Decreasing Returns to Scale, Fund Flows, and Performance

Campbell R. Harvey

Duke University, Durham, NC 27708 USA
National Bureau of Economic Research, Cambridge, MA 02138 USA

Yan Liu*

Purdue University, West Lafayette, IN 47906 USA

Current version: June 21, 2021

Abstract

Theoretical models imply fund size and performance should be negatively linked. However, empiricists have failed to uncover consistent support for this negative relation. Using a new econometric framework which structurally models fund-specific sensitivities to decreasing returns to scale, we find a both economically and statistically significant negative relation between fund size and performance. Exploiting fund heterogeneity to decreasing returns to scale, we show that investors direct flows to those funds with low sensitivity to decreasing returns to scale. Interestingly, investors appear to over-allocate capital to these low sensitivity funds leading to significantly negative excess performance.

Keywords: Hedge funds, Mutual funds, Performance evaluation, EM algorithm, Fixed effects, Random effects, Scale, Multiple testing, Alpha.

* Current Version: June 21, 2021. Previously circulated with the title "Does Scale Impact Skill?". Send correspondence to Yan Liu, Krannert School of Management, Purdue University, West Lafayette, IN 47906, E-mail: liu2746@purdue.edu. We appreciate the comments of Martijn Cremers, Zheng Lu, Ľuboš Pástor, Min Zhu, and seminar participants at the Western Finance Associate (WFA) meetings in Whistler, Canada. We thank Martijn Cremers for supplying us with the data on active share. All errors are our own.

1 Introduction

Are there decreasing returns to scale in the asset management business? If yes, can we identify the managers that get greedy and take more assets than they can handle and the others that resist the temptation to dilute investor returns? Are investors able to sort this out and invest accordingly? The answers to these questions have profound implications for the fund management business. Currently, even the most basic question (decreasing returns to scale) is unanswered due to conflicting findings. For instance, while Chen et al. (2004) document decreasing returns to scale for mutual funds using Fama-MacBeth regressions, Pastor, Stambaugh and Taylor (PST, 2015) find insignificant fund-level decreasing returns to scale using a fixed effects panel regression approach.[1]

We argue that many challenges we face when evaluating mutual fund decreasing returns to scale are similar to those faced by the development economics literature that studies cross-country income growth. Borrowing insights from this literature, we propose a new structural approach (i.e., a random coefficient model) that features heterogeneous decreasing returns to scale, while at the same time permitting the inference on the effect population.

When evaluating the impact of scale, we pay particular attention to the bias of the OLS estimate when unexpected fund returns and the change in fund size are contemporaneously positively correlated, i.e., the Stambaugh (1999) bias. PST and Zhu (2018) show the presence of Stambaugh bias in a panel regression setup and propose methods to adjust for this bias. We show that our framework is much less affected by the Stambaugh bias. In particular, instead of using the individual fund size as a portion of the equity market as in PST, we first adjust individual fund size by the industry size, capturing the percentage wealth of a fund relative to the industry. We argue that our metric, by purging out variation in industry size, is a more intuitive measure of individual fund size. Additionally, in contrast to the fixed effects OLS framework in PST, our random coefficient model applies a different weighting scheme to the cross-section of funds. It underweights information provided by funds with a smaller sample size, for which the Stambaugh bias is particularly severe. Similar to PST, we also include in our regression industry size (divided by the overall wealth of the equity market) as a separate explanatory variable. We show through simulations that our estimators on both measures of scale (i.e., industry size and individual fund size) are largely unbiased and perform well when the data generating process (DGP) of the simulated data resembles the DGP of the actual data.

[1] Other recent papers that examine decreasing returns to scale include Pástor and Stambaugh (2012), Berk and van Binsbergen (2015), Zhu (2018), Magkotsios (2018), Pástor, Stambaugh, and Taylor (2020), van Binsbergen, Kim, and Kim (2020), and Dahlquist, Ibert, Wilke (2021), and Harvey, Liu, Tan, and Zhu (2021).

Our results point to a large impact of scale at the individual fund level. In particular, for an average fund in the cross-section that doubles its size relative to industry in one year, its alpha drops by around 20bp per annum. The impact of scale is significant both statistically and economically. We reconcile our finding with PST, who document a much smaller and insignificant impact of individual fund scale using a similar dataset. First, our definition of fund size contributes to the difference in results. While PST, by measuring scale with the dollar TNA (adjusted for market equity), try to estimate the impact on alpha per unit change in dollar TNA, our definition of scale allows us to estimate the impact on alpha per percentage change in dollar TNA (controlling for the change in industry size). Given the extreme differences in the magnitude of the dollar TNA in the cross-section, we think that our way of measuring scale has some advantages in that it helps standardize the magnitudes of the cross-section of the response coefficients to a change in scale, allowing us to pool information from the cross-section to accurately estimate the impact of scale on an average fund.[2] Second, our structural approach automatically draws less information from funds with a short return history, which are precisely the ones that are more affected by the Stambaugh bias.

The strong evidence we find on individual fund scale therefore lends considerable support to the Berk and Green (2004) model. It is also consistent with recent papers that estimate the impact of scale based on alternative research designs and find a significant impact of individual fund scale, such as Golez and Shive (2015), Zhu (2018), and McLemore (2019). Different from these papers, our framework provides a systematic approach to evaluate the impact of scale by allowing for fund fixed effects, and cross-sectional heterogeneity in the response to scale and controlling for fund specific exposures to benchmark risk factors. It therefore can potentially provide a more accurate measure of the impact of scale.

We also find a significant impact of industry size, consistent with PST. We estimate that a 1% increase in industry size (at the monthly level) implies a 5bp drop in alpha (per annum) for the average fund. Our estimate of the impact of industry-level scale is therefore higher but similar in magnitude to what PST find. In addition, thanks to our framework that allows heterogeneous loadings on scale, we discover an interesting U-shaped pattern for the time evolution of the impact of industry size. We argue that this U-shaped pattern is driven by the interaction between two effects: the dilution effect (new capital dilutes existing capital) and the diminishing alpha effect (the profitability of investment ideas deteriorating through time). Our finding has important implications for the overall capacity of the fund industry.

Equipped with our model, we construct long-short portfolios based on our fund-specific estimates for the degree of decreasing returns to scale. Exploiting decreasing returns to industry scale, a long-short portfolio that takes a long (short) position in funds with a lower degree of decreasing returns to scale generates an economically

[2]Indeed, we also find that the impact of individual fund scale is roughly homogeneous across different size groups, further supporting our regression setup.

significant positive alpha, controlling for fund size and past performance. Our result highlights the importance of identifying the differential exposure to industry size for the cross-section of funds, making it a profitable strategy to take a long (short) position in funds that are less (more) sensitive to industry growth.

Turning to fund level decreasing returns to scale (i.e., the log of the fund's TNA divided by the size of the industry), we find that, contrary to the case for industry level decreasing returns to scale, a long-short portfolio that takes a long (short) position in funds with a higher (lower) degree of decreasing returns to scale generates a large and positive alpha. To interpret this finding, we study the relation between decreasing returns to scale and future fund flows. We find that the degree of decreasing returns to scale predicts future fund flows, controlling for variables (e.g., past performance) that are documented by the existing literature. Moreover, decreasing returns to scale is able to explain the convex relation between past performance and future fund flows: holding past performance constant and assuming it is positive, funds that display the lowest level of decreasing returns to scale attract a disproportionately large amount of capital. Therefore, investors respond to decreasing returns to scale by rewarding funds with a lower degree of decreasing returns to scale with much more capital, which reduces the performance of these funds in the future. This explains our results on portfolio sorts based on fund level decreasing returns to scale.

We interpret our findings in the context of theoretical models such as Berk and Green (2004). Our results lend considerable support to Berk and Green in two aspects. First, we document a significant impact of decreasing returns to scale, both at the industry level and at the individual fund level. Second, we find that investors favor funds with a low degree of decreasing returns to scale, which is consistent with Berk and Green's main insight in that, since investors supply funds competitively, more capital should flow to funds that are better at absorbing new capital without reducing performance, that is, funds with a lower degree of decreasing returns to scale. However, our results on portfolio sorts using fund level decreasing returns to scale also suggest that investors allocate an excessive amount of capital to funds with a low degree of decreasing returns to scale, to the extent that these funds perform worse in the future than funds with a high degree of decreasing returns to scale.

Finally, the fact that the long-short strategies we construct produce excess returns validates the basic assumptions for our estimation framework, that is, the degree of decreasing returns to scale is both fund specific and persistent.

Our paper is organized as follows. In the second section, we provide an economic foundation for our model by drawing on the development economics literature. In the third section, we propose a new econometric framework to estimate the impact of scale and discuss a comprehensive simulation study. In the fourth section, we discuss the data we use and present summary statistics. In the fifth section, we show our main results on the estimation of the impact of scale. In the sixth section, we present some additional results, including the evaluation of the time-varying impact of scale and the construction of profitable investment strategies that exploit the cross-

sectional difference in resisting decreasing returns to scale. Some concluding remarks are offered in the final section.

2 Economic Foundation

We provide a new framework to evaluate the relation between scale and performance. While (dis)economies of scale has been the focus of several theoretical papers on investment management (e.g., Berk and Green, 2004, Pastor and Stambaugh, 2012), we still lack a full-fledged theory that can describe the dynamics of the cross-section of fund returns. In contrast, there is a large literature in development economics where macroeconomic models attempt to explain cross-country income growth. Our strategy is to borrow some of the insights from the growth literature, which, in many ways, faces similar challenges.

Admittedly, important differences exist between the two strands of research. For example, individual funds are better treated as micro units while countries are macro units. While the Solow (1956) growth model provides a strong theoretical basis for the empirical growth literature, we do not have such luxury in the realm of investment management. In addition, while the dependent variables in growth regressions are well defined and readily available, alphas are usually unobservable, creating an additional difficulty when evaluating the impact of amount of assets under management. Despite these differences, by studying the evolution of growth regressions, we learn important lessons on how growth econometrics accommodate both theoretical concerns and empirical practice, shedding light on what is a good way to carry out "scale" regressions.

More specifically, we make three modeling choices motivated by the development economics literature. We provide a summary of these choices below and discuss them in detail in Appendix A.

First, an important strand of growth regressions use country fixed effects to allow for time-invariant idiosyncratic growth components.[3] In our context, as pointed out by PST, the use of fund fixed effects allows us to identify the impact of scale through the time-series dynamics, which helps address the endogeneity concern that arises when performing a cross-sectional regression of fund alpha on fund size because funds with a large size are more likely to fall into capable hands. We propose a dynamic panel regression approach that allows for fund fixed effects.

Second, one benefit of having the Solow neo-classical growth model to guide empirical explorations is that it guarantees that all variables are properly scaled, so regression coefficients correspond to the structural parameters in the model and have straightforward economic interpretations. Although we do not have such a benchmark

[3]See, e.g., Islam (1995).

model for fund size and returns, we strive to achieve the same objective by properly scaling variables related to fund size. In particular, we introduce a new measure for fund size—defined as a fund's *total net assets* (TNA) relative to the size of the fund industry—and advocate taking a logarithmic transformation of such a measure in our regression model.[4] We show our measure helps standardize the cross-section of funds that have vastly different levels of size and allows a comparable economic interpretation of the regression coefficients in scale regressions.

Third, and most importantly, we follow the more recent growth literature by building a structural model that allows heterogenous regression coefficients in scale regressions for the cross-section of funds.[5] We believe parameter heterogeneity is important because, just as with manager skill, the ability of a manager to resist decreasing returns to scale should also be manager specific. Besides capturing heterogeneity, our framework also provides estimates for the cross-sectionally averaged impact of scale, making it possible to interpret the impact of assets under management in general.

Distilling the insights of the growth literature, we propose a *random coefficient* framework to evaluate the impact of economies of scale on fund performance. Our framework is not specific to scale regressions. It can be used in a general context to evaluate cross-sectional alpha prediction models. Given the inconsistent results in the literature,[6] our dynamic panel regression framework may be useful in resolving many unanswered questions.

Our framework builds on the insights of the standard random effects panel regression model. As shown in Searle, Casella, and McCulloch (1992), effects should be random if there is interest in the underlying population. Stoker (1993) also points out that effects should be treated as random if one wishes to make a statement about macrorelationships based on micro estimates from a subpopulation of data. Stoker's insight seems particularly relevant for our application since we only have partial coverage of the universe of mutual funds and the TNA's of most funds are very small relative to the GDP so it makes more sense to treat them as micro units than macro units. While our framework allows heterogeneous fund loadings on an alpha predictor, a population perspective (i.e., aggregating the fund specific coefficients) should help us understand the overall economic impact of a predictive variable. Moreover, a refined inference on the distribution of loadings aids the inference on individual funds, which is often difficult given the high level of noise and the limited sample size at the individual fund level.

Our random coefficient framework parametrically models the population of regression coefficients, thereby extending the standard random effects panel regression

[4]Besides the literature on economic growth, see Backus, Kehoe, and Kehoe (1992) for another application that introduces a similar measure for the scale of the economy.

[5]See, e.g., Kevin, Pesaran, and Smith (1998), Durlauf, Kourtellos, and Minkin (2001), Banerjee and Duflo (2003), and Phillips and Sul (2007).

[6]See Jones and Mo (2021) for a summary of proposed variables that help predict fund alphas and the out-of-sample evaluation of their performance.

model. The advantage is that, unlike the standard random effects model, we are able to make inference on individual funds by utilizing information from the loadings population. This is important as it allows us to identify fund managers that exercise discipline and resist diseconomies of scale. Our framework is also different from papers in the growth literature that incorporate parameter heterogeneity (e.g., Banerjee and Duflo (2003), Durlauf, Kourtellos, and Minkin (2001), and Kevin, Pesaran, and Smith (1998)). While these papers explicitly model parameter heterogeneity through instrumental variables, our framework does not rely on pre-specified instruments (see Solow (2001)). We only use funds' return time-series to identify the impact of an alpha predictor. Phillips and Sul (2007) is an exception from the growth literature that also does not rely on pre-specified instruments to model parameter heterogeneity. While they focus on the time-series convergence of the cross-sectional distribution of the loadings on independent variables, our framework uses a time-invariant distribution to model the cross-section of loadings on an alpha predictor.

Equation-by-equation OLS is often used to assess alpha predictability or the timing ability of funds, being it market timing or liquidity timing.[7] Statistical evidence on alpha predictability or timing ability is often established by showing that a certain fraction of parameter estimates for individual funds are statistically significant. We show that this is an ill-advised practice. The limited sample size for most funds makes the inference for individual funds unreliable. In addition, given the large cross-section of funds, certain funds may exist that generate extreme test statistics that appear to exceed the significance threshold, even after imposing a multiple testing threshold. However, such funds may tell us little about the overall economic impact of a predictive variable as the variable may have a negligible impact on the average fund. On the other hand, it is also inappropriate to discard funds that generate extreme estimates as this may bias our estimate of the effect population. Our framework provides a structural approach to draw information from the cross-section to make inference on a particular fund, while minimizing the extreme and implausible estimates for certain funds. Harvey and Liu (2018) apply a similar idea to make inference on the underlying population of alphas.

[7] See, e.g., Treynor and Mazuy (1966), Henriksson and Merton (1981), Ferson and Schadt (1996), Chen et al. (2013).

3 Method

3.1 Model

Suppose fund excess returns can be decomposed as:

$$r_{i,t} = \underbrace{\alpha_i + \sum_{\ell=1}^{L} \gamma_{i,\ell} g_{i,\ell,t}}_{\alpha_{i,t}} + \underbrace{\sum_{j=1}^{K} \beta_{ij} f_{j,t}}_{F_{i,t}} + \varepsilon_{i,t},\ i = 1, \ldots, N;\ t = 1, \ldots, T, \quad (1)$$

where $r_{i,t}$ is the excess return for fund i in period t, α_i is fund i's time-invariant alpha, $\gamma_{i,\ell}$ is fund i's loading on characteristic ℓ (i.e., $g_{i,\ell,t}$ which, in general, is fund specific and therefore depends on subscript i), β_{ij} is fund i's time-invariant risk loading on the j-th factor $f_{j,t}$, and $\varepsilon_{i,t}$ is the residual. For simplicity, we assume a balanced panel. But this is not required for either the exposition or the estimation of our model.

Our formulation offers a three-way decomposition of fund returns: $\alpha_{i,t}$ is the time-varying alpha that could depend on fund characteristics, $F_{i,t}$ captures the exposure to benchmark risk factors, and $\varepsilon_{i,t}$ is the residual noise component.[8]

Next, we need to determine the set of parameters that we want to focus on. These will be the parameters that are treated as random effects, whose estimation will draw on information from both the cross-section and time-series. Given our focus on fund characteristics (i.e., $g_{i,\ell,t}$) that help predict alphas, we assume that fund i's loading vector on fund characteristics (i.e., $[\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{iL}]'$) is randomly drawn from a multivariate probability distribution that is parameterized by Λ.

Harvey and Liu (2018) treat funds' unconditional alphas as random effects and seek to estimate the underlying alpha distribution. In this paper, we focus on fund characteristics and assume that both funds' time-invariant alphas (i.e., α_i) and betas (i.e., β_{ij}) are fixed effects in our main specification. To the extent that treating either time-invariant alphas or betas as random effects may improve our model estimates, we also explore alternative model specifications.

To write down the likelihood function of the model, we introduce some notation. Let $R_i = [r_{i,1}, r_{i,2}, \ldots, r_{i,T}]'$ be the vector of excess returns for fund i and $\mathcal{R} = [R_1, R_2, \ldots, R_T]'$ be the panel of excess returns. Let $\beta_i = [\alpha_i, \beta_{i1}, \beta_{i2}, \ldots, \beta_{iK}]'$ be the vector of the time-invariant alpha and risk loadings for fund i and $\mathcal{B} = [\beta_1, \beta_2, \ldots, \beta_N]$ be the panel of time-invariant alphas and risk loadings. Notice that for simplicity we treat α_i as the risk loading on a constant of one. Let $\gamma_i = [\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{iL}]'$ be the

[8] Although we do not allow time-varying betas in our presentation in Eq. (1), time-varying betas as captured by instrumented variables can be straightforwardly modeled in our framework by introducing interactions between benchmark factors and instrumented variables.

vector of loadings on characteristics for fund i and $\mathbf{\Gamma} = [\gamma_1, \gamma_2, \ldots, \gamma_N]$ be the panel of loadings. Let the standard deviation for the residuals of fund i be σ_i and we collect the cross-section of residual standard deviations into $\mathbf{\Sigma} = [\sigma_1, \sigma_2, \ldots, \sigma_N]'$.

