the in sample t values to help sort mutual funds. No other back tests or filters are imposed. The overall results are similar to that of the benchmark case in Table II. Among the four models tested, the four-factor OLS model is the only one for which t-statistic back testing produces above market fund-of-fund returns with any confidence. However, even its estimated performance

Table XIV. Out of sample performance sorted by *t*-statistics (no parameter forecast filters, no back testing)

For all domestic equity mutual funds having at least 5 years of monthly return data, both the Kalman and the OLS models are used to forecast a fund's alpha. Starting from January 1970, the 1-factor and 4-factor OLS models (OLS 1 and OLS 4) and 1-factor and 4-factor Kalman models (Kal 1 and Kal 4) independently sort funds into equally weighted portfolios based on the in-sample *t*-statistics of forecasted alphas. Decile 10 contains funds with the highest *t*-statistics. The top *X* portfolios contain an equally weighted portfolio of the *X* funds with the highest alphas. Each portfolio is held for 1 month and then rebalanced until December 2002. There are no restrictions on forecasted alphas and betas. Panel A reports the 4-factor (MKT, SMB, HML, and MOM) adjusted monthly returns realized during the entire period and the corresponding *t*-statistics. In Panel B, the portfolio returns are risk-adjusted by both the four factors and 1-month-lagged 4-factor returns (8 factors in total).

Decile	OLS 1	OLS 4	Kal 1	Kal 4	OLS 1	OLS 4	Kal 1	Kal 4
	A1. Four-factor monthly alpha				A2. *T*-ratio			
1	0.0004	−0.0006	−0.0007	−0.0015	(0.76)	(−1.34)	(−1.28)	(−2.59)
2	−0.0001	−0.0006	−0.0003	−0.0012	(−0.11)	(−1.31)	(−0.65)	(−2.16)
3	−0.0007	−0.0007	−0.0003	−0.0007	(−1.38)	(−1.38)	(−0.55)	(−1.40)
4	−0.0005	−0.0012	−0.0000	−0.0005	(−1.00)	(−2.39)	(−0.06)	(−1.04)
5	−0.0005	−0.0008	−0.0007	−0.0005	(−1.04)	(−1.67)	(−1.79)	(−1.18)
6	−0.0004	−0.0004	−0.0008	0.0002	(−0.88)	(−0.87)	(−1.92)	(0.47)
7	−0.0003	−0.0004	−0.0002	−0.0003	(−0.61)	(−0.82)	(−0.38)	(−0.63)
8	−0.0006	0.0002	−0.0007	0.0005	(−1.25)	(0.37)	(−1.38)	(1.07)
9	−0.0005	0.0011	−0.0005	0.0004	(−1.03)	(2.25)	(−0.89)	(0.77)
10	0.0001	0.0020	−0.0003	−0.0001	(0.19)	(4.20)	(−0.46)	(−0.15)
Top 20	0.0005	0.0021	0.0001	0.0005	(0.83)	(3.67)	(0.14)	(0.87)
Top 10	0.0009	0.0027	0.0002	0.0004	(1.35)	(4.16)	(0.39)	(0.67)
Top 5	0.0010	0.0034	0.0015	0.0007	(1.27)	(4.10)	(2.15)	(0.86)
	B1. Eight−factor monthly alpha				B2. *T*-ratio			
1	0.0002	−0.0005	−0.0008	−0.0015	(0.47)	(−1.15)	(−1.32)	(−2.44)
2	0.0001	−0.0004	−0.0005	−0.0010	(0.23)	(−0.79)	(−1.04)	(−1.68)
3	−0.0007	−0.0005	−0.0001	−0.0005	(−1.38)	(−0.94)	(−0.26)	(−0.87)
4	−0.0002	−0.0009	0.0001	−0.0006	(−0.37)	(−1.79)	(0.12)	(−1.12)
5	−0.0002	−0.0007	−0.0005	−0.0004	(−0.41)	(−1.39)	(−1.38)	(−0.84)
6	−0.0002	−0.0004	−0.0008	0.0004	(−0.31)	(−0.90)	(−1.84)	(0.76)
7	−0.0001	−0.0003	0.0000	−0.0002	(−0.22)	(−0.56)	(0.09)	(−0.28)
8	−0.0006	0.0000	−0.0006	0.0007	(−1.13)	(0.01)	(−1.05)	(1.22)
9	−0.0006	0.0013	−0.0005	0.0004	(−1.12)	(2.52)	(−0.78)	(0.77)
10	0.0002	0.0019	−0.0000	−0.0001	(0.42)	(3.72)	(−0.08)	(−0.09)
Top 20	0.0004	0.0020	0.0002	0.0003	(0.76)	(3.27)	(0.32)	(0.60)
Top 10	0.0009	0.0023	0.0004	0.0002	(1.25)	(3.35)	(0.62)	(0.35)
Top 5	0.0011	0.0028	0.0019	0.0009	(1.23)	(3.31)	(2.56)	(1.08)

