Detecting Repeatable Performance

Campbell R. Harvey

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

Yan Liu*

Texas A&M University, College Station, TX 77843 USA

Current version: January 21, 2018

Abstract

Past fund performance does a poor job of predicting future outcomes. The reason is noise. Using a random effects framework, we reduce the noise by pooling information from the cross-sectional alpha distribution to make density forecasts for each individual fund's alpha. In simulations, we show that our method generates parameter estimates that outperform alternative methods, both at the population and at the individual fund level. An out-of-sample forecasting exercise also shows that our method generates improved alpha forecasts.

Keywords: Performance evaluation, Mutual funds, Hedge funds, EM algorithm, Fixed effects, Random effects, Regularization, Multiple testing, Bayesian

* Current Version: January 21, 2018. First posted on SSRN: November 19, 2015. Send correspondence to: Campbell R. Harvey, Fuqua School of Business, Duke University, Durham, NC 27708. Phone: +1 919.660.7768, E-mail: cam.harvey@duke.edu , and Yan Liu, Mays Business School, Texas A&M University, College Station, TX 77843. Phone: +1 919.428.1118. A discussion with Neil Shephard provided the genesis for this paper — we are grateful. We appreciate the comments of the Editor, Andrew Karolyi, as well as two anonymous referees. We have also benefited from the comments of Andrew Ang, Laurent Barras, Jonathan Berk, Svetlana Bryzgalova, Yong Chen, Ian Dew-Becker, Wayne Ferson, Christopher Jones, Juhani Linnainmaa, Stefan Nagel, David Ng, Ľuboš Pástor, Andrew Patton, Olivier Scaillet, Robert Stambaugh, Luke Taylor and Russ Wermers as well as seminar participants at the 2016 SFS Finance Cavalcade Conference at the University of Toronto, the 2016 WFA meeting in Park City, the 2017 AFA meetings in Chicago, Hong Kong Polytechnic University, University of Ottawa, Man Quant Conference, Research Affiliates, Norges Bank, and APG. All errors are our own.

1 Introduction

In a method reaching back to Jensen (1969), most applications of performance evaluation run separate regressions to obtain alpha estimates and standard errors. Following this fund-by-fund approach, it is very difficult to detect repeatable performance. That is, there is little or no information that there is a positive relation between past performance and future performance. The reason is simple: the individual fund alpha is plagued by noise.

The traditional fund-by-fund Jensen's alpha is analogous to the fixed effects model in panel regressions where a non-random intercept is assumed for each fund. We advocate a "random effects" counterpart, which we term the *noise reduced alpha* (NRA) model. In particular, we assume that fund i's alpha, α_i, is drawn independently from a common cross-sectional distribution.

Why is our approach relevant for performance evaluation? First, our NRA model provides a structural approach to study the distribution of fund alphas. It allows us to learn from the entire cross-section of funds when making inference on a particular fund's alpha. It not only provides estimates for quantities that are economically important (e.g., the 5th percentile of alphas, the fraction of positive alphas), but also provides standard errors for these estimates by taking into account various sources of parameter uncertainty, in particular the uncertainty in the estimation of alphas. Second, the fund data that researchers use (particularly, hedge fund data) are likely to cover only a fraction of the entire population of funds. Therefore, with the usual caveats about sample selection in mind, it makes sense to make inference on this underlying population rather than just focusing on the available fund data. This is one of the situations where a random effects setup is preferred over a fixed effects procedure in panel regression models.[1]

Traditionally, performance evaluation involves fund-by-fund regressions in the first stage and hypothesis tests are performed in the second stage. There are several problems with this approach when it comes to making inference on the cross-sectional distribution of fund alphas. First, while fund-specific hypothesis testing may be useful to search for outperformers and underperformers, its use is limited when it comes to

[1] See, for example, Greene (2003) and Maddala (2001). Searle, Casella, and McCulloch (1992) explore the distinction between a fixed effects model and a random effects model in more details.

making precise statements about the properties of the population of alphas. Consider one obvious economically important question: what fraction of mutual funds or hedge funds generate a positive alpha? Under the usual fund-by-fund testing framework, one candidate answer is the fraction of funds that generate a significant and positive alpha. However, this answer is likely biased given the limited power of the test of an individual fund's alpha. In essence, the fund-by-fund approach ignores valuable information in the cross-section that can potentially improve the inference on the cross-sectional distribution of fund alphas.

Second, recent papers have adjusted fund-by-fund hypothesis tests for test multiplicity (Barras, Scaillet, and Wermers 2010; Fama and French 2010; Ferson and Chen 2015; and Harvey and Liu 2017a). The idea of correcting for multiple hypothesis tests involves choosing a test statistic based on the cross-section of alphas (or the t-statistics of alphas) and balancing the trade-off between Type I and Type II error. For instance, while Fama and French (2010) focus on extreme t-statistic percentiles, Barras, Scaillet, and Wermers (2010) and Ferson and Chen (2015) focus on the false discovery rate. In contrast, our framework relies on the likelihood function of the panel of fund returns, which allows an efficient weighting of cross-sectional and time-series information.

Recent research focuses on the cross-sectional distribution of the alphas. Barras, Scaillet, and Wermers (2010) and Ferson and Chen (2015), who model the alphas as drawn from a discrete distribution, can be thought of as simplified versions of methods that model the cross-sectional distribution of alphas. But several other papers—including ours—have taken a structural approach to provide a more general and detailed description of the alpha population. This new initiative allows us to address the standard approach's unanswered questions. For example, by modeling the cross-sectional distribution of alphas, it is possible to answer the question of how many managers outperform. In addition, inference on the performance of any individual manager is enhanced by taking cross-sectional information into account. Furthermore, various sources of uncertainty are directly incorporated into the inference. In order to understand our contribution, consider the three paths that this recent research has taken.

The first path involves initially running fund-level ordinary least squares (OLS) and then trying to estimate the distribution of the fitted alphas. By doing this, it is possible to make inference on the alpha population. Chen, Cliff, and Zhao (CCZ;

2015) provide a variant of this approach by proposing a two-stage estimation procedure that takes fund-level alpha uncertainty into account. However, this approach uses only fund-specific information to estimate regression parameters that govern return dynamics (i.e., factor loadings and residual standard deviations) and ignores the information in the alpha population. Such information is important given the high level of estimation uncertainty for individual funds since many have limited histories. In addition, their two-stage estimation procedure is inconsistent from the perspective of decomposing fund returns into luck and skill, as what is identified as skill in the first-stage estimation may be attributable to luck in the second-stage estimation, for which cross-sectional information is taken into account. Our approach simultaneously estimates parameters that govern the alpha population and parameters that govern individual fund return dynamics, providing a unified framework to incorporate individual fund time-series uncertainty and cross-sectional alpha uncertainty.

The second path applies Bayesian methods to learn about the alpha population. For example, Jones and Shanken (JS; 2005) impose a prior on the mean and standard deviation of the normal distribution from which the cross-section of fund alphas are drawn from.[2] Conceptually, their approach is closely related to ours in that we also try to make inference on the alpha population. However, there are important differences. We build on the frequentist approach and do not need to impose a prior on the parameters that govern the alpha population. We also allow fund alphas to be drawn from several subpopulations, which enriches the structure of the alpha population.[3] Later, we provide a detailed discussion of Bayesian methods and contrast them with our approach.

The third path incorporates information other than return performance. By using portfolio holdings data, Cohen, Coval, and Pastor (2005) infer a manager's skill from the skill of managers that have similar portfolio holdings. Intuitively, if two managers have similar time series of holdings, their alpha estimates should be close to each other. Cohen, Coval, and Pastor (200 5) weight the cross-section of historical alpha estimates by the current portfolio holdings to refine the alpha estimate of a particular fund. Their idea of learning from the cross-section of managers is similar to ours. However, while their method learns through current portfolio holdings, we learn about skill

[2]Other papers that apply Bayesian methods to study fund performance include Avramov and Wermers (2005), Baks, Metrick, and Wachter (2001), Busse and Irvine (2006), Kosowski, Naik, and Teo (2007), Pástor and Stambaugh (2002a,b), and Stambaugh (2003).

[3]See Barras, Scaillet, and Wermers (2010), Ferson and Chen (2015), and Chen, Cliff, and Zhao (2015).

from the cross-section of alpha estimates (adjusted for the uncertainty in the alpha estimation), which reflect and summarize the entire history of holdings. Our method thus relies on the return data alone and is applicable to hedge fund performance evaluation, for which we do not have holdings data for most funds.

The common element among these three paths is the use of cross-sectional information to refine funds' alpha estimates. All three methods imply a certain degree of shrinkage of the OLS alpha estimates.[4] Our approach also implies shrinkage but differs from existing methods in that we present a new way to model the shrinkage target (i.e., the alpha population) as well as the optimal degree of shrinkage. We achieve this by trying to answer three fundamental questions. First, what is the best way to estimate the shrinkage target? We adopt a frequentist approach to estimate the underlying alpha population, and this distinguishes our method from the Bayesian approach. Second, how can we use the shrinkage target to improve the inference on individual funds, both for their alpha estimates and for other OLS parameters that govern return dynamics? We derive the optimal inference on individual funds, conditional on a given shrinkage target. Our solution features the revision of all OLS parameters according to the shrinkage target, and this distinguishes our method from CCZ, who only update the alpha estimates. Third, how do the previous two questions interact with each other in that refined inference on individual funds (the second question) can also improve our inference on the underlying alpha population (the first question)? We use an algorithm that allows us to solve the maximum-likelihood estimate iteratively. Our solution features the sequential updates of OLS parameters and parameters that govern the underlying alpha population, capturing the interaction between the two sets of parameters.

Our empirical work begins with a simulation study that takes many realistic features of the mutual fund data into account. We show that our method generates parameter estimates that achieve both a low finite-sample bias and standard error, dominating those that are generated under OLS and CCZ. The superior performance of our model applies to the alpha population as well as the individual funds.

While our research contribution is mainly methodological, we offer an application to a sample of mutual fund returns. We perform an out-of-sample exercise by estimat-

[4] "Shrinkage" in our context refers to the phenomenon that a standard estimator (e.g., the fund specific OLS alpha estimate) is shrank or pulled toward a certain value that is deemed useful in improving the performance (e.g., in terms of mean squared estimation error) of the estimator. See Cosemans et al. (2015), Karolyi (1993), and Vasicek (1973) for finance applications of shrinkage estimators.

ing our model in-sample and forecasting the alphas of individual funds out-of-sample. We show that our method provides a substantial improvement over the CCZ method and a marginal improvement over the JS method with respect to forecasting accuracy.

We are also able to determine the proportion of funds that outperform. While previous research suggests only 0-1% outperform, our results suggest 10%. The very low proportion found in previous research is due to the high level of estimation uncertainty associated with a fund-by-fund analysis. Our framework provides a more powerful procedure to identify funds with small positive alphas by directly modeling the underlying alpha population.

We also propose a new procedure to efficiently estimate our structural model. It extends the standard expectation-maximization (EM) algorithm, which sequentially learns about fund alphas (which are treated as missing observations) and estimate model parameters.[5] Our method allows us to capture the heterogeneity in fund characteristics in the cross-section. While we focus on performance evaluation, the procedure has a number of immediate extensions. For example, fund attributes can be incorporated to sharpen inference, and macroeconomic data may also be useful in characterizing how the cross-sectional distribution evolves through time (see Harvey and Liu 2017c). It is also possible to use the technique in other applications, such as choosing the set of factors with significant risk premia (see Harvey and Liu 2017b).

2 Model

2.1 The Likelihood Function

For ease of exposition, suppose we have a $T \times N$ balanced panel of fund returns, with T denoting the number of monthly periods and N denoting the number of funds in the cross-section. Importantly, balanced data is not required in our framework. As we shall see later, both our model and its estimation can be easily adjusted for unbalanced panel data.

[5] See Dempster, Laird, and Rubin (1977).

Suppose we are evaluating fund returns against a set of K benchmark factors. Fund excess returns are modeled as

$$r_{i,t} = \alpha_i + \sum_{j=1}^{K} \beta_{ij} f_{j,t} + \varepsilon_{i,t}, \quad i = 1, \ldots, N;\ j = 1, \ldots, K;\ t = 1, \ldots, T, \tag{1}$$

where $r_{i,t}$ is the excess return (i.e., actual return minus the one-month Treasury bill rate) for fund i in period t, α_i is the alpha, β_{ij} is fund i's risk loading on the j-th factor $f_{j,t}$, and $\varepsilon_{i,t}$ is the residual.

To simplify the exposition, let us introduce some notation. Let $R_i = [r_{i,1}, r_{i,2}, \ldots, r_{i,T}]'$ be the excess return time series for fund i. The panel of excess returns can be expressed as $\mathcal{R} = [R_1, R_2, \ldots, R_N]'$. Let $\beta_i = [\beta_{i,1}, \beta_{i,2}, \ldots, \beta_{i,K}]'$ be the risk loadings for fund i. We collect the cross-section of risk loadings into the vector $\mathcal{B} = [\beta_1', \beta_2', \ldots, \beta_N']'$. Similarly, we collect the cross-section of alphas into the vector $\mathcal{A} = [\alpha_1, \alpha_2, \ldots, \alpha_N]'$. Let the residual covariance matrix be $\mathbf{\Sigma}$. Finally, let θ be the parameter vector that describes the population distribution of the elements in $\mathcal{A}$.

Under the model assumptions, the likelihood function of the model is

$$\begin{aligned} f(\mathcal{R}|\theta, \mathcal{B}, \mathbf{\Sigma}) &= \int f(\mathcal{R}, \mathcal{A}|\theta, \mathcal{B}, \mathbf{\Sigma}) d\mathcal{A} &(2)\\ &= \int f(\mathcal{R}|\mathcal{A}, \mathcal{B}, \mathbf{\Sigma}) f(\mathcal{A}|\theta) d\mathcal{A}, &(3) \end{aligned}$$

where $f(\mathcal{R}, \mathcal{A}|\theta, \mathcal{B}, \mathbf{\Sigma})$ is the complete data likelihood function (that is, the joint likelihood of both returns $\mathcal{R}$ and alphas $\mathcal{A}$), $f(\mathcal{R}|\mathcal{A}, \mathcal{B}, \mathbf{\Sigma})$ is the conditional likelihood of returns given the cross-section of alphas and model parameters, and $f(\mathcal{A}|\theta)$ is the conditional density of the cross-section of alphas given the parameters that govern the alpha distribution.[6]

Notice that by assumption the likelihood function of the model does not depend on the cross-section of alphas (i.e., $\mathcal{A}$). This is because, in our approach, $\mathcal{A}$ is treated as missing data and needs to be integrated out of the complete likelihood function $f(\mathcal{R}, \mathcal{A}|\theta, \mathcal{B}, \mathbf{\Sigma})$. However, once we obtain the estimates of the model parameters, the conditional distribution of $\mathcal{A}$ can be obtained through Bayes' law:

$$f(\mathcal{A}|\mathcal{R}, \hat{\theta}, \hat{\mathcal{B}}, \hat{\mathbf{\Sigma}}) \propto f(\mathcal{R}|\mathcal{A}, \hat{\mathcal{B}}, \hat{\mathbf{\Sigma}}) f(\mathcal{A}|\hat{\theta}). \tag{4}$$

[6]While we focus on alphas in our paper, we later discuss extensions of our model that can be applied to alternative performance metrics such as information ratios.

This enables to us to evaluate the performance of each individual fund. Our approach to making inference on individual funds is distinctively different from current frequentist methods. Existing approaches, as mentioned previously, draw their inference based on either the time-series likelihood (i.e., $f(\mathcal{R}|\mathcal{A},\mathcal{B},\mathbf{\Sigma})$) as in Barras, Scaillet, and Wermers (2010), Fama and French (2010), and Ferson and Chen (2015), or the cross-sectional likelihood (i.e., $f(\mathcal{A}|\theta)$) as in Chen, Cliff, and Zhao (2015). Our method, as shown in Equation (4), combines information from both types of likelihoods, leading to a more informative inference.

Assuming that the residuals (i.e., $\varepsilon_{i,t}$'s) are independent both across funds and across time, the likelihood function can be written as:

$$f(\mathcal{R}|\theta,\mathcal{B},\mathbf{\Sigma}) = \int \prod_{i=1}^{N} f(R_i|\alpha_i,\beta_i,\sigma_i) f(\alpha_i|\theta) d\mathcal{A}, \tag{5}$$

$$= \prod_{i=1}^{N} \int f(R_i|\alpha_i,\beta_i,\sigma_i) f(\alpha_i|\theta) d\alpha_i, \tag{6}$$

where we have simplified the residual covariance matrix into a diagonal matrix given cross-sectional independence—that is, $\mathbf{\Sigma} = [\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2]'$, where σ_i is the residual standard deviation for the i-th fund. Our goal is to find the maximum-likelihood estimate of θ, which is the focus of the paper, along with other auxiliary parameters (i.e., $\mathcal{B}$ and $\mathbf{\Sigma}$) that govern the return dynamics of each individual fund. To obtain an explicit expression for the likelihood function, we assume that the residuals are normally distributed.

Residual independence is not a key assumption for our model. When there is residual dependency, the model will be misspecified. The likelihood function becomes a quasi-likelihood function. Our quasi-maximum likelihood estimate (QMLE) still makes sense, as the parameters governing the dependency structure are treated as auxiliary parameters with respect to the goal of our analysis. Despite the model misspecification, in theory, the QMLE is still consistent in that it gives asymptotically unbiased estimates. It will be less efficient compared with the maximum likelihood estimate (MLE) of a correctly specified model. In our simulation study, we consider residual dependency and quantify the loss in efficiency.

2.2 The Specification of the Alpha Distribution

The density of the alphas, Ψ, needs to be flexible enough to capture the true underlying distribution of alpha. For instance, two groups of fund managers could exist, one group consisting of skilled managers, and the other consisting of unskilled managers. Rather than two groups, we might choose five categories similar to the Morningstar system. As such, the density of Ψ should be able to display a multimodal pattern.

On the other hand, having a flexible distribution does not mean that the distribution should be complicated. In fact, the very principle of regularization in statistics is to use parsimonious models to avoid overfitting.[7] Hence, without sacrificing too much flexibility, we use a distribution that is simple and interpretable: Gaussian mixture distribution (GMD), a weighted sum of Gaussian distributions. A one-component GMD is just a standard Gaussian distribution. The two-component GMD is a mixture of two Gaussian distributions and allows for considerable heterogeneity but is characterized by only five parameters: two means, two variances, and a mixing parameter.[8] In our context, the model has a simple interpretation in which one component of the GMD represents unskilled managers and the other component skilled managers.

To achieve model identification, we assume the mean of the low type distribution is less than the high type. In Appendix B, we discuss the generalization of the two-component GMD to a multicomponent GMD, as well as the identifiability and interpretability of a general GMD.