Given the model structure, the joint likelihood function (i.e., $\mathcal{L}(\mathcal{R}|\Lambda, \mathcal{B}, \mathbf{\Sigma})$) can be written down as:

$$\begin{aligned} \mathcal{L}(\mathcal{R}|\Lambda, \mathcal{B}, \mathbf{\Sigma}) &= \int \mathcal{L}(\mathcal{R}, \mathbf{\Gamma}|\Lambda, \mathcal{B}, \mathbf{\Sigma}) d\mathbf{\Gamma} \quad (2) \\ &= \int \mathcal{L}(\mathcal{R}|\mathbf{\Gamma}, \mathcal{B}, \mathbf{\Sigma}) \mathcal{L}(\mathbf{\Gamma}|\Lambda) d\mathbf{\Gamma}, \quad (3) \end{aligned}$$

where $\mathcal{L}(\mathcal{R}|\mathbf{\Gamma}, \mathcal{B}, \mathbf{\Sigma})$ is the conditional likelihood function assuming all model parameters are known and $\mathcal{L}(\mathbf{\Gamma}|\Lambda)$ is the density function of the loadings on fund characteristics. Hence, the joint likelihood function integrates out the loadings on fund characteristics from the conditional likelihood function (i.e., $\mathcal{L}(\mathcal{R}|\mathbf{\Gamma}, \mathcal{B}, \mathbf{\Sigma})$).

Assuming that the residuals are independent across funds and across time, the joint likelihood function can be written as:

$$\begin{aligned} \mathcal{L}(\mathcal{R}|\Lambda, \mathcal{B}, \mathbf{\Sigma}) &= \int \prod_{i=1}^{N} L(R_i|\gamma_i, \beta_i, \sigma_i) L(\gamma_i|\Lambda) d\gamma_i, \quad (4) \\ &= \prod_{i=1}^{N} \int L(R_i|\gamma_i, \beta_i, \sigma_i) L(\gamma_i|\Lambda) d\gamma_i. \quad (5) \end{aligned}$$

Our goal is to estimate the structural parameters (i.e., Λ) that govern the population of loadings on characteristics as well as the parameters that govern fund return dynamics, e.g., β_i and σ_i. Notice that the only difference between the likelihood function in our model and the likelihood function for the traditional equation-by-equation OLS (i.e., $\prod_{i=1}^{N} L(R_i|\gamma_i, \beta_i, \sigma_i)$) is that in our model, the likelihood for each fund's return dynamics (i.e., $L(R_i|\gamma_i, \beta_i, \sigma_i)$) is weighted by the density function of γ_i (i.e., $L(\gamma_i|\Lambda)$). Hence, we draw on information from the cross-sectional distribution of γ_i to make inference on a particular fund. This helps alleviate the small sample problem that we often face when evaluating alpha predictors on a fund-by-fund basis.

To have a fully specified likelihood function, we further assume that both the innovations in fund returns and the cross-sectional distributions of the elements in γ_i follow normal distributions. In particular, for each element in γ_i (e.g., $\gamma_{i,\ell}, \ell \in \{1, 2, \ldots, L\}$), we assume that it follows a normal distribution with its own mean $\mu_{\gamma,\ell}$ and standard deviation $\sigma_{\gamma,\ell}$. For simplicity, we also assume that elements in γ_i are drawn independently from their respective distributions ex ante. However, ex post, the individual fund data may as well suggest correlations among the loadings as the independent variables for the fund (i.e., fund characteristics) may be correlated through time. Notice that the assumption of a normal distribution on loadings is not

necessary in our framework. We can generalize it by using more flexible distributions such as a Gaussian-mixture distribution. However, for our application, we think a normal distribution suffices as it succinctly captures the average and dispersion of the loadings population.[9]

Notice that we derive our model estimates under residual independence. However, residual independence is not needed to guarantee the consistency of our model estimates. As we show in simulations, our estimation produces consistent parameter estimates even when residuals are correlated in the cross-section. However, the level of estimation uncertainty becomes higher when residuals are correlated. We evaluate the impact of residual correlation in detail in our simulation study.

3.2 Estimation Procedure

We rely on the well-known Expectation-Maximization (EM) algorithm to efficiently estimate our model. The idea of the algorithm is to treat parameters that follow a certain population structure (e.g., γ_i's in our framework) as missing observations and iteratively update these missing observations and other model parameters. Harvey and Liu (2018) and Chen, Cliff, and Zhao (2015) apply the EM algorithm to make inference on the underlying alpha population, providing a new approach to performance evaluation.[10] Our innovation in this paper is to apply the EM algorithm to uncover the underlying distribution of the loadings on alpha predictors, offering a systematic approach to study the loadings population (e.g., what is the impact of industry size on an average fund) as well as to refine the estimates of individual loadings, which are difficult to estimate using fund-specific information alone.

The algorithm, adapted to our framework, proceeds as follows.

Step I Let $\mathcal{G} = \{\Lambda, \mathcal{B}, \mathbf{\Sigma}\}$ denote the collection of parameters to be estimated. We start at some parameter value $\mathcal{G}^{(0)}$. A sensible initial choice is the collection of parameter estimates obtained through the equation-by-equation OLS. In particular, the equation-by-equation OLS directly generates estimates for $\mathcal{B}$ and $\mathbf{\Sigma}$ as well as the cross-section of loadings on characteristics. The mean and standard deviation estimates for the cross-section of loadings then provide estimates for parameters in Λ.

Step II After the k-th iteration of the algorithm, suppose the model parameters are estimated as $\mathcal{G}^{(k)}$ ('k' is the generic indicator for the round of iteration. Note

[9] Harvey and Liu (2018) show the necessity of using a Gaussian-mixture distribution to capture the non-normal features of the alpha population. Different from their framework, we do not attempt to model the alpha population. As such, alphas in (1) are freely estimated and are not subject to any population distribution.

[10] See Harvey and Liu (2018) and the references therein for a detailed description of the EM algorithm.

we start from $k = 0$ as in *Step I.*). We calculate the expected value of the log complete likelihood function, with respect to the conditional distribution of $\mathbf{\Gamma}$ given the current parameter values and $\mathcal{R}$, i.e.,

$$L(\mathcal{G}|\mathcal{G}^{(k)}) = E_{\mathbf{\Gamma}|\mathcal{R},\mathcal{G}^{(k)}}[\log L(\mathcal{R}, \mathbf{\Gamma}|\mathcal{G})], \quad (6)$$

$$= E_{\mathbf{\Gamma}|\mathcal{R},\mathcal{G}^{(k)}}[\sum_{i=1}^{N} \log L(R_i|\gamma_i, \beta_i, \sigma_i)L(\gamma_i|\Lambda)]. \quad (7)$$

Step III We maximize $L(\mathcal{G}|\mathcal{G}^{(k)})$ and update the parameter estimates as $\mathcal{G}^{(k+1)}$.

Step IV With the new parameter estimate $\mathcal{G}^{(k+1)}$, we return to *Step II* to start the $(k+1)$-th iteration. We iterate between *Step II* and *Step III* until the parameter estimates converge.

In our setup, fortunately, we have closed-form formulas for each step of the algorithm, as we show in Appendix B.

3.3 Model Discussion

Our model features the evaluation of alpha predictors by drawing on cross-sectional information. At the same time, we allow fund-specific loadings to capture the heterogeneity of the impact of an alpha predictor. Both features of our model make it appealing for empirical applications. For example, when measuring economies of scale, we would like to know the impact of economies of scale for an average fund in the cross-section. To have an estimate of this impact, we should look at the underlying effect population, instead of the collection of noisy equation-by-equation OLS estimates, to make inference. Our model directly targets the estimation of the effect population. On the other hand, when individual funds' responses to an alpha predictor is of interest (e.g., we may want to identify funds that have the ability to resist decreasing returns to scale), we would like to have a good estimate of the impact of the alpha predictor for each individual fund. Such an estimate is impossible using fund-specific information alone, given the limited time-series data for many funds in the mutual fund sample. Our model provides an estimate for an individual fund's response to the alpha predictor by drawing on information in the cross-section. Similar ideas have been applied in Jones and Shanken (2005), Chen et al. (2015), and Harvey and Liu (2018) to estimate fund alphas.

To appreciate the two features of our model, we later show through simulations that the equation-by-equation OLS estimates of the mean loadings on the two metrics for scale, especially the metric on individual fund scale, are not only substantially more noisy than our model's estimates, but are also severely biased. Moreover, the

equation-by-equation OLS generates even more noisy estimates for the loadings of individual funds than the estimates for the mean loadings. In contrast, by using information in the loadings population, our model provides much more precise and economically meaningful estimates for the loadings of individual funds.

To explore how our method works, we take a closer look at the steps of our estimation procedure. Assuming the model parameters are known, in *Step II* of the EM algorithm, we estimate the loadings on fund characteristics for each fund. In particular, when there is only one alpha predictor, the estimated loading on the single fund characteristic follows a normal distribution with mean (m_i) and variance (v_i), where m_i and v_i are given by:

$$m_i \equiv \frac{(\sum_{t=1}^{T} g_{i,t}(r_{i,t} - \beta_i' f_t)/\sum_{t=1}^{T} g_{i,t}^2)(\sum_{t=1}^{T} g_{i,t}^2/\sigma_i^2) + \mu_\gamma/\sigma_\gamma^2}{\sum_{t=1}^{T} g_{i,t}^2/\sigma_i^2 + 1/\sigma_\gamma^2}, \tag{8}$$

$$v_i \equiv \frac{1}{\sum_{t=1}^{T} g_{i,t}^2/\sigma_i^2 + 1/\sigma_\gamma^2}. \tag{9}$$

Hence, the variance v_i is a harmonic average of the usual time-series variance (i.e., $\sigma_i^2/\sum_{t=1}^{T} g_{i,t}^2$) and the cross-sectional variance σ_γ^2. As a result, it takes into account both the time-series and the cross-sectional uncertainty in estimating a fund's loading on the characteristic. At the same time, the mean estimate m_i weights the time-series estimate (i.e., $\sum_{t=1}^{T} g_{i,t}(r_{i,t} - \beta_i' f_t)/\sum_{t=1}^{T} g_{i,t}^2$) and the cross-sectional estimate (i.e., μ_γ) by their respective precisions (i.e., the reciprocal of the variance), allowing us to pool information from the cross-section to refine the time-series estimate of a fund's loading on the characteristic. Note when T becomes large, $\sum_{t=1}^{T} g_{i,t}^2/\sigma_i^2$ dominates $\mu_\gamma/\sigma_\gamma^2$, implying that our model will assign a high weight to fund i's time-series information. From a population perspective, funds with a longer time series are weighted more heavily in inferring the population parameters. This insight is important in understanding our model's superior performance in the simulation study.

After we obtain an estimate for each fund's loading on the characteristic, in *Step III* of the EM algorithm, we re-estimate the other OLS parameters (i.e., fund alpha, loadings on benchmark factors, and residual standard deviation) for each fund as well as parameters that govern the cross-section of loadings on the characteristic (i.e., Λ). We show Appendix B that the MLEs for both types of parameters have closed-form solutions and appeal to intuition.

When there are multiple characteristics, the formulas in *Step II* are more complex. In particular, the loadings on characteristics for a particular fund are not independent of each other since the time-series of characteristics are in general correlated. However, the basic intuition for the case with a single characteristic still applies to the case with multiple characteristics. We also derive analytical expressions for the latter in Appendix B.

In essence, the EM algorithm iteratively updates missing observations (i.e., the cross-section of loadings $\mathbf{\Gamma}$) and model parameters $\mathcal{G}$, which include the OLS parameters other than the loadings on characteristics and parameters that govern the cross-sectional distributions of the loadings. In particular, in *Step II*, given our current estimates of the model parameters (i.e., $\mathcal{G}^{(k)}$), we back out the missing observations in $\mathbf{\Gamma}$. Subsequently, fixing the missing observations at their estimates in *Step II*, we update the model parameters in $\mathcal{G}$ in *Step III* and obtain a new set of model parameters (i.e., $\mathcal{G}^{(k+1)}$). We iterate between *Step II* and *Step III* until the structural parameters in $\mathcal{G}$ converge.

3.4 Measuring Scale

We use two measures of scale.

The first metric is industry-level scale, which we denote as $IndusSize$. At the beginning of each month, we add up the TNA's across funds and then divide by the aggregate market capitalization of the stock market ($AggStock$). Our metric is the same as the one used in PST, who are the first to examine the impact of industry-level scale. $IndusSize$ is the weight of the mutual fund industry relative to the entire equity market.[11]

The other metric is fund-level scale, which we denote as $FundSize$. We construct it in several steps. First, at the beginning of each month, we divide a fund's TNA by the aggregate TNA of the mutual fund industry, creating a variable that measures the scale of an individual fund relative to the size of the industry. Second, we take a log transformation of the relative scale metric that is defined in the previous step, creating the time-series of the log of the relative size for each fund. Finally, we subtract the first observation of this time-series from the entire time-series, essentially adjusting the time-series for the initial fund size. The time-series of $FundSize$ is then taken to be this adjusted time-series of the log of the relative scale of each fund.

There are several reasons for us to define $FundSize$ in this way.

First, we believe it is important to control for industry size while measuring fund-specific scale. Intuitively, a \$100 million fund in 1991 (the beginning of our sample) should be treated differently from a \$100 million fund in 2011 (the end of our sample) given the mutual fund industry has grown substantially during this period. Notice that this difference is not picked up by the industry scale variable (i.e., $IndusSize$) because $IndusSize$ is a single time series and is not fund specific. Suppose that the aggregate equity market increases tenfold and that the aggregate industry size is a constant proportion of the aggregate equity market throughout our sample (i.e.,

[11] To be consistent with PST, we do not take the log of $IndusSize$. Notice that, different from the case for fund size, we do not need to take the log of $IndusSize$ since $IndusSize$ is the same for every fund in the cross-section, allowing the loadings population to be homogenous.

$IndusSize$ is a constant). Suppose that the \$100 million fund stays at \$100 million throughout our sample. Then, intuitively, the fund's impact on both the fund industry and the equity market at the end of our sample becomes one-tenth of its impact at the beginning of our sample, effectively reflecting a shrinkage of its relative size. In addition, since $IndusSize$ remains constant, it cannot pick up this change in impact and the associated change in alpha for this particular fund.

On the other hand, suppose that the size of the equity market stays constant (\$10 trillion) and the industry size changes from one trillion to two trillion. Suppose a fund has a constant size of 100 million throughout. Should we consider the size of the fund as constant as in PST (since the size of the equity market stays constant), or being smaller as in our definition? First of all, since the industry size doubles relative to the equity market, due to decreasing returns to scale at the industry level, the additional one trillion dollars of assets may earn a lower return compared to the initial one trillion, reducing the overall profitability of the industry. This is the industry effect that is captured by $IndusSize$, as defined in PST and our paper. Purging out the industry effect, the additional one trillion dollars should be considered as equally profitable as the initial one trillion. However, since the equity pool (i.e., equities that are managed by the industry) gets larger, the effective size of the fund should decline, much in the same way as how PST define $IndusSize$ as the industry size relative to the size of the equity market. This decline in effective size, together with the assumption of decreasing returns to scale at the fund level, implies better performance for the fund, which is consistent with the intuition that the combination of two equally profitable sets of equities should benefit the fund in the original set, despite the increased competition from new entrants generated by the additional one trillion dollars. This is similar to the idea that investing internationally benefits investors from all the countries.

To put it differently, if we apply a log transformation to PST's definition of $FundSize$, we can decompose it as

$$\log \frac{TNA_{i,t}}{MKT_t} = \log \frac{TNA_{i,t}}{ITNA_t} + \log \frac{ITNA_t}{MKT_t},$$

where MKT_t is the size of the equity market at time t and $ITNA_t$ is the size of the industry at time t. Notice that $\frac{ITNA_t}{MKT_t}$ is simply $IndusSize$. Hence, our measure of $FundSize$ (i.e., $\log \frac{TNA_{i,t}}{ITNA_t}$) purges out the variation of $IndusSize$ from PST's definition of $FundSize$, allowing us to evaluate the impact of fund size that is independent of industry size.

Overall, compared to PST, we think our definition of $FundSize$ can potentially better disentangle the impact of industry size and fund size on fund performance.

Another benefit of scaling an individual fund's size by the industry size, from a technical perspective, is that innovations in a fund's TNA are no longer mechanically related to its return, alleviating the finite-sample bias when regressing fund returns

on lagged TNA, as shown in PST. Additionally, our framework allows us to pool information from the entire cross-section of funds to estimate the impact of fund size, further reducing the reliance on any particular fund's time-series to make inference. As we show in our simulation study, our estimation procedure performs well, producing essentially unbiased estimates for the means of the population of loadings on the two scale proxies (i.e., $IndusSize$ and $FundSize$).

It is also important to take a log transformation of the industry-adjusted fund's TNA, as we discussed previously. This ensures that the regression coefficient on $FundSize$ represents the change in alpha if the log of the industry-adjusted TNA goes up by 100% (which is equivalent to a growth of 171.8% in a fund's TNA since $\log(2.718) = 1$), regardless of the initial level of the TNA. Considering the very large differences in the levels of TNA for the cross-section of funds, the log transformation is necessary to obtain roughly homogeneous regression coefficients on $FundSize$ in the cross-section, allowing us to pool information from the cross-section of funds to accurately estimate the impact of $FundSize$.

Finally, our last step of defining $FundSize$ (i.e., adjusting the time-series of the log of the industry-adjusted fund scale for its initial observation) is not essential for our results as adjusting the time-series of regressors by a constant has no impact on the estimation of the regression coefficient. However, it allows us to interpret the alpha estimate as the estimate that corresponds to the initial TNA of the fund. We adopt this to standardize our interpretation of funds' alphas.

To summarize, our analysis will examine the impact of two scale metrics on fund performance. Using the our notation from previous sections, we have $g_{i,1,t} = IndusSize_{i,t}$ and $g_{i,2,t} = FundSize_{i,t}$.

3.5 A Simulation Study

We perform a comprehensive simulation study to examine the performance of our model, paying particular attention to the finite-sample bias issue in Stambaugh (1999) and PST. We provide details of the simulation study in Appendix B. Below we summarize the main findings of our simulation study.