degrades somewhat compared to the earlier results in Table IV. Oddly though, on the negative side, the four factor OLS model no longer reliably identifies funds with future below market performance. In fact other than the four factor Kalman model, the low decile t-statistic based portfolios in both Panels A and B fail to exhibit statistically significant persistent performance. Overall this table further confirms that selection schemes based on in sample diagnostics perform worse than a simpler back test that does not.

Intuitively, the reason in sample diagnostics are of so little value in producing accurate forecasts is that they apparently suffer from the same problems as the estimates they are supposed to diagnose. If a model is in any way incorrectly specified (whether it be the underlying distribution of the error terms, or the functional form of the model itself) then the resulting test diagnostics will be of limited or no value. Not surprisingly, then, if an incorrectly specified model produces large parameter values then it is not much of a stretch to imagine that it will produce large diagnostics (like t-statistics) as well. As Table XIV demonstrates this is apparently what occurs when using mutual funds data.

8. Conclusion

There is a substantial academic literature examining the question of whether or not it is possible to identify mutual fund managers who can produce predictable above market returns based only upon their past returns. However, these studies have used a single statistical model to describe every fund in their database. This is asking a lot of a single model. Fund managers follow a wide variety of strategies (see, e.g., Brown and Goetzmann (1997)) and it is unlikely that any one statistical model will accurately capture such a wide variety of dynamics. As a result, if the assumptions behind a model fit a particular fund's dynamics poorly, the resulting parameter estimates will suffer from misspecification error. In particular, a fund's alpha may be poorly estimated causing any sorts on this parameter to place it either at the top or bottom of any predicted performance list. Of course, going forward misspecification errors cannot aid in the prediction of fund returns and as a result there then appears to be little relation between forecasted and realized alphas.

This paper has documented that, indeed, the standard procedure of using a 5 year rolling window to estimate fund alphas and betas tends to place funds with questionable parameter estimates in the extreme deciles. Sorting funds by their predicted betas induces an inverse sort on alphas. This indicates that the poor beta estimates are feeding back into poor alpha estimates.

When the OLS model underestimates beta it tries to make up for this by overestimating alpha. Of course, none of this helps predict future returns or even future factor loadings. The latter can be seen by looking at the stability of the out of sample beta estimates. As the paper has shown, very high and low alpha sorted deciles produce the least accurate beta forecasts going forward.

This paper suggests that it may be useful to depart from the research methodology in which a single model is used across all funds. Instead one might wish to see whether a model has done a good job of predicting a particular fund's excess returns in the past before using that model on that fund in the future. This naturally leads to another possibility; using a variety of different models might allow for the identification of yet more funds whose performance can be predicted out of sample. Indeed, the results presented here indicate that an alternative model specification that allows for time variation in a fund manager's alpha and beta can be gainfully employed towards precisely this end.

The simplest system for back checking a model is to require it to accurately predict the sign of a fund's abnormal return in month $t-1$ before allowing it to make predictions about the fund in month t. Remarkably even this very crude back test is enough to produce alpha sorted portfolios for which the top *two* deciles yield positive and statistically significant above market returns. Depending on the model used, one can obtain risk adjusted (using a four factor model for the out-of-sample risk adjustment) above market returns of about 3.5 to 7.0% per year. These results also hold up under an eight-factor model designed to control for stale pricing issues. The resulting portfolios picked by the different models are also very diverse, with a common fund overlap of only about a third. This means that for every 100 funds selected by either model for inclusion in its top decile only about 33 of these funds will appear on both lists.