The recent literature on investment fund performance evaluation attempts to group funds into different categories. For example, Barras, Scaillet, and Wermers (2010) and Ferson and Chen (2015) assume that funds are drawn from a few subpopulations, with "good" and "bad" managers coming from distinct subpopulations. Our parameterization of Ψ also bears this simple interpretation of a multipopulation structure for the alpha distribution. However, different from the approach of Barras, Scaillet, and Wermers (2010) and Ferson and Chen (2015), our structural estimation approach allows us to take various sources of estimation risk into account when we classify funds into distinct performance groups. Our empirical results show that our approach can lead to important differences in the classification outcome.

[7] See, for example, Bickel and Li (2006), Fan and Lv (2010), and Vidaurre, Bielza, and Larrañaga (2013).

[8] For applications of the Gaussian mixture distribution in finance, see Bekaert and Harvey (1995) and Gray (1996).

2.3 Model Discussion

The traditional OLS fund-by-fund hypothesis testing framework poses a number of challenges with respect to making inference on the population of fund alphas. While hypothesis testing may be useful when we want to test the significance of a single fund, we need to make adjustment for test multiplicity when the same test is performed on many funds.[9] This framework is less useful when we try to make inference on the entire alpha population. By testing against a common null hypothesis (e.g., alpha equals zero), it essentially treats fund alphas as dichotomous variables, while, more realistically, they should be continuous. Our model assumes that the true alpha is a continuous variable and provides density estimates that can be used to evaluate each individual fund as well as the cross-sectional alpha population.

The traditional approach also places too much weight on the statistical significance of individual alphas and overlooks their economic significance from a population perspective. For example, suppose we have two funds that both have a t-statistic of 1.5. One has an alpha of 20% (per annum), and the other has an alpha of 2% (per annum). Should we treat them the same? We think not. The 20% alpha, albeit volatile, tells us more about the plausible realizations of alphas in the cross-section than the 2% alpha.[10] Following the standard OLS approach, we not only ignore the difference in magnitude between the two alphas, but we also classify both funds as zero-alpha funds, causing an unnecessary loss of information regarding the cross-sectional distribution of alphas.

We now discuss the details of our model.[11] To see how our method takes estimation uncertainty into account, we focus on the likelihood function in Equation (6) (that is, $\prod_{i=1}^{N} \int f(R_i|\alpha_i, \beta_i, \sigma_i) f(\alpha_i|\theta) d\alpha_i$). Suppose we already have an estimate of $\mathcal{B}$ and $\mathbf{\Sigma}$ (e.g., the OLS estimate) and seek to find the estimate for θ. Notice that $f(R_i|\alpha_i, \beta_i, \sigma_i)$, the likelihood function of the returns of fund i, can be viewed as a probability density on α_i. In particular, under normality of the residuals, we have

$$f(R_i|\alpha_i, \beta_i, \sigma_i) \equiv w(\alpha_i) \propto \exp\{-\frac{[\alpha_i - \frac{\sum_{t=1}^{T}(r_{it} - \beta_i' f_t)}{T}]^2}{2\sigma_i^2/T}\}, \tag{7}$$

[9] For recent papers on investment fund performance evaluation that emphasize multiple hypotheses testing, see Barras, Scaillet, and Wermers (2010), Fama and French (2010), and Ferson and Chen (2015).

[10] While some investment funds can use leverage to amplify gains and losses, they also face leverage constraints. Therefore, 20% tells us more about the tails of the alpha distribution than 2%.

[11] Our online Appendix IE provides two examples that highlight the intuition behind our approach.

where $f_t = [f_{1,t}, f_{2,t}, \ldots, f_{K,t}]'$ is the vector of factor returns at time t. Viewing in this way, $\int f(R_i|\alpha_i, \beta_i, \sigma_i) f(\alpha_i|\theta) d\alpha_i = \int w(\alpha_i) f(\alpha_i|\theta) d\alpha_i$ is a weighted average of $f(\alpha_i|\theta)$, with the weights (i.e., $w(\alpha_i)$) given in Equation (7).

When $\sigma_i/\sqrt{T}$ is small, that is, when there is little uncertainty in the estimation of α_i, $w(\alpha_i)$ has most of its mass on (roughly) the OLS estimate of alpha—that is, $\frac{\sum_{t=1}^{T}(r_{it}-\beta_i' f_t)}{T}$. In fact, when $\sigma_i \to 0, i = 1, \ldots, N$ and when $\mathcal{B}$ and $\mathbf{\Sigma}$ are set at their OLS estimates, the likelihood function in Equation (6) converges to $\prod_{i=1}^{N} f(\hat{\alpha}_i^{OLS}|\theta)$ — the likelihood function when the alphas are exactly set at their OLS estimates. Therefore, ignoring the time-series uncertainty in the estimation of the alphas, the likelihood function collapses to the likelihood function constructed under the traditional approach, that is, running equation-by-equation OLS first and then estimating the distribution for the fitted alphas. Simple as it is, this is what many do when trying to summarize fund performance in the cross-section. Our approach, by using a weighting function $w(\alpha_i)$ that depends on $\sigma_i/\sqrt{T}$, allows us to take the time-series uncertainty in the estimation of the alpha into account.

Moreover, the weighting function $w(\alpha_i)$ is fund specific—that is, $w(\alpha_i)$ depends on the particular level of estimation uncertainty for α_i (i.e., $\sigma_i/\sqrt{T}$). Therefore, the likelihood function in Equation (6) allows different weighting functions for different funds. This is important given the cross-sectional heterogeneity in estimation uncertainty across funds, in particular across investment styles.

Our approach offers more than just taking the estimation uncertainty for α_i (i.e., $\sigma_i/\sqrt{T}$) into account. As shall become clear later, our estimates of both α_i and σ_i^2 not only rely on fund i's time series, but also use information from the cross-sectional distribution of the alphas. Hence, in our framework, the OLS t-statistic is not an appropriate metric to summarize the significance of fund alphas. Both its numerator and denominator need to incorporate information from the alpha population. In contrast, the approach in Chen, Cliff, and Zhao (2015) relies on the OLS t-statistics to estimate the cross-sectional distribution of the alphas. As a result, they fail to use information in the alpha population to adjust OLS t-statistics, and their approach yields biased and inefficient estimates of the cross-sectional distribution of the alphas, as we will show in our simulation study.

On the other hand, our knowledge about the alpha population helps refine our estimates of the risk loadings and the residual variances. Suppose we already have an estimate of θ and seek to estimate $\mathcal{B}$ and $\mathbf{\Sigma}$. We again focus on the likelihood func-

tion $\int f(R_i|\alpha_i, \beta_i, \sigma_i) f(\alpha_i|\theta) d\alpha_i$, but instead view $f(\alpha_i|\theta)$ as the weighting function. $f(\alpha_i|\theta)$ tells us how likely it is to observe a certain α_i from a population perspective. If α_i is unlikely to occur over a certain range, the likelihood function will downweigh this range relative to other ranges over which the occurrence of alpha is more plausible. In the extreme case when we have perfect knowledge about the alpha of a certain fund (say, $\hat{\alpha}_i^0$), the likelihood function becomes $f(R_i|\hat{\alpha}_i^0, \beta_i, \sigma_i)$, essentially the likelihood function for a linear regression model when the intercept is fixed. In general, the MLE of β_i and σ_i will be different from their unconstrained OLS estimates, reflecting our knowledge about the alpha population. This is again different from the approach in Chen, Cliff, and Zhao (2015), in which risk loadings and residual variances are fixed at their OLS estimates.[12]

2.4 Context

Our paper is related to several strands of literature.

A recent paper by Chen, Cliff, and Zhao (CCZ; 2015) also implements the EM algorithm to extract the underlying alpha population. In particular, they employ a two-stage estimation procedure to first run equation-by-equation OLS to obtain the fitted alphas and standard errors, and then feed them into an EM framework to estimate the underlying alpha distribution. We show their method is problematic in several aspects. Instead of relying on fund-specific information to estimate the OLS parameters (i.e., factor loadings and residual standard deviations) as in CCZ, our structural approach uses information in the entire cross-section to estimate individual funds' OLS parameters. We show through simulations that our model represents a substantial improvement over CCZ in estimating both the structural parameters and important summary statistics of the alpha population.[13] We provide a detailed comparison of the two models in the next section.

[12] Much of the intuition of our model can also be understood from the perspective of empirical Bayes methods, as we discuss in greater detail in the next section.

[13] On a deeper level, CCZ's approach is inconsistent from the perspective of decomposing fund performance into luck and skill. In their first-stage regression, by running OLS, they are forcing the luck component for each fund (i.e., return residuals) to sum up to zero. However, in their second-stage regression, in which one updates individual funds alphas by drawing on information from the cross-section, the luck components no longer sum up to zero. This inconsistency does not exist in our framework: our model finds the MLE for all parameters simultaneously, making sure that structural parameters that govern the alpha population are compatible with fund-specific OLS parameters.

Bayesian methods have also been applied to performance evaluation.[14] Our approach relies on MLE and is therefore inherently a frequentist approach. By trying not to add to the long-standing debate between frequentist and Bayesian approaches, we provide some simulation results to highlight the superior performance of our method in comparison with Bayesian methods. Given the widespread use of frequentist methods in general, our framework at the very least offers a competing approach to Bayesian performance evaluation.[15]

Besides the full-blown Bayesian approach, another way to learn from the cross-section of funds is through multiple shrinkage.[16] In our context, multiple shrinkage amounts to a method that first partitions the cross-section of funds into meaningful groups (e.g., groups classified by certain fund characteristics) and then uses group information to refine the alpha estimate based on fund-specific information.[17] While this method is useful when there are fund attributes (e.g., fund size) that are believed to affect performance a priori, our framework is agnostic about instrumental variables that can potentially help predict fund returns, and we extract the shrinkage target — the underlying alpha population — given it is based only on returns data. Harvey and Liu (2017c) show how to extend our current framework to evaluate alpha predictors.[18]

3 Estimation

3.1 A New Expectation-Maximization Framework

A direct maximization of Equation (6) is difficult. The size of the parameter space is large and the likelihood function involves high-dimensional integrals. We offer a new implementation of the well-known expectation-maximization (EM) algorithm to facilitate the computation.

[14] See, e.g., Baks, Metrick, and Wachter (2001), Busse and Irvine (2006), Jones and Shanken (2005), Kosowski, Naik, and Teo (2007), and Pástor and Stambaugh (2002).

[15] We provide a more detailed discussion of the Bayesian performance evaluation literature in Appendix C.

[16] One can also think of our approach as an empirical Bayes method by casting it into an empirical Bayes framework in which alphas are assumed to be drawn from an underlying distribution that is characterized by a few population parameters (also known as hyperparameters). For finance applications of empirical Bayesian methods, see Frost and Savarino (1986) and Karolyi (1993).

[17] George (1986a,b) provides statistical foundations for this approach. For finance applications, Cosemans et al. (2015), Karolyi (1993), and Vasicek (1973) are examples that use cross-sectional information to obtain better estimates for risks.

[18] In Appendix C, we provide additional details on the related literature.

The idea of the EM algorithm is to treat the cross-section of alphas as missing observations and iteratively update our knowledge of the alpha distribution and the model parameters. With this approach, parameter estimates and learning about the missing observations can be done sequentially. In the context of our application, manager skill is missing observations (i.e., alphas). In the "expectation" step of the EM algorithm, for a given set of parameter values,[19] we fill in the missing observations with random draws from the conditional distribution of alphas given the parameter values. We calculate the averaged value of the likelihood function across these random draws. Essentially, at this step, we learn about manager skill to the best of our knowledge of the model parameters and update the likelihood function accordingly. In the "maximization" step of the algorithm, we maximize the updated likelihood function, which takes into account our recently updated information about manager skill. We obtain a new set of parameter estimates for factor loadings and residual standard deviations. These parameter estimates are subsequently fed into another "expectation" step to start a new round of estimation. The "expectation" step and the "maximization" step are performed iteratively to arrive at the MLE.[20]

3.2 Estimation Procedure

We discuss the basic idea of the algorithm in the main text and describe the details in Appendix A. The following steps describe the procedure of the EM algorithm:

Step I Let $\mathcal{G} = [\theta', \mathcal{B}', \mathbf{\Sigma}']'$ denote the collection of parameters to be estimated. We start at some parameter value $\mathcal{G}^{(0)}$. A sensible initial choice is the equation-by-equation OLS estimate for $\mathcal{B}$ and $\mathbf{\Sigma}$, and the MLE for θ based on the fitted OLS alphas.

[19] In our model, parameter values refer to fund-specific factor loadings, residual standard deviations, and parameters that govern the alpha population. The given set of parameter values could be the initial set of parameters to start the entire algorithm, for which a reasonable choice is the factor loadings and residual standard deviations from the equation-by-equation OLS estimates. It could also be the optimization outcome following the intermediate step (i.e., the "maximization" step) of the algorithm.

[20] Our online Appendix IF details the difference between the usual EM model and our implementation with fund heterogeneity.

Step II After the k-th iteration of the algorithm,[21] suppose the model parameters are estimated as $\mathcal{G}^{(k)}$. We calculate the expected value of the log complete likelihood function, with respect to the conditional distribution of $\mathcal{A}$ given the current parameter values and $\mathcal{R}$, that is,

$$L(\mathcal{G}|\mathcal{G}^{(k)}) = E_{\mathcal{A}|\mathcal{R},\mathcal{G}^{(k)}}[\log f(\mathcal{R},\mathcal{A}|\mathcal{G})], \tag{8}$$

$$= E_{\mathcal{A}|\mathcal{R},\mathcal{G}^{(k)}}[\sum_{i=1}^{N} \log f(R_i|\alpha_i,\beta_i,\sigma_i) f(\alpha_i|\theta)]. \tag{9}$$

It is very likely that $L(\mathcal{G}|\mathcal{G}^{(k)})$ will not have a closed-form expression. But a variant of the EM algorithm — named the Monte Carlo EM algorithm — recommends replacing the expectation with the sample mean, for which the sample is generated by simulating from the distribution of $\mathcal{A}|\mathcal{R},\mathcal{G}^{(k)}$.[22] We draw $M(=100)$ $\mathcal{A}$'s from the distribution $\mathcal{A}|\mathcal{R},\mathcal{G}^{(k)}$ and approximate the expectation in Equation (9) by its sample counterpart:[23]

$$\widehat{L}(\mathcal{G}|\mathcal{G}^{(k)}) = \frac{1}{M}\sum_{m=1}^{M}[\sum_{i=1}^{N} \log f(R_i|\alpha_i^m,\beta_i,\sigma_i) f(\alpha_i^m|\theta)]. \tag{10}$$

Step III We need to find parameter values that maximize $\widehat{L}(\mathcal{G}|\mathcal{G}^{(k)})$ and update the parameter estimate as $\mathcal{G}^{(k+1)}$. This is usually not easy if the dimension of the parameter space is high. However, in our context, there is a simple solution. An inspection of Equation (10) shows that $(\mathcal{B}',\mathbf{\Sigma}')$ and θ can be updated separately. More specifically, Equation (10) can be written as

$$\widehat{L}(\mathcal{G}|\mathcal{G}^{(k)}) = \sum_{i=1}^{N}\frac{1}{M}\sum_{m=1}^{M} \log f(R_i|\alpha_i^m,\beta_i,\sigma_i) + \frac{1}{M}\sum_{m=1}^{M}\sum_{i=1}^{N} \log f(\alpha_i^m|\theta). \tag{11}$$

[21] In our algorithm, an iteration refers to one round of updates for all model parameters, including parameters that characterize the alpha population, as well as fund-specific factor loadings and residual standard deviations.

[22] See Booth and Hobert (1999), Greg, Wei, and Tanner (1990), and McCulloch (1997).

[23] A larger number of M gives us a closer approximation to the expectation in Equation (9). However, it also increases the computational burden. We find that $M = 100$ gives us an estimate of θ (notice that the estimates of $\mathcal{B}$ and $\mathbf{\Sigma}$ do not depend on M, as shown in Appendix A) that is very close to that under, say, $M = 1,000$. This is because we have a large cross-section of alphas, so an insufficient sampling of the alpha distribution for individual funds does not have a large impact on the optimization outcome. We therefore set $M = 100$ to save computational time.

Notice that $\widehat{L}(\mathcal{G}|\mathcal{G}^{(k)})$ splits into two parts, one involving $\mathcal{B}$ and $\mathbf{\Sigma}$, and the other involving θ. This allows us to maximize $\widehat{L}(\mathcal{G}|\mathcal{G}^{(k)})$ by separately maximizing the two parts.[24]

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

The EM algorithm provides a tractable approach to find the MLE. It breaks the multidimensional optimization problem into smaller steps that are manageable. In theory, the EM estimator is guaranteed to converge to at least a local optimum of the likelihood function.[25] It has been successfully applied to panel regression models with random effects when the the random effects do not follow a standard distribution.[26] However, our model falls out of the realm of the standard application of the EM algorithm to panel regression models in that we allow heterogeneous risk loadings across funds. Therefore, it is an open question as to whether the algorithm performs well in our application. We provide a detailed simulation study to evaluate the performance of our EM algorithm.

We pay particular attention to the local optimum issue and construct a sequential estimation procedure to maximize the chance that our estimator converges to the global optimum. In particular, we first try a large number of randomly generated vectors of parameters to start the algorithm. Under a mild convergence threshold, we obtain many sets of initial parameter estimates. Some of these estimates correspond to a local optimum. We then select the top performers among these estimates and apply tougher convergence thresholds to sequentially identify the global optimum. Our online Appendix ID provides the details of the implementation of our algorithm.

The steps of the EM algorithm are intuitive. They build on the idea that our knowledge about the cross-section of alphas and the model parameters can be sequentially updated. In *Step I*, we start with some initial parameter estimates, possibly the standard OLS estimates. In *Step II*, given our starting estimates of the model parameters, we calculate the expected value of the log likelihood function conditional

[24]The fact that the log likelihood function that involves fund-specific parameters permits a closed-form solution also tremendously simplifies the computational burden in our framework. In Appendix F, we show that an extension of our framework that directly models information ratios no longer has this property.

[25]See Wu (1983) for the convergence properties of the EM algorithm.

[26]See Chen, Zhang, and Davidian (2002).

on the distribution of the alphas. An intuitive way to think about this step is to replace $\mathcal{A}|\mathcal{R}, \mathcal{G}^{(k)}$ with the best estimate of $\mathcal{A}$ given $\mathcal{R}$ and $\mathcal{G}^{(k)}$.[27] By doing this, we are trying to come up with our best guess of the missing alphas given the return data and the model parameters. This is the step in which we update our knowledge about the cross-section of alphas given our current estimates of the model parameters. In *Step III*, pretending that the estimated alphas in *Step II* are the true alphas, we have complete data and can easily estimate the model parameters. This is the step in which we update our knowledge about the risk loadings and the residual variances (i.e., $\mathcal{B}$ and $\mathbf{\Sigma}$). It is through the iterations between *Step II* and *Step III* that our estimates of the model parameters get refined. In Appendix D, we provide a more detailed discussion of the EM algorithm by specifying the parametric distribution Ψ as a GMD.