Several features mark our simulation study. First, we allow heterogeneous loadings on characteristics as well as factor returns to provide a realistic data generating process. Second, we explicitly model the endogenous relation between fund size and fund returns. Third, heterogeneous loadings on characteristics (in particular, the two scale variables) are drawn from normal distributions. Our goal in the estimation is to recover both population parameters (i.e., population means and variances) and fund-level parameters. Fourth, we also allow cross-sectional dependence in fund idiosyncratic risks as well as factor returns to study how these features in the data

affect our estimation, even though our estimation procedure does not take residual dependence into account.

Our main results can be summarized as follows. While the usual equation-by-equation OLS is shown to be severely biased when estimating fund-level decreasing returns to scale, our model performs much better in estimating the population parameters: both the mean and the variance for the loading population are estimated with a substantially lower bias compared with the equation-by-equation OLS. We dissect our results further by examining the performance of our model across different groups of funds as sorted by sample length.[12] We find, as expected, funds that exist for a shorter period of time display a larger bias in the estimation. However, these funds are precisely the ones that are downweighted in our random effects framework, making our estimator largely unbiased when estimating population parameters.

We highlight two factors that contribute to the better performance of our model compared with PST. First, our definition of industry-adjusted individual fund size dampens the mechanical contemporaneous correlation between a fund's return and its *FundSize*, alleviating the Stambaugh bias. Second, our random effects framework automatically (and optimally) overweights funds with a larger number of observations, further mitigating the small-sample-induced Stambaugh bias.

4 Data

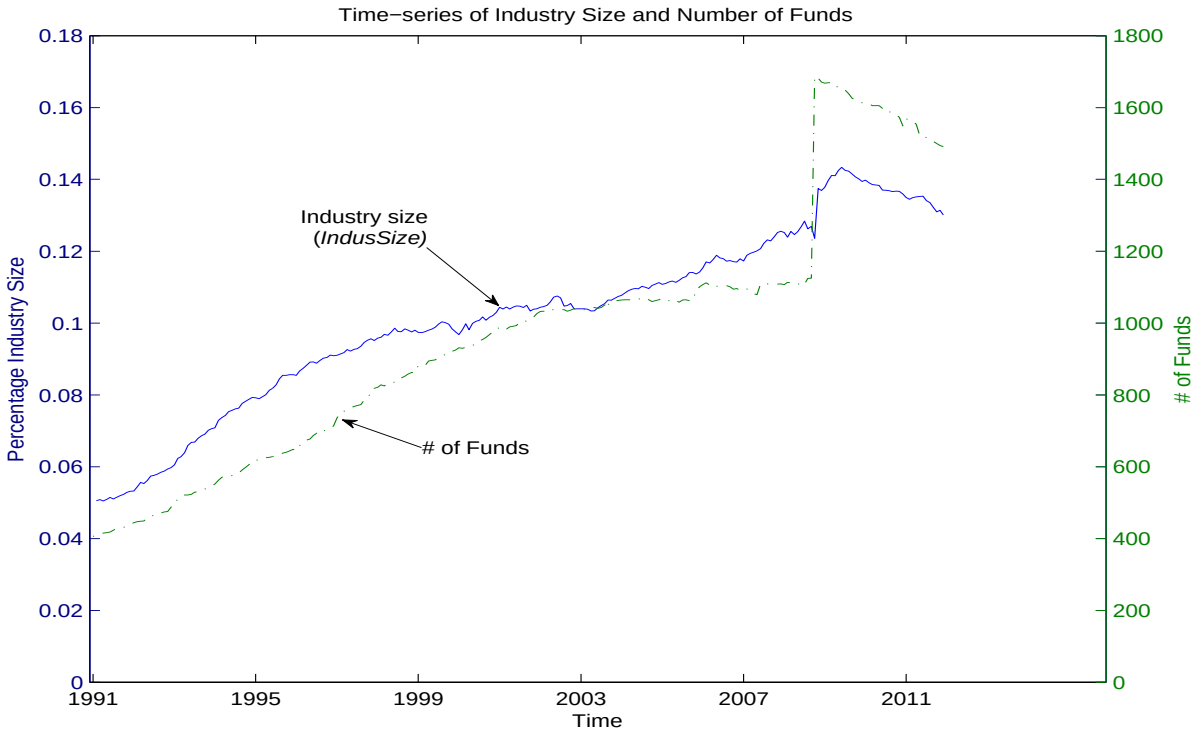
We obtain the mutual fund data from the Center for Research in Security Prices (CRSP) Mutual Fund database. We focus on active, domestic equity funds covering the 1991-2011 period. We start from 1991 as many funds do not have monthly updates on their TNA's before 1991. We end at 2011 to facilitate our comparison with the results in PST. To mitigate omission bias (Elton, Gruber and Blake, 2001) and incubation and back-fill bias (Evans, 2010), we apply some screening procedures. We only keep funds that have a TNA above \$10 million and have more than 80% of their holdings in stocks. We also combine multiple share classes. As we mentioned before, we require that a fund has at least 18 non-missing monthly observations to enter our test since our fund-level regression has six regressors. This leaves us with

[12] One may wonder whether estimates from equation-by-equation OLS are a strawman comparison for our approach. While from a bias perspective methods such as Zhu (2018) should lead to a better performance compared to the equation-by-equation OLS, the finite-sample Stambaugh only constitutes a small fraction to the overall estimation uncertainty for the equation-by-equation OLS, mainly because the equation-by-equation estimates, without using information from the population, are highly noisy. As such, removing the Stambaugh bias from equation-by-equation OLS will not have a large impact on estimation uncertainty. Our exercise, by comparing with the equation-by-equation OLS, suffices to highlight two results: 1. Our estimates for the population parameters are largely unbiased (similar to Zhu (2018)), whereas equation-by-equation OLS leads to severely biased estimates; and 2. Equation-by-equation OLS also leads to much noisier estimates, regardless of bias adjustment.

3,623 mutual funds for the 1991-2011 period. We use the four-factor model in Fama and French (1993) and Carhart (1997) as our benchmark model.

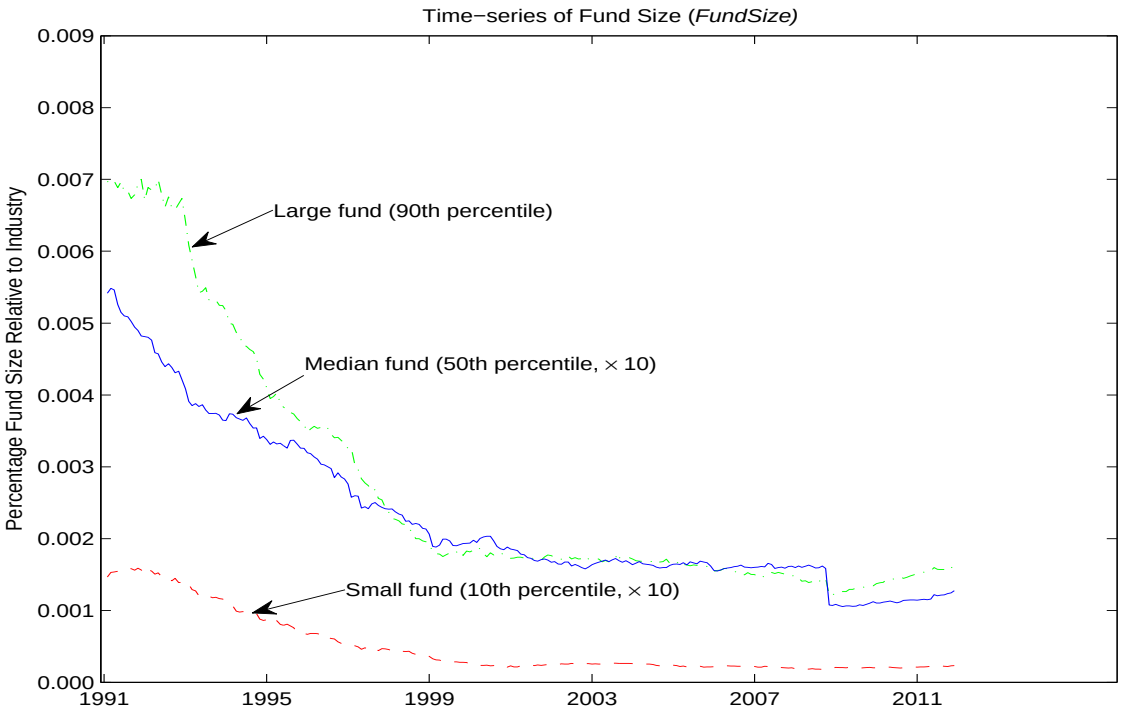
Figure 1 plots the time-series of industry size (as a proportion of the size of the equity market) and the number of funds in our sample. Figure 2 shows the evolution of the cross-sectional distribution of $FundSize$. There is a jump in the number of funds and industry size in September 2008 as many funds in our database (exceeding the $ 10 million size cutoff) start reporting returns right after the 2008 financial crisis and hence enter into our sample.

Figure 1: **Industry Size and Number of Funds**



Time-series of industry size and number of funds. Industry size $IndusSize$ is defined as the ratio between the TNA of the mutual fund industry and the aggregate market capitalization of the equity market. Missing TNA's for living funds are imputed using past returns and TNA's following Pastor et al. (2015). We also report the total number of funds that enter our regression analysis. A fund needs to satisfy two criteria to enter our analysis. First, it needs to have an initial TNA of at least \$10 million. Second, it needs to have at least 18 non-missing monthly observations for all variables in our regression analysis.

Figure 2: **Fund Size for the Cross-Section of Funds**



Time-series of individual fund size. At each point in time, we sort funds based on individual fund size $FundSize$, defined as the ratio of the TNA of an individual fund and the total TNA of the industry. We report the 10th, 50th, and 90th percentiles of the distribution of $FundSize$ at each point in time. We multiply the 10th and 50th percentiles by 10 so the three series are roughly on the same scale.

5 Results

5.1 Measuring Decreasing Returns to Scale

Table 1 reports the estimates for the parameters that govern the population of loadings on $IndusSize$ and $FundSize$. In our framework, they can be used to measure the impact of scale for an average fund in the cross-section.

To interpret the parameter estimates, let's first provide some summary statistics on $IndusSize$ and $FundSize$.[13] In our sample, the 50th and 90th percentile for the monthly change in $IndusSize$ are 0.43% and 1.5% per annum. For individual funds, the 50th and 90th percentile for the monthly change in the industry-scaled TNA are -1.3% and 86% per annum. Therefore, to have an apples-to-apples comparison, it seems fair to compare the economic magnitudes of the impact of scale for a change of 1.5% per annum in $IndusSize$ with a change of 86% per annum in $FundSize$, since both changes happen with a probability of 10%.

Based on Table 1, for $IndusSize$, a 1.5% annual increase in $IndusSize$ results in a decrease in alpha of 0.08% (=$0.0525 \times 1.5\%$) per annum for the average fund. For $FundSize$, an 86% annual increase in $FundSize$ results in a decrease in alpha of 0.18% (=$\log(1.86) \times 0.283$). Hence, although both $IndusSize$ and $FundSize$ have an economically significant impact on fund alpha, the impact of $FundSize$ is more than twice as large as the impact of $IndusSize$.

Moreover, the impact of $FundSize$ is estimated with a much higher precision than the impact of $IndusSize$. As shown in Table 1, the 90% confidence band for the mean loading on $FundSize$ (relative to the magnitude of the point estimate) is much narrower than the confidence band for the mean loading on $IndusSize$. Taking estimation uncertainty into account, the 90% confidence intervals for the impact on alpha corresponding to the aforementioned changes in $IndusSize$ and $FundSize$ are $[-12.2\text{bp}, -3.8\text{bp}]$ and $[-18.7\text{bp}, -16.8\text{bp}]$, respectively. Hence, in terms of the lower bound of the impact of scale, the difference between $FundSize$ and $IndusSize$ (i.e., -16.8bp vs. -3.8bp) is even more dramatic than the difference based on point estimates.

[13] Summary statistics are available upon request.

Table 1: **Model Estimates**

Parameter vector (Λ) for the estimated model. μ_γ and σ_γ are the mean and standard deviation of the normal distribution from which $\gamma_{i,1}$'s ($\gamma_{i,2}$'s) are drawn from. $\gamma_{i,1}$ and $\gamma_{i,2}$ are defined in (C.1) and denote the loading on industry size (i.e., $IndusSize_t$) and fund size (i.e., $FundSize_{i,t}$), respectively. $\rho = 0$ corresponds to the specification where residual correlation is set at zero. "Empirical ρ" corresponds to the correlation specification described in Table B.2.

	Loadings on $IndusSize_t$ ($\gamma_{i,1} \times 100$)	Loadings on $FundSize_{i,t}$ ($\gamma_{i,2} \times 100$)
μ_γ	-5.253	-0.283
[*p(5)*, *p(95)*] ($\rho = 0$)	[-7.943, -2.585]	[-0.359, -0.217]
[*p(5)*, *p(95)*] (Empirical ρ)	[-8.120, -2.452]	[-0.302, -0.270]
σ_γ	77.341	0.284
[*p(5)*, *p(95)*] ($\rho = 0$)	[74.901, 79.327]	[0.242, 0.346]
[*p(5)*, *p(95)*] (Empirical ρ)	[75.317, 78.031]	[0.246, 0.303]

To put our findings into context, we compare our results with PST. First of all, there are important differences between PST and our framework that are unrelated to the estimation of the impact of scale. For instance, while PST focus on fund alphas relative to the index-based benchmark,[14] we regress fund returns on benchmark factors (in particular, the four-factor model). While PST argue that index-based benchmarks better explain the cross-section of mutual fund returns,[15] we believe that it is important to control for standard risk factors to make sure that the impact of scale we are picking up for $FundSize$ corresponds to variations in $FundSize$ that are independent of the movements in standard risk factors. Another difference is our choice of data. While PST use the cross-validated dataset that reconciles the CRSP with Morningstar databases, we focus on the CRSP (with the same time sample), similar to many existing papers on mutual fund research.

The differences in implementation aside, PST find a much smaller and insignificant impact of $FundSize$. In particular, using their instrumental variables approach to adjust for the Stambaugh bias, they estimate an annual change in alpha in the range of 1.3 to 2.5bp for a $100 million increase in fund size.[16] In our sample, the median TNA is $122 million, so a $100 million inflow corresponds to a 82% change in $FundSize$, which would result in an annual change in alpha of 17bp, much higher than what PST estimate.

As we discussed before, this difference in results is attributable to several reasons. First, our definition of $FundSize$ is different from PST's definition. While PST's main specification uses the dollar TNA (divided by the size of the equity market), we take the log transformation of the TNA divided by the size of the fund industry. Simple as it is, this changes the interpretation of the regression coefficient: while PST try to estimate the impact on alpha per unit change in dollar TNA, we are estimating the impact on alpha per percentage change in TNA relative to the industry. Given the very large differences in dollar TNA across funds, we believe our transformation has the advantage of standardizing the magnitudes of the cross-section of regression coefficients on scale, allowing us to pool information from the cross-section to estimate the impact of scale for the average fund.

Second, while PST divide the dollar TNA by the size of the equity market, we divide the dollar TNA by the size of the mutual fund industry. As we discussed extensively in section 3.4, our definition allows us to purge out the variation of $IndusSize$ from PST's definition, leaving us with a measure of $FundSize$ that is independent of the movement of industry size.

[14] A fund is matched with a Morningstar designated benchmark. Fund alphas are calculated by simply subtracting the benchmark return from the fund return. There is no additional risk adjustment.

[15] See Cremers, Petajisto, and Zitzewitz (2013).

[16] Applying PST's approach as well as their variable definitions (but still using our benchmark adjustment), we find an estimate of 3.5bp for a $100 million increase in fund size, which is roughly consistent with PST's estimate given the fund samples have some differences even though the time sample is identical.

Putting the interpretation of the definition of $FundSize$ aside, one additional benefit in using our definition of $FundSize$ is that it helps kill the mechanical contemporaneous correlation between a fund's return and the growth rate of $FundSize$, alleviating the Stambaugh bias, as we discussed in the simulation study.

Lastly, our random coefficient framework has a different weighting scheme of information provided by the cross-section of funds than the fixed effects OLS model used in PST. While PST equally weight the cross-section of funds, our model down-weights information provided by funds with high residual standard deviations (as well as those with shorter samples) as seen from (8) and (9), which include smaller funds for which the inference on decreasing returns to scale should be more difficult. While this difference in the weighting scheme may not lead to a difference in the estimation of the population mean (as we shall see later in Table 5, the loadings on $FundSize$ are roughly homogenous across different size groups), it will have an impact on the statistical significance of the population mean. As a result, while we find a negative and significant population mean for the impact of $FundSize$, PST estimate the population mean to be negative but statistically insignificant.

Our estimate of the impact of $IndusSize$ is similar in magnitude to what PST find. For a 1.5% annual change in $IndusSize$, PST estimate a decrease in alpha (per annum) of 4.9bp,[17] which is similar in magnitude to our estimate of 8bp.[18] The difference between the estimates likely comes from the difference in the benchmark models we use. While PST use index-based benchmarks, we use the four-factor model. Our evidence suggests that the industry scale effect is even stronger after we purge out the variations in standard risk factors.

Overall, we find strong evidence for the impact of scale at the individual fund level, consistent Chen et al. (2004), Yan (2008) and Bris et al. (2007).[19] In contrast to these studies, our structural estimation framework that features fund fixed effects allows us to make a precise statement about the impact of scale for the cross-section of funds. We also find evidence consistent with the impact of scale at the industry level, supporting the findings in PST.

[17] Based on Table 3 in PST, a 1.5% annual change in $IndusSize$ results in a change of annual alpha of $0.049\% = (1.5\% \times 0.0326\%)$.

[18] Applying PST's approach (but using our benchmark adjustment), we find an estimate of 6.5bp for a 1.5% increase in $IndusSize$, which is in the same ballpark as PST's estimate.

[19] Zhu (2018) modifies the PST framework and shows how their narrative changes with her new estimation technique. We confirm Zhu's (2018) results by using a different framework. Different from Zhu (2018), our model features a new industry-adjusted fund size definition and a random effects model, both of which are important to drive our estimation results. Our heterogeneous-coefficient model also highlights the large degree of heterogeneity in fund-level decreasing returns to scale, which we later explore to predict future fund performance.

5.2 Dissecting the Impact of Scale

5.2.1 The Heterogeneous Impact of Scale: Cross-section

Given that our framework allows for heterogeneous loadings on two measures of scale, we are able to study how the impact of scale varies in both the cross-section and in the time-series. We first focus on the cross-sectional evidence.