The results in this paper indicate that if one is looking to see whether managerial talent can produce above market returns it may pay to experiment with a number of statistical models. Mutual funds can, at times, follow dynamically complex strategies. It is unreasonable to expect a single model to accurately track the resulting changes in factor loadings without producing large misspecification errors. By using a variety of models and requiring them to prove their accuracy before employing them on a particular fund, it may be possible to identify far more funds with either positive or negative out-of-sample performance than has previously been thought.

Appendix

A. A Dynamic Kalman Filter Model of Mutual Fund Portfolios

A detailed derivation of the Kalman filter model presented here can be found in Mamaysky et al. (2007). To conserve space this Appendix presents just the necessary details for carrying out the estimation procedure.

Returning to Equation (1), for uninformed individuals the α_t terms always equals zero. However, for investors possessing information the α_t's may at times be positive or negative. This view of asset returns is in line with heterogeneous information models such as Admati (1985). Thus, to be precise the α_t terms should have a subscript indicating the individual and his information. For notational simplicity these indicators are suppressed, but one should bear in mind that the return equations are conditional on an investor's information set.

Even if the factor loadings for stocks and bonds do not change over time it is unlikely that this will be true of any actively managed portfolio containing these same instruments.[15] Let w_{it} represent the fraction of the portfolio in security i at time t, and W_t the $I \times 1$ vector containing the I individual weights. Then the portfolio's time t return equals the weighted average of the returns from the underlying I assets:

$$r_{Pt} - r_{ft} = W_{t-1}'(\alpha_t + \beta'(r_{mt} - r_{ft}) + \epsilon_t) - k_t. \tag{2}$$

The β term represents a matrix with I columns containing the vectors β_i. The k_t term equals the transactions costs incurred by the portfolio, which for mathematical tractability are assumed to be proportional to the funds under management.

Absent information about a fund's holdings, as well as the alphas and betas of the underlying assets, the parameters in Equation (2) cannot be estimated. However, these problems can be overcome by adding some additional assumptions that obviate the need to know the underlying portfolio's composition.

Let F_t represent some signal (normalized to have an unconditional mean of zero) that a particular fund uses to trade. Now assume that the signal's value

[15] Many studies like those of Ferson and Harvey (1991, 1993), and Ferson and Korajczyk (1995) question whether or not individual security loadings are constant (the β_i term in (1)). However, this will not qualitatively alter this paper's conclusion that fund loadings change over time. If anything, intertemporal variation in the underlying securities will only add to the importance of allowing for time variation in the mutual funds themselves.

follows the AR(1) process

$$F_t = \gamma_F F_{t-1} + \eta_t \tag{3}$$

through time. The $\gamma_F \in [0, 1)$ coefficient measures the degree to which the signal's value persists over time, and η_t represents an i.i.d. innovation.

If the signal F has value then one expects it to influence both the fund's holdings, and future expected stock returns. Statistically, these dual impacts can be represented by assuming that the portfolio weights follow:

$$w_{it} = \overline{w}_i + l_i F_t, \tag{4}$$

and that stock alphas equal

$$\alpha_{it} = \overline{\alpha}_i F_{t-1}. \tag{5}$$

Here $\overline{w}_i$ represents the steady-state fraction of the strategy invested in a given security. Alternatively, $\overline{w}_i$ can depend upon any set of observable variables, in which case it may be time dependent. The variable l_i is stock i's loading on a common unobservable factor F_t, which shifts the portfolio weights from their steady-state values. One can view this formulation as an empirical application of Admati's (1985) general equilibrium asset pricing model with asymmetric information. It is also generally consistent with Blake et al's. (1999) finding of mean reversion in fund weightings across securities among UK pension funds. Finally, $\overline{\alpha}_i$ represents the degree to which a stock's expected return is predictable by the signal F. If the signal has no value then all of the $\overline{\alpha}_i$ terms equal zero. Also, the present specification ensures that the steady-state alpha values equal zero.