One concern about our model estimation is the large number of parameters. Indeed, since we allow heterogeneity in fund risk loadings and residual variances, the number of parameters grows almost proportionally with the number of funds in the cross-section. However, the set of parameters that grow with the number of funds are auxiliary parameters that govern the time-series dynamics of each individual fund. The key parameter set of interest — θ that parameterizes Ψ — does not change with the size of the cross-section. Intuitively, each additional fund added to the cross-section, while creating a new set of parameters to estimate for its time-series dynamics, will provide additional information for us to estimate θ. We show in the simulation study that θ is accurately estimated when we have a large cross-section.

3.3 A Simulation Study

3.3.1 Simulation Design

We detail a comprehensive simulation study to examine the performance of the NRA model and compare it with existing models.

We use mutual fund data as an example.[28] For our simulation study, we require a fund to have at least eight months of return observations. This allows us to have enough time series to estimate the factor model and is consistent with the existing

[27] See Neal and Hinton (1998) for a more rigorous interpretation of the EM algorithm.

[28] For a detailed description of these data, see the next section, in which we apply our method to the universe of mutual funds.

literature (e.g., Fama and French 2010; Ferson and Chen 2015). Imposing this constraint, we have 3,619 funds in the cross-section covering the 1983–2011 period. We obtain monthly returns for these funds. Except for the restriction on sample length, we do not impose any further restrictions on the data, and we use all the funds in the data for our simulation study.

With this sample of mutual funds, we run equation-by-equation OLS based on the full sample to obtain the estimates for $\mathcal{B}$ and $\mathbf{\Sigma}$ (i.e., $\mathcal{B}^*$ and $\mathbf{\Sigma}^*$). Factor loadings in $\mathcal{B}$ are based on the four-factor model in Carhart (1997). To make sure that the parameters in θ are representative of the parameter space that governs the alpha population, we use the set of parameters (θ^*) that correspond to the optimal parameter estimates for our mutual fund application in the next section. We collect all parameter estimates into $\mathcal{G}^* = [\theta^*, \mathcal{B}^*, \mathbf{\Sigma}^*]'$. $\mathcal{G}^*$ will be the true underlying parameter vector that governs the data-generating process. Special attention is paid to funds that do not have enough data to cover the entire sample period. In our simulations, we make sure that the simulated returns for these funds cover the same time periods as the original fund data.[29] In our online Appendix IA, we provide additional results under alternative parameterizations of θ.

Table 1 reports the summary statistics of the parameter vector θ^*. The two-component GMD separates the alpha cross-section into two groups. The first group has a mean that is very negative (-2.28%, per annum) and a large standard deviation (1.51%), and the second group has a mean that is slightly negative (-0.69%, per annum) and a smaller standard deviation (0.59%). It is less frequent for an alpha to fall into the first group, as its drawing probability is around 28%.

Based on $\mathcal{G}^*$, we simulate D (=100) panels of fund returns, each one having the same size as the original data panel. In particular, for each fund i, we randomly generate its alpha based on the GMD that is parameterized by θ^*. We then generate n_i random numbers independently from $\mathcal{N}(0, (\sigma_i^2)^*)$, where n_i is the sample size for fund i in the original data. These random numbers will be the simulated return

[29] We need to make a choice for the number of component distributions for the GMD in our simulation study. A one-component GMD (i.e., a single normal distribution) is obviously the simplest GMD one can specify, but it is too special for a simulation study. We therefore specify a two-component GMD — the simplest multicomponent GMD, which also turns out to be the preferred model for our mutual fund application, as shown in Section 4. Our simulation results do not depend on the chosen structure of the GMD. We also try a three-component GMD. The results are qualitatively similar in that NRA provides both more accurate (i.e., less biased) and less volatile (as measured by the root-mean-square error [RMSE]) estimates for both model parameters and implied summary statistics for the alpha population than alternative models.

Table 1: **Parameter Vector (θ^*) for the Simulated Model**

Parameter vector (θ^*) for the simulated model. It is the same as the estimated parameter vector for the mutual fund application in Section 4. μ_l and σ_l are the (annualized) mean and the (annualized) standard deviation for the l-th component normal distribution, and π_l is the probability for drawing from the l-th component, $l = 1, 2$.

	First component ($l = 1$)	Second component ($l = 2$)
$\mu_l(\%)$	-2.277	-0.685
$\sigma_l(\%)$	1.513	0.586
π_l	0.283	0.717

residuals. Together with the randomly generated alpha and the factor loadings β_i^*, these residuals enable us to construct the simulated return series for fund i.

To examine how residual correlation affects our results, we experiment with two correlation choices. One scheme is to allow the cross-section of residuals to be contemporaneously correlated with a common correlation coefficient of ρ. The other scheme, which is more realistic, is to calibrate a parametric model to match salient features of the collection of pairwise correlations among funds in the cross-section, capturing the heterogeneous cross-sectional dependency in the actual data. We provide simulation results for both approaches. We describe the details of the second scheme in our online Appendix IA.

In our simulation study, while we pay particular attention to the comparison between our method and the equation-by-equation OLS in our analysis, recent papers (e.g., Busse and Irvine 2006, Cohen, Coval, and Pástor 2005) on performance evaluation are aware of the pitfalls of the OLS approach. Instead of relying on the OLS alpha estimates, these papers use the fund-by-fund OLS estimates as the inputs for the second-stage analysis, in which funds with a similar performance are grouped together to diversify some of the noise that results from the fund-by-fund alpha estimation. However, our paper uses the equation-by-equation OLS as the main benchmark model for two reasons. First, it is routine to use the OLS alphas in the practice of management. Second, investors are primarily interested in the evaluation of a particular fund's performance — not a portfolio of funds. Nonetheless, besides

the equation-by-equation OLS, we also compare our framework to recent models such as Chen, Cliff, and Zhao (2015) and Jones and Shanken (2005).

3.3.2 The Alpha Population

For each of the simulated return panels, we estimate our model, thereby obtaining D sets of estimates. Table 2 summarizes these estimates and compares them with the estimates of two alternative models.[30] The first is the standard OLS model (that is, we first run equation-by-equation OLS and then fit a GMD for the cross-section of alpha estimates). Simple as it is, OLS is widely used for performance evaluation. Researchers often run equation-by-equation OLS first and then provide summary statistics on fitted alphas. The second is the model in CCZ, which offers a two-step procedure that first estimates the cross-section of fund alphas and their associated standard errors through the equation-by-equation OLS, and then fits a GMD based on these estimates. It improves on the standard OLS model by taking into account the equation-by-equation estimation uncertainty for the alphas. Since Jones and Shanken (2005) use a single normal distribution to model the alpha population, we do not compare their estimates with our model in Table 2.

Based on the results in Table 2, the NRA model stands out as superior to the two alternative models. In particular, its finite sample biases are uniformly smaller (in absolute value) than those under alternative models.

We focus on the case when there is zero correlation among return residuals—that is, $\rho = 0$. We first examine the means of the component normal distributions. For the first component, for which the group mean is very negative (-2.28%) and the drawing probability is relatively small (28%), the bias is 0.16% for NRA, 0.37% for CCZ, and 0.46% for the OLS model. Moreover, the estimation uncertainty (RMSE) for NRA (0.19%) is about half of that of CCZ (0.39%), both of which are substantially smaller than the RMSE for OLS (2.59%). For the second normal component, which happens more frequently than the first group (drawing probability is 72%), the bias in the mean estimate is small across all three models. Although both OLS and CCZ are inferior to NRA by making less precise and more noisy alpha estimates for individual funds (as we shall see later), when we pool the cross-section of funds together to

[30]Note that $\pi_1 + \pi_2 = 1$. However, we present the estimates for both parameters for completeness. Summary statistics for π_1 and π_2 in general will not sum up to one as we are averaging over the simulations.

estimate the overall population mean, the noise at the individual fund level largely cancels out. As a result, OLS and CCZ do not seem to provide significantly worse parameter estimates for the population means than NRA. This is particularly the case for the second normal component, as we have more observations (72%) that fall into this component, so the diversification effect is stronger. For the first component, for which we have fewer observations (28%) and more extreme alphas, NRA substantially improves on OLS and CCZ in estimating the population mean.

Table 2: **A Simulation Study: Parameter Estimates for the Alpha Population**

Model estimates in a simulation study. We fix the model parameters at $\mathcal{G}^*$ (Table 1) and generate D sets of data sample. For each set of data sample, we estimate our model using the proposed noise reduced alpha model (NRA), the standard equation-by-equation OLS (OLS), and the model in Chen, Cliff, and Zhao (CCZ; 2015). ρ is the assumed level of pairwise correlation for the correlation model that assumes an equal correlation for each pair of residual series. *Data depen.* corresponds to the correlation model that resembles the realized correlations for the actual data. 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 $\mathcal{G}^*$. *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—that is, $\sqrt{\sum_{d=1}^{D}(\gamma_d - \gamma)^2/D}$. *p(10)* reports the 10th percentile of the parameter estimates, and *p(90)* reports the 90th percentile of the parameter estimates. μ_l and σ_l are the (annualized) mean and the (annualized) standard deviation for the l-th component normal distribution, and π_l is the probability for drawing from the l-th component, $l = 1, 2$. Bold faced denotes lowest bias and RMSE.

		$\rho = 0$			$\rho = 0.2$			Data depen.		
		NRA	OLS	CCZ	NRA	OLS	CCZ	NRA	OLS	CCZ
μ_1(%)	Bias	**0.160**	0.455	0.367	**0.178**	0.937	0.442	**0.126**	0.317	0.382
(True = −2.277)	RMSE	**0.187**	2.592	0.394	**0.507**	3.333	0.778	**0.340**	2.949	0.540
	p(10)	−2.233	−4.457	−2.097	−2.655	−4.755	−2.638	−2.478	−4.485	−2.355
	p(90)	−2.007	0.481	−1.749	−1.307	2.350	−0.837	−1.832	0.197	−1.371
σ_1(%)	Bias	**0.046**	13.255	0.564	**−0.020**	13.920	0.496	**−0.003**	15.167	0.521
(True = 1.513)	RMSE	**0.081**	16.501	0.598	**0.217**	18.486	0.597	**0.193**	20.047	0.591
	p(10)	1.486	2.594	1.898	1.261	3.887	1.602	1.303	4.340	1.701
	p(90)	1.631	27.107	2.237	1.826	32.800	2.459	1.719	29.454	2.368
π_1(%)	Bias	**0.023**	−0.128	0.084	**0.006**	−0.141	0.067	**0.013**	−0.146	0.077
(True = 0.283)	RMSE	**0.029**	0.324	0.089	**0.071**	0.315	0.087	**0.099**	0.318	0.115
	p(10)	0.284	0.016	0.341	0.205	0.017	0.289	0.197	0.017	0.284
	p(90)	0.324	0.979	0.398	0.392	0.584	0.435	0.445	0.556	0.483
μ_2(%)	Bias	**−0.012**	0.066	−0.015	**0.017**	0.078	0.019	0.017	−0.318	**0.012**
(True = −0.685)	RMSE	**0.027**	1.549	0.051	0.310	1.603	**0.265**	0.247	0.785	**0.200**
	p(10)	−0.729	−1.180	−0.756	−1.047	−1.664	−0.998	−0.966	−1.554	−0.927
	p(90)	−0.675	1.870	−0.647	−0.232	−0.057	−0.263	−0.332	−0.586	−0.392
σ_2(%)	Bias	**0.009**	5.752	0.055	**−0.023**	4.914	0.033	**−0.001**	4.719	0.053
(True = 0.586)	RMSE	**0.018**	13.429	0.059	**0.046**	11.429	0.048	**0.083**	10.120	0.086
	p(10)	0.579	2.204	0.618	0.514	2.114	0.576	0.504	2.128	0.569
	p(90)	0.608	23.765	0.669	0.608	11.795	0.657	0.672	13.857	0.733
π_2(%)	Bias	**−0.023**	0.128	−0.084	**−0.006**	0.141	−0.067	**−0.013**	0.146	−0.077
(True = 0.717)	RMSE	**0.029**	0.324	0.089	**0.071**	0.315	0.087	**0.099**	0.318	0.115
	p(10)	0.676	0.022	0.602	0.608	0.416	0.565	0.555	0.444	0.517
	p(90)	0.719	0.984	0.659	0.795	0.983	0.711	0.803	0.983	0.716

Turning to the estimates of the variances of the two component distributions, the contrast in model performance is starker. First of all, OLS provides variance estimates that are severely biased. This is not surprising since, by ignoring the uncertainty in

the estimation of fund alphas, OLS attributes all the variation in the cross-section of fitted alphas to the variation of the underlying alpha population, thereby exaggerating the level of alpha dispersion for the alpha population. This result suggests that it is an ill-advised practice to first run equation-by-equation OLS and then provide summary statistics on fitted alphas.

NRA also provides variance estimates that are significantly better than CCZ. For example, for the first normal component, CCZ overestimates the standard deviation by 37% (= 0.564/1.513) and has an RMSE of 0.60%. Under the NRA model, the percentage bias is only 3% (= 0.046/1.513) and the RMSE is 0.08%. The reason for the underperformance of CCZ, relative to the NRA model, is that it takes the first-stage OLS parameter estimates (i.e., factor loadings and residual standard deviations) as given and fails to use the estimated alpha population in the second stage to update the first-stage estimates. Intuitively, if we know what a fund's alpha is (more precisely, in our context, the underlying population from which the alpha is drawn), we should be able to use this information to obtain better estimates for the fund's factor loadings and residual standard deviation.

When return residuals are correlated, the estimates for model parameters in general become more variable for all three methods. This is expected, as we have less information in the cross-section compared with the case when residuals are independent. Across different correlation specifications, our model still performs well. In particular, both bias and RMSE are small relative to the magnitude of the true parameters. For example, compared with the case of $\rho = 0$, when $\rho = 0.2$ and for σ_1, the estimation bias (= -0.02%) does not seem to go up (in absolute value), while RMSE goes up from 0.08% to 0.22%. Both remain small compared with the magnitude of the true σ_1 (= 1.51%), dominating the performance of CCZ, which is quite upwardly biased in estimating σ_1. The increased estimation uncertainty (RMSE) is the price we have to pay for misspecifying the likelihood function by not taking the correlation structure into account. However, the increase seems small for reasonable levels of residual correlations. Barras, Scaillet, and Wermers (2010) document that the average pairwise correlation among the four-factor model residuals is 0.08, which is consistent with our estimate of 0.06 (see our online Appendix IA). We think that $\rho = 0.2$ is a conservative upper bound for the average level of residual correlation. Moreover, when we take the heterogeneity in residual dependence into account by using our second correlation specification, our model performs even better than the case with $\rho = 0.2$. Overall, residual correlation does not substantially alter our model's

performance, and neither does it alter the relative performance of the three models. Hereafter, we use $\rho = 0$ as the benchmark correlation specification. We will also rely on our heterogeneous correlation model to provide robust standard errors.

Our results in Table 2 suggest that the OLS model, by first running equation-by-equation OLS regressions to obtain estimated alphas and then fitting a parametric distribution to these alphas, is massively biased in estimating the parameters that govern the cross-sectional alpha distribution. CCZ provides improvements over OLS by taking the estimation uncertainty in alpha into account by using fund-specific time-series information. However, in the context of our structural model, where fund alphas are intrinsically linked in the cross-section, using fund-specific information alone is insufficient. In particular, information about the cross-sectional alpha distribution can be used to update the OLS parameter estimates for individual funds, which in turn leads to, first, a more accurate adjustment for estimation uncertainty in alpha for individual funds, and second, a better estimate for the underlying alpha distribution. Simulation results show that the NRA model, by utilizing information from the entire cross-section, presents consistent improvements over CCZ, in terms of both estimation bias and estimation uncertainty.

We have shown that the NRA model produces superior parameter estimates for the alpha population in comparison with OLS and CCZ. Based on these parameter estimates, we can calculate several important statistics that summarize the alpha population. The NRA model produces more accurate and less volatile estimates for these statistics than alternative models, as shown in Table 3. We also include Jones and Shanken (2005) under diffuse priors in Table 3. To save space, we only present results for $\rho = 0$, that is, assuming there is no cross-sectional residual correlation. Alternative correlation specifications lead to qualitatively similar results. We leave them to our online Appendix IA.

Table 3: **A Simulation Study: Population Statistics**

This table shows population statistics based on the model estimates in a simulation study. We fix the model parameters at $\mathcal{G}^*$ (Table 1) and generate D sets of data sample. For each set of data sample, we estimate our model using the proposed noise reduced alpha model (NRA), the standard equation-by-equation OLS (OLS), the model in Chen, Cliff, and Zhao (CCZ; 2015), and the Bayesian model with diffuse priors in Jones and Shanken (JS; 2005). We then calculate summary statistics for the alpha population for both models based on the estimated model parameters. *Mean* is the mean of the alpha distribution. *Stdev.* is the standard deviation of the alpha distribution. *Iqr.* is the inter-quartile range of the alpha distribution. *p10* is the *10*th percentile of the alpha distribution. The other percentiles are similarly defined. *True* reports the population statistics based on the true model. *Estimate* reports the averaged estimate of the population statistics across the D sets of simulations. *RMSE* reports the square root of the mean squared estimation error—that is, $\sqrt{\sum_{d=1}^{D}(s_d - s)^2/D}$, where s is the true statistic and s_d is the estimated statistic based on the d-th simulated sample. Residual correlation is set at zero. Bold denotes lowest bias and RMSE.