For each fund, we calculate the median level of $FundSize$, which proxies for the average $FundSize$ throughout the lifetime of a fund. We then group the cross-section of funds into quintiles. Table 2 shows the average loadings on $IndusSize$ and $FundSize$ for each group.[20]

For the loadings on $FundSize$, there does not seem to be much cross-sectional variation. All loadings on $FundSize$ fall closely around the population estimate, i.e., the mean loading for the normal distribution from which the cross-section of loadings are drawn from. This, as we mentioned before, is attributable to the way we measure $FundSize$. By using log TNA, we are measuring the impact of scale per percentage change in a fund's TNA. Therefore, in spite of the difference in the dollar TNA across funds, the impact of scale seems homogenous across funds.

For the loadings on $IndusSize$, interestingly, we document a decreasing impact of scale when the median $FundSize$ is increasing. Hence, larger funds imply a milder response to changes in industry-level scale than smaller funds. The difference in impact of scale between large funds and small funds seems large from an economic standpoint. In particular, the impact of industry-level scale for very small funds (i.e., bottom 20% in terms of median $FundSize$) almost doubles that for very large funds (i.e., top 20% in terms of median $FundSize$).[21]

One possible explanation is to use the fact that small funds may trade illiquid stocks and large funds focus on liquid stocks and execute in large blocks, as shown in Chen et al. (2004) and Busse et al. (2016). Suppose the overall size of the fund industry relative to the aggregate equity market doubles and this affects each type of fund along the size spectrum proportionally (this assumes that there is no change in the composition of small vs. large funds, which is largely the case for the post-2000 periods, see Figure 2). Under this assumption, small funds grow by 100%. However, given the limited supply of small and illiquid stocks in the market, it becomes more difficult to invest in such stocks. Small funds are forced to invest in large and liquid stocks, which may not reflect their expertise. As a result, there is a decline in alpha.

[20] Alternatively, we can group funds into quintiles at each point in time based on the cross-section of fund sizes. We calculate the average loadings for each quintile, and then take the time-series average. Our results are similar.

[21] Note our random coefficients framework, by construction, assumes all loadings are continuous and therefore non zero. We therefore do not perform hypothesis testing for each fund. See extensive discussion in Harvey and Liu (2016) for the interpretation of model parameters in a random coefficients model.

In contrast, for large funds, even if their size also grows by 100%, since the market has a much larger capacity for large and liquid stocks, they may still be able to find new investment opportunities. As a result, they are not hurt as much as small funds, implying a milder response to an increase in industry size than the response of small stocks.[22]

Table 2: **Impact of Scale: Cross-section**

Impact of scale for the cross-section of funds. For each fund in our sample, we calculate its median $FundSize$. We group funds into different groups based on their median $FundSize$, each group covering 20% funds in our sample. We calculate the (cross-sectionally) averaged mean loadings on $IndusSize$ and $FundSize$ ("Avg. Estimate"). "Std. Err." reports the standard error

		Loadings on $IndusSize_t$ $\times 100$		Loadings on $FundSize_{i,t}$ $\times 100$	
	$FundSize$ quintiles	Avg. Estimate	Std. Err.	Avg. Estimate	Std. Err.
Small	$<= 0.029$ (20% of funds)	-8.699	2.931	-0.294	0.025
	$(0.029, 0.064]$ (20% of funds)	-5.711	3.067	-0.294	0.021
	$(0.064, 0.136]$ (20% of funds)	-4.970	2.451	-0.291	0.018
	$(0.136, 0.327]$ (20% of funds)	-4.955	2.790	-0.300	0.015
Large	> 0.327 (20% of funds)	-4.615	2.237	-0.281	0.014
Overall		-5.253	2.834	-0.283	0.016

[22] We do not think that this story is the only possibility. There are other plausible explanations. More empirical work (possibly based on holdings data) can help us better identify the source of the difference in the impact of industry size between small and large funds. We leave this to future research.

5.2.2 The Heterogeneous Impact of Scale: Time-series

We next examine the time-series variation of the impact of scale. In particular, at each point in time, we look at the cross-section of funds that are available. We obtain the loadings of these funds based on our full-sample estimates. We then calculate the cross-sectionally averaged loadings on $IndusSize$ and $FundSize$ for these funds. Figure 3 plots the time-series of these cross-sectionally averaged loadings.

For the impact of individual fund size (i.e., $FundSize$), there seems to be some time-series variation. For example, around 2001, right after the dot-com bubble bursts, the impact of $FundSize$ reaches its all-time high. This makes sense as when the market is bearish, there may be limited investment opportunities so a growth in a fund's size may have a large negative impact. Overall, the time-series of the impact of $FundSize$ seem to cluster around the population estimate, consistent with our evidence shown in the previous section that the cross-sectional variation in the impact of $FundSize$ is small.

For the impact of industry size (i.e., $IndusSize$), there appears to be a U-shaped pattern: the average loading on $IndusSize$ first declines, reaches its lowest around 2000, and then bounces back in 2011, reaching a level close to zero. We believe that this U-shaped pattern is not a coincidence. It provides information about the overall capacity of the mutual fund industry. In particular, we believe that the U-shaped pattern can be rationalized by the interaction between the dilution effect and the diminishing alpha effect, as we shall explain below.[23]

At the beginning of time t, suppose the amount of capital (in dollars) in the industry is C_t. The average fund is generating an alpha of α_0. By the end of the year, the industry generates a profit of $C_t\alpha_0$ (in dollars). Suppose the industry returns all the profits to investors, only keeping the initial capital (i.e., C_t). At the same time, there is a capital inflow of N_t. So in total the amount of capital is $C_t + N_t$ at the end of time t. However, due to the scarcity of investment ideas, for each dollar of the new capital coming in, funds only expect to generate an alpha of $\alpha_0 d_t$, where $d_t \in (0, 1)$ measures the diminishing alpha effect. Under these assumptions, the total amount of profits between time $t+1$ and $t+2$ is $C_t\alpha_0 + N_t\alpha_0 d_t$. Hence, the alpha generated by the average fund between $t+1$ and $t+2$ is:

$$\alpha_1 \quad = \quad \frac{C_t\alpha_0 + N_t\alpha_0 d_t}{C_t + N_t}.$$

[23] Another observation from Figure 3 is that there appears to be a sudden drop in the impact of industry size around 2008–2009. This can be explained by the influx of relatively small funds around the same period, as we see in Figure 1. Relatively small funds display a larger impact of $IndusSize$ as shown in Table 2. As such, a larger fraction of smaller funds would imply a temporarily lower averaged decreasing returns to scale.

The drop in alpha between the two periods is calculated to be:

$$\begin{aligned} \Delta\alpha &= \alpha_1 - \alpha_0, \\ &= \underbrace{(1-d_t)}_{dimishing\ alpha} \times \underbrace{\frac{N_t/C_t}{1+N_t/C_t}}_{dilution} \alpha_0 \end{aligned} \tag{10}$$

Note that we can use $\Delta\alpha$ to approximate the loading on $IndusSize$. (10) shows that $\Delta\alpha$ is driven by two effects. The *diminishing alpha* effect captures the idea that when good ideas are exhausted, we can only explore the not-so-good ideas (see, e.g., Chen et al., 2004) and hence experience a drop in alpha by $(1-d_t)\alpha_0$. The *dilution* effect captures how new capital dilutes existing capital. Notice that the dilution effect is the strongest when N_t/C_t is large, that is, when the amount of new capital is large relative to the amount of existing capital.

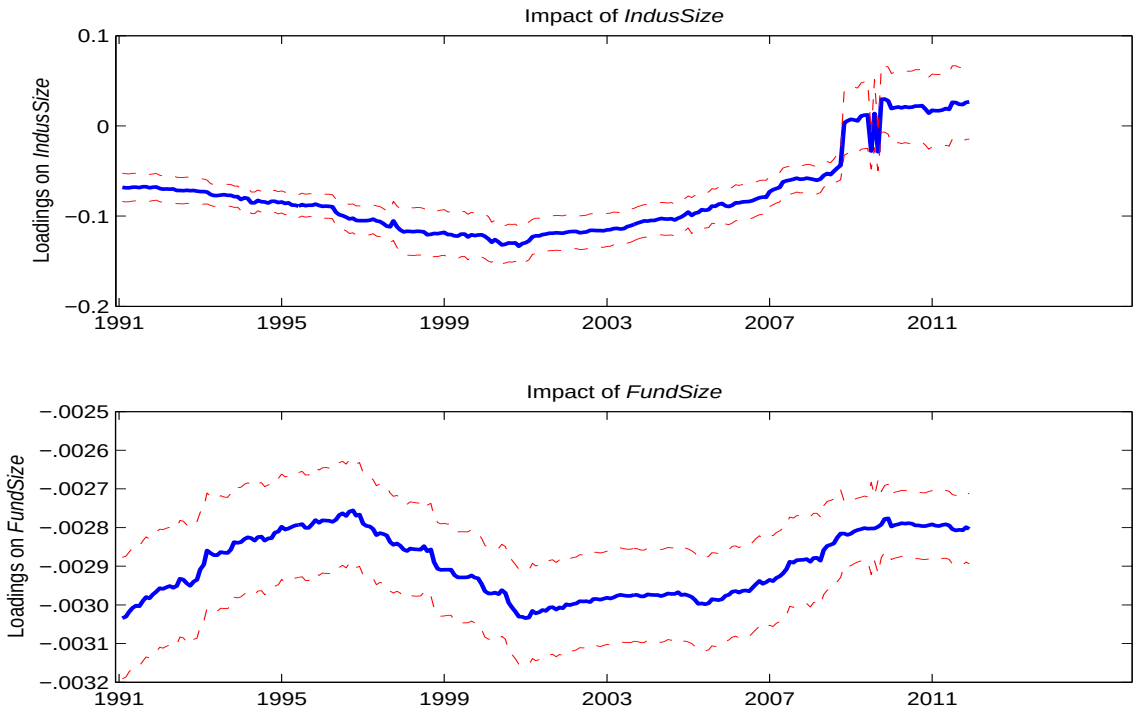
Using (10), we can account for the U-shaped pattern for the time-series of the response to $IndusSize$. Initially, C_t is small, so the amount of existing capital is low. This implies a strong dilution effect. Meanwhile, d_t is close to one as there are still plenty good investment ideas. In the extreme case, when $d_t = 1$ so each dollar of new capital comes in with a new and equally profitable investment idea as existing ideas, then the response to $IndusSize$ should be zero. Indeed, this explains why we see a small response to $IndusSize$ in 1991.

As more capital flows into the industry, C_t becomes large, so the dilution effect becomes smaller. At the same time, d_t goes down ($1-d_t$ goes up) as the quality of investment ideas deteriorates. As a result, the dilution effect and the diminishing alpha effect (i.e., $1-d_t$) work against each other. Overall, the diminishing alpha effect dominates, explaining the downward movement in the response to $IndusSize$ between 1991 and 2000. $\Delta\alpha$ reaches its peak in 2000 when the increase in $1-d_t$ exactly offsets the decrease in the dilution effect.

Finally, after 2000, the dilution effect takes over and remains the dominating effect in driving the response to $IndusSize$. This is consistent with the upward movement in the response to $IndusSize$ after 2000. In the extreme case, when there is already a large amount of capital (i.e., C_t) in the industry, the dilution effect of a capital inflow of \$1 should be close to zero. At the same time, $1-d_t$ is bounded above by one since in the worst case scenario, new capital comes in with no investment ideas (it is unlikely that new capital comes in with ideas that destroy alphas). As a result, $\Delta\alpha$ is close to zero, consistent with what we see for the response to $IndusSize$ around 2011.[24]

[24]Again, we are offering is a simple plausible story to explain the time-series dynamics of the response to $IndusSize$. Alternative explanations may be available. To distinguish among different ideas, we need to calibrate model parameters, in particular d_t that quantifies the impact of diminishing alpha. Such a calibration may require detailed trading data and is therefore beyond the scope of our paper.

Figure 3: **Impact of Scale: Time-series**



Time-series of cross-sectionally averaged loadings on scale. At each point in time, we identify funds that are available in our database. We obtain the loadings of these funds based on our full-sample estimate. We then calculate the cross-sectionally averaged loadings on $IndusSize$ and $FundSize$ for these funds. We plot the time-series of these averaged loadings. The dashed lines show the 95% confidence intervals.

5.3 Do Investors Respond to Decreasing Returns to Scale?

As with manager skill, the ability of a manager to resist decreasing returns to scale should also be heterogeneous. We explore the implications of this heterogeneity by creating long-short portfolios that sort the cross-section of funds by the estimates of their loadings on $IndusSize$ and $FundSize$ using past information.

In particular, we estimate our model using a rolling five-year window.[25] At time t, we look back five years to obtain the loadings on $IndusSize$ and $FundSize$ for each fund. We then create long-short portfolios of funds based on these loadings.

Table 3 presents the results on double sorts based on the loadings on $IndusSize$ and fund sizes.[26] We also present results on triple sorts based on the loadings on $IndusSize$, fund sizes and past performances in Appendix D. By taking a long position in funds with a larger loading on $IndusSize$ (i.e., less sensitive to industry decreasing returns to scale) and a short position in funds with a smaller loading (i.e., more sensitive to industry decreasing return to scale), we generate a positive and significant average return (both statistically and economically) across funds with different industry sizes and past performances.[27]

[25] Similar to our full-sample estimates, a fund needs to have at least 18 monthly observations to be considered. Our results are similar if we use an expanding window, that is, if we use all the information in the past to estimate our model. While five years may be considered too short to have a reliable estimate on decreasing returns to scale, fund portfolio sorts allow us to diversify away some estimation error for the fund-level parameter estimate and still generate meaningful signals to predict fund returns in the future.

[26] We use conditional sorts based on fund sizes and loadings to create our portfolios. In particular, we first sort funds into quintiles based on fund size. Within each quintile, we further sort funds into quintiles based on loadings. Our results are similar if we use independent sorts, that is, sorts that are based on cutoffs for fund sizes and loadings that are calculated independently of each other.

[27] There are five long-short strategies. After correcting for test multiplicity (see Harvey et al., 2016), it is likely that some of these strategies are statistically significant, even under Bonferroni's correction.

Table 3: **Portfolio Sorts Based on Loadings on $IndusSize$ and Fund Sizes**

Annualized alphas for strategies that sort the cross-section of funds based on loadings on $IndusSize$ and fund sizes. Using a rolling five-year window, we estimate our model to obtain the cross-section of loadings on $IndusSize$. We then sort funds into different groups based on fund sizes and the loadings. We use conditional sorts to first sort funds into size quintiles, and then sort funds into five groups based on the loadings. A low value of loading means the fund is sensitive to decreasing returns to scale. Our sample is from 1991 to 2011. We start sorting in 1996 to have the initial five-year window to estimate our model. Panel A subtracts the market excess return (i.e., no beta adjustment) from the fund excess return. Panel B reports fund alphas based on the Carhart (1997) four-factor model.

Panel A: Fund Returns Adjusted by Market Return

	Fund's TNA					
Loadings	*Small*	2	3	4	*Large*	All
Low	−0.346	−0.923	−0.393	−0.789	−0.722	−0.635
2	0.121	−0.204	−1.016	−1.526	−2.074	−0.940
3	0.073	−0.738	−1.474	−1.814	−1.253	−1.041
4	0.562	0.113	−0.385	−0.527	−0.500	−0.147
High	1.346	2.826	2.402	1.714	0.314	1.720
High - Low	1.692	3.749	2.795	2.503	1.037	2.355
(t-stat)	(1.00)	(3.73)	(2.68)	(2.27)	(0.92)	(2.65)

Panel B: Fund Returns Adjusted by 4-Factor Model

	Fund's TNA					
Loadings	*Small*	2	3	4	*Large*	All
Low	−1.000	−1.677	−1.318	−1.752	−1.068	−1.363
2	−0.434	−1.134	−1.620	−2.097	−2.150	−1.487
3	−0.393	−1.356	−1.798	−2.324	−1.219	−1.418
4	−0.449	−0.271	−0.897	−1.088	−0.778	−0.697
High	0.548	1.763	0.829	0.645	−0.364	0.684
High - Low	1.548	3.440	2.148	2.397	0.704	2.047
(t-stat)	(1.02)	(3.77)	(2.23)	(2.15)	(0.67)	(2.53)

We offer an explanation for the patterns in the long-short portfolio returns based on a return decomposition. Suppose the future return of a fund can be decomposed into $\alpha_i+\gamma_i^{ind}\Delta s^{ind}+\gamma_i^{fund}\Delta s^i$, where α_i is the alpha, γ_i^{ind} is the loading on $IndusSize$, Δs^{ind} is the log growth for industry size, γ_i^{fund} is the loading on $FundSize$, and Δs^i is the log growth for the size of individual fund i. Notice that we are assuming that idiosyncratic fund returns are zero since we are looking at well-diversified fund portfolios. Also, α_i should be thought of as the alpha corresponding to a given benchmark factor model. Under these assumptions, the long-short portfolio return (LS^{ind}) can be expressed as:

$$\begin{aligned}
LS^{ind} &= \frac{1}{N_h^{ind}}\sum_{\gamma_i^{ind}\geq\gamma_h^{ind}}[\alpha_i+\gamma_i^{ind}\Delta s^{ind}+\gamma_i^{fund}\Delta s^i]-\frac{1}{N_l^{ind}}\sum_{\gamma_i^{ind}\leq\gamma_l^{ind}}[\alpha_i+\gamma_i^{ind}\Delta s^{ind}+\gamma_i^{fund}\Delta s^i],\\
&= \underbrace{(\frac{1}{N_h^{ind}}\sum_{\gamma_i^{ind}\geq\gamma_h^{ind}}\alpha_i-\frac{1}{N_l^{ind}}\sum_{\gamma_i^{ind}\leq\gamma_l^{ind}}\alpha_i)}_{\alpha^{ind}}\\
&\quad+\underbrace{(\frac{1}{N_h^{ind}}\sum_{\gamma_i^{ind}\geq\gamma_h^{ind}}\gamma_i^{ind}-\frac{1}{N_l^{ind}}\sum_{\gamma_i^{ind}\leq\gamma_l^{ind}}\gamma_i^{ind})\Delta s^{ind}}_{S^{ind}}\\
&\quad+\underbrace{\frac{1}{N_h^{ind}}\sum_{\gamma_i^{ind}\geq\gamma_h^{ind}}\gamma_i^{fund}\Delta s^i-\frac{1}{N_l^{ind}}\sum_{\gamma_i^{ind}\leq\gamma_l^{ind}}\gamma_i^{fund}\Delta s^i}_{S_{fund}^{ind}},
\end{aligned}$$

where γ_h^{ind} and γ_l^{ind} are the loading cutoffs for the top quintile and bottom quintile, respectively, and N_h^{ind} and N_l^{ind} are the number of funds within the top quintile and bottom quintile, respectively.