Now use (4), and (5) in (2) to produce the Kalman observation equation:

$$\begin{aligned} r_{Pt} - r_{ft} &= (\overline{W} + l F_{t-1})'(\overline{\alpha} F_{t-1} + \beta'(r_{mt} - r_{ft} + \epsilon_t) - k_t \\ &= l'\overline{\alpha} F_{t-1}^2 - k_t + \overline{W}'\beta'(r_{mt} - r_{ft}) \\ &\quad + (\overline{W}'\overline{\alpha} + l'\beta'(r_{mt} - r_{ft}))F_{t-1} + (\overline{W} + l F_{t-1})'\epsilon_t \\ &= b_P F_{t-1}^2 - k_t + \overline{\beta}_P(r_{mt} - r_{ft}) \\ &\quad + (\overline{a}_P + c_P(r_{mt} - r_{ft}))F_{t-1} + \epsilon_{Pt}. \end{aligned} \tag{6}$$

In this equation b_P equals $l'\overline{\alpha}$, $\overline{\beta}_P$ equals $\overline{W}'$, $\overline{\alpha}_P$ equals $\overline{W}'\overline{\alpha}$, c_P equals $l'\beta'$, and ϵ_{Pt} equals $(\overline{W} + l F_{t-1})'\epsilon_t$.

The $\overline{\alpha}_i$, $\overline{\alpha}_P$, and b_P each play a unique economic role in the analysis. In Equation (5), $\overline{\alpha}_i \neq 0$ implies that a given fund's signal has a systematic relationship with the instantaneous excess returns of individual stocks in the economy. Therefore, one can alternatively write $\overline{\alpha}_{iP}$ to indicate that this coefficient is both stock *and* fund dependent. The point, though, of having nonzero $\overline{\alpha}_i$'s is to allow the fund's α_P to depend on the fund's trading strategy F. This dependence comes about through a linear term, the $\overline{\alpha}_P$, and a quadratic term b_P. There is no constant alpha term in α_P because in the long-run all alphas are assumed to be zero (their unconditional value). The linear term $\overline{\alpha}_P$ simply measures the degree to which a given fund's strategy is actually related to the instantaneous alphas of individual stocks. Since F can be positive or negative, a nonzero α_P does not indicate either under or overperformance. The quadratic term b_P, on the other hand, does indicate exactly this—it measures the degree to which a fund is able to systematically go long (short) positive (negative) alpha stocks.[16] Note that this is a sufficient, though not necessary, condition for a given fund to exhibit occasional (as opposed to systematic) risk-adjusted outperformance. A weaker and necessary condition is that a fund's α_P is persistent and occasionally positive (which obtains when $\overline{\alpha}_P \neq 0$ and when $\gamma_F > 0$).

The empirical model derived above is very flexible. For example, if one assumes that η_t has a variance of zero, or that γ_F equals zero, the Ferson and Schadt (1996) specification can be reproduced. Importantly, however, even absent these assumptions the model can still be estimated. Also note that nowhere does the econometrician need data on the actual portfolio weights used to produce the observed returns.

Due to the F_{t-1}^2 term in (6) the standard Kalman filtering techniques will fail as the conditional variance of $r_P(t) - r(t)$ will no longer be independent of the estimated values of F_{t-1}. The standard solution is to use a first-order Taylor expansion around the conditional expectation of F_{t-1}, or

$$F_{t-1}^2 \approx 2\,\mathbb{E}\Big[F_{t-1}\Big|r_{P,t-1} - r_{t-1}, F_{t-2}\Big]F_{t-1} - \mathbb{E}\Big[F_{t-1}\Big|r_{P,t-1} - r_{t-1}, F_{t-2}\Big]^2 \tag{7}$$

to replace the F_{t-1}^2 term in Equation (6) where $\mathbb{E}$ is the expectations operator.[17] Equation (3) then forms the state equation.[18] Note, the vector c_P has n elements

[16] Intuitively, b_P can be thought of as the covariance between a fund's security weights ($f(t)$) and the underlying security alphas.