		NRA	OLS	CCZ	JS
Mean(%)	Estimate	**−1.133**	−1.140	−1.142	−1.110
(True = −1.136)	RMSE	**0.034**	0.061	0.039	0.039
Stdev.(%)	Estimate	**1.191**	3.908	1.577	1.243
(True = 1.187)	RMSE	**0.031**	2.886	0.294	0.074
Iqr.(%)	Estimate	**1.163**	3.385	1.362	1.669
(True = 1.144)	RMSE	**0.050**	2.250	0.198	0.529
p5(%)	Estimate	−3.642	−5.499	−4.270	−3.030
(True = −3.700)	RMSE	0.108	1.806	0.481	0.528
p10(%)	Estimate	−2.815	−4.420	−3.150	−2.606
(True = −2.832)	RMSE	0.091	1.595	0.341	0.131
p50(%)	Estimate	−0.892	−1.120	−0.882	−1.110
(True = −0.860)	RMSE	0.044	0.285	0.043	0.279
p90(%)	Estimate	0.041	2.148	0.244	0.387
(True = 0.008)	RMSE	0.051	2.145	0.243	0.453
p95(%)	Estimate	0.307	3.174	0.675	0.811
(True = 0.284)	RMSE	0.054	2.893	0.407	0.629

Three of the methods (i.e., NRA, OLS, and CCZ) generate similar results regarding the overall population mean of the alpha distribution. This is not surprising, as different methods that feature shrinkage do not change the overall mean of the shrinkage target—that is, the population mean. It is the degree of shrinkage, which is characterized by the cross-sectional dispersion of the alpha population, that distinguishes among alternative shrinkage models. For this metric, the NRA model has a far better performance than both OLS and CCZ. In particular, OLS severely overestimates the standard deviation of the alpha population. CCZ overestimates it by 33% ($= (1.58 - 1.19/1.19)$, while the NRA model implies a bias of 0.3% ($= (1.191 - 1.187)/1.187$). Hence, CCZ considerably overstates the cross-sectional dispersion of the alpha population. Consequently, it provides biased estimates for different alpha percentiles. For example, for underperformers that rank at the 5th percentile of the alpha distribution, CCZ underestimates their alphas by 15% ($= (4.17 - 3.70)/3.70$), while NRA implies a bias of -2% ($= (3.64 - 3.70)/3.70$).

The JS method fits the standard deviation well, but it falls short in capturing the tails of the alpha distribution. This is not surprising given that JS uses a single normal distribution to describe the alpha population. When JS is correctly specified (i.e., the alpha population is indeed captured by a single normal distribution), we show in Appendix E that our model still dominates.

One message from our simulation study is that it is a mistake to infer the cross-sectional distribution of skill from the OLS alphas. When we test the performance of thousands of funds, extreme performers will exist and may influence our perception about the cross-sectional distribution. However, since skill is estimated with error, it is incorrect to equate the cross-sectional distribution of the OLS alphas with the cross-sectional distribution of true skill. One has to take estimation uncertainty into account to extract the underlying alpha distribution. Our model provides a systematic framework to achieve this.

3.3.3 Individual Funds

Having discussed the simulation results regarding the alpha population, we now turn to the inference of each individual fund. As mentioned previously, our method allows us to make inference on each individual fund through Equation (4). More specifically, given a set of parameter estimates, the density forecast of an individual fund is given by Equations (19)—(21) in Appendix D.

To evaluate relative model performance, we need to choose a few statistics that summarize the forecasting accuracy at the individual fund level. We concentrate on two statistics. The first focuses on the point estimates. In particular, the absolute deviation (AD) calculates the absolute distance between the alpha estimate and the true alpha value. The second reflects estimation uncertainty. We calculate the length of the confidence interval that is constructed to cover the true alpha value with a certain probability. Notice that the usual t-statistic is not an appropriate metric for model comparisons in our simulations since, by assumption, fund alphas are nonzero. For example, suppose the true alpha is 5% per annum for a certain fund and the point estimates based on the NRA model and the OLS are 4% and 7%, respectively. In addition, suppose the standard errors for the two models are the same. Clearly, the NRA model is a better model as it provides a more accurate point estimate without raising the standard error. However, the OLS t-statistic will be higher than that based on the NRA model, suggesting a more significant finding under the OLS. This is misleading. We therefore avoid the use of the t-statistic and separately show the improvement of our model over alternative methods for the numerator and the denominator of the t-statistic—that is, the point estimate and the length of the confidence interval, both of which can be easily obtained through the density forecast of the NRA model. Ideally, a better performing model will imply both a more accurate point estimate and a narrower confidence interval.

Table 4 reports the results. In terms of both point estimates and confidence intervals, OLS is dominated by the other models by having a larger mean absolute deviation in point estimates and a wider confidence interval for a given confidence level. Between the NRA model and CCZ, the mean absolute deviation by CCZ is 31% $(= (0.80 - 0.61)/0.61)$ higher than that of the NRA model. In terms of estimation uncertainty, both methods generate confidence intervals that roughly achieve the pre-specified coverage rate (i.e., the probability for the confidence interval to contain the true alpha value) of 90% and 95%. However, the length of the confidence interval generated under CCZ is larger than that generated by the NRA model. For instance, under 90% significance, the median length is 3.23% for CCZ, which is 15%$(= (3.23 - 2.81)/2.81)$ larger than that of the NRA model.

Between the NRA model and JS, the NRA model generates more accurate alpha estimates. Interestingly, JS has a smaller mean absolute deviation than CCZ. Given that CCZ is correctly specified while JS is not, this result suggests that the mis-specified simpler model (i.e., JS) may yield more accurate estimates than the more

complex model (i.e., CCZ), if the more complex model does not achieve the MLE. Hence, it highlights the advantage of our model over CCZ in correctly solving the MLE.

Table 4: **A Simulation Study: Individual Funds**

Summary statistics on model performance at the individual fund level. We fix the model parameters at $\mathcal{G}^*$ (Table 1) and generate D sample sets of data. For each sample set of data, we estimate our model using the proposed noise-reduced alpha model (NRA); the standard equation-by-equation OLS (OLS); the model in Chen, Cliff, and Zhao (CCZ; 2015), and the Bayesian model with diffuse priors in Jones and Shanken (JS, 2005). For NRA and CCZ, given the parameter estimates, we use Equations (19)-(21) in Appendix D to first construct the density forecast for each individual fund, and then obtain the point estimate and the confidence interval. For OLS, its point estimate is the estimate for the intercept, and its confidence interval is constructed using the point estimate and the standard error for the intercept. *Mean absolute deviation* is the average (across simulations) mean absolute distance between the estimated alpha and the true alpha for the cross-section of funds. *Stdev. of mean absolute deviation* is the average (across simulations) standard deviation of the absolute distance between the estimated alpha and the true alpha for the cross-section of funds. *Length, p* reports the averaged (across simulations) p-th percentile of the length of the 90% (or 95%) confidence intervals for the cross-section of funds. *Coverage probability* reports the averaged (across simulations) probability for the 90% (or 95%) confidence intervals to cover the true alpha values for the cross-section of funds. Other variables are similarly defined. Residual correlation is set at zero. Bold denotes lowest mean absolute deviation and standard deviation of mean absolute deviation.

		NRA	OLS	CCZ	JS
	Mean absolute deviation (%)	**0.611**	1.853	0.803	0.751
	Stdev. of mean absolute deviation (%)	0.659	3.187	0.724	**0.653**
90% confidence interval	Length, *p10*(%)	1.975	3.297	2.169	2.544
	Length, *p50*(%)	2.812	6.163	3.232	3.379
	Length, *p90*(%)	4.095	12.487	4.770	3.876
	Coverage probability	0.890	0.893	0.907	0.907
95% confidence interval	Length, *p10*(%)	2.433	3.928	2.680	3.033
	Length, *p50*(%)	3.545	7.343	4.083	4.028
	Length, *p90*(%)	4.822	14.878	5.677	4.623
	Coverage probability	0.943	0.943	0.952	0.945

Overall, our results suggest that the NRA model dominates the equation-by-equation OLS, CCZ, and JS, in terms of both estimating the alpha cross-section and making inference on a particular fund's alpha. Hence, under the assumption that fund alphas can be viewed as coming from an underlying population, we advocate the use of the NRA model for performance evaluation.

3.3.4 Further Comparisons with CCZ and JS

We further contrast our approach with CCZ and JS in Appendix E. In particular, we discuss the consequence of the methodological shortcut of CCZ relative to our method. We argue that CCZ yields inconsistent alpha estimates from the perspective of decomposing fund returns into luck and skill. We also show that our method seems to perform better than JS even when the cross-sectional alpha distribution follows a normal distribution—that is, when JS is correctly specified. We believe that the diffuse priors imposed on individual funds' OLS parameters (i.e., factor loadings and residual standard deviations) may have a non-negligible impact on the inference of the alpha population. We provide simulation-based evidence.

4 Results

4.1 Mutual Funds Example

We now provide an empirical application of our method to a sample of mutual funds.[31]

We obtain the mutual fund data used in Ferson and Chen (2015). Their fund data is from the Center for Research in Security Prices Mutual Fund database. They focus on active domestic equity funds covering the 1984—2011 period. To mitigate omission bias (Elton, Gruber, and Blake, 2001) and incubation and backfill bias (Evans 2010), they apply several screening procedures. They limit their tests to funds that have initial total net assets (TNA) above $10 million and have more than 80% of their holdings in stock. They also combine multiple share classes. We require that a fund has at least eight months of return observations to enter our test. This leaves us with

[31] We choose this application given that the data are relatively clean compared with hedge fund data. However, we recognize that there are disadvantages to applying our method to mutual funds, with the primary one being the distinct lack of evidence that any mutual fund manager has skill (see, e.g., Berk and Green 2004). That is, if we are proposing a method to reduce the noise to better measure skill, then it will be more challenging to apply to data for which there is very little evidence of skill in the first place.

a sample of 3,619 mutual funds for the 1984-2011 period.[32] We use the four-factor model in Fama and French (1993) and Carhart (1997) as our benchmark model.[33]

4.2 Parameter Estimates and Model Selection

A central issue is how we choose the number of components for the GMD that models the alpha distribution in the cross-section. A more complex model (i.e., a model with more component distributions) can potentially provide a better approximation to the underlying alpha distribution, but may overfit, leading to a model that has inferior forecasts out of sample. Standard model selection criteria (e.g., the Akaike information criterion or the Bayesian information criterion) may not work well in our context, as their derivations usually rely on asymptotic approximations within standard regression models (e.g., the linear regression model). In our application, since the number of parameters grows with the number of funds in the cross-section, it is unclear what size of the cross-section would be regarded as large enough to warrant asymptotic approximations. To have a rigorous model selection framework that takes many aspects of our application into account (e.g., unbalanced panel, large number of model parameters), we use a simulation-based model selection approach.[34]

Consider two nested models M_0 and M_1, with M_1 being the bigger model. For example, in our context, a GMD with a single component distribution will be nested within a two-component GMD specification, as, by setting the drawing probability for one of the component distributions to zero, the latter collapses to the former. To distinguish between M_0 and M_1, we need a metric that evaluates relative model performance. Given that our estimation relies on the MLE, a natural choice is the likelihood-ratio statistic, which measures the difference in likelihoods between the two candidate models. The likelihood-ratio statistic is also a key ingredient for many popular model selection criteria. In particular, let L_0 (L_1) be the value of the likelihood function evaluated at the model estimates for M_0 (M_1). The likelihood-ratio (LR) statistic is defined as:

$$\text{LR} = -2(\log L_0 - \log L_1). \tag{12}$$

[32] We thank Yong Chen for providing us with the mutual fund data used in Ferson and Chen (2015).

[33] Both the Fama-French factors and the momentum factor are obtained from Ken French's online data library. We report results on our model estimation using alternative benchmark factor models in the online Appendix IC.

[34] For a similar approach that bootstraps likelihood ratios to test the number of components in a GMD, see Feng and McCulloch (1996).

When the bigger model (i.e., M_1) provides a substantial improvement over the smaller model (i.e., M_0), LR will be large and positive. Therefore, a large likelihood-ratio statistic provides evidence against the smaller model.

We use simulations to find the cutoff value for LR. We first estimate M_0 and obtain its parameter estimates. Assuming M_0 is the true model, we simulate normally distributed return innovations to generate $D = 100$ return panels, similar to what we do in the simulation study in the previous section. For each panel, we estimate both M_0 and M_1, and calculate the LR statistic. The *95*th percentile of these LR statistics will be used as the cutoff for the LR statistic.

We incrementally select the best-performing parsimonious model. We first estimate a one-component and a two-component model. Based on the parameter estimates, the LR statistic between the two models is calculated to be 45.60. Assuming that the one-component model is true and simulating the model based on its parameter estimates, the *95*th percentile of the LR statistic is found to be 3.64,[35] which is smaller than the realized likelihood statistic. Therefore, the two-component model presents a significant improvement over the one-component model.

Next, we estimate a three-component model. The LR statistic between the two-component model and the three-component model is calculated to be 4.70. This time, assuming that the two-component model is true and simulating the model based on its parameter estimates, the *95*th percentile of the LR statistic is 14.30.[36] Hence, the realized LR statistic is less than the simulated LR cutoff, suggesting that we do not have enough evidence to discard the simpler two-component model.[37] Given the rejection of the three-component model, we do not need to further consider the four-component model, as its incremental contribution to the three-component model is likely to be even smaller than the incremental contribution of three-component model to the two-component model. We therefore select the two-component model

[35]The *90*th and *99*th percentiles of the LR statistic are 2.57 and 5.83, respectively. Hence, the two-component model is superior than the one-component model even at the 1% level.

[36]The *90*th and *99*th percentiles of the LR statistic are 12.46 and 17.47, respectively. Hence, the three-component model is not significant, even at the 10% level.

[37]The results reported are based on the benchmark correlation specification in which residual correlation is set at zero. We also perform the model selection exercise using our correlation model that mimics the dependence structure in the actual data. Between the one-component model and the two-component model, the simulated *90*th, *95*th, and *99*th percentiles of the LR statistic are 9.24, 12.08, and 33.73, respectively. Hence, we have strong evidence to reject the one-component model. Between the two-component model and the three-component model, the simulated *90*th, *95*th, and *99*th percentiles of the LR statistic are 16.16, 21.72, and 46.80, respectively. Hence, we do not have enough evidence to reject the two-component model.

as the final model. It is the most parsimonious model that still provides an adequate description of the cross-sectional distribution of fund alphas.

Our finding of a two-group categorization of mutual fund managers is consistent with the recent literature on mutual fund performance evaluation. For example, Barras, Scaillet, and Wermers (2010) use the false discovery approach to control for multiple testing and find that 75% of the funds are zero-alpha funds and 24% are unskilled (i.e., significantly negative).[38] The remaining 1% appear to be skilled but are statistically indistinguishable from zero. We also find that a two-group classification is sufficient to describe the universe of fund managers. In particular, unlike for underperformers, we do not need a third component distribution to model outperformers.

4.3 Evaluating the Population of Fund Performance

Table 5, Panel A, shows the parameter estimates for the GMD that describes the alpha population. Panel B reports the estimates for several important population statistics. We also report the corresponding standard errors for both parameter estimates and the estimates for population statistics. Notice that the parameter estimates in Panel A are the same as those in Table 1 since we deliberately set the model parameters at their MLE to evaluate our model's performance in the simulation study.[39]

Our estimates are related to but different from previous findings in the literature. For example, Fama and French (2010) document that the left tail of the alpha distribution should be more dispersed than the right tail.[40] To make inference, they propose an informal approach that assumes that alphas for the two tails are drawn from two normal distributions with the same mean (i.e., a mean of zero) as alphas but different variances. They calibrate their model by matching extreme percentiles

[38]Barras, Scaillet, and Wermers (2010) study 2,076 funds covering the 1975–2006 period. So their sample is somewhat different from ours. However, given the 23 years of overlap between our samples, we believe their estimates should roughly apply to our sample as well.

[39]Notice that the standard errors in Table 5 are different from those reported in the simulation study. In the simulation study, we simply set factor loadings and residual standard deviations at their equation-by-equation OLS estimates. In contrast, the results in Table 5 are based on the MLE, which are different from the OLS estimates.

[40]While the main goal of Fama and French (2010) is to test the overall null of zero outperformance, they do propose an informal "plug-in" approach to estimate the underlying alpha population. Our framework extends their method in two ways. First, we flexibly model the underlying alpha population, which, as they acknowledged, might be necessary to capture the tails of the distribution of the cross-section of alphas. Second, while they provide inference by matching certain t-statistic percentiles of the actual data to the simulated data, we rely on the likelihood function to provide more rigorous and efficient inference.

Table 5: **The Alpha Population: Mutual Funds**

Model estimates and population statistics for mutual funds. For a cross-section of 3,619 mutual funds covering the 1983–2011 period, we estimate our model, which is based on a two-component GMD specification for the alpha population. Assuming the estimated model is the true underlying model, we simulate to find the percentiles of both the parameter estimates and the population statistics. Panel A reports the parameter estimates for the model. μ_l and σ_l are the (annualized) mean and the (annualized) standard deviation for the l-th component normal distribution, and π_l is the probability for drawing from the l-th component, $l = 1, 2$. Panel B reports the estimated population statistics for the alpha distribution. *Mean* is the mean of the alpha distribution. *Standard deviation* is the standard deviation of the alpha distribution. *Interquartile range* is the interquartile range of the alpha distribution. *10th percentile* is the *10*th percentile of the alpha distribution. The other percentiles are similarly defined. For both Panels A and B, *p(5)* and *p(95)* report the *5*th and *95*th percentiles of the variable of interest across simulations, respectively.