We interpret the three parts (i.e., α^{ind}, S^{ind}, S_{fund}^{ind}) in the return decomposition separately. First, α^{ind} tends to be close to zero due to several reasons. Since there is evidence for alpha persistence in the short-run, then α_i can be thought of as the alpha estimate based on past data. However, we find that past loadings on $IndusSize$ (i.e., γ_i^{ind}) are approximately independent of past alpha estimates. Hence, a sort based on past loadings on $IndusSize$ does not generate an alpha differential, either for the past or for the future. As a result, α^{ind} is close to zero.[28]

The second component S^{ind} tends to be positive. This is due to two facts. First, the sort based on past loadings on $IndusSize$ makes the coefficient on industry growth (i.e., Δs^{ind}) positive. Second, Δs^{ind} is on average positive in our sample since the fund industry has been expanding. Taken together, S^{ind} tends to be positive since we take a long position on funds with a larger (and negative) loading on $IndusSize$ (i.e., more resistant to industry decreasing returns to scale) and a short position in funds

[28]In Appendix D, we show that our results still hold if we sort on past alphas.

with a smaller (and negative) loading in $IndusSize$ (i.e., less resistant to industry decreasing returns to scale), and that the industry is expanding on average. Notice that S^{ind} is a directional bet in the sense that it is positive only if the industry is expanding. If we expect the industry to shrink in the future, then we need to take the opposite position to generate a positive alpha.

For the third component S^{ind}_{fund} and for ease of exposition, let's make the assumption that Δs^i is homogeneous within the group $\gamma^{ind}_i \geq \gamma^{ind}_h$ ($\Delta s^i = \Delta s^h$) and $\gamma^{ind}_i \leq \gamma^{ind}_l$ ($\Delta s^i = \Delta s^l$) which simplifies S^{ind}_{fund} as:[29]

$$S^{ind}_{fund} \approx \Delta s^h \left(\frac{1}{N^{ind}_h} \sum_{\gamma^{ind}_i \geq \gamma^{ind}_h} \gamma^{fund}_i\right) - \Delta s^l \left(\frac{1}{N^{ind}_l} \sum_{\gamma^{ind}_i \leq \gamma^{ind}_l} \gamma^{fund}_i\right).$$

We also find that the past loadings on $IndusSize$ and $FundSize$ are approximately uncorrelated in the cross-section. This suggests that a sort on the loadings on $IndusSize$ will not generate a large differential for the loadings on $FundSize$. Hence, assuming $\frac{1}{N^{ind}_h} \sum_{\gamma^{ind}_i \geq \gamma^{ind}_h} \gamma^{fund}_i = \frac{1}{N^{ind}_l} \sum_{\gamma^{ind}_i \leq \gamma^{ind}_l} \gamma^{fund}_i = \bar{\gamma}^{fund}$, which is approximately the population mean of γ^{fund}_i, we have:

$$S^{ind}_{fund} = \bar{\gamma}^{fund}(\Delta s^h - \Delta s^l).$$

Notice that $\bar{\gamma}^{fund} < 0$ given our previous estimate. We later show that loadings on $IndusSize$ predict future fund flows in the cross-section in that funds that are more resistant to industry decreasing returns to scale attract more capital in the future. Hence, $\Delta s^h - \Delta s^l > 0$. Taken together, $S^{ind}_{fund} < 0$.

Adding up all three components, the fact that sorts based on the loadings on $IndusSize$ generate a positive return on average implies that S^{ind} (positive) dominates S^{ind}_{fund} (negative). When we sort funds based on the differential decreasing returns to scale at the industry level, we make a positive return through the on-average positive growth of the fund industry and by taking a long (short) position in funds that are more (less) resistant to industry decreasing returns to scale. However, funds that are more resistant to industry decreasing returns to scale attract more capital in the future (as we show later), which hurts their performance through decreasing returns to scale at the fund level. Balancing the positive return generated by industry growth with the negative return driven by individual fund growth, the overall effect is positive for the long-short portfolio.

[29] Our argument does not rely on our simplifying assumption that Δs^i is homogeneous within the two groups sorted by γ^{ind}_h and γ^{ind}_l. As long as one recognizes that Δs^i is on average higher within $\gamma^{ind}_i \geq \gamma^{ind}_h$ than within $\gamma^{ind}_i \leq \gamma^{ind}_l$ (as we show later through regressions based on fund flows), our argument goes through.

To summarize, we make a profit from the long-short strategies that are based on the loadings on $IndusSize$ through the differential exposure of the long and the short position to the growth of the mutual fund industry. Given the growth in the overall size of the mutual fund industry (scaled by the size of the equity market), funds that are less sensitive to $IndusSize$ suffer less than funds that are more sensitive, yielding the positive average return for the long-short strategies in Table 3.

Table 4 presents the results for double sorts based on the loadings on $FundSize$ and fund sizes. In Appendix D, we also present results for triple sorts based on the loadings, fund sizes, and past performances. Interestingly, notice that one needs to take a long (short) position in funds with a higher (lower) degree of decreasing returns to scale at the individual fund level to generate a positive return, contrary to our strategy based on the loadings on $IndusSize$.

Why do funds with a higher degree of decreasing returns to scale at the individual fund level (which should make them appear less "attractive") earn on average a higher return than funds with a smaller degree of decreasing returns to scale? To answer this question, we take a closer look at how the loadings on both $IndusSize$ and $FundSize$ correlate with future fund flows.

Table 5 reports the results of cross-sectional regressions that project future fund flows onto the loadings and existing variables that may explain future fund flows. While Table 5 shows that past loadings help predict future flows above and beyond the contribution of existing variables, Table 6 offers deeper insights into this predictability by exploring the nonlinear relationship between past loadings and future flows.

Table 6 sorts funds into portfolios based on past performances and the loadings, and calculates the average percentage flow for each portfolio. Previous literature documents that past performance helps predict future flows. Moreover, this relationship is convex in that the best past performers attract a disproportionate amount of capital in the future.[30] Our results in Table 6 show that the degree of decreasing returns to scale (both at the industry level and at the individual fund level) seems to be the omitted variable that is driving this convex relationship between past performance and future flows. For example, in Panel B, within funds with the best past performance (i.e., top 20%), funds with a large loading on $FundSize$ (i.e., top 20%) have an inflow of 53%, which almost triples the average inflow across the other funds that also have a great past performance.[31] A similar pattern holds for the loadings on $IndusSize$ as in Panel A. In addition, the relationship between loadings and future flows seems to be monotonic across different quintiles of past performance: funds with a higher loading on either $IndusSize$ or $FundSize$ (i.e., a lower degree of decreasing

[30]For the related literature on flow-performance sensitivity and the convex relation between past performances and future flows, see, e.g., Ippolito (1992), Gruber (1996), Chevalier and Ellison (1997), Sirri and Tufano (1998), Spiegel and Zhang (2013), Franzoni and Schmalz (2017), Starks and Sun (2016), and Harvey and Liu (2019).

[31]Based on Table 6 and Panel B, the average inflow across the other funds $= \frac{1}{4} \times (11.8\% + 16.3\% + 24.5\% + 22.8\%) = 18.9\%$.

returns to scale) attract more capital in the future than funds with a lower loading (i.e., a higher degree of decreasing returns to scale).

Table 4: **Portfolio Sorts Based on Loadings on $FundSize$ and Fund Size**

Annualized alphas for strategies that sort the cross-section of funds based on loadings on $FundSize$ and fund sizes. Using a rolling five-year window, we estimate our model to obtain the cross-section of loadings on $FundSize$. We then sort funds into different groups based on fund sizes and the loadings. We use conditional sorts to first sort funds into size quintiles, and then sort funds into five groups based on the loadings. A low value of loading means the fund is sensitive to decreasing returns to scale. Our sample is from 1991 to 2011. We start sorting in 1996 to have the initial five-year window to estimate our model. Panel A subtracts the market return from the fund excess return. Panel B reports fund alphas based on the Carhart (1997) four-factor model.

Panel A: Fund Returns Adjusted by Market Return

	Fund's TNA					
Loadings	*Small*	2	3	4	*Large*	All
Low	1.559	1.679	1.453	0.929	0.801	1.284
2	−0.930	−0.334	0.585	−0.258	−0.377	−0.263
3	−0.439	0.885	−0.073	−0.643	−0.844	−0.223
4	1.507	−0.494	−1.342	−2.130	−1.538	−0.799
High	0.182	−0.653	−1.447	−0.811	−2.275	−1.001
Low - High	1.378	2.332	2.900	1.739	3.076	2.285
(t-stat)	(1.03)	(2.48)	(2.72)	(1.92)	(3.32)	(3.02)

Panel B: Fund Returns Adjusted by 4-Factor Model

	Fund's TNA					
Loadings	*Small*	2	3	4	*Large*	All
Low	0.878	0.855	0.362	−0.137	0.562	0.504
2	−1.788	−1.498	−0.446	−1.242	−0.998	−1.194
3	−0.973	−0.091	−0.988	−1.434	−1.381	−0.973
4	0.804	−0.993	−2.062	−2.713	−1.634	−1.320
High	−0.434	−0.922	−1.639	−1.064	−2.125	−1.237
Low - High	1.312	1.776	2.001	0.927	2.687	1.741
(t-stat)	(1.08)	(2.16)	(2.28)	(1.19)	(3.00)	(2.79)

Table 5: **Forecasting Future Fund Flows**

Results for Fama-MacBeth regressions that regress the cross-section of fund flows on explanatory variables. Using a rolling five-year window, we estimate our model to obtain the cross-section of loadings on $IndusSize$ and $FundSize$. We then use these loadings and other control variables to explain the cross-section of future fund flows, which is defined as the one-year percentage growth in a fund's TNA. *Alpha (long-run)* is the average fund excess return over the past two years. *Alpha (short-run)* is the average fund excess return over the past quarter. $log(TNA)$ is the logarithm of the fund's TNA. *Fund age* is the total number of months that a fund exists up to the forecasting period. $IdioVol$ is the standard deviation for fund excess returns over the past two years. R-square reports the average cross-sectional R-square. Standard errors in parentheses.

Variables	1	2	3	4	5
Loadings on $IndusSize$		3.472		3.596	3.455
		(3.94)		(3.95)	(4.08)
Loadings on $FundSize$			18.484	19.189	12.796
			(5.31)	(4.98)	(4.09)
Alpha (long-run)	18.212	15.161	18.692	15.929	21.441
	(9.76)	(10.39)	(9.87)	(10.72)	(12.15)
Alpha (short-run)	3.117	3.086	3.283	3.208	2.074
	(6.61)	(6.07)	(6.75)	(6.27)	(3.73)
$log(TNA)$					−0.106
					(−6.09)
Fund age					−0.001
					(−3.49)
$IdioVol$					−1.306
					(−2.64)
Avg. R-square	0.040	0.050	0.045	0.053	0.077

Table 6: **Fund Flows for Portfolio Sorts Based on Loadings on $IndusSize$, $FundSize$ and Past Performance**

Percentage fund flows for portfolio sorts based on the loadings on $IndusSize/FundSize$ and past performances. Using a rolling five-year window, we estimate our model to obtain the cross-section of loadings on $IndusSize/FundSize$. We then sort funds into different groups based on past performances and the loadings. We use conditional sorts to first sort funds into performance quintiles, and then sort funds into five groups based on loadings. A low value of loading means the fund is sensitive to decreasing returns to scale. Past performance is measured as the average fund excess return in the past two years. Our sample is from 1991 to 2011. We start sorting in 1996 to have the initial five-year window to estimate our model.

Panel A: Flow Sorts Based on Loadings on $IndusSize$ and Past Performance

Loadings ($IndusSize$)	Past Performance					All
	Worst	2	3	4	*Best*	
Low	−0.125	−0.082	−0.035	0.047	0.176	−0.004
2	−0.127	−0.086	−0.031	0.090	0.152	−0.001
3	−0.127	−0.026	−0.025	0.073	0.178	0.015
4	0.080	−0.038	0.025	0.086	0.205	0.072
High	0.014	0.021	0.064	0.203	0.555	0.171
High - Low	0.139	0.103	0.098	0.156	0.379	0.175
(t-stat)	(4.08)	(8.67)	(10.11)	(3.91)	(3.78)	(9.17)

Panel B: Flow Sorts Based on Loadings on $FundSize$ and Past Performance

Loadings ($FundSize$)	Past Performance					All
	Worst	2	3	4	*Best*	
Low	−0.091	−0.053	−0.023	0.037	0.118	−0.002
2	−0.099	−0.054	−0.002	0.038	0.163	0.009
3	−0.128	−0.061	−0.028	0.082	0.245	0.022
4	0.010	−0.008	0.012	0.088	0.228	0.066
High	0.028	−0.031	0.042	0.261	0.526	0.165
High - Low	0.119	0.022	0.065	0.224	0.409	0.168
(t-stat)	(1.84)	(1.61)	(6.47)	(4.76)	(3.89)	(6.34)

Based on the results in Table 5 and 6, we offer an explanation for our results in Table 4, where we show that funds with a higher loading on $FundSize$ (greater exposure to decreasing returns to scale) perform better than funds with a lower loading. Funds with a higher loading on $FundSize$ (i.e., more resistant to individual fund decreasing returns to scale), controlling for past performances, appear to be more attractive to investors than funds with a lower loading. As a result, investors reward funds with a higher loading on $FundSize$ with a disproportionately larger amount of capital in the future. However, these funds cannot absorb these capital without sacrificing future performance (Berk and Green (2004); see more detailed discussion later), due to decreasing returns to scale. Consequently, the performance of these funds drops relative to funds with a lower loading on $FundSize$, who experience either a much smaller capital inflow or even a capital outflow.

One missing piece for the argument above is whether there is a substantial difference in the out-of-sample degree of decreasing returns to scale between funds with a high in-sample loading on $FundSize$ and those with a low in-sample loading on $FundSize$. If this were the case, then the large difference in decreasing returns to scale may justify the large difference in future flows between funds with a high loading on $FundSize$ and funds with a low loading, to the extent that there is no difference in future performances, contrary to what we see in Table 4 and Appendix Table C.2. In untabulated analyses, we show that this is not the case.[32] While past loadings on $FundSize$ are indicative of the future loadings, i.e., decreasing returns to scale is persistent, the cross-sectional variation in the future loadings is not large enough to wipe out the differential impact of future flows on future performances.

To summarize, our results highlight the dynamic relationships between past performances, decreasing returns to scale, and future flows. While past alphas are important in driving future flows, what is also important is how these alphas are earned. Funds that earn a high alpha while at the same time display abilities to resist decreasing returns to scale attract a disproportionately large amount of capital in the future, which reduces their performance in the future. Equivalently, funds that earn a low alpha while at the same time are sensitive to decreasing returns to scale experience a disproportionately large amount of capital outflow, which causes their performances to rebound in the future. Relative to the previous literature, the heterogeneous decreasing returns to scale we document appears to be an important conditioning variable to further our understanding of the link between past performance and future performances.

We now interpret our findings in the context of theoretical models such as Berk and Green (2004) that rationalize flow-performance relationship through diseconomies of scale. First of all, consistent with Berk and Green, we find that decreasing returns to scale, both at the industry level and at the individual fund level, is an integral

[32]By sorting funds based on the in-sample loadings estimated over a five-year window and calculating the average out-of-sample loadings over the following five years, we find that the average out-of-sample loading on $FundSize$ is -0.90% for funds that are ranked the top 20% in terms of in-sample loadings, and is -1.13% for the bottom 20%.

part of fund performance. The trading strategies we construct exploit the interaction between decreasing returns to scale and future fund flows. What is also consistent with Berk and Green but is not explicitly modeled in their framework is the heterogeneity in the degree of decreasing returns to scale. In particular, consistent with the idea that investors supply funds to managers competitively, we find that funds with a lower degree of decreasing returns to scale—hence a higher capacity in absorbing new capital without reducing performance—attract more capital than funds whose performance will likely suffer if additional capital is accepted. However, our results on portfolio sorts show that investors seem to overact to decreasing returns to scale in the sense that holding alphas constant, investors reward funds with a lower degree of decreasing returns to scale with a disproportionately larger amount of capital, to the extent that the future performances of these funds become lower (due to decreasing returns to scale) than those with a higher degree of decreasing returns to scale, which is at odds with the theoretical prediction of Berk and Green that funds should offer the same competitive market return in the future.

How plausible is our interpretation? We believe that decreasing returns to scale should be of first-order importance in the decision to allocate to a particular fund. First of all, investment advisors such as Morningstar explicitly pay attention to the capacities of funds. A basic search of key words related to fund size and performance on Morningstar generates thousands of results related to the size and performance relation.[33] Second, conditional on having a good record, fund managers are likely to boast about their increase in assets, if there is any. This may further signal the quality of the fund and help attract future flows.[34] This is consistent with our empirical analysis (Table 6) that there is a positive correlation between a low degree of decreasing returns to scale (i.e., achieving a good record despite an increase in size) and inflows in the future. Finally, theoretical models such as Berk and Green (2004) and Pastor and Stambaugh (2012) hinge on the first-order importance of decreasing returns to scale in generating the flow-performance relation. In their model economies, investors must be aware of the degree of decreasing returns to scale to allocate their capital efficiently, and, consistent with our empirical findings, investors respond to decreasing returns to scale.

Finally, our results on portfolio sorts highlight two important facts for decreasing returns to scale (both at the industry level and at the individual fund level). First, there is considerable cross-sectional heterogeneity in the degree of decreasing returns to scale, allowing us to create long-short portfolios that exploit this heterogeneity. Second, the degree of decreasing returns to scale estimated by our model is persistent, making it possible for us to use past information to predict future performance.[35] Both

[33]A search of "fund size, performance" on morningstar.com results in about 12,000 hits, many of which are related to the size and performance relation.

[34]In the hedge fund industry, contrary to mutual funds, a good hedge fund may signal its quality by closing some of its funds to new investment.