[17] For details about extended Kalman filtering see Harvey (1989).

[18] The estimated dynamic Kalman filter model bears some philosophical resemblance to the Bayesian approaches found in Baks et al. (2001), and Pástor and Stambaugh (2002).

(one for each risk factor) but only $n - 1$ degrees of freedom. Thus, in the scalar case (as in the CAPM) it can be normalized to one when estimating the model. In the case where n is greater than one, at least one element's value must be fixed or some other normalization must be applied. The other fact needed for estimation is that the variance of $\epsilon_p(t)$, conditional on time $t - 1$ information, is given by

$$Var_{t-1}\left(\epsilon_{Pt}\right) = \sum_{i=1}^{I} w_{i,t-1}^2 Var_{t-1}\left(\epsilon_{it}\right).$$

This follows from $\epsilon_{Pt} = W'_{t-1}\epsilon_t$, and from the fact that all ϵ_{it}'s are independent.

The system specified in Equations (6) and (3) imbeds an important timing convention. The alphas and betas that determine time t returns are known at time $t - 1$ (assuming that k_t is deterministic). Therefore any covariance that exists between a portfolio's time t alphas and time t market returns indicates an ability of the portfolio manager to make investment decisions at time $t - 1$ that successfully anticipate market returns at time t. Similarly for time t betas and time t market returns.

To estimate the Kalman filter model one needs to initialize it. To do so the estimates in this paper use the OLS regression parameters. Given this starting point, the parameter values are perturbed and a parameter search is conducted to maximize the likelihood function. This process is repeated ten times. The parameters from the run that produce the highest likelihood function are then used to forecast mutual fund performance in the next period.

References

Admati, A. (1985) A noisy rational expectations equilibrium for multi-asset securities markets, *Econometrica* **53**, 629–657.

Avramov, D., and Wermers, R. (2006) Investing in mutual funds when returns are predictable, *Journal of Financial Economics* **81**, 339–377.

Baker, M., Litov, L., Wachter, J., and Wurgler, J. (2004) Can Mutual Fund Managers Pick Stocks? Evidence from their Trades Prior to Earnings Announcements, working paper, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=570381.

Baks, K., Metrick, A., and Wachter, J. (2001) Should investors avoid all actively managed funds? A study in bayesian performance evaluation, *Journal of Finance* **56**, 45–86.

Berk, J. (1995) A critique of size-related anomalies, *Review of Financial Studies* **8**, 275–286.

Berk, J., and Green, R. (2004) Mutual fund flows and performance in rational markets, *Journal of Political Economy* **112**, 1269–1295.

Berk, J., and Xu, J. (2004) Persistence and Fund Flows of the Worst Performing Mutual Funds, working paper, U.C. Berkeley http://faculty.haas.berkeley.edu/berk/persist.html.