Panel A: Parameter estimates for the alpha population

	Estimate	ρ=0		Data depen.	
		p(5)	*p(95)*	*p(5)*	*p(95)*
μ_1(%)	−2.277	−2.301	-1.948	−2.543	-1.739
σ_1(%)	1.513	1.424	1.654	1.221	1.730
π_1	0.283	0.280	0.330	0.173	0.496
μ_2(%)	−0.685	−0.748	-0.894	−0.993	-0.269
σ_2(%)	0.586	0.569	0.615	0.468	0.724
π_2	0.717	0.670	0.720	0.504	0.827

Panel B: Population statistics for the alpha population

	Estimate	ρ=0		Data depen.	
		p(5)	*p(95)*	*p(5)*	*p(95)*
Mean (%)	−1.135	−1.189	-1.075	−1.596	-0.717
Standard deviation (%)	1.185	1.121	1.247	0.972	1.456
Interquartile range (%)	1.142	1.085	1.234	0.824	1.716
*5*th percentile (%)	−3.689	−3.803	-3.445	−4.257	-3.038
*10*th percentile (%)	−2.862	−2.966	-2.652	−3.544	-2.104
*50*th percentile (%)	−0.894	−0.935	-0.851	−1.354	-0.447
*90*th percentile (%)	0.012	−0.016	0.096	−0.327	0.504
*95*th percentile (%)	0.287	0.222	0.390	−0.091	0.811
Fraction of positive alphas	0.106	0.095	0.123	0.038	0.261

of the cross-section of t-statistics. They estimate a dispersion of 1.25% to 1.50% for the left tail and 1.25% for the right tail of the alpha distribution. We generalize their insights in two ways. First, there is little reason to believe that alphas drawn

from the left tail should have the same mean as alphas drawn from the right tail. We allow a flexible two-component GMD to model the underlying alpha distribution. Second, we estimate our model through the joint likelihood function that combines information from the cross-section and time series. Our estimates are materially different from their estimates. For example, we find a much thinner right tail (dispersion $= 0.59\%$).[41]

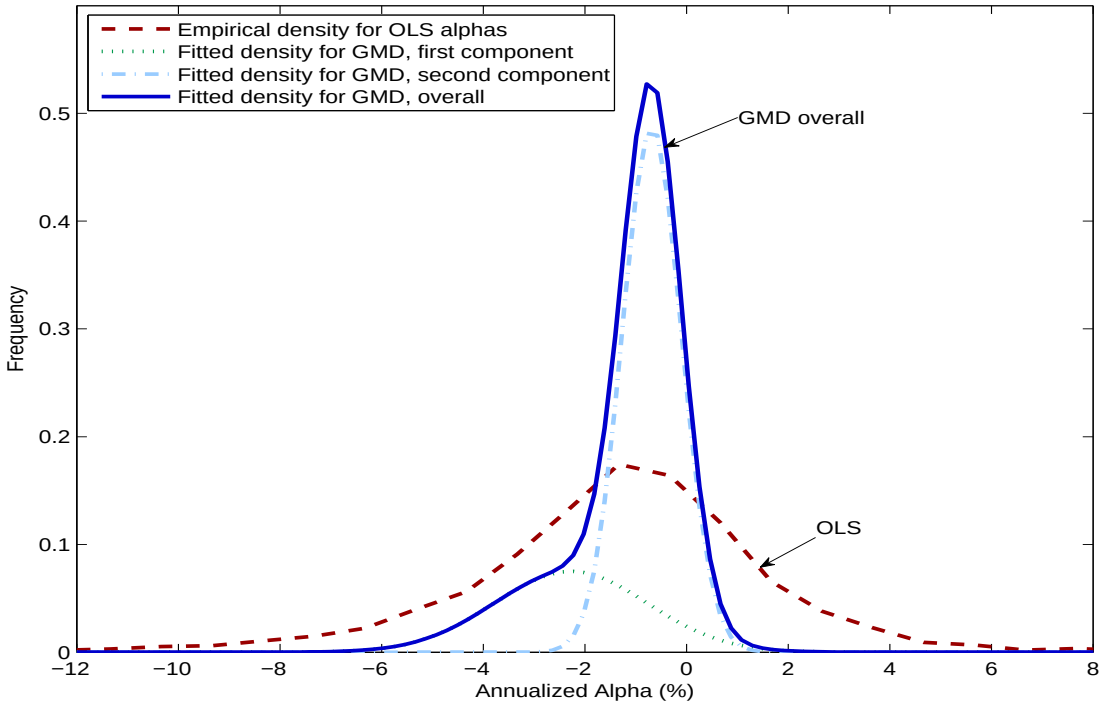
Figure 1 plots the density for the estimated alpha distribution as well as the empirical density for the OLS estimates. The density for the OLS fitted alphas is left skewed, indicating that there are more managers with large negative alphas than there are managers with large positive alphas. Our model estimation picks this up by having a separate component distribution that mostly covers negative alpha values. Allowing multiple component distributions gives our model the flexibility to capture the departure from normality in the data. Our results on model selection also show that it is both necessary (i.e., statistically significant) and sufficient to have this separate component distribution.

Another important observation from Figure 1 is that our method does not try to fit the OLS alphas. In fact, the overall density for the estimated GMD is more concentrated around its population mean than the empirical density for the OLS alphas. This is because our method downweighs noisy alpha estimates of individual funds when making inference on the alpha population. Extreme alpha estimates based on OLS are more likely to happen for funds with a short sample, more variable risk loadings, and/or more noisy return residuals. Our structural approach allows us to take these sources of estimation risk into account.

Our method allows us to make inference on important population characteristics by deviating from the usual fund-by-fund hypothesis testing framework. For example, we estimate the fraction of funds generating positive alphas to be 10.6%. In contrast, Barras, Scaillet, and Wermers (2010) use the multiple testing approach and find that fewer than 1% of funds generate a positive yet statistically insignificant alpha. To interpret the difference between our results and those in Barras, Scaillet, and Wermers (2010), it is important to bear in mind the difference between our method and the usual fund-by-fund hypothesis testing. By testing against the null hypothesis that fund alphas are zero, the traditional approach places more prominence on alpha equaling zero than alternative values. Our method assumes that the alpha distribution is

[41] We provide a more detailed discussion of our empirical findings and relate them to the literature in our online Appendix IG.

Figure 1: **Alpha Distribution for the Mutual Fund Population**



Density plots for the alpha population. For a cross-section of 3,619 mutual funds covering the 1984–2011 period, we estimate our model, which is based on a two-component GMD specification for the alpha population. The solid line shows the density for the estimated GMD. The dotted line shows the density for the first component of the GMD that has a negative mean. The dash-dotted line shows the density for the second component of the GMD that has a positive mean. We also estimate the equation-by-equation OLS. The dashed line shows the empirical density for the fitted OLS alphas.

continuous and tries to back out this distribution. It is therefore more appropriate to provide inference on population characteristics. We will likely have more power in identifying alphas with a small magnitude in our framework than the fund-by-fund approach, provided that our parametric assumption of the alpha distribution is a good approximation of reality.[42]

As expected, our estimates of extreme percentiles of the alpha distribution are substantially lower in magnitude than the estimates based on the equation-by-equation OLS. For example, the estimate for the 5th percentile is -3.69% under NRA and -5.48% under OLS. The estimate for the 95th percentile is 0.29% under NRA and

[42]See Harvey, Liu, and Zhu (2016) for a discussion on test power in the context of multiple testing.

3.07% under OLS. This stems from the shrinkage effect that we mentioned previously. Since the median fund generates an alpha of around -0.89%, cross-sectional learning forces us to pull alphas that are different from the population mean toward the population mean. Notice that the shrinkage effect seems to be stronger for large positive alphas than for large negative alphas. This is because large positive alphas are usually generated with a higher level of residual standard deviation than large negative alphas with the same magnitude. For example, the mean residual standard deviation for funds with alphas above the 95th percentile (i.e., 3.07%) is 7.4% (per annum), whereas the mean residual standard deviation for funds with alphas below 3.07% is 5.9%. Intuitively, in a competitive market, it is more difficult to generate a positive alpha than a negative alpha of the same magnitude. As a result, our method downweights the time-series information of funds with positive alphas more aggressively than funds with negative alphas with the same magnitude. These two features reinforce each other and generate the large discounts for positive alphas within our structural framework.

4.4 Individual Funds Evaluation: In-sample

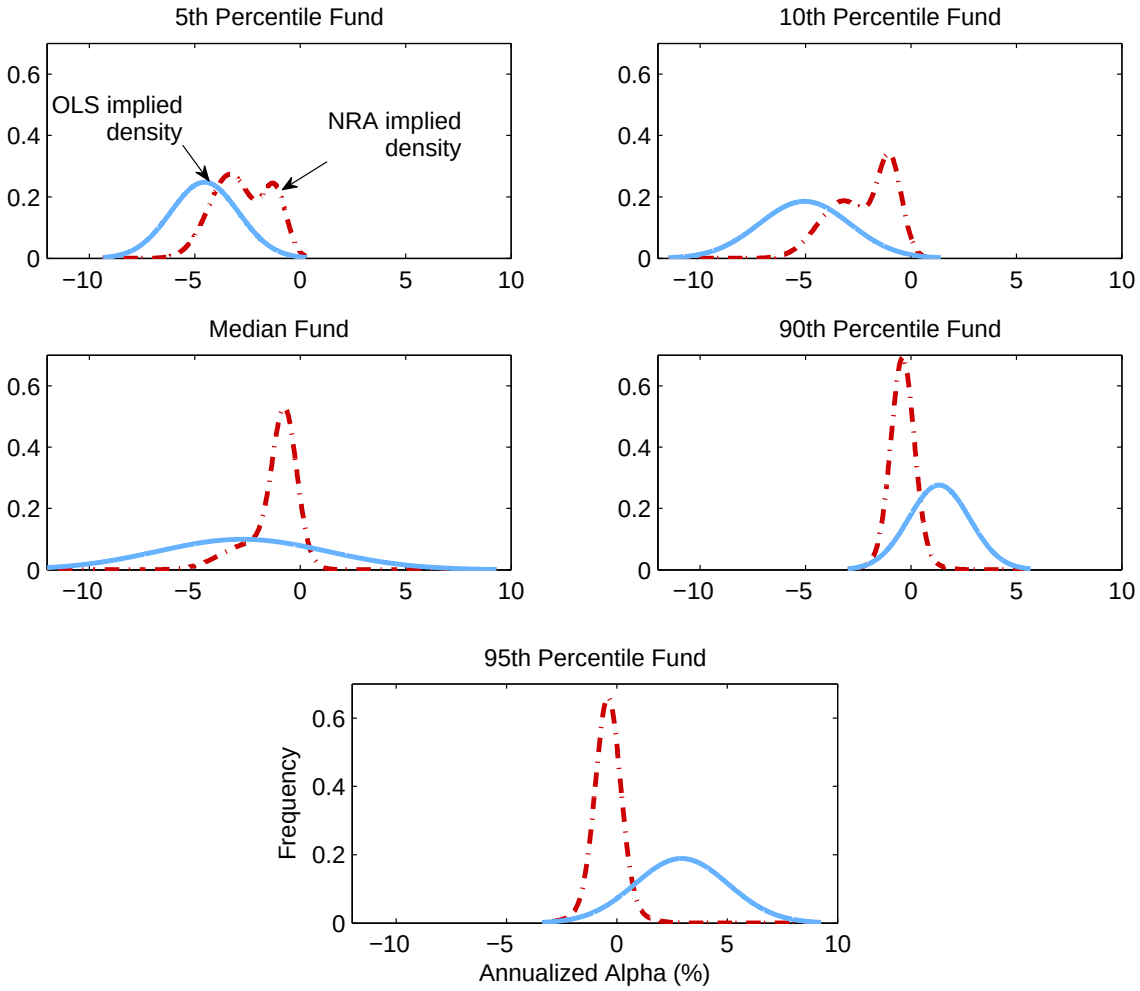
We use our estimated model to make inference on the alphas of individual funds. Given a set of parameter estimates, which use the information from the cross-section of funds, we are able to refine the alpha estimate of an individual fund that is based on time-series information alone, providing a more informative alpha estimate for an individual fund.

The formulas that provide density forecasts for individual funds are given in Equations (19)-(21) in Appendix D. We compare our model with the equation-by-equation OLS from both an in-sample fit and an out-of-sample forecasting perspective. We also compare our model with JS and CCZ later in forecasting alphas out-of-sample.

Focusing on in-sample fitting, Figure 2 shows the density forecasts based on our model for several exemplar funds. In particular, we rank funds by the t-statistics of their OLS alpha estimates and choose several funds that represent different percentiles of the cross-section of t-statistics.

We see several noticeable differences between our density forecasts and the forecasts based on OLS. First, there is a shrinkage effect in which the means of our forecasts pull the OLS means toward the overall population mean. This is the cross-

Figure 2: **Alpha Distributions for Individual Mutual Funds**



Density plots for individual funds. For a cross-section of 3,619 mutual funds covering the 1984–2011 period, we estimate our model, which is based on a two-component GMD specification for the alpha population. We also estimate the equation-by-equation OLS. We rank the cross-section of funds based on the t-statistics of their OLS alpha estimates and choose five funds whose t-statistics are the closest to the 5th, 10th, 50th, 90th, and 95th percentiles of the cross-section of t-statistics. Based on our model estimate, we plot the density estimates for these funds using Equations (19)-(21) in Appendix D. We also plot the density estimates for the OLS alphas.

sectional learning effect that we mentioned previously. Knowing the alpha distribution of other funds helps us make better inference on the alpha of a particular fund. The OLS alpha estimate uses exclusively fund-specific information and our approach

augments with cross-sectional information. The shrinkage effect seems particularly strong for funds with large positive OLS alphas. This is because we are more likely to observe a negative alpha than a positive alpha for the alpha population. In addition, as we mentioned previously, large positive alphas are usually associated with a larger residual standard deviation than negative alphas with the same magnitude. The cross-sectional learning effect therefore shrinks a positive alpha toward the population mean by more than what it shrinks a negative alpha with the same magnitude toward the population mean.

Second, the dispersion for the density forecast of our model is uniformly lower than that based on the OLS density forecast. This is consistent with our simulation study, in which we show that the average length of the confidence interval based on our method is substantially lower than that based on the OLS. Intuitively, our density forecast combines information from both the cross-section and the time series so it is less disperse than the OLS density forecast, which uses only the time-series information. Equation (20) makes this intuition more precise. Suppose we have a single-component distribution for the GMD—then the variance of a fund's alpha estimate following our approach is always smaller than its variance based on time-series information alone.

Finally, our density forecasts display non-normality, especially for funds with a negative mean estimate for alpha. For funds with a positive mean estimate, although the density looks unimodal, it is still a mixture distribution of two normal densities. This shows the flexibility of the GMD specification to capture different shapes of a probability density function. It also makes sense to have a non-normal density forecast for individual funds if the the underlying distribution for the alpha population is non-normally distributed. If this underlying distribution is more heavy-tailed and skewed than the normal distribution, then the density forecasts for individual funds should be able to reflect these non-normal features for the alpha population.

Table 6 summarizes the differences in both point estimates and confidence intervals between our model and the OLS. We group funds into different categories based on their OLS t-statistics and calculate the average difference between our model estimates and the OLS model estimates.

Focusing on the mean estimates, we see the differential impact of the shrinkage effect across different t-statistic groups. For example, for funds with an OLS t-statistic below -2.0, on average our model pulls the OLS alpha estimate closer to

Table 6: **Differences in Density Forecasts between OLS and the NRA Model**

Differences in density forecasts between the OLS and the noise-reduced alpha model. For a cross-section of 3,619 mutual funds covering the 1983–2011 period, we estimate our model, which is based on a two-component GMD specification for the alpha population. We also estimate the equation-by-equation OLS. We group funds into several groups based on the t-statistics of their OLS alpha estimates (denoted as t_α^{OLS}). We calculate the average difference in point estimates and confidence intervals between the NRA model and the OLS model. *Diff. in mean* reports the average difference in the mean forecast between our model and the OLS. *% diff. in CI(90)* and *% diff. in CI(95)* report the percentage differences in the length of the 90% and 95% confidence intervals between our model and OLS, respectively. *# of funds* reports the number of funds for each t-statistic category.

t_α^{OLS}	Diff. in mean (%)	% diff. in CI (90)	% diff. in CI (95)	# of funds
< -2.0 (worst)	3.391	−30.8%	−32.7%	523
$[-2.0, -1.5)$	2.352	−42.1%	−42.5%	391
$[-1.5, 0)$	0.688	−54.9%	−53.3%	1,640
$[0, 1.5)$	−2.052	−63.8%	−61.7%	906
$[1.5, 2.0)$	−3.774	−63.2%	−61.0%	98
> 2.0 (best)	−5.722	−64.5%	−61.8%	61

zero by 3.4% per annum. At the other extreme, for funds with significantly positive OLS alpha estimates (i.e., OLS t-statistic > 2.0), we on average move their alpha estimates closer to zero by 5.7% per annum. The shrinkage effect seems to be more pronounced for funds with large positive alpha estimates. This is attributable to, as we mentioned previously, the differential treatment of positive and negative alphas by the cross-sectional learning effect since we are more likely to observe a negative alpha than a positive alpha for the alpha population and a large positive alpha is usually generated with more uncertainty than a negative alpha with the same magnitude.

For confidence intervals, our model is able to shrink the 90% and 95% confidence intervals by at least 30% of the corresponding OLS confidence intervals. The reductions in estimation uncertainty seem substantial and are consistent with our results in the simulation study (see Table 4), in which we show that the reduction in the length of the confidence interval is not accompanied by a loss in the coverage rate. In fact, we are able to achieve a prespecified coverage rate (i.e., 90% or 95%) with a much narrower confidence interval.

The difference between Table 6 and Table 4 is that, unlike in the simulation study, we no longer observe the true alpha for each individual fund. To better assess the power of our approach, we perform an out-of-sample forecasting exercise in the next section.

4.5 Individual Funds Evaluation: Out-of-sample

We perform an out-of-sample analysis of our method by splitting our data into an in-sample estimation period and an out-of-sample holdout period. One way to interpret our results is to assume that someone tries to assess the predictive power of our model by following a simple strategy. She estimates our model at the end of the in-sample period and uses the model estimates to forecast risk-adjusted returns for the out-of-sample period. We try to evaluate such a strategy from a historical perspective.

It should be noted that our out-of-sample analysis is for illustration only. There are many issues related to our design of the out-of-sample analysis.[43] For example, assuming a certain degree of alpha persistence, what is the best sample size to estimate the model in-sample and what is the best horizon out-of-sample to forecast alphas? Answers to these questions will likely depend on the particular model we use. As another example, what should be the objective function for the out-of-sample analysis? Papers that study alpha persistence often look at the cross-sectional correlation of alphas for sorted portfolios.[44] However, due to the lack of diversification across funds from an investor's perspective, it seems that calculating the forecasting error in alpha may be more informative. In this section, we choose a very specific setup to illustrate the out-of-sample forecasting performance of our model. Our main point is to show the poor performance of metrics that are routinely used in the practice of management relative to our model, such as the equation-by-equation OLS and simple averages across funds. We leave a more detailed study of the out-of-sample performance of our model and shrinkage methods in general to future research.

Our sample runs from 1984 to 2011. We partition our sample into two parts, with the first two-thirds as the estimation period and the last one-third as the out-of-

[43]The key problem for the out-of-sample analysis is survivorship bias: we have selected mutual funds that exist both in-sample and out-of-sample, which may induce a large survivorship bias. The out-of-sample forecasting errors would be higher if funds selected in-sample drop out in the out-of-sample. However, since all models suffer from the same survivorship bias, our analysis is still valid from the perspective of model comparison.

[44]See, e.g., Elton, Gruber, and Blake (1996).

sample testing period. This way of partitioning the sample makes sure that we have a long enough in-sample period to have a reasonable model estimate.

For the in-sample period (i.e., 1984-2001), we estimate our model, the equation-by-equation OLS, and a third model that simply takes the in-sample mean of OLS alphas across all funds, CCZ, and JS (with a diffuse prior). Based on our model estimates, we construct a density forecast for each fund's alpha and use the mean of this density forecast to predict fund alpha in the future. For OLS, we use its in-sample alpha estimate to forecast its alpha in the future. For the in-sample mean model, we use a constant across all funds, which is the in-sample mean of OLS alphas, to predict fund alphas in the future. The implementations of CCZ and JS follow Chen, Cliff, and Zhao (2015) and Jones and Shanken (2005). The future alpha for a fund is obtained by running equation-by-equation OLS for the out-of-sample period (i.e., 2002-2011). Importantly, the out-of-sample alpha may not represent the true alpha.