[35]Consistent with the literature on mutual funds that creates long-short trading strategies to demonstrate the usefulness of a certain signal to predict future fund performance, we also use long-short trading strategies to highlight the implications of decreasing returns to scale. However, one

facts help validate the underlying assumptions for our estimation framework: 1. The degree of decreasing returns to scale is fund specific; and 2. Our main specification assumes a constant loading on industry/fund size for each fund, consistent with the persistence in the degree of decreasing returns to scale.[36]

6 Conclusion

Berk and Green (2004) pave a new way for us to think about active portfolio management. However, there is mixed evidence in the literature for one of the key assumptions in their model, that is, diseconomies of scale for assets under management. We develop a new structural framework to estimate the impact of scale for mutual funds. Our framework allows for fund fixed effects and the heterogeneous impact of scale in the cross-section. Our model also does not suffer from the Stambaugh (1999) bias that plagues predictive regressions that include price-scaled variables as regressors. Importantly, we show that the way we measure fund size plays a key role in estimating the impact of scale.

The panel regression model we propose is a dynamic heterogeneous coefficients panel regression—a model where model parameters are mainly identified through time-series dynamics and where heterogenous loadings on regressors are allowed. In fact, one can think about our framework as first running separate time-series regressions, and then applying a shrinkage estimator to the cross-section of parameter estimates. This is a different approach than the standard fixed effects panel approach. The literature in growth economics proposes models that feature similar ideas.

We find strong evidence for decreasing returns to scale at the individual fund level. We also present evidence for decreasing returns to scale at the industry level, although its economic significance is smaller. By allowing heterogeneous loadings for funds, we also find that the impact of industry scale is decreasing in fund's size while the impact of fund scale is fairly homogeneous among different size groups.

Finally, we highlight the implications of heterogeneous decreasing returns to scale by creating long-short portfolios that exploit the differential sensitivities to decreasing returns to scale for the cross-section of funds. We show that both industry level and

should interpret these long-short trading strategies with caution. In our context, while we use long-short trading strategies to highlight the importance of persistence and heterogeneity in decreasing returns to scale in generating significant alpha differentials, such strategies are complicated by high transaction costs and the inability to short mutual funds. As such, the average returns for these trading strategies should be thought of as an indication of the expected alpha in choosing a fund based on the estimated decreasing returns to scale, rather than as the actual alpha earned by investing in a diversified portfolio of funds.

[36] One can extend our framework to incorporate time-varying decreasing returns to scale by modeling the loading parameters as functions of time-varying fund-level characteristics (e.g., age, TNA, etc.). We leave these extensions to future research.

fund level decreasing returns to scale can be used to construct long-short portfolios that generate a sizable alpha. To interpret our findings, we discover that decreasing returns to scale is the omitted variable that drives the convex relation between past performance and future flows: funds with the best performance attract a disproportionate amount of capital only if they display a low sensitivity to decreasing returns to scale.

References

Avramov, D., R. Kosowski, N. Y. Naik, and M. Teo. 2011. Hedge funds, managerial skill, and macroeconomic variables. *Journal of Financial Economics 99: 672–692.*

Banerjee, A. V. and E. Duflo. Inequality and growth: What can the data say? *Journal of Economic Growth 8:267–299.*

Barras, L., O. Scaillet, and R. Wermers. 2010. False discoveries in mutual fund performance: Measuring luck in estimated alphas. *Journal of Finance 65, 179-216.*

Barro, R. 1991. Economic growth in a cross-section of countries. *Quarterly Journal of Economics 106:407-443.*

Barro, R. and X. Sala-i-Martin. .1992. Convergence. *Journal of Political Economy 100:223-251.*

Backus, D. K., P. J. Kehoe, and T. J. Kehoe. 1992. In search of sacle effects in trade and growth. *Journal of Economic Theory 58, 377-409.*

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

Berk, J. B., and van Binsbergen, J. H. 2015. Measuring skill in the mutual fund industry. *Journal of Financial Economics 118, 1–20.*

van Binsbergen, Jules H., Jeong Ho John Kim, and Soohun Kim. 2020. Capital allocation and the market for mutual funds: Inspecting the mechanism. Georgia Tech Scheller College of Business Research Paper 3462749 (2020).

Bris, A., H. Gulen, P. Kadiyala, and P. Raghavendra, 2007, Good stewards, cheap talkers, or family men? The impact of mutual fund closures on fund managers, flows, feed, and performance. *Review of Financial Studies 20, 953–982.*

Carhart, M. M. 1997. On persistence in mutual fund performance. *Journal of Finance 52, 57–82.*

Chen, J., H. Hong, H. Ming, and J. D. Kubik, 2004. Does fund size erode mutual fund performance? The role of liquidity and organization. *American Economic Review 94, 1276–1302.*

Chen, Y., C. Cao, B. Liang, and A. Lo. 2013. Can hedge funds time market liquidity? *Journal of Financial Economics 109, 493-516.*

Chen, Y., M. Cliff, and H. Zhao. 2015. Hedge funds: The good, the bad, and the lucky. *Journal of Financial and Quantitative Analysis.*

Chevalier, J. A., and G. D. Ellison, 1997, Risk taking by mutual funds as a response to incentives. *Journal of Political Economy 105, 1167–1200.*

Cremers, M., A. Petajisto, and E. Zitzewitz. 2013. Should benchmark indices have alpha? Revisiting performance evaluation. *Critical Finance Review 2: 1-48.*

Dahlquist, M., M. Ibert, and F. Wilke. 2021. Expectations of active mutual fund performance. *Working Paper.*

Durlauf, S., A. Kourtellos, and A. Minkin. 2001. The local Solow growth model. *European Economic Review 45:928–940.*

Elton, E. J., M. J. Gruber, and C. R. Blake. 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.*

Evans, R. B. 2010. Mutual fund incubation. *Journal of Finance 65, 1581–1611.*

Fama, E. F., and K. R. French. 2010. Luck versus skill in the cross-section of mutual fund returns. *Journal of Finance 65, 1915-1947.*

Ferson, W., and R. Schadt. 1996. Measuring fund strategy and performance in changing economic conditions. *Journal of Finance 51, 425-460.*

Ferson, W., and Y. Chen. 2015. How many good and bad fund managers are there, really? *Working Paper.*

Franzoni, F. A., and M. C. Schmalz. 2017. Fund flows and market states. *Review of Financial Studies 30: 2621–2673.*

Glode, V., B. Hollifield, M. Kacperczyk, and S. Kogan. 2012. Time-varying predictability in mutual fund returns. *Working Paper.*

Golez, B., and S. Shive. 2015. When fund flows take the fun (alpha) away. *Working Paper.*

Gruber, M. J., 1996, Another puzzle: The growth in actively managed mutual funds. *Journal of Finance 51, 783–810.*

Harberger, A. 1987. Comment. *Macroeconomics Annual 1987,* Stanley Fischer, ed., Cambridge: MIT Press.

Harvey, C. R., Y. Liu, and H. Zhu. 2016. ... and the cross-section of expected returns. *Review of Financial Studies 29, 5–72.*

Harvey, C. R., and Y. Liu. 2016. Detecting repeatable performance. *Review of Financial Studies 31: 2499—2552.*

Harvey, C. R., and Y. Liu. 2019. Cross-sectional alpha dispersion and performance evaluation. *Journal of Financial Economics 134: 273–296.*

Harvey, C. R., Y. Liu, E. K. M. Tan, and M. Zhu. 2021. Crowding: Evidence from fund managerial structure. *Working Paper.*

Henriksson, R., and R. Merton. 1981. On market timing and investment performance II: Statistical procedures for evaluating forecasting skills. *Journal of Business 54, 513-534.*

Islam, N. 1995. Growth Empirics: A panel data approach. *Quarterly Journal of Economics 110:1127-1170.*

Ippolito, R. A., 1992. Consumer reaction to measures of poor quality: Evidence from the mutual fund industry. *Journal of Law and Economics 35, 45–70.*

Jones, C., and J. Shanken. 2005. Mutual fund performance with learning across funds. *Journal of Financial Economics 78, 507–552.*

Jones, C., and H. Mo. 2021. Out-of-sample performance of mutual fund predictors. *Review of Financial Studies 34: 149—193.*

Lee, K., M. H. Pesaran, and R. Smith. 1998. Growth empirics: A panel data appraoch — A comment. *Quarterly Journal of Economics: 319–323.*

Magkotsios, G. 2018. Industry-level returns to scale and investor flows in asset management. *Working Paper.*

Mankiw, N. G., D. Romer, and D. Weil. 1992. A contribution to the theory of economic growth. *Quarterly Journal of Economics 107:407-437.*

McLemore, P. 2019. Do mutual funds have decreasing returns to scale? Evidence from fund mergers. *Journal of Financial and Quantitative Analysis 54, 1683–1711.*

Pástor, L., R. Stambaugh. 2012. On the size of the active management industry. *Journal of Political Economy 120, 740–781.*

Pástor, L., R. Stambaugh, and L. A. Taylor. 2015. Scale and skill in active management. *Journal of Financial Economics 116, 23–45.*

Pástor, L., R. Stambaugh, and L. A. Taylor. 2020. Fund tradeoffs. *Journal of Financial Economics 138, 614–634.*

Phillips, P. C. B. and D. Sul. 2007. Transition modeling and econometric convergence tests. *Econometrica 75, 1771-1855.*

Searle, S. R., G. Casella, and C. E. McCulloch. 1992. Variance components. John Wiley & Sons, New York.

Sirri, E. R., and P. Tufano, 1998, Costly search and mutual fund flows. *Journal of Finance 53, 1589–1622.*

Solow, R. 1956. A contribution to the theory of economic growth. *Quarterly Journal of Economics 70:65-84.*

——. 1994. Perspectives on growth theory. *Journal of Economic Perspectives 8:45-54.*

——. 2001. Applying growth theory across countries. *World Bank Economic Review 15:283–288.*

Spiegel, M., and H. Zhang, 2013, Mutual fund risk and market share-adjusted fund flows, *Journal of Financial Economics 108, 506–528.*

Stambaugh, R. F. 1999. Predictive regressions. *Journal of Financial Economics 54: 375-421.*

Starks, L. T., and S. Y. Sun, 2016, Economic policy uncertainty, learning and incenetives: Theory and evidence on mutual funds. *Working Paper.*

Stoker, T. M. 1993. Empirical approaches to the problem of aggregation over individuals. *Journal of Economic Literature 31, 1827-1874.*

Swan, T. 1956. Economic growth and capital accumulation. *Economic Record 32:334-361.*

Treynor, J., and K. Mazuy. 1966. Can mutual funds outguess the market? *Harvard Business Review 44, 131–136.*

Yan, X., 2008. Liquidity, investment style, and the relation between fund size and fund performance. *Journal of Financial and Quantitative Analysis 43, 741–767.*

Zheng, S., A. W. Wang, and L. Zheng. 2014. Hedge fund performance persistence over different market conditions. *Working Paper.*

A Economic Growth and Decreasing Returns to Scale

First, while early work on economic growth focuses on cross-sectional regressions,[37] directly following the seminal work by Solow (1956) and Swan (1956), later generations of growth regressions use country fixed effects to allow for time-invariant idiosyncratic growth components.[38] We face a similar issue when running scale regressions. As pointed out by PST, funds that have a large size are more likely to fall into capable hands, creating an endogeneity bias if we simply run a cross-sectional regression of fund alpha on fund size. Following PST, we propose a dynamic panel regression approach that allows for fund fixed effects.

Second, one benefit of having the Solow neo-classical growth model to guide empirical explorations is that it guarantees that all variables are properly scaled, so regression coefficients correspond to the structural parameters in the model and have straightforward economic interpretations. For example, in a standard growth regression, income growth is regressed on log per capita GDP, where the slope coefficient β is the key variable of interest. By doing this, from a time-series perspective, β measures the change in income growth if log GDP per capita goes up by one unit (i.e., current GDP per capita grows by 100% relative to previous GDP per capita), regardless of the levels of GDP per capita across countries, whose distribution is rather dispersed in the cross-section. This is important as it makes it plausible to have a common slope that applies to all countries. In contrast, as implied by the neoclassical growth theory, it would be inappropriate to directly use the levels of GDP per capita. In growth economics, the use of log to scale level variables is so natural that researchers barely mention the reason to do so. This practice also applies to research outside the area of growth. For example, Backus, Kehoe and Kehoe (1992) study the impact of the size of the economy on trade and growth. They explicitly take log transformations of the various measures of scale that they study.

As we mentioned previously, the lack of a benchmark theoretical model such as the Solow growth model creates a challenge for empirical research on the impact of scale. For instance, PST directly use the level of the *total net assets* (TNA) — adjusted for the aggregated value of the equity market — to measure scale and study its impact on manager skill.[39] A common coefficient γ is assumed to apply to the cross-section of funds to pick up the impact of scale. In light of our previous discussion on the use of log GDP in growth research, it is challenging to interpret the γ coefficient. For example, for a small fund that has an initial TNA of \$10 million, doubling its size would imply a change of alpha of $\gamma \times$ \$10 million. For a large fund that has an initial

[37] See, e.g., Barro (1991), Barro and Sala-i-Martin (1992), and Mankiw, Romer, and Weil (1992).

[38] For an early influential paper, see Islam (1995).

[39] To be clear, PST also try the log of the TNA in one specification in their supplementary analysis. We advocate the use of the log of the TNA in our main analysis as this makes sure that the cross-section of regression parameters are comparable economically.

TNA of \$10 billion, an inflow of \$10 million would imply the same change of alpha in the PST framework. However, \$10 million is only 0.1% of the initial TNA of the large fund and thus should have a much smaller impact on its alpha than the impact on the small fund.

In measuring decreasing returns to scale, we consider two metrics. The first is the size of the fund industry as a whole relative to the total market capitalization of the equity market. The second is a fund specific measure of size where we look at a fund's size relative to the size of the fund industry. By doing this, the impact of scale, as captured by the regression coefficient, can be thought of as approximately homogeneous across funds, making it possible to be estimated through panel regressions by pooling information from the cross-section of funds. We show that this specification of TNA is essential to the estimation of the impact of scale and can imply dramatically different estimates than what existing papers find. We also consider a second measure of size — the size of the fund industry relative to the size of the stock market, similar to PST.

Third, one of the assumptions underlying the basic version of growth regressions is a common data generating process across all countries. One implication of this assumption is a common regression coefficient (i.e., γ) in the cross-section. However, this simple setup may be too restrictive to capture γ heterogeneity in the cross-section, as argued in an influential paper by Harberger (1987). In fact, Solow (1994, 2001) himself expresses the concern that different countries do not represent random draws from a common growth model. Heterogeneity should be taken into account to adjust for the difference in slopes. Recent papers that address the issue of parameter heterogeneity include Banerjee and Duflo (2003), Durlauf, Kourtellos, and Minkin (2001), Kevin, Pesaran, and Smith (1998), and Phillips and Sul (2007).

When it comes to the impact of assets under management, we believe that the heterogeneous impact of scale could be even more important. This is because, unlike Solow's growth model where we have well-specified and micro-founded (to a certain extent) production functions to characterize countries' income growth processes, there is a large amount of model uncertainty — both in terms of the business models that fund managers use to generate returns and the statistical models that econometricians use to make inference on alphas — for the data generating process of fund returns. Model uncertainty makes it unlikely that a common regression coefficient (i.e., γ) is sufficient to describe the cross-sectional impact of scale. Moreover, in the context of performance evaluation, just as with manager skill, the ability of a manager to resist decreasing returns to scale should also be manager specific. We propose a framework that captures this heterogeneity while at the same time producing an estimate for the cross-sectionally averaged impact of scale, which makes it possible to interpret the impact of assets under management in general.

Importantly, as remarked by Solow (2001) in the context of growth regressions, one has to recognize that parameter heterogeneity is unlikely captured by control variables, which include the hundreds of variables that have been proposed to ex-

plain cross-country growth differences. In our context, while we can use fund level characteristics as instruments to capture parameter heterogeneity, we will probably never know whether a given list of characteristics is exhaustive or what the consequences are of omitting several instruments. However, our model does not rely on a pre-specified list of characteristics. Instead, it uncovers the regression coefficient (i.e., the loading on fund size or industry size) of an individual fund by combining cross-sectional information with the fund's time-series information.

B Model Estimation

B.1 Characterizing $f(\mathbf{\Gamma}|\mathcal{R}, \mathcal{G}^{(k)})$: *Step II*

Using Bayes' law, we have:

$$f(\mathbf{\Gamma}|\mathcal{R}, \mathcal{G}^{(k)}) \propto f(\mathcal{R}|\mathbf{\Gamma}, \mathcal{G}^{(k)}) f(\mathbf{\Gamma}|\mathcal{G}^{(k)}). \tag{B.1}$$

Given the independence of the residuals and the γ_i's, the right-hand side of (B.1) is the product of the likelihoods of all funds, i.e.:

$$f(\mathcal{R}|\mathbf{\Gamma}, \mathcal{G}^{(k)}) f(\mathbf{\Gamma}|\mathcal{G}^{(k)}) \;\; = \;\; \prod_{i=1}^{N} f(R_i|\gamma_i, \mathcal{G}^{(k)}) f(\gamma_i|\mathcal{G}^{(k)}).$$

To characterize $f(\mathbf{\Gamma}|\mathcal{R}, \mathcal{G}^{(k)})$, it is sufficient to determine $f(R_i|\gamma_i, \mathcal{G}^{(k)}) f(\gamma_i|\mathcal{G}^{(k)})$ for each fund. To avoid the cluster of notations, we use $\mathcal{G}$ and $\mathcal{G}^{(k)}$ interchangeably to denote the known parameters at the k-th iteration.

Under normality and for the simple case where γ_i's are scalars (that is, there is only one fund characteristic that affects returns), it can be shown that

$$f(R_i|\gamma_i, \mathcal{G}) f(\gamma_i|\mathcal{G}) \propto \exp\{-\frac{[\gamma_i - \frac{\sigma_\gamma^2 \sum_{t=1}^{T} g_{i,t}(r_{i,t} - \beta_i' f_t) + \sigma_i^2 \mu_\gamma}{\sigma_\gamma^2 \sum_{t=1}^{T} g_{i,t}^2 + \sigma_i^2}]^2}{2\frac{\sigma_\gamma^2 \sigma_i^2}{\sigma_\gamma^2 \sum_{t=1}^{T} g_{i,t}^2 + \sigma_i^2}}\},$$

where μ_γ and σ_γ are the mean and the standard deviation of the population of γ_i's, that is, $\gamma_i|\mathcal{G} \sim \mathcal{N}(\mu_\gamma, \sigma_\gamma^2)$.