Blake, D., Lehmann, B., and Timmermann, A. (1999) Asset allocation dynamics and pension fund performance, *Journal of Business* **72**, 429–461.
Bollen, N., and Busse, J. (2004) Short–term persistence in mutual fund performance, *Review of Financial Studies* **18**, 569–597.
Bossaerts, P., and Hillion, P. (1999) Implementing statistical criteria to select return forecasting models: what do we learn? *Review of Financial Studies* **12**, 405–428.
Brown, S., and Goetzmann, W. (1995) Performance persistence, *Journal of Finance* **50**, 679–698.
Brown, S., and Goetzmann, W. (1997) Mutual fund styles, *Journal of Financial Economics* **43**, 373–399.
Busse, J., and Irvine, P. (2005) Bayesian alphas and mutual fund persistence, *Journal of Finance* **61**, 2251–2288.
Carhart, M. (1995) Survivor bias and persistence in mutual fund performance, PhD dissertation, *University of Chicago Graduate School of Business*, http://wwwlib.umi.com/dissertations/dlnow/9609921.
Carhart, M. (1997) On persistence in mutual fund performance, *Journal of Finance* **52**, 57–82.
Chalmers, J., Edelen, R., and Kadlec, G. (2001) On the perils of financial intermediaries setting security prices: the mutual fund wild card option, *Journal of Finance* **56**, 2209–2236.
Chen, H., Jegadeesh, N., and Wermers, R. (2000) The value of active mutual fund management: an examination of the stockholdings and trades of fund managers, *Journal of Financial and Quantitative Analysis* **35**, 343–368.
Cohen, R., Coval, J., and Pástor, L. (2005) Judging fund managers by the company they keep, *Journal of Finance* **60**, 1057–1096.
Elton, E., Gruber, M., and Blake, C. (2001) A first look at the accuracy of the CRSP mutual fund database and a comparison of the CRSP and morningstar mutual fund databases, *Journal of Finance* **56**, 2415–2430.
Ferson, W., and Harvey, C. (1991) The variation of economic risk premiums, *Journal of Political Economy* **99**, 385–415.
Ferson, W., and Harvey, C. (1993) The risk and predictability of international equity returns, *Review of Financial Studies* **6**, 527–566.
Ferson, W., and Korajczyk, R. (1995) Do arbitrage pricing models explain the predictability of stock returns? *Journal of Business* **68**, 309–349.
Ferson, W., and Schadt, R. (1996) Measuring fund strategy and performance in changing economic conditions, *Journal of Finance* **51**, 425–462.
Ferson, W., and Khang, K. (2002) Conditional performance measurement using portfolio weights: evidence for pension funds, *Journal of Financial Economics* **65**, 249–282.
Goetzmann, W., Ivkovic, Z., and Rouwenhorst, G. (2001) Day trading international mutual funds: evidence and policy solutions, *Journal of Financial and Quantitative Analysis* **36**, 287–310.
Goyal, A., and Welch, I. (2006) A comprehensive look at the empirical performance of equity premium prediction, *Review of Financial Studies*, forthcoming.
Greene, J., and Hodges, C. (2002) The dilution impact of daily fund flows on open–end mutual funds, *Journal of Financial Economics* **65**, 131–158.
Grinblatt, M., and Titman, S. (1989) Portfolio performance evaluation: old issues and new insights, *Review of Financial Studies* **2**, 396–422.
Grinblatt, M., and Titman, S. (1994) A study of monthly mutual fund returns and performance evaluation techniques, *Journal of Financial and Quantitative Analysis* **29**, 419–444.
Harvey, A. (1989) *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, Cambridge.

Hendricks, D., Patel, J., and Zeckhauser, R. (1993) Hot hands in mutual funds: short–run persistence of relative performance, 1974–1988, *Journal of Finance* **48**, 93–130.
Kacperczyk, M., Sialm, C., and Zheng, L. (2005a) On the industry concentration of actively managed equity mutual funds, *Journal of Finance* **60**, 1983–2011.
Kacperczyk, M., Sialm, C., and Zheng, L. (2005b) Unobserved Actions of Mutual Funds, *Review of Financial Studies*, forthcoming.
Kosowski, R., Timmermann, A., Wermers, R., and White, H. (2007) Can mutual fund stars really pick stocks? New evidence from a bootstrap analysis, *Journal of Finance* **61**, 2551–2595. Available at SSRN: http://ssrn.com/abstract=855425.
Mamaysky, H., Spiegel, M., and Zhang, H. (2007) Estimating the dynamics of mutual fund alphas and betas, *Review of Financial Studies*, forthcoming.
Pástor, L., and Stambaugh, R. (2002) Investing in equity mutual funds, *Journal of Financial Economics* **63**, 351–380.
Pesaran, H., and Timmermann, A. (2005) Real-time econometrics, *Econometric Theory* **21**, 212–231.
Timmermann, A., and Granger, C. (2004) Efficient market hypothesis and forecasting, *International Journal of Forecasting* **20**, 15–27.
White, H. (2000) A reality check for data snooping, *Econometrica* **68**, 1097–1126.
Zitzewitz, E. (2003) Who cares about shareholders? Arbitrage–proofing mutual funds, *Journal of Law Economics & Organization* **19**, 245–280.