For the in-sample period (i.e., 1984-2001), similar to our requirement for the full-sample estimate, a fund needs to have at least eight monthly observations to be considered in our estimation. This leaves us with 1,765 funds. Additionally, in order to have a valid alpha proxy for the out-of-sample period, we again require a fund to have at least eight monthly observations for the out-of-sample period. This further requirement leaves us with 1,488 funds for the out-of-sample period. To sum up, our in-sample estimation is based on 1,765 funds. Among these funds, 1,448 will be used in out-of-sample testing.

Table 7, Panel A, shows the in-sample model estimates, and Panel B shows the out-of-sample forecasting performance. Focusing on Panel A, there are noticeable differences between the parameter estimates for the 1984-2001 period and for the full sample period (see Table 5). Compared with the estimates in Table 5, it is less likely (drawing probability = 1.2%) to draw the alpha from the group with a very negative mean. However, conditional on drawing from this group, the alpha dispersion (15.15%) is much higher than the corresponding dispersion in Table 5 (1.51%). For the group with a mildly negative mean, its mean (-0.35%) is higher than the corresponding mean in Table 5 (-0.69%). At least two factors contribute to these differences in model estimates. First, the average fund return (and OLS alpha) is significantly higher for the in-sample period than for the full sample period. Second, compared with the full sample estimation, we have fewer funds for the in-sample estimation. This implies a lesser degree of learning across funds and may cause

a larger estimate for the dispersion of the alpha distribution. Despite these differences between the subsample and the full sample estimation, it remains interesting to see how our model performs out-of-sample.

Table 7: **Out-of-sample Forecasts for Mutual Funds**

In-sample model estimates (1984-2001) and out-of-sample forecasts (2002-2011) are based on the NRA model, the equation-by-equation OLS, and the simple mean model for which we take the average in-sample OLS alpha estimates across all funds (i.e., MEAN). We partition the mutual fund data into two parts and use the first part (1984-2001) for in-sample model estimation and the second part (2002-2011) for out-of-sample testing. For the in-sample period, we require a fund to have at least eight monthly observations. This leaves us with 1,765 funds. We estimate both our model and the equation-by-equation OLS based on these 1,765 funds. Panel A shows the parameter estimates for the NRA model. μ_l and σ_l are the (annualized) mean and the (annualized) standard deviation for the l-th component normal distribution, and π_l is the probability for drawing from the l-th component, $l = 1, 2$. For out-of-sample testing, we additionally require a fund to have at least eight monthly observations for the out-of-sample period, and 1,448 out of the 1,765 funds satisfy this additional requirement. We evaluate the out-of-sample forecasting performances of models based on these 1,448 funds. In particular, based on the in-sample estimates for our model, we construct a density forecast for each fund's alpha and use the mean of this density forecast to predict fund alpha in the future. For OLS, we use its in-sample alpha estimate to forecast its alpha in the future. For MEAN, we take the in-sample average OLS alpha across funds that are available in-sample. For CCZ, we follow the model proposed in Chen, Cliff, and Zhao (2015). For JS, we implement Jones and Shanken (2005) with a diffuse prior. The future alpha for each fund is obtained by running equation-by-equation OLS for the out-of-sample period. Panel B shows the forecasting results for the NRA model and the equation-by-equation OLS. t_α^{OLS} denotes the in-sample t-statistic for the alpha estimate of the OLS model. NRA (%), OLS (%), MEAN (%), CCZ (%), and JS (%) calculate the average absolute forecasting error (i.e., the alpha forecast based on the in-sample model minus the out-of-sample OLS alpha estimate) for the NRA model, the equation-by-equation OLS, and the simple mean model, CCZ, and JS, respectively, within a group of funds. Bold denotes lowest mean absolute forecasting error.

Panel A: In-sample model estimates, 1984–2001

Parameters	Estimate
$\mu_1(\%)$	-2.935
$\sigma_1(\%)$	15.146
π_1	0.012
$\mu_2(\%)$	-0.354
$\sigma_2(\%)$	1.065
π_2	0.988

Panel B: Out-of-sample forecasting error, 2002–2011

In-sample, t_α^{OLS}	NRA (%)	OLS (%)	MEAN (%)	CCZ (%)	JS (%)	# of funds
< -2.0 (worst)	3.286	6.613	3.881	4.751	**2.979**	64
$[-2.0, -1.5)$	**3.089**	3.699	4.377	2.857	3.205	75
$[-1.5, 0)$	2.748	2.916	3.723	**2.631**	2.820	565
$[0, 1.5)$	**2.606**	5.542	3.305	2.943	2.677	610
$[1.5, 2.0)$	**2.381**	10.469	2.829	4.223	2.504	87
> 2.0 (best)	2.766	12.022	**2.264**	7.441	2.466	87
Overall	**2.710**	5.165	3.454	3.236	2.748	1,488

Panel B shows the out-of-sample forecasting results. We again group funds based on their in-sample OLS t-statistics and present the average forecast error for each group. Compared with the equation-by-equation OLS estimates, our model provides a better alpha forecast across all groups of funds. The improvement of our model over the OLS is substantial. For example, for the 610 funds that have an in-sample t-statistic between zero and 1.5, our model is able to reduce the average forecast error from 5.54% to 2.61% (per annum). The reduction in forecast error is more pronounced for funds with large (absolute) OLS t-statistics. This is consistent with our finding based on the full sample estimation that the shrinkage effect is stronger for funds with large (absolute) OLS t-statistics. Across all groups of funds, the average percentage reduction in forecast error is 48% $(= (5.17\% - 2.71\%)/5.17\%)$. Therefore, our model is able to provide substantially better out-of-sample alpha forecasts compared with the OLS model.

On the other hand, the simple shrinkage model that takes the average of the OLS alphas across all funds also substantially improves on the equation-by-equation OLS estimates. However, our model seems to perform better than the simple shrinkage model across all but one group of funds. The overall reduction in forecasting error of our model compared with the simple shrinkage model is 21% $(= |2.71 - 3.45|/3.45)$. If the simple shrinkage model is taken to be the benchmark model that displays no alpha persistence, then the fact that our model implies a smaller forecasting error provides some evidence for alpha persistence.

Finally, comparing our model with CCZ and JS, our model presents a substantial improvement over CCZ and a mild improvement over JS. Looking into the forecasting performance across different fund groups, between our model and CCZ, our model substantially improves on CCZ in forecasting the alphas of funds with extreme in-sample alpha estimates. This can be explained by the fact that CCZ tends to overestimate the dispersion of the alpha population (as shown in Table 3 in our simulation study), resulting in an insufficient amount of shrinkage that leads to suboptimal alpha forecasts. On the other hand, between JS and our model, JS implies a smaller forecasting error for funds with extreme in-sample alpha estimates and a larger forecasting error for the majority of funds with in-sample alpha estimates that are not so extreme. We believe that a contributing factor for this result is the long holdout sample that we entertain for our out-of-sample forecasting exercise. While a relatively long holdout sample is desirable to obtain accurate alpha estimates for the out-of-sample, it may lead to alpha instability or performance reversion, especially for

funds with extreme in-sample performance. As a result, it favors methods that imply a larger amount of shrinkage (such as JS) than what is optimal under the assumption of alpha persistence. Alternative designs of the out-of-sample forecasting exercise, especially those that feature a shorter out-of-sample forecasting horizon, may suggest a better performance of our method in comparison with JS in forecasting alphas for funds with extreme in-sample alpha estimates.

In general, there are two important and related questions raised by the literature on performance evaluation: (i) how to obtain a good estimate of fund alpha? and (ii) how persistent are fund alphas? To obtain a good estimate of fund alpha, we have to assume that fund alphas are constant or at least persistent over a certain time window. On the other hand, to evaluate alpha persistence, we need to rely on the alpha estimates provided by methods that can answer the first question.

Our framework focuses on the first question—that is, obtaining a good estimate of fund alpha. In order to achieve this, we need to assume that fund alphas are constant and therefore persistent during the period over which we estimate the model. This is also the implicit assumption underlying the many papers that take the fund-by-fund hypothesis testing approach (see, e.g., Barras, Scaillet, and Wermers 2010; Fama and French 2010; Ferson and Chen 2015). However, a method that answers the first question has implications for the second question. Therefore, we use our method to refine the alpha estimates for individual funds and show that there seems to exist some persistence. The fact that our framework outperforms the commonly used benchmark model in forecasting alphas out-of-sample highlights the practical relevance of our approach. One extension of our forecasting exercise is to estimate our model over shorter horizons to allow for time variation in model parameters, similar to what people often do when evaluating performance persistence. We leave this extension to future research.

5 Other Issues

In Appendix F, we consider five additional issues relevant for performance evaluation: (i) misspecification of the factor model (we argue that the concern about model risk is to some extent alleviated by considering the noise-reduced alpha model); (ii) sample selection bias (we discuss the implications of both survivorship bias and reverse-survivorship bias on our model); (iii) multiple testing vs. NRA (we compare our model

with the multiple hypothesis testing approach that has been applied to performance evaluation); (iv) time-varying alphas (we discuss extensions of our framework that can deal with time-varying alphas); and (v) modeling information ratios rather than alphas (we demonstrate that our framework can be applied to alternative performance evaluation metrics such as the information ratio).

6 Conclusions

Is past performance indicative of future performance? Currently it is very difficult to detect repeatable performance because past returns are so noisy. Viewing fund alphas as coming from an underlying population, our structural model first backs out the distribution of the alpha population and then uses this distribution to refine the alpha estimate for each individual fund. By drawing on information from the cross-section of alphas, we show that our model is able to generate more accurate alpha estimates, both in-sample and out-of-sample, than current methods.

Methodologically, we propose a panel regression framework that allows us to make inference on the underlying population. We allow fund-specific regression coefficients to capture cross-sectional heterogeneity. Facing a large number of parameters to estimate, we adapt the EM algorithm to provide efficient inference based on the MLE.

Our approach is likely useful for a number of finance applications. Essentially, when there is cross-sectional heterogeneity and when it is appropriate to view the effects as coming from a certain population, we can apply our model to make inference on both the population and the individual effects. Our use of the GMD is also flexible enough to approximate a variety of parametric distributions for the population.

Our framework can be extended along several important directions. First, we can allow both time-varying alphas and risk loadings. Second, instead of shrinking alphas in the cross-section, we can pool information from the cross-section to estimate the loadings on important alpha predictors, such as fund size (see Harvey and Liu 2017c) and other fund characteristics. We leave these extensions to future research.

References

Andrikogiannopoulou, A., and F. Papakonstantinou. 2016. Estimating mutual fund skill: A new approach. *Working Paper.*

Avramov, D., and R. Wermers. 2005. Investing in mutual funds when returns are predictable. *Journal of Financial Economics 81, 339–377.*

Baks, K., A. Metrick, and J. Wachter. 2001. Should investors avoid all actively managed mutual funds? A study in Bayesian performance evaluation. *Journal of Finance 56, 45–85.*

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.*

Bekaert, G., and C. R. Harvey. 1995. Time-varying world market integration. *Journal of Finance 50, 403–444.*

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

Bickel, P. J., and B. Li. 2006. Regularization in statistics. *Test 15, 271–344.*

Bilmes, J. A. 1998. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute 4(510), 126.*

Booth, J. G. and J. P. Hobert. 1999. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society. Series B. 61, 265–285.*

Brown, S. J., W. Goetzmann, R. G. Ibbotson, and S. A. Ross. 1992. Survivorship bias in performance studies. *Review of Financial Studies 5, 553–580.*

Busse, J. A., and P. J. Irvine. 2006. Bayesian alphas and mutual fund persistence. *Journal of Finance 61, 2251–2288.*

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

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

Chen, Y., M. Cliff, and H. Zhao. 2015. Hedge funds: The good, the bad, and the lucky. *Journal of Financial and Quantitative Analysis 52, 1081–1109.*

Chen, J., D. Zhang, and M. Davidian. 2002. A Monte Carlo EM algorithm for generalized linear mixed models with flexible random effects distribution. *Biostatistics 3, 347–360.*

Christopherson, J. A., W. E. Ferson, and A. L. Turner. 1999. Performance evaluation using conditional alphas and betas. *Journal of Portfolio Management 26, 59-72.*

Cohen, A. C. 1967. Estimation in mixtures of two normal distributions. *Technometrics 9, 15–28.*

Cohen, R. B., J. D. Coval, and L. Pástor. 2005. Judging fund managers by the company they keep. *Journal of Finance 60, 1057–1096.*

Cosemans, M., R. Frehen, P. C. Schotman, and R. Bauer. 2015. Estimating security betas using prior information based on firm fundamentals. *Review of Financial Studies: hhv131.*

Day, N. E. 1969. Estimating the components of a mixture of normal distributions. *Biometrika 56, 463–474.*

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society 39, 1–38.*

Elton, E. J., M. J. Gruber, and C. R. Blake. 1996. The persistence of risk-adjusted mutual fund performance, *Journal of Business 69, 133–157.*

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.*

Fan, J., and J. Lv. 2010. A selective overview of variable selection in high dimensional feature space. *Statistical Sinica 20, 101–148.*

Feng, Z. D., and C. E. McCulloch. 1996. Using bootstrap likelihood ratios in finite mixture models. *Journal of the Royal Statistical Society. Series B, 609–617.*

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

Ferson, W., S. Sarkissian, and T. Simin. 2008. Asset pricing models with conditional alphas and betas: The effects of data snooping and spurious regression. *Journal of Financial and Quantitative Analysis 43, 331-354.*

Figueiredo, M. A., and A. K. Jain. 2002. Unsupervised learning of finite mixture models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 24, 381–396.*

French, K. 2008. Presidential address: The cost of active investing. *Journal of Finance 63, 1537–1573.*

Frost, P. A., and J. E. Savarino. 1986. An empirical bayes approach to efficient portfolio selection. *Journal of Financial and Quantitative Analysis 21, 293–305.*

George, E. I. 1986a. Combining minimax shrinkage estimators. *Journal of the American Statistical Association 81, 437–445.*

George, E. I. 1986b. Minimax multiple shrinkage estimation. *Annals of Statistics 14, 188–205.*

Gray, S. F. 1996. Modeling the conditional distribution of interest rates as a regime-switching process. *Journal of Financial Economics 42, 27–62.*

Greene, W. H. 2003. *Econometric analysis*. Pearson Education India, Delhi.

Greg, C., G. Wei, and M. A. Tanner. 1990. A monte carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association 85, 699–704.*

Harvey, C. R., and Y. Liu. 2017a. Luck vs. skill and factor selection. In J. Cochrane and T. J. Moskowitz (Eds.), *The Fama portfolio*, 250–60. Chicago: University of Chicago Press.

Harvey, C. R., and Y. Liu. 2017b. Lucky factors. *Working Paper.* Available at http://ssrn.com/abstract=2528780.

Harvey, C. R., and Y. Liu. 2017c. Decreasing returns to scale, fund flows, and performance. *Working Paper* Available at https://ssrn.com/abstract=2872385.

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

Hsiao, C. 2014. *Analysis of panel data*. Cambridge: Cambridge University Press.

Huij, J., and M. Verbeek. 2007. Cross-sectional learning and short-run persistence in mutual fund performance. *Journal of Banking & Finance 31, 973–997.*

Ishwaran, H., and M. Zarepour. 2002. Dirichlet prior sieves in finite normal mixtures. *Statistica Sinica 12, 941–963.*

Jensen, M. C. 1968. The performance of mutual funds in the period 1945-1964. *Journal of Finance 23, 389–416.*

Jensen, M. C. 1969. Risk, the pricing of capital assets, and the evaluation of investment portfolios. *Journal of Business 42, 167–247.*

Jones, C., and H. Mo. 2016. Out-of-sample performance of mutual fund predictors. *Working Paper.*

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

Karolyi, G. A. 1992. Predicting risk: Some new generalizations. *Management Science 38, 57–74.*

Karolyi, G. A. 1993. A Bayesian approach to modeling stock return volatility for option valuation. *Journal of Financial and Quantitative Analysis 28, 579–594.*

Kass, R. E., and L. Wasserman. 1996. The selection of prior distributions by formal rules. *Journal of the American Statistical Association 91, 1343–1370.*

Kosowski, R., N. Y. Naik, and M. Teo. 2007. Do hedge funds deliver alpha? A Bayesian and bootstrap analysis. *Journal of Financial Economics 84, 229–264.*

Kosowski, R., A. Timmermann, R. Wermers, and H. White. 2006. Can mutual fund "stars" really pick stocks? New evidence from a bootstrap analysis. *Journal of Finance 61, 2551–2595.*

Linnainmaa, J. T. 2013. Reverse survivorship bias. *Journal of Finance 68, 789–813.*

Maddala, G. S. 2001. Introduction to econometrics, John Wiley and Sons. *West Sussex, England.*

McCulloch, C. E. 1997. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association 92, 162–170.*

McLachlan, G., and T. Krishnan. 2007. *The EM algorithm and extensions*, 2nd ed. John Wiley & Sons, 2007.

Neal, R., and G. Hinton. 1998. A view of the EM algorithm that justifies incremental, sparse, and other variants. In Jordan, M., editor, *Learning in Graphical Models.* Kluwer Academic Press.

Pástor, L., and R. Stambaugh. 2002a. Mutual fund performance and seemingly unrelated assets. *Journal of Financial Economics 63, 315–349.*

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

Sastry, R. 2013. The cross-section of investing skill. *Working Paper.*

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

Stambaugh, R. 2003. Inference about survivors. Unpublished working paper. Wharton School, University of Pennsylvania.

Vasicek, O. A. 1973. A note on using cross-sectional information in Bayesian estimation of security betas. *Journal of Finance 28, 1233-1239.*

Verbeke, G., and E. Lesaffre. 1996. A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association 91, 217–221.*

Vidaurre, D., C. Bielza, and P. Larrañaga. 2013. A survey of L_1 regression. *International Statistical Review 81, 361–387.*

Wooldridge, J. M. 2013. Random effects estimation. *Introductory Econometrics: A Modern Approach, 474–478.*

Wu, C. F. J. 1983. On the convergence properties of the EM algorithm. *Annals of Statistics 11, 95–103.*

A Implementing the EM Algorithm

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

Using Bayes' law, we have:

$$f(\mathcal{A}|\mathcal{R}, \mathcal{G}^{(k)}) \propto f(\mathcal{R}|\mathcal{A}, \mathcal{G}^{(k)}) f(\mathcal{A}|\mathcal{G}^{(k)}). \tag{13}$$

Given the independence of the residuals and the α_i's, the right-hand side of Equation (13) is the product of the likelihoods of all funds—that is,

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

Therefore, to characterize $f(\mathcal{A}|\mathcal{R}, \mathcal{G}^{(k)})$, it is sufficient for us to determine $f(R_i|\alpha_i, \mathcal{G}^{(k)}) f(\alpha_i|\mathcal{G}^{(k)})$ for each fund i. For ease of exposition, we use $\mathcal{G}$ and $\mathcal{G}^{(k)}$ interchangeably to denote the known parameters at the k-th iteration.