Hence, let

$$\begin{aligned} m_i &\equiv \frac{\sum_{t=1}^{T} g_{i,t}(r_{i,t} - \beta_i' f_t)/\sigma_i^2 + \mu_\gamma/\sigma_\gamma^2}{\sum_{t=1}^{T} g_{i,t}^2/\sigma_i^2 + 1/\sigma_\gamma^2}, \\ v_i &\equiv \frac{1}{\sum_{t=1}^{T} g_{i,t}^2/\sigma_i^2 + 1/\sigma_\gamma^2}, \end{aligned}$$

we have

$$\gamma_i | \mathcal{R}, \mathcal{G} \sim \mathcal{N}(m_i, v_i).$$

In general, when there are more than one fund characteristics that affect returns, the formulas are more involved as loadings on characteristics are generally not independent of each other. We present the formulas for the case with two characteristics, corresponding to our main application in the paper. More general cases can be similarly derived.

For the case with two characteristics, let $g_{i,1} = [g_{i,1,t=1}, g_{i,1,t=2}, \ldots, g_{i,1,t=T}]'$ and $g_{i,2} = [g_{i,2,t=1}, g_{i,2,t=2}, \ldots, g_{i,2,t=T}]'$ be the two column vectors of fund characteristics for fund i. We combine $g_{i,1}$ and $g_{i,2}$ into the $T \times 2$ matrix G_i of fund characteristics, i.e., $G_i = [g_{i,1}, g_{i,2}]$.

Next, let $b_1 = [1, 0]'$ and $b_2 = [0, 1]'$ be two basis vectors. Let $res_i = [r_{i,t=1} - \beta_i' f_{t=1}, r_{i,t=2} - \beta_i' f_{t=2}, \ldots, r_{i,t=T} - \beta_i' f_{t=T}]'$ be the column vector of residuals.

Define

$$\begin{aligned} VAR_i &= (G_i' G_i/\sigma_i^2 + b_1 b_1'/\sigma_{\gamma,1}^2 + b_2 b_2'/\sigma_{\gamma,2}^2)^{-1}, \\ CPROD_i &= (G_i' res_i)/\sigma_i^2 + b_1 \mu_{\gamma,1}/\sigma_{\gamma,1}^2 + b_2 \mu_{\gamma,2}/\sigma_{\gamma,2}^2, \\ MEAN_i &= VAR_i \cdot CPROD_i, \end{aligned}$$

where $(\mu_{\gamma,1}, \sigma_{\gamma,1})$ and $(\mu_{\gamma,2}, \sigma_{\gamma,2})$ are the mean and standard deviation for the population of $\gamma_{i,1}$'s and $\gamma_{i,2}$'s, respectively.

Then $\gamma_i = [\gamma_{i,1}, \gamma_{i,2}]'$ follows a bivariate normal distribution with mean vector $MEAN_i$ and variance matrix VAR_i, that is,

$$\gamma_i | \mathcal{R}, \mathcal{G} \sim \mathcal{N}(MEAN_i, VAR_i).$$

B.2 Finding the MLE: *Step III*

We again first focus on the univariate case, that is, there is a single fund characteristic that is driving fund returns. To find the MLE, we first need to calculate $E_{\mathbf{\Gamma}|\mathcal{R},\mathcal{G}^{(k)}}[\sum_{i=1}^{N}\log f(R_i|\gamma_i,\beta_i,\sigma_i)f(\gamma_i|\Lambda)]$, which can be decomposed as

$$\begin{aligned} & E_{\mathbf{\Gamma}|\mathcal{R},\mathcal{G}^{(k)}}[\sum_{i=1}^{N}\log f(R_i|\gamma_i,\beta_i,\sigma_i)f(\gamma_i|\Lambda)], \\ = \; & E_{\mathbf{\Gamma}|\mathcal{R},\mathcal{G}^{(k)}}[\sum_{i=1}^{N}\log f(R_i|\gamma_i,\beta_i,\sigma_i)] + E_{\mathbf{\Gamma}|\mathcal{R},\mathcal{G}^{(k)}}[\sum_{i=1}^{N}\log f(\gamma_i|\Lambda)]. \end{aligned}$$

We first focus on $E_{\mathbf{\Gamma}|\mathcal{R},\mathcal{G}^{(k)}}[\sum_{i=1}^{N}\log f(R_i|\gamma_i,\beta_i,\sigma_i)]$ and try to find the MLE for $\{\beta_i\}_{i=1}^{N}$ and $\{\sigma_i\}_{i=1}^{N}$. Note that $f(R_i|\gamma_i,\beta_i,\sigma_i)$ is a normal density and its kernel, which is the only part that involves γ_i, is the sum of squares of the regression residuals, i.e., $\sum_{t=1}^{T}(r_{i,t}-\beta_i'f_t-\gamma_i g_{i,t})^2$. Taking the expectation of $\log f(R_i|\gamma_i,\beta_i,\sigma_i)$ with respect to $\gamma_i|\mathcal{R},\mathcal{G}^{(k)}$, whose distribution is given in the previous section, we have:

$$E_{\gamma_i|\mathcal{R},\mathcal{G}^{(k)}}\log f(R_i|\gamma_i,\beta_i,\sigma_i) = -\frac{T}{2}\log(2\pi\sigma_i^2) - \frac{1}{2\sigma_i^2}[\sum_{t=1}^{T}(r_{i,t}-\beta_i'f_t-m_i g_{i,t})^2 + v_i\sum_{t=1}^{T}g_{i,t}^2]. \tag{B.2}$$

Treating $\{r_{i,t}-m_i g_{i,t}\}_{t=1}^{T}$ as the characteristic adjusted returns, the MLE for β_i and σ_i^2 can be found as

$$\beta_i^{MLE} = (F'F)^{-1}F'R_i^{adj}, \tag{B.3}$$

$$(\sigma_i^2)^{MLE} = \frac{1}{T}[\sum_{t=1}^{T}(r_{i,t}-(\beta_i^{MLE})'f_t-m_i g_{i,t})^2 + v_i\sum_{t=1}^{T}g_{i,t}^2], \tag{B.4}$$

where

$$F_{(T\times(K+1))} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_T \end{bmatrix}, R_{i(T\times 1)}^{adj} = \begin{bmatrix} r_{i,1}-m_i g_{i,1} \\ r_{i,2}-m_i g_{i,2} \\ \vdots \\ r_{i,T}-m_i g_{i,T} \end{bmatrix}.$$

We now focus on $E_{\mathbf{\Gamma}|\mathcal{R},\mathcal{G}^{(k)}}[\sum_{i=1}^{N}f(\gamma_i|\Lambda)]$ and try to find the MLE for $\Lambda=(\mu_\gamma,\sigma_\gamma^2)'$. Taking the expectation of $\sum_{i=1}^{N}\log f(\gamma_i|\Lambda)$ with respect to $\mathbf{\Gamma}|\mathcal{R},\mathcal{G}^{(k)}$, we have:

$$E_{\mathbf{\Gamma}|\mathcal{R},\mathcal{G}^{(k)}}[\sum_{i=1}^{N}\log f(\gamma_i|\Lambda)] = -\frac{N}{2}\log(2\pi\sigma_\gamma^2) - \frac{1}{2\sigma_\gamma^2}[\sum_{i=1}^{N}(m_i-\mu_\gamma)^2 + \sum_{i=1}^{N}v_i]. \tag{B.5}$$

The MLE for μ_γ and σ_γ^2 can be found as

$$\mu_\gamma^{MLE} = \frac{1}{N}\sum_{i=1}^{N} m_i, (\sigma_\gamma^2)^{MLE} = \frac{1}{N}\sum_{i=1}^{N}(m_i - \mu_\gamma^{MLE})^2 + \frac{1}{N}\sum_{i=1}^{N} v_i. \tag{B.6}$$

For the bivariate case, let $R_i^{adj,bi} = R_i - G_i \cdot MEAN_i$. The MLE for β_i is

$$\beta_i^{MLE,bi} = (F'F)^{-1}F'R_i^{adj,bi}. \tag{B.7}$$

Let $res_i^{bi} = R_i^{adj,bi} - F\beta^{MLE,bi}$. The MLE for σ_i^2 is

$$(\sigma_i^2)^{MLE,bi} = ((res_i^{bi})' res_i^{bi} + tr(G_i' G_i VAR_i))/T. \tag{B.8}$$

The MLE for $(\mu_{\gamma,1}, \sigma_{\gamma,1}^2)$ and $(\mu_{\gamma,2}, \sigma_{\gamma,2}^2)$ are given by:

$$\begin{aligned}
\mu_{\gamma,1}^{MLE,bi} &= \frac{1}{N}\sum_{i=1}^{N} MEAN_{i,1}, \\
(\sigma_{\gamma,2}^2)^{MLE,bi} &= \frac{1}{N}\sum_{i=1}^{N}(MEAN_{i,1} - \mu_{\gamma,1}^{MLE,bi})^2 + \frac{1}{N}\sum_{i=1}^{N} VAR_{i,11}, \\
\mu_{\gamma,2}^{MLE,bi} &= \frac{1}{N}\sum_{i=1}^{N} MEAN_{i,2}, \\
(\sigma_{\gamma,2}^2)^{MLE,bi} &= \frac{1}{N}\sum_{i=1}^{N}(MEAN_{i,2} - \mu_{\gamma,2}^{MLE,bi})^2 + \frac{1}{N}\sum_{i=1}^{N} VAR_{i,22},
\end{aligned}$$

where $MEAN_{i,1}$ and $MEAN_{i,2}$ are the first and second element of $MEAN_i$, and $VAR_{i,11}$ and $VAR_{i,22}$ are the upper-left and lower-right element of VAR_i.

C A Simulation Study

We detail a comprehensive simulation study to examine the performance of our model, paying particular attention to the finite-sample bias issue in Stambaugh (1999) and PST.

Given a $T \times N$ panel of fund returns, we obtain summary statistics and parameter estimates that later will be used to generate random return panels. First, we record the first months in which funds report a TNA. These will be the entry months for funds. Next, for each fund, we run OLS to estimate loadings on characteristics ($IndusSize_t$ and $FundSize_{i,t}$) as well as on benchmark factors, i.e.,

$$r_{i,t+1} = \alpha_i + \gamma_{i,1} IndusSize_t + \gamma_{i,2} FundSize_{i,t} + \sum_{j=1}^{K} \beta_{ij} f_{j,t+1} + \varepsilon_{i,t+1}, \quad \text{(C.1)}$$

where $r_{i,t+1}$ is the excess return (i.e., actual return minus the one-month U.S. Treasury bill rate) for fund i in period $t+1$, α_i is the unconditional alpha, $\gamma_{i,1}$ and $\gamma_{i,2}$ are the loadings on characteristics, $\{\beta ij\}_{j=1}^{K}$ includes exposures to benchmark factors, and ε_{t+1} is the return residual. We store the estimates of alpha and loadings on benchmark factors in $\beta_i = [\alpha_i, \beta_{i1}, \beta_{i2}, \ldots, \beta_{iK}]'$ and the estimate of the residual standard deviation in σ_i. Since we allow up to four benchmark factors (Carhart, 1997) and thus in total six independent variables, we require that a fund has at least 18 non-missing monthly observations for all variables in the above regression to enter our simulation.

For each fund, we also run OLS to determine how its growth in TNA depends on contemporaneous fund returns. In particular, we estimate

$$\frac{TNA_{i,t+1}}{TNA_{i,t}} - 1 = c_{0i} + c_{1i}(r_{i,t+1} + r_{f,t}) + \eta_{i,t+1}, \quad \text{(C.2)}$$

and record c_{0i}, c_{1i} and $\sigma_{\eta,i}$, where $\sigma_{\eta,i}$ is the estimate of the residual standard deviation in the above regression. Notice that the Stambaugh bias (Stambaugh, 1999) arises in the above setup as $r_{i,t+1}$ is driving the growth in TNA so it is positively correlated with innovations in TNA. As shown in Stambaugh, this creates downward bias in the regression slope coefficient if one were to regress returns on lagged TNA. PST also implement the above regressions in their simulation study but fix the cross-section of c_{0i}'s, c_{1i}'s, and $\sigma_{\eta,i}$'s at (roughly) their cross-sectional averages to simulate the path of TNA's for each fund. We deviate from their framework by using fund specific parameter values to generate each fund's path of TNA's. However, estimates for fund specific parameters can be extremely large, due to small samples for some funds. We therefore first require that a fund has at least 18 non-missing monthly observations for all variables in the above regression to enter our simulation. In addition, we

winsorize c_{0i}'s, c_{1i}'s, and $\sigma_{\eta,i}$'s at their respective 10th and 90th percentiles of the cross-sectional estimates.[40]

With the above parameter estimates, we are ready to simulate random return panels. Given the interdependence between returns and TNA (in particular, TNA depends on the contemporaneous return through (C.2); return depends on lagged TNA through (C.1)), we generate the return panel recursively.

First of all, to take time-series uncertainty into account, we generate random samples of factor realizations and the aggregate market capitalization for stocks $AggStock$. Since there is little persistence in factor realizations, we resample the time periods to generate random samples of factor returns. The aggregate market capitalization for stocks is persistent. We therefore first fit an AR(1) model on the time-series of $\log(AggStock)$ and then bootstrap the residuals. Importantly, similar to Fama and French (2010), we keep the cross-section intact when we resample the time periods, making sure that the cross-sectional dependency among factor returns and innovations to $AggStock$ is preserved.[41] We resample and obtain one bootstrapped sample of factor returns $\{\hat{f}_t\}_{t=1}^{T}$ and aggregate market capitalization $\{\widehat{AggStock}\}_{t=0}^{T-1}$.[42]

To generate a random return and TNA panel, we first randomly draw the cross-section of $\gamma_{i,1}$'s and $\gamma_{i,2}$'s from two normal distributions: $\mathcal{N}(\mu_{\gamma,1}, \sigma^2_{\gamma,1})$ and $\mathcal{N}(\mu_{\gamma,2}, \sigma^2_{\gamma,2})$. We collect the parameters into $\Lambda = [\mu_{\gamma,1}, \sigma_{\gamma,1}, \mu_{\gamma,2}, \sigma_{\gamma,2}]'$ and will later set Λ at values that are consistent with our estimates based on the actual data. These will be the structural parameters that we try to make inference on through our estimation procedure.

To generate a random return and TNA panel, we start at time zero, which is the beginning of the first month. We first generate size-related variables, that is, TNA's, $IndusSize$, and $FundSize$. We look at the actual data on TNA and find funds that report TNA at time zero. We calculate the aggregate TNA across funds and generate $FundSize$. With the bootstrapped market capitalization at time zero (i.e., $\widehat{AggStock}_0$), we can generate $IndusSize$.

Moving forward to the next step, we simulate fund returns that are realized between time 0 and $t = 1$. In particular, we first find funds that have non-missing $IndusSize$ and $FundSize$ in the previous step (i.e., time zero). We then apply (6) to generate returns for these funds, using $IndusSize$ and $FundSize$ from the previous step, factor returns from the bootstrapped sample, and an independent normal shock that has the same standard deviation as the fund in the actual data. Factor

[40] The particular way we winsorize the parameters does not affect our simulations results.

[41] Our simulation results do not change if we do not resample $AggStock$. However, since the market factor in the four-factor model is correlated with innovations in $AggStock$, in the same way as how fund returns correlate with innovations in TNA in (7), we believe it makes more sense to resample both $AggStock$ and factor returns.

[42] Since we fit an AR(1) on $AggStock_t$, the first observation in the bootstrapped sample will always be $AggStock_0$ in the actual data.

loadings are the same as the estimated factor loadings for the actual data. Loadings on characteristics are randomly generated previously.

Next, we simulate size-related variables at the beginning of the second month. Similar to before, we look at the actual data to identify funds that report data on TNA for the first time. These will be the new entries to the fund industry. Different from before, we also have incumbent funds that survive the first month. For these funds, their TNA evolution, as emphasized in PST, will follow (7) in that innovations in TNA are correlated with contemporaneous fund returns. We therefore follow (7) to generate TNA for these funds at the beginning of the second month, using fund returns from the previous step (i.e., returns realized between time zero and $t = 1$) and an independent normal shock that has the same standard deviation as the estimated equation (7) in the actual data. The regression coefficients in (7) are also the same as their estimates for the actual data. We record TNA's for both new entries and incumbent funds. We generate $FundSize$ for these funds. With the bootstrapped market capitalization at $t = 1$ (i.e., $\widehat{AggStock}_1$), we can generate $IndusSize$.[43]

We next simulate fund returns that are realized between $t = 1$ and $t = 2$. This is the same as the previous step where we simulate fund returns between $t = 0$ and $t = 1$. After this, we simulate size-related variables at the beginning of the third month, exactly following the previous step where we simulate these variables at the beginning of the second month. Hence, we follow the above steps recursively to fill in the entire panel of fund returns and TNA.

Several features mark our simulation process. First, we follow PST to explicitly model the dynamics of TNA through (7). As argued in PST, this is the main channel that creates bias for traditional inference. Second, we take into account the heterogeneity in parameters that are fund specific. For example, we allow factor loadings to mimic their cross-sectional distributions for the actual data. As another example, we also allow c_{0i} and c_{1i} to be fund specific, in contrast to PST. Third, our simulated sample provides a realistic description of both the return panel and the TNA panel if history repeats. In particular, we keep track of fund entries, which would be difficult to describe if one were to model fund entries as a stochastic process.[44] Overall, we believe that our simulation process incorporates important features of the actual data while recognizing the sources of uncertainty in the return and TNA generating process, providing a fair setup to evaluate model performance.

Table B.1 reports our choice of Λ that governs the loadings on fund characteristics. Our choice of Λ follows our model estimates on the actual data (which we haven't

[43]Notice that the simulated $IndusSize$ in our simulation study mimics the time trend and autocorrelation of the time-series of $IndusSize$ in the actual data. We show that our estimates are still consistent.