Under normality, we have

$$\begin{aligned} f(R_i|\alpha_i, \mathcal{G}^{(k)}) &\propto \exp\{-\frac{\sum_{t=1}^{T}(r_{it} - \alpha_i - \beta_i' f_t)^2}{2\sigma_i^2}\}, \\ &\propto \exp\{-\frac{[\alpha_i - \frac{\sum_{t=1}^{T}(r_{it} - \beta_i' f_t)}{T}]^2}{2\sigma_i^2/T}\}, \end{aligned}$$

which can be viewed as the probability density for α_i. Moreover, this is a normal density with mean $\bar{a}_i \equiv \sum_{t=1}^{T}(r_{it} - \beta_i' f_t)/T$ and variance σ_i^2/T, that is, $\mathcal{N}(\bar{a}_i, \sigma_i^2/T)$.

By assumption, $f(\alpha_i|\mathcal{G}^{(k)})$ is the density for a GMD that is parameterized by $\theta = (\{\pi_l\}_{l=1}^N, \{\mu_l\}_{l=1}^N, \{\sigma_l^2\}_{l=1}^N)$. It can be shown that $f(R_i|\alpha_i, \mathcal{G}^{(k)}) f(\alpha_i|\mathcal{G}^{(k)})$ — the product of a normal density (i.e., $\mathcal{N}(\bar{a}_i, \sigma_i^2/T)$) and the density for a GMD — is also a density for a GMD, whose parameters are given by

$$\begin{aligned} \tilde{\mu}_{i,l} &= (\frac{\sigma_l^2}{\sigma_l^2 + \sigma_i^2/T})\bar{a}_i + (\frac{\sigma_i^2/T}{\sigma_l^2 + \sigma_i^2/T})\mu_l, \\ \tilde{\sigma}_{i,l}^2 &= \frac{1}{1/\sigma_l^2 + 1/(\sigma_i^2/T)}, \\ \tilde{\pi}_{i,l} &= \frac{\pi_l \phi(\bar{a}_i - \mu_l, \sigma_l^2 + \sigma_i^2/T)}{\sum_{l=1}^{L} \pi_l \phi(\bar{a}_i - \mu_l, \sigma_l^2 + \sigma_i^2/T)}, \; l = 1, 2, \ldots, L, \end{aligned}$$

where $\phi(\mu, \sigma^2)$ is the density of the normal distribution $\mathcal{N}(0, \sigma^2)$ evaluated at μ.

Therefore, $f(\mathcal{A}|\mathcal{R},\mathcal{G}^{(k)})$ can be characterized as the density for N independent variables. The i-th variable follows a GMD that is parameterized by

$$\tilde{\theta}_i = (\{\tilde{\pi}_{i,l}\}_{l=1}^{L}, \{\tilde{\mu}_{i,l}\}_{l=1}^{L}, \{\tilde{\sigma}_{i,l}^2\}_{l=1}^{L}).$$

A.2 Maximizing $\sum_{i=1}^{N} \frac{1}{M} \sum_{m=1}^{M} \log f(R_i|\alpha_i^m, \beta_i, \sigma_i)$ (*Step III*)

For α_i^m, m denotes the m-th random draw from the marginal distribution of α_i obtained from the previous step, and i denotes the i-th fund in the cross-section. Given the independence of the residuals, we can find the MLE of $\mathcal{B}$ and $\mathbf{\Sigma}$ fund-by-fund. In particular, the log-likelihood for fund i is given by

$$\frac{1}{M}\sum_{m=1}^{M} \log f(R_i|\alpha_i^m, \beta_i, \sigma_i) = \frac{1}{M}\sum_{m=1}^{M}\sum_{t=1}^{T} \log f(r_{it}|\alpha_i^m, \beta_i, \sigma_i), \tag{14}$$

through which we can find the MLE of β_i and σ_i. Under the normality assumption, it can be shown that the right-hand side of Equation (14) can be written as

$$\frac{1}{M}\sum_{m=1}^{M}\sum_{t=1}^{T} \log f(r_{it}|\alpha_i^m, \beta_i, \sigma_i) = -\frac{T}{2}\log(2\pi\sigma_i^2) - \frac{1}{2\sigma_i^2}[\sum_{t=1}^{T}(r_{it} - \beta_i' f_t - \bar{\alpha}_i)^2 + T(\overline{\alpha_i^2} - \bar{\alpha}_i^2)], \tag{15}$$

where $\bar{\alpha}_i$ and $\overline{\alpha_i^2}$ are defined as:

$$\bar{\alpha}_i = \frac{1}{M}\sum_{m=1}^{M}\alpha_i^m, \quad \overline{\alpha_i^2} = \frac{1}{M}\sum_{m=1}^{M}(\alpha_i^m)^2.$$

An inspection of Equation (15) shows that the MLE of β_i and σ_i can be found sequentially. We find the MLE for β_i first. Notice that the MLE $\hat{\beta}_i$ is essentially the estimates of the slope coefficients for the OLS that regresses the time series of $\{r_{it} - \bar{\alpha}_i\}_{t=1}^{T}$ on $\{f_t\}_{t=1}^{T}$. As a result, we have

$$\hat{\beta}_i = (F'F)^{-1}F'Y_i,$$

where

$$F_{(T\times K)} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_T \end{bmatrix}, \quad Y_{i(T\times 1)} = \begin{bmatrix} r_{i,1} - \bar{\alpha}_i \\ r_{i,2} - \bar{\alpha}_i \\ \vdots \\ r_{i,T} - \bar{\alpha}_i \end{bmatrix}.$$

Fixing β_i at its MLE, we take the first-order derivative of Equation (15) with respect to σ_i^2 to obtain the MLE for σ_i^2, that is,

$$\hat{\sigma}_i^2 = \frac{1}{T}\sum_{t=1}(r_{it} - \hat{\beta}_i' f_t - \bar{\alpha}_i)^2 + (\overline{\alpha_i^2} - \bar{\alpha}_i^2).$$

Define $\overline{\hat{\varepsilon}_i^2} \equiv \frac{1}{T}\sum_{t=1}(r_{it} - \hat{\beta}_i' f_t - \bar{\alpha}_i)^2$ and $\widehat{Var}(\alpha_i) = (\overline{\alpha_i^2} - \bar{\alpha}_i^2)$. The MLE of σ_i^2 can be expressed as

$$\hat{\sigma}_i^2 = \overline{\hat{\varepsilon}_i^2} + \widehat{Var}(\alpha_i). \tag{16}$$

Note that $\{\alpha_i^m\}_{m=1}^M$ are simulated data. When the size of the simulated data is large, the sample moments in Equation (16) will be close to the population moments. We therefore replace the sample moments with their population moments. This helps us obtain the exact analytical solutions for β_i and σ_i when the conditional distribution of $\mathcal{A}$ is given in Section 1 of Appendix A. In particular, the exact MLE for β_i is:

$$\breve{\beta}_i = (F'F)^{-1}F'\tilde{Y}_i,$$

where $\tilde{Y}_i = [r_{i,1} - m(\alpha_i), r_{i,2} - m(\alpha_i), \ldots, r_{i,T} - m(\alpha_i)]'$ and $m(\alpha_i) = E_{\mathcal{A}|\mathcal{R},\mathcal{G}^{(k)}}(\alpha_i) = \sum_{i=1}^L \tilde{\pi}_{i,l}\tilde{\mu}_{i,l}$ (here "m" denotes mean). The exact MLE for σ_i^2 is:

$$\breve{\sigma}_i^2 = \frac{1}{T}\sum_{t=1}(r_{it} - \breve{\beta}_i' f_t - m(\alpha_i))^2 + var(\alpha_i),$$

where

$$\begin{aligned} var(\alpha_i) &\equiv Var_{\mathcal{A}|\mathcal{R},\mathcal{G}^{(k)}}(\alpha_i), \\ &= \sum_{l=1}^L \tilde{\pi}_{i,l}[(\tilde{\mu}_{i,l} - m(\alpha_i))^2 + \tilde{\sigma}_{i,l}^2]. \end{aligned}$$

The parameter values in $\tilde{\theta}_i = (\{\tilde{\pi}_{i,l}\}_{l=1}^L, \{\tilde{\mu}_{i,l}\}_{l=1}^L, \{\tilde{\sigma}_{i,l}^2\}_{l=1}^L)'$ can be found in Section 1 of Appendix A.

A.3 Maximizing $\sum_{m=1}^M \sum_{i=1}^N \log f(\alpha_i^m|\theta)$ (*Step III*)

To optimize $\sum_{m=1}^M \sum_{i=1}^N \log f(\alpha_i^m|\theta)$, we need to invoke the EM algorithm. Our goal is to find the MLE of θ when MN observations are assumed to be drawn from the GMD that is parameterized by θ. For ease of exposition, we replace the subscript in α_i^m with j so that $\{\alpha_i^m\}_{(i=1,\ldots,N;\ m=1,\ldots,M)} = \{\alpha_{ij}\}_{(i=1,\ldots,N;\ j=1,\ldots,M)}$. The starting value of θ is obtained from $\mathcal{G}^{(k)}$.

- Suppose the initial parameter vector is $\hat{\theta} = (\{\hat{\pi}_l\}_{l=1}^L, \{\hat{\mu}_l\}_{l=1}^L, \{\hat{\sigma}_l^2\}_{l=1}^L)$.
- Expectation Step: Compute the expected value of the indicator variable that indicates which population (e.g., the population of skilled or unskilled managers) α_{ij} is drawn from:

$$\begin{aligned} \hat{p}_{ijl} &= \widehat{Pr}(\alpha_{ij} \text{ comes from Group } l) \\ &= \frac{\hat{\pi}_l \phi(\alpha_{ij}; \hat{\mu}_l, \hat{\sigma}_l^2)}{\sum_{l=1}^L \hat{\pi}_l \phi(\alpha_{ij}; \hat{\mu}_l, \hat{\sigma}_l^2)},\ i = 1, \ldots, N;\ j = 1, \ldots, M;\ l = 1, \ldots, L, \end{aligned}$$

where $\phi(\,\cdot\,;\mu,\sigma^2)$ is the density of the normal distribution $\mathcal{N}(\mu,\sigma^2)$.

- Maximization Step: Compute the weighted means and variances, with weights obtained from the Expectation Step:

$$\begin{aligned}\tilde{\mu}_l &= \frac{\sum_{ij}\hat{p}_{ijl}\alpha_{ij}}{\sum_{ij}\hat{p}_{ijl}},\ \tilde{\sigma}_l^2 = \frac{\sum_{ij}\hat{p}_{ijl}(\alpha_{ij}-\tilde{\mu}_l)^2}{\sum_{ij}\hat{p}_{ijl}},\\ \tilde{\pi}_l &= \frac{\sum_{ij}\hat{p}_{ijl}}{MN},\ l=1,\ldots,L.\end{aligned}$$

- Iterate between the Expectation Step and the Maximization Step until convergence.

A.4 The Value of the Likelihood Function

We derive the value of the likelihood function given in Equation (3). This is used to evaluate relative model performance.

Under the model assumptions, the overall likelihood function can be decomposed as

$$\begin{aligned}L(\mathcal{G}|\mathcal{R}) &\equiv f(\mathcal{R}|\theta,\mathcal{B},\Sigma), &(17)\\ &= \prod_{i=1}^{N}\int f(R_i|a_i,\mathcal{G})f(a_i|\mathcal{G})da_i, &(18)\end{aligned}$$

where $\mathcal{G}$ is the model MLE. Therefore, to obtain the overall likelihood, we only need to calculate the component likelihood, that is, $\int f(R_i|a_i,\mathcal{G})f(a_i|\mathcal{G})$. Under the model assumptions, the integrand of the component likelihood can be written as:

$$
\begin{aligned}
& f(R_i|a_i,\mathcal{G})f(a_i|\mathcal{G}) \\
= & \prod_{t=1}^{T}(2\pi\sigma_i^2)^{-1/2}\exp[-\frac{(r_{it}-\alpha_i-\beta_i'f_t)^2}{2\sigma_i^2}]\times\sum_{l=1}^{L}\pi_l(2\pi\sigma_l^2)^{-1/2}\exp[-\frac{(a_i-\mu_l)^2}{2\sigma_l^2}], \\
= & (2\pi\sigma_i^2)^{-T/2}\sum_{l=1}^{L}\pi_l(2\pi\sigma_l^2)^{-1/2}\exp[-\frac{\sum_{i=1}^{T}(r_{it}-\alpha_i-\beta_i'f_t)^2}{2\sigma_i^2}-\frac{(\alpha_i-\mu_l)^2}{2\sigma_l^2}], \\
= & (2\pi\sigma_i^2)^{-T/2}\sum_{l=1}^{L}\pi_l(2\pi\sigma_l^2)^{-1/2} \\
& \times\exp\{-\frac{(\sigma_l^2+\sigma_i^2/T)}{2\sigma_l^2(\sigma_i^2/T)}[\alpha_i-\frac{\frac{\sum_{t=1}^{T}(r_{it}-\beta_i'f_t)}{T}\sigma_l^2+\mu_l\frac{\sigma_i^2}{T}}{\sigma_l^2+\sigma_i^2/T}]^2 \\
& +\frac{(\frac{\sum_{t=1}^{T}(r_{it}-\beta_i'f_t)}{T}\sigma_l^2+\mu_l\frac{\sigma_i^2}{T})^2}{2(\sigma_l^2+\sigma_i^2/T)\sigma_l^2(\sigma_i^2/T)}-\frac{(\frac{\sum_{t=1}^{T}(r_{it}-\beta_i'f_t)^2}{T}\sigma_l^2+\mu_l^2\frac{\sigma_i^2}{T})}{2\sigma_l^2(\sigma_i^2/T)}\}, \\
= & (2\pi\sigma_i^2)^{-T/2}\sum_{l=1}^{L}\pi_l(2\pi\sigma_l^2)^{-1/2}\times\sqrt{2\pi(\sigma_i^2/T)\sigma_l^2/(\sigma_l^2+\sigma_i^2/T)} \\
& \times\underbrace{\frac{1}{\sqrt{2\pi(\sigma_i^2/T)\sigma_l^2/(\sigma_l^2+\sigma_i^2/T)}}\exp\{-\frac{(\sigma_l^2+\sigma_i^2/T)}{2\sigma_l^2(\sigma_i^2/T)}[\alpha_i-\frac{\frac{\sum_{t=1}^{T}(r_{it}-\beta_i'f_t)}{T}\sigma_l^2+\mu_l\frac{\sigma_i^2}{T}}{\sigma_l^2+\sigma_i^2/T}]^2\}}_{\phi(\alpha_i;\mu_{0i},\sigma_{0i}^2)} \\
& \times\exp\{\frac{(\frac{\sum_{t=1}^{T}(r_{it}-\beta_i'f_t)}{T}\sigma_l^2+\mu_l\frac{\sigma_i^2}{T})^2}{2(\sigma_l^2+\sigma_i^2/T)\sigma_l^2(\sigma_i^2/T)}-\frac{(\frac{\sum_{t=1}^{T}(r_{it}-\beta_i'f_t)^2}{T}\sigma_l^2+\mu_l^2\frac{\sigma_i^2}{T})}{2\sigma_l^2(\sigma_i^2/T)}\},
\end{aligned}
$$

where $\phi(\alpha_i;\mu_{0i},\sigma_{0i}^2)$ is the density function for a normal distribution parameterized by:

$$
\begin{aligned}
\mu_{0i} &= \frac{\frac{\sum_{t=1}^{T}(r_{it}-\beta_i'f_t)}{T}\sigma_l^2+\mu_l\frac{\sigma_i^2}{T}}{\sigma_l^2+\sigma_i^2/T}, \\
\sigma_{0i}^2 &= (\sigma_i^2/T)\sigma_l^2/(\sigma_l^2+\sigma_i^2/T).
\end{aligned}
$$

Therefore, by integrating over a_i, the part involving the normal density becomes one, and we have:

$$
\begin{aligned}
\int f(R_i|a_i,\mathcal{G})f(a_i|\mathcal{G})da_i &= (2\pi\sigma_i^2)^{-T/2}\sum_{l=1}^{L}\pi_l\sqrt{(\sigma_i^2/T)/(\sigma_l^2+\sigma_i^2/T)} \\
& \times\exp\{\frac{(\frac{\sum_{t=1}^{T}(r_{it}-\beta_i'f_t)}{T}\sigma_l^2+\mu_l\frac{\sigma_i^2}{T})^2}{2(\sigma_l^2+\sigma_i^2/T)\sigma_l^2(\sigma_i^2/T)}-\frac{(\frac{\sum_{t=1}^{T}(r_{it}-\beta_i'f_t)^2}{T}\sigma_l^2+\mu_l^2\frac{\sigma_i^2}{T})}{2\sigma_l^2(\sigma_i^2/T)}\}.
\end{aligned}
$$

Define

$$\begin{aligned}
\hat{\alpha}_i &= \frac{\sum_{t=1}^{T}(r_{it}-\beta_i' f_t)}{T}, \\
\widehat{\alpha_i^2} &= \frac{\sum_{t=1}^{T}(r_{it}-\beta_i' f_t)^2}{T}, \\
w_{l,i}^c &= \frac{\sigma_l^2}{\sigma_l^2+\sigma_i^2/T}, \\
w_{l,i}^t &= 1-w_{l,i}^c,
\end{aligned}$$

then the component likelihood can be written as

$$\begin{aligned}
\int f(R_i|a_i,\mathcal{G})f(a_i|\mathcal{G})da_i &= (2\pi\sigma_i^2)^{-T/2}\sum_{l=1}^{L}\pi_l\sqrt{w_{l,i}^t} \\
&\quad \times \exp\{\frac{(\hat{\alpha}_i w_{l,i}^c+\mu_l w_{l,i}^t)^2-(\widehat{\alpha_i^2}w_{l,i}^c+\mu_l^2 w_{l,i}^t)}{2[1/(1/\sigma_l^2+1/(\sigma_i^2/T))]}\}.
\end{aligned}$$

The overall likelihood can be calculated as the product of the component likelihoods of the cross-section of funds, as given in Equation (18).

B Multi-component GMD: Identifiability and Interpretability

The two-component model can be easily generalized to multi-component models. For a general L-component GMD, we order the means of its component distributions in ascending order (i.e., $\mu_1 < \mu_2 < \cdots < \mu_L$) and parameterize the probabilities of drawing from each component distribution as

$$\pi = (\pi_1, \pi_2, \ldots, \pi_L)', \quad \sum_{l=1}^{L}\pi_l = 1.$$

With enough components in the model, the GMD is able to approximate every density with arbitrary accuracy, the fact of which partly explains its popularity. However, the model becomes more difficult to identify when the number of components gets large.[45] Therefore, between two models that produce similar likelihood values, we prefer the parsimonious model. We rely on our simulation framework to perform formal hypothesis testing on the candidate models and to select the best model.[46]

[45] See, for example, Figueiredo and Jain (2002) for a discussion on the identifiability problem for a GMD and a potential solution.