[44]In our simulated samples, the history of fund entries is always the same as the actual data while we independently sample the innovations for the aggregate stock market capitalization (through bootstrap). This implicitly assumes that fund's entry is independent of the innovations in the aggregate stock market capitalization. However, this is not key for our simulation results. Our model still performs well if we do not resample the aggregate stock market capitalization.

discussed yet, see section 3). To offer some economic interpretations of the model parameters, $\mu_{\gamma,1} = -0.05$ means that a 1% annual increase in $IndusSize$ (that is, the percentage of the size of the mutual fund industry relative to the aggregate market capitalization of all stocks goes up by 1%) implies a 0.05% ($= 0.05 \times 1\%$) decrease in alpha (per annum) for the average fund in the cross-section, while $\mu_{\gamma,2} = -0.003$ implies a 0.2% ($= \log(2^{1/12}) \times 0.003 \times 12$) decrease in alpha (per annum) if the average fund in the cross-section doubles its size over a year. Both effects are economically significant. However, the impact of scale at the individual fund level is higher than at the industry level. We offer more detailed interpretations of our model parameters in the results section.

Table B.1: **Parameter Vector (Λ^*) for the Simulated Model**

Parameter vector (Λ^*) for the simulated model. We choose parameter values in Λ that are similar to our estimates based on the actual data (see Table 4). μ_γ and σ_γ are the mean and standard deviation of the normal distribution from which $\gamma_{i,1}$'s ($\gamma_{i,2}$'s) are drawn from.

	Loadings on $IndusSize_t$ ($\gamma_{i,1}$)	Loadings on $FundSize_{i,t}$ ($\gamma_{i,2}$)
μ_γ	-0.05	-0.003
σ_γ	0.80	0.003

We can also allow residual correlation in fund returns in our simulations. We experiment with two correlation schemes. The first scheme assumes a common contemporaneous pairwise correlation (i.e., ρ) for the cross-section of return residuals. We set ρ at 0.2, which we think is a reasonable upper bound on the average pairwise correlation based on the evidence in Barras et al. (2010) and Harvey and Liu (2018). The other scheme, which is more realistic, calibrates a structural model to match key features of the cross-sectional distribution of pairwise residual correlations. We follow the parameterization of the structural model in Harvey and Liu (2018) to realistically capture residual correlation.

Table B.2 reports the results of our simulation study. For the loadings on the common variable $IndusSize$, both our model labeled AP (alpha predictor) and the equation-by-equation OLS perform reasonably well. For example, when residuals are uncorrelated (i.e., $\rho = 0$), our model estimate of the population mean of the loadings on $IndusSize$ has a bias of 0.5% (=0.026/5.0) relative to the magnitude of the true value and the equation-by-equation OLS has a bias of -0.8% (=-0.042/5.0). On the other hand, for the standard deviation of the population of the loadings, our model has a bias of -0.7% (=-0.524/80) relative to the magnitude of the true value and the equation-by-equation OLS has a bias of 1.9% (=1.518/80). Hence, while our model performs better than the equation-by-equation OLS under all specifications,

both models seem to perform well. This is not surprising since, unlike $FundSize$, innovations in $IndusSize$ are not strongly correlated with fund returns ($IndusSize$ is a common variable) so the Stambaugh bias does not affect the estimation under either our model or the equation-by-equation OLS. In addition, there is a large amount of cross-sectional variation in the loadings on $IndusSize$ as $\sigma_{\gamma,1}$ is large relative to the mean loading (i.e., $\mu_{\gamma,1}$). As a result, by (8) and (9), cross-sectional information plays a limited role in helping refine the equation-by-equation OLS estimates. This explains why our model performs similarly to the equation-by-equation OLS.

Turning to the loadings on $FundSize$, the story is very different. Our model generates largely unbiased estimate for the mean of the loadings population while the equation-by-equation OLS is severely biased. For instance, when the simulated residual correlations mimic the dependence structure among the residuals for the actual data, our model estimate of the mean of the loadings on $FundSize$ has a bias of -0.3% ($=-0.008/3.0$) relative to the magnitude of the true value. In contrast, the equation-by-equation OLS has a bias of -14% ($=-0.412/3.0$). As such, it would be a mistake to use the equation-by-equation OLS to make inference on the impact of $FundSize$.

Table B.2: **A Simulation Study: Parameter Estimates for the Loadings Population**

Model estimates in a simulation study. We fix the model parameters at Λ^* (Table B.1) and generate D sets of data sample. For each set of data sample, we estimate our model using our *alpha predictor* (AP) model and the usual equation-by-equation OLS. ρ is the assumed level of pairwise correlation for the correlation model that assumes an equal correlation for each pair of residual series. For a given parameter γ, let γ_d be the model estimate based on the d-th simulation run, $d = 1, 2, \ldots, D$. "True" reports the assumed true parameter value given in Λ^*. "Bias" reports the difference between the average of the simulated parameter estimates and the true value, that is, $(\sum_{d=1}^{D} \gamma_d)/D - \gamma$. "RMSE" reports the square root of the mean squared estimation error. "*p(10)*" reports the 10th percentile of the parameter estimates and "*p(90)*" reports the 90th percentile of the parameter estimates. $(\mu_{\gamma,1}, \sigma_{\gamma,1})$ and $(\mu_{\gamma,2}, \sigma_{\gamma,2})$ are the mean and standard deviation of the population of $\gamma_{i,1}$'s and $\gamma_{i,2}$'s, respectively. ρ is the assumed pairwise correlation for the cross-section of return residuals. "Empirical ρ" corresponds to the parameterization in Harvey and Liu (2018) that models the cross-sectional distribution of return residuals.

		$\rho = 0$		$\rho = 0.2$		Empirical ρ	
		AP	OLS	AP	OLS	AP	OLS
$IndusSize_t$							
$\mu_{\gamma,1}$ ($\times 10^2$)	Bias	0.026	−0.042	−0.128	−0.231	−0.183	−0.224
(True = −5.0)	RMSE	1.737	1.738	3.068	3.168	2.081	2.074
	p(10)	−6.714	−6.850	−8.825	−9.151	−7.990	−7.966
	p(90)	−2.602	−2.691	−2.038	−1.620	−2.503	−2.734
$\sigma_{\gamma,1}$ ($\times 10^2$)	Bias	−0.524	1.518	−0.532	1.711	−0.234	1.921
(True = 80)	RMSE	1.257	2.042	1.394	2.306	1.238	2.444
	p(10)	77.989	79.877	77.807	80.256	78.365	80.244
	p(90)	81.118	83.231	81.153	83.501	81.236	83.801
$FundSize_{i,t}$							
$\mu_{\gamma,2}$ ($\times 10^3$)	Bias	−0.017	−0.420	−0.022	−0.393	−0.008	−0.412
(True = −3.0)	RMSE	0.108	0.527	0.173	0.863	0.124	0.725
	p(10)	−3.135	−3.841	−3.283	−4.232	−3.175	−4.191
	p(90)	−2.888	−3.021	−2.784	−2.347	−2.843	−2.697
$\sigma_{\gamma,2}$ ($\times 10^3$)	Bias	−0.192	11.628	−0.177	11.536	−0.182	11.402
(True = 3.0)	RMSE	0.218	11.667	0.220	11.580	0.246	11.456
	p(10)	2.683	13.436	2.652	13.225	2.616	13.016
	p(90)	3.106	15.914	3.018	15.957	3.062	15.903

Why is our model unbiased while the equation-by-equation OLS is biased? To provide insight, we compare model performance by looking at the fund specific loadings on $IndusSize$ and $FundSize$.

Table B.3 reports the results. Focusing on Panel A, not surprisingly, our model provides more accurate estimates for the loadings on $IndusSize$ and $FundSize$. The improvement of our model over the equation-by-equation OLS is substantial for the loadings on $FundSize$. For instance, while our model implies a mean absolute de-

viation that is 60% (=1.814/3.0) relative to the mean loading on $FundSize$ (i.e., $\mu_{\gamma,2}$), the equation-by-equation OLS generates a mean absolute deviation of 241% (=7.236/3).

We next divide the cross-section of funds into groups based on the number of monthly time periods that are available for each fund, as shown in Panel B (each group consists of 20% of funds in our sample). Focusing on the loadings on $FundSize$ corresponding to the equation-by-equation OLS, we see that there is a very large negative bias for funds that exist for a short time period (e.g., $T < 36$). The bias is much smaller (in magnitude) for funds that exist for a longer time period (e.g., $T > 126$). This is related to the Stambaugh bias. In PST, assuming a balanced panel, they show that the percentage bias for their OLS estimator with fixed effects ranges from -73% to -8%, depending on the magnitude of the true parameter value. In our framework, by assuming an unbalanced panel that is consistent with the actual data, all the funds in our sample have a smaller sample size than what PST assume. As a result, the Stambaugh bias, which arises in small samples, should be even more pronounced in our framework. However, this is not what we see from Table B.3. In particular, for the 40% of funds in our sample that have a sample size no less than 81 months, the percentage bias is lower than -1% (=$-0.036/3.0$), much smaller in magnitude than what PST show in their simulation study. This is because we define $FundSize$ as the log of the industry adjusted fund size, not the dollar size of each fund (adjusted by the size of the equity market). By doing this, our definition helps kill the mechanical contemporaneous correlation between a fund's return and the growth rate in $FundSize$, thereby alleviating the Stambaugh bias.[45] In fact, for funds that have more than 81 monthly observations (which account for 40% of our sample), the equation-by-equation OLS has a smaller percentage bias compared to the proposed IV method in PST, which has a percentage bias in the range of -2% and -3%.

For funds that exist for a short period time, the variation in industry size is small and likely swamped by the variation in individual fund size, leading to a strong Stambaugh bias for the equation-by-equation OLS. However, funds that have a shorter sample also have a larger standard error for the parameter estimates. For instance, as shown in Panel B of Table B.3, the averaged OLS standard error for the loading on $FundSize$ is 18.9×10^{-3} for funds that satisfy $T < 36$ and is 2.4×10^{-3} for funds that satisfy $T \geqslant 171$. In our framework, cross-sectional learning down weights the importance of funds with more noisy OLS parameter estimates, alleviating the bias in the OLS estimates for these funds. Following the previous example, based on (8) and (9), the weight we assign to funds that have $T \geqslant 171$ is 62 (=$(18.9/2.4)^2$) times the weight we assign to funds that have $T < 36$. Hence, our framework substantially (and optimally) over weights funds with a long sample, which, as we discussed before, are less affected by the Stambaugh bias. Panel B of Table B.3 shows that our model

[45]For example, in our data for funds with no greater than 36 monthly observations (for which the Stambaugh bias should be the most severe), the median contemporaneous correlation between a fund's return and the growth rate of $FundSize$ is 3.5% for our definition of $FundSize$, and is 17.9% under PST's definition.

implies roughly unbiased estimate for the loading on $FundSize$ across all groups of funds categorized by sample size. In fact, the percentage bias for the group of funds with $T < 36$, the worst case scenario in our framework, is -1% ($=-0.028/3.0$), still lower (in magnitude) than the bias for the IV approach in PST.

Table B.3: **A Simulation Study: Parameter Estimates for Individual Fund Loadings**

Model estimates in a simulation study. We fix the model parameters at Λ^* (Table B.1) and generate D sets of data sample. For each set of data sample, we estimate our model using our *alpha predictor* (AP) model and the usual equation-by-equation OLS. Panel A reports results on estimation accuracy for the loadings for individual funds. "Mean absolute deviation" is the averaged (across simulations) mean absolute distance between the estimated loading and the true loading for the cross-section of funds. $p10$, $p50$, and $p90$ report the averaged (across simulations) 10-th, 50-th, and 90-th percentile of the mean absolute distance between the estimated loading and the true loading for the cross-section of funds. Panel B looks at estimation bias and standard errors for loadings estimates for individual funds. T denotes the number of monthly observations that is available for a fund. For funds that have a T that falls within a certain range, "Bias" reports the averaged (across simulations) mean difference (across funds) between the estimated loading and the true loading; "Avg. Std. Err." reports the averaged (across simulations) mean (across funds) standard error for the estimated loading. All simulations are run under the "Data Depen." specification in Table B.2.

Panel A: Estimation Accuracy					
		$IndusSize_t$ $\gamma_{i,1}$ ($\times 10^2$)		$FundSize_{i,t}$ $\gamma_{i,2}$ ($\times 10^3$)	
		AP	OLS	AP	OLS
Mean absolute deviation		7.659	9.317	1.814	7.236
$p10$		0.741	0.798	0.244	0.377
$p50$		4.574	5.078	1.407	2.787
$p90$		18.129	21.675	3.965	19.370
Panel B: Bias and Standard Errors for Number of Observations					
Bias	$18 \leqslant T < 36$	0.173	−0.222	−0.028	−1.385
	$36 \leqslant T < 81$	0.205	0.224	0.009	−0.258
	$81 \leqslant T < 126$	−0.016	−0.098	−0.022	−0.116
	$126 \leqslant T < 171$	0.059	0.050	−0.012	−0.036
	$171 \leqslant T$	−0.018	−0.012	−0.011	−0.023
Avg. Std. Err.	$18 \leqslant T < 36$	2.247	18.881	2.641	18.922
	$36 \leqslant T < 81$	3.029	14.523	2.222	8.661
	$81 \leqslant T < 126$	2.412	8.127	1.867	4.278
	$126 \leqslant T < 171$	2.084	6.032	1.605	2.782
	$171 \leqslant T$	1.996	4.053	1.489	2.364

D Additional Results

D.1 Portfolio Sorts

Table C.1: **Portfolio Sorts Based on Loadings on $IndusSize$, Fund Sizes and Past Performances**

Annualized alphas for strategies that sort the cross-section of funds based on loadings on $IndusSize$, fund sizes, and past performances. Using a rolling five-year window, we estimate our model to obtain the cross-section of loadings on $IndusSize$. We then sort funds into different groups based on the loadings, fund sizes, and past performances. We use conditional sorts to first sort funds into size terciles. Within each size tercile, we sort funds into three groups based on past performances. Finally, within each group for past performance, we sort funds into three groups based on the loadings. A low value of loading means the fund is sensitive to decreasing returns to scale. Past performance is measured as the average fund excess return in the past two years. Our sample is from 1991 to 2011. We start sorting in 1996 to have the initial five-year window to estimate our model. Panel A subtracts the market excess return from the fund excess return. Panel B reports fund alphas based on the Carhart (1997) four-factor model.

Panel A: Fund Returns Adjusted by Market Return

	Fund's TNA									
	Small			*Median*			*Large*			
Loadings	α_{low}	α_{med}	α_{high}	α_{low}	α_{med}	α_{high}	α_{low}	α_{med}	α_{high}	All
Low	−3.895	0.091	1.940	−1.523	−0.053	0.624	−1.440	−1.054	0.719	−0.510
2	−2.340	−0.424	2.303	−2.668	−1.240	1.014	−3.208	−1.424	0.192	−0.866
High	−1.012	1.042	4.558	−1.265	0.047	3.487	−2.475	−0.563	2.189	0.668
High - Low	2.883	0.951	2.619	0.258	0.099	2.863	−1.036	0.491	1.469	1.178
(t-stat)	(1.89)	(0.91)	(2.17)	(0.26)	(0.12)	(2.77)	(-1.34)	(0.58)	(1.36)	(1.93)

Panel B: Fund Returns Adjusted by 4-Factor Model

	Fund's TNA									
	Small			*Median*			*Large*			
Loadings	α_{low}	α_{med}	α_{high}	α_{low}	α_{med}	α_{high}	α_{low}	α_{med}	α_{high}	All
Low	−4.277	−0.238	0.673	−1.708	−0.763	−1.075	−1.338	−1.656	−0.728	−1.234
2	−2.784	−0.599	0.707	−2.983	−1.396	−0.534	−2.981	−1.348	−1.228	−1.461
High	−1.522	0.545	3.065	−1.597	−0.388	1.867	−2.130	−0.281	0.836	0.044
High - Low	2.755	0.782	2.392	0.111	0.375	2.942	−0.792	1.376	1.564	1.278
(t-stat)	(1.77)	(0.75)	(2.13)	(0.12)	(0.44)	(2.87)	(-1.02)	(2.00)	(1.44)	(2.03)

Table C.2: **Portfolio Sorts Based on Loadings on $FundSize$, Fund Sizes and Past Performances**

Annualized alphas for strategies that sort the cross-section of funds based on loadings on $FundSize$, fund sizes, and past performances. Using a rolling five-year window, we estimate our model to obtain the cross-section of loadings on $FundSize$. We then sort funds into different groups based on the loadings, fund sizes, and past performances. We use conditional sorts to first sort funds into size terciles. Within each size tercile, we sort funds into three groups based on past performances. Finally, within each group for past performance, we sort funds into three groups based on the loadings. A low value of loading means the fund is sensitive to decreasing returns to scale. Past performance is measured as the average fund excess return in the past two years. Our sample is from 1991 to 2011. We start sorting in 1996 to have the initial five-year window to estimate our model. Panel A subtracts the market excess return from the fund excess return. Panel B reports fund alphas based on the Carhart (1997) four-factor model.

Panel A: Fund Returns Adjusted by Market Return

	Fund's TNA									
	Small			*Median*			*Large*			
Loadings	α_{low}	α_{med}	α_{high}	α_{low}	α_{med}	α_{high}	α_{low}	α_{med}	α_{high}	All
Low	−2.971	0.467	4.303	−0.734	0.534	2.783	−1.728	0.034	1.789	0.497
2	−2.251	−0.118	2.528	−2.170	−0.772	1.043	−2.907	−1.112	1.045	−0.524
High	−1.874	0.454	1.912	−2.597	−1.039	1.307	−2.488	−1.969	0.279	−0.668
Low - High	−1.097	0.013	2.390	1.863	1.573	1.476	0.760	2.003	1.511	1.166
(t-stat)	(-0.93)	(0.01)	(1.81)	(2.17)	(1.94)	(1.63)	(1.21)	(3.05)	(1.92)	(2.42)

Panel B: Fund Returns Adjusted by 4-Factor Model

	Fund's TNA									
	Small			*Median*			*Large*			
Loadings	α_{low}	α_{med}	α_{high}	α_{low}	α_{med}	α_{high}	α_{low}	α_{med}	α_{high}	All
Low	−3.597	0.132	2.695	−0.957	−0.217	0.829	−1.316	−0.310	0.225	−0.280
2	−2.496	−0.357	0.876	−2.558	−1.266	−0.530	−2.973	−1.254	−0.558	−1.235
High	−2.363	0.016	0.807	−2.823	−1.097	−0.035	−2.156	−1.728	−0.776	−1.128
Low - High	−1.234	0.117	1.888	1.866	0.880	0.863	0.840	1.418	1.001	0.849
(t-stat)	(-1.02)	(0.12)	(1.40)	(2.17)	(1.22)	(1.06)	(1.36)	(2.55)	(1.30)	(1.84)