[46] Another benefit in using the GMD is that it reduces the computational burden for the estimation of our model. In particular, when the components in $\mathcal{A}$ follow a GMD and the returns $\mathcal{R}$ follow a

The idea of using a mixture distribution to model the cross-section of fund alphas has also been explored by the recent literature on performance evaluation—for example, Chen, Cliff, and Zhao (2015). However, we offer a new approach that takes the various sources of estimation uncertainty into account.

Alternatively, we can think of Ψ as a parametric density to approximate the distribution of the population of fund alphas. The GMD is a flexible and widely used parametric family to approximate unknown densities. As in most density estimation problems, we are facing a trade-off between accuracy and overfitting. In our application, we pay special attention to the overfitting issue. In particular, we perform a simulation-based model selection procedure to choose a parsimonious model. This allows us to use the simplest structure — provided that it adequately models the alpha distribution — to summarize the alpha population. This also makes it easier to interpret the composition of the alpha population.

To think about the identification of Ψ in our model, we first focus on an extreme case. Suppose we have an infinitely long time series for each fund so that there is no estimation uncertainty in alpha. In this case, our model will force Ψ to approximate the cross-section of "true" alphas. Suppose the left tail of the alpha distribution is very different from the right tail. The single-component GMD will fail to capture this asymmetry.[47] A two-component GMD may be a better candidate. Intuitively, we can first fit a normal distribution for the alpha observations that fall below a certain threshold and another normal distribution for the alpha observations that fall above a certain threshold (these two thresholds are not necessarily equal). We then mix these two distributions in a way that the mixed distribution approximates the middle part of the alpha distribution well—that is, the alpha distribution that covers the non-extreme alphas.

In practice, we have a finite return time series. This introduces estimation uncertainty in both the alphas and the other OLS parameters. As a result, instead of fitting the cross-section of "true" alphas, our method tries to fit the cross-section of the distributions of the alphas, each distribution corresponding to the estimation problem of the alpha of an individual fund and capturing estimation risk. However, our previous discussion on the identification of Ψ when "true" alphas are available is still valid. In particular, the parameters in Ψ are identified by capturing the departure of the alpha distribution from a single normal distribution, only that this time the alpha distribution is no longer the distribution of "true" alphas but a mixed distribution of the estimated distributions of the alphas.

normal distribution conditional on $\mathcal{A}$, we show in Appendix A that the conditional distribution of the components in $\mathcal{A}$ given $\mathcal{R}$ is also a GMD. This makes it easy for us to simulate from the conditional distribution of $\mathcal{A}$ given $\mathcal{R}$, which is the key step for the implementation of the EM algorithm that we use to estimate our model.

[47] Fama and French (2010) find that the left tail of the alpha distribution is indeed more dispersed than the right tail, consistent with our findings when we apply our model to mutual funds.

More rigorously, the parameters in Ψ can be shown to be identified through high-order moments of the alpha population. For example, for a two-component GMD, its five parameters can be estimated by matching the first five sample moments of the data with the corresponding moments of the model.[48] Despite its intuitive appeal, the moments-based approach cannot weight different moments efficiently. Our likelihood-based approach is able to achieve estimation efficiency. In our simulation study, where we experiment with a two-component GMD, the model parameters seem to be well identified and accurately estimated.

C Related Literature

Our paper is also related to the literature on latent factor models and the EM algorithm.[49] Standard EM algorithms apply to situations where we need to fit a mixture distribution (e.g., a GMD) to the data. Since we do not know which component distribution of the mixture model that an observation falls into, the EM algorithm provides an efficient way to sequentially classify observations into the components and estimate the density function for each component distribution. In our application, importantly, we do not observe fund alphas. They are defined through the assumed return dynamics that involve unknown parameters such as factor loadings and residual standard deviations for each individual fund. As such, our innovation is to embed the EM framework into a panel regression model that allows heterogeneous regression coefficients in the cross-section. We analytically derive key formulas that allow us to implement the EM algorithm and illustrate its performance through simulations.

Bayesian methods have also been applied to performance evaluation. Both Bayesian methods and our approach imply shrinkage. However, while we explicitly estimate the shrinkage target — the underlying alpha population, some Bayesian methods imply shrinkage toward a prespecified target. For example, Baks, Metrick, and Wachter (2001) use informative priors to show how prior beliefs about investment opportunities affect people's investment decisions. Jones and Shanken (2005) use several intuitive priors, including both informative and non-informative priors, to summarize information in the cross-section. When diffuse priors are used, Bayesian methods arguably can minimize the impact of the prior specification and let the data speak. However, the choice of a particular distributional family (e.g., normal vs. non-normal distributions) for the prior still has a substantial impact on the posterior inference, as we show in simulations. While such a decision is likely to be an issue for any type of parametric inference, we try to minimize the impact of this choice by using a more flexible distributional family than what is used in Jones and Shanken (2005). Our extension is important given the documented asymmetric tail behavior of the alpha population (see, e.g., Fama and French 2010) that is crucial to identify extreme

[48] See Cohen (1967) and Day (1969) for the derivation of a two-component GMD based on the method of moments approach.

[49] See Bilmes (1998), Dempster, Laird, and Rubin (1977), McLachlan and Krishnan (2007).

performers, as well as the recent attempt to classify funds into broad performance groups (Barras, Scaillet, and Wermers 2010, Chen, Cliff, and Zhao 2015, Ferson and Chen 2015, Kosowski et al. 2006).

While Bayesian methods and our approach share some common features (e.g., shrinkage), we provide some simulation results to address the Bayesian critique that one can generate the frequentist estimate of the alpha distribution by specifying the "correct" prior. In particular, we choose a few of the prior specifications in Jones and Shanken (2005) and show that the associated Bayesian estimates are much different from our estimates, both for the alpha population and for individual fund alphas. Since our estimate for the underlying alpha distribution is unbiased (as we show in simulations), Bayesian methods may lead to biased inference on the alpha population, echoing the findings in Busse and Irvine (2006) that the prior specification greatly affects the predictive accuracy of Bayesian alphas. Given that our goal is to have an objective assessment of the underlying alpha population and, through which, to evaluate individual fund performance, our framework at the very least offers a competing approach to Bayesian performance evaluation.

Our approach features the use of a mixture distribution to model the underlying alpha population. This gives us the flexibility to capture the non-normal distribution of fund alphas, as emphasized by recent findings of the literature.[50] With Bayesian methods, such flexibility is challenging; therefore, conjugate priors are imposed for analytical tractability. As shown in Verbeke and Lesaffre (1996), the population parameters for random effects may be badly estimated under the normality assumption in a random effects model. We confirm this in our simulation study: the Jones and Shanken (2005) specification leads to biased inference on the alpha population when the alpha distribution features non-normality. Sastry (2013) extends Jones and Shanken (2005) by proposing a Bayesian approach that incorporates non-normal priors and shows improvement over Jones and Shanken (2005) in capturing funds with extreme alphas. However, since funds with extreme alphas are infrequently observed, seemingly non-informative priors on key parameters of the non-normal prior often yield inconsistent estimates of the model, as shown in Ishwaran and Zarepour (2012).[51] In addition, by imposing a GMD with a prespecified number of component distributions, we are not able to evaluate the incremental contribution of alternative

[50] See, e.g., Barras, Scaillet, and Wermers (2010), Chen, Cliff, and Zhao (2015), Ferson and Chen (2015), and Kosowski et al. (2006). The non-normality of the distribution of alphas should be more pronounced for hedge fund and venture capital returns, making our framework an appealing candidate for performance evaluation for these alternative investment vehicles.

[51] In particular, Ishwaran and Zarepour (2012) show that naive use of the non-informative Dirichlet prior for the drawing probabilities of the mixture model leads to inconsistent estimates of the mixture density. Besides the issue with the prior on the mixture model, Sastry (2013) also imposes diffuse but proper priors (i.e., conjugate priors with large variances) on the means and variances of the mixture model, as well as on the cross-section of factor loadings. However, our simulation experience with Jones and Shanken (2005) is that one needs diffuse and improper priors on factor loadings in order to achieve consistent parameter estimates when the mixture model is composed of a single normal component. Hence, we do not expect Sastry (2013) to be able to generate consistent parameter estimates for the cross-sectional distribution of alphas. See Kass and Wasserman (1996) for a further

GMDs statistically, which may lead to biased inference of the alpha population. Andrikogiannopoulou and Papakonstantinou (2016) further generalize Sastry (2013) to model each tail of the alpha distribution as following a separate normal-mixture distribution and estimate the model using Bayesian methods. However, using likelihood ratio tests, we show that such extensions are unlikely to be necessary and may lead to model overfitting. Overall, our frequentist framework provides estimates that are not driven by prior specifications and are shown to be consistent in simulations. At the same time, we rely on the likelihood ratio statistic to test for model adequacy, overcoming the dilemma of being forced to choose a prior distribution in the Bayesian framework.

D Our Extension of the EM Algorithm

More insight can be gained into the EM algorithm by specifying the parametric distribution Ψ. In *Step II*, assuming a Gaussian Mixture Distribution, Appendix A shows that the conditional distribution of $\mathcal{A}$ given the current parameter values (denoted as $\hat{\mathcal{G}}$) and $\mathcal{R}$ can be characterized as the distribution for N independent variables, with the i-th variable α_i following a fund-specific GMD that is parameterized by $\tilde{\theta}_i = (\{\tilde{\pi}_{i,l}\}_{l=1}^L, \{\tilde{\mu}_{i,l}\}_{l=1}^L, \{\tilde{\sigma}_{i,l}^2\}_{l=1}^L)$:

$$\tilde{\mu}_{i,l} = (\frac{\hat{\sigma}_l^2}{\hat{\sigma}_l^2 + \hat{\sigma}_i^2/T})\bar{a}_i + (\frac{\hat{\sigma}_i^2/T}{\hat{\sigma}_l^2 + \hat{\sigma}_i^2/T})\hat{\mu}_l, \tag{19}$$

$$\tilde{\sigma}_{i,l}^2 = \frac{1}{1/\hat{\sigma}_l^2 + 1/(\hat{\sigma}_i^2/T)}, \tag{20}$$

$$\tilde{\pi}_{i,l} = \frac{\hat{\pi}_l \phi(\bar{a}_i - \hat{\mu}_l, \hat{\sigma}_l^2 + \hat{\sigma}_i^2/T)}{\sum_{l=1}^L \hat{\pi}_l \phi(\bar{a}_i - \hat{\mu}_l, \hat{\sigma}_l^2 + \hat{\sigma}_i^2/T)}, \ l = 1, 2, \ldots, L, \tag{21}$$

where

$$\bar{a}_i \equiv \sum_{t=1}^T (r_{it} - \hat{\beta}_i' f_t)/T,$$

and $\phi(\mu, \sigma^2)$ is the density of the normal distribution $\mathcal{N}(0, \sigma^2)$ evaluated at μ.

We can think of $\bar{a}_i$ as the fitted alpha when β_i is fixed at $\hat{\beta}_i$. It would be the OLS estimate of alpha if $\hat{\beta}_i$ were the OLS estimate of β_i. The variance of the time-series residuals is fixed at $\hat{\sigma}_i^2$. Taken together, $\bar{a}_i$ and $\hat{\sigma}_i^2/T$ can be interpreted as the alpha estimate and its variance based on time-series information. On the other hand, $\hat{\theta}_i = (\{\hat{\pi}_{i,l}\}_{l=1}^L, \{\hat{\mu}_{i,l}\}_{l=1}^L, \{\hat{\sigma}_{i,l}^2\}_{l=1}^L)$ is the current parameter vector governing the GMD for the cross-sectional distribution of the alphas. Therefore, Equations (19), (20), and (21) update our estimates of the alphas by combining time-series and cross-sectional information.

discussion on how small samples (i.e., time series) may lead to inconsistent estimation of individual fund regression coefficients under different prior specifications (including diffuse priors).

We start with an L-component GMD specification for the alpha population. The updated alpha distribution for each individual fund is also a GMD with the same number of components. However, the parameters that govern the GMD will be different across funds. For each of the L component distributions for the fund-specific GMD, the mean (i.e., $\tilde{\mu}_{i,l}$) is a weighted average of the fitted time-series alpha and the original mean for the GMD, the variance (i.e., $\tilde{\sigma}^2_{i,l}$) is the harmonic average of the time-series variance and the original variance for the GMD, and the drawing probability (i.e., $\tilde{\pi}_{i,l}$) weights the original probability by $\phi(\bar{\mu}_i - \hat{\mu}_l, \hat{\sigma}^2_l + \hat{\sigma}^2_i/T)$, which depends on the distance between $\bar{\mu}_i$ and $\hat{\mu}_l$ (i.e., $|\bar{\mu}_i - \hat{\mu}_l|$) and the average of the variances $\hat{\sigma}^2_l + \hat{\sigma}^2_i/T$.

Holding everything else constant, a lower time-series variance (i.e., $\hat{\sigma}^2_i/T$) pulls both the updated mean and variance closer to their time-series estimates, thereby overweighing time-series information relative to cross-sectional information. On the other hand, a smaller distance between $\bar{\mu}_i$ and $\hat{\mu}_l$ implies a higher drawing probability (i.e., $\tilde{\pi}_{i,l}$), which means that compared with the original GMD, we are now more likely to draw from the component distribution that has a mean that is closer to $\bar{\mu}_i$. Hence, we revise our estimate of the cross-sectional distribution based on time-series information. The expressions in Equations (19), (20), and (21) bear intuitive interpretations as to how we update the alpha estimates based on both time-series and cross-sectional information. This synthesis of information is important as it allows us to obtain the most informative estimate of the $\mathcal{A}$ distribution, which is then used to evaluate the likelihood function, as in *Step II* of the EM algorithm. It also distinguishes our method from existing approaches that rely on only one source of information, either cross-sectional or time series.

Another way to interpret the formulas in Equations (19)-(21) is to consider the extreme case and assume that we have a single-component GMD (that is, $L = 1$), and moreover, its mean is zero (that is, $\hat{\mu}_1 = 0$). In this case, we link the t-statistic of fund i's alpha (defined as $\tilde{\mu}_i/\tilde{\sigma}_i$) with its OLS t-statistic (defined as $\bar{a}_i/\sqrt{\hat{\sigma}^2_i/T}$) through:

$$\frac{\tilde{\mu}_i}{\tilde{\sigma}_i} = \frac{\bar{a}_i}{\sqrt{\hat{\sigma}^2_i/T}} \times \sqrt{\frac{\hat{\sigma}^2_1}{\hat{\sigma}^2_1 + \hat{\sigma}^2_i/T}}. \tag{22}$$

Notice that $\sqrt{\frac{\hat{\sigma}^2_1}{\hat{\sigma}^2_1+\hat{\sigma}^2_i/T}} < 1$, and the larger the time-series variance (that is, $\hat{\sigma}^2_i/T$) is relative to the cross-sectional variance (that is, $\hat{\sigma}^2_l$), the smaller this number becomes. Therefore, when the average alpha is zero in the population, we discount the OLS t-statistic with a discount factor that equals $\sqrt{\frac{\hat{\sigma}^2_1}{\hat{\sigma}^2_1+\hat{\sigma}^2_i/T}}$. More time-series uncertainty results in a harsher discount.

The idea of discounting the OLS t-statistic is consistent with the idea of multiple testing adjustment, which has recently gained attention in both performance

evaluation and asset pricing in general.[52] However, the mechanism in our model to deflate t-statistics is different from standard multiple testing approaches. Our model, by treating the alpha of an investment fund as random, takes into account the cross-sectional uncertainty in alpha from a population perspective. Multiple testing methods, by treating the alpha as a fund-specific variable (that is, a fixed effect), adjust t-statistics by having a more stringent Type I error threshold. Despite the methodological difference, these two fundamentally different approaches arrive at the same conclusion — we need to apply a "haircut" to the individual t-statistics of fund alphas.

In *Step III*, we update our parameter estimates based on the conditional distribution of the alphas. We first update the OLS parameters except for the regression intercepts, and then update θ — the parameter vector that governs the alpha population.

For the update of the OLS parameters (see Appendix A), we derive analytical expressions for the MLE of β_i and σ_i^2. In particular, let $m(\alpha_i) = E_{\mathcal{A}|\mathcal{R},\mathcal{G}^{(k)}}(\alpha_i)$ and $var(\alpha_i) = Var_{\mathcal{A}|\mathcal{R},\mathcal{G}^{(k)}}(\alpha_i)$ be the conditional mean and variance of α_i. The MLE of β_i can be found as the regression coefficients obtained by projecting the return time series (i.e., $\{r_{i,t}\}_{t=1}^T$) onto the factor time series (i.e., $\{f_t\}_{t=1}^T$), fixing the regression intercept at $m(\alpha_i)$. Therefore, the MLE of β_i in our model differs from the usual OLS estimate in that the regression intercept is forced to equal $m(\alpha_i)$, the mean of α_i given our current knowledge about the alpha population (i.e., $\mathcal{A}|\mathcal{R},\mathcal{G}^{(k)}$).

The MLE of σ_i^2 can be found by fixing β_i at its MLE (i.e., $\breve{\beta}_i$). In particular, define

$$\overline{\varepsilon_i^2} \equiv \frac{1}{T}\sum_{t=1}(r_{it} - \breve{\beta}_i' f_t - m(\alpha_i))^2, \tag{23}$$

as the fitted residual mean squared error. Then the MLE of σ_i^2 is given by

$$\breve{\sigma}_i^2 = \overline{\varepsilon_i^2} + var(\alpha_i). \tag{24}$$

Notice that if we use $(\sigma_i^2)^{MLE}$ to denote the MLE of the residual variance for the standard regression model that projects the time series of excess returns (i.e., $\{r_{i,t}\}_{t=1}^T$) onto $\{f_t\}_{t=1}^T$, then we must have:

$$\overline{\varepsilon_i^2} \geq (\sigma_i^2)^{MLE},$$

since the standard regression model seeks to minimize the sum of squared residuals without any parameter constraints. Therefore, two effects make the MLE of the residual variance (i.e., $\breve{\sigma}_i^2$) in our model larger than the standard-model MLE (i.e., $(\sigma_i^2)^{MLE}$). First, $\overline{\varepsilon_i^2}$ is no less than $(\sigma_i^2)^{MLE}$ because we are considering a regression model whose intercept is fixed at $m(\alpha_i)$. Second, there is uncertainty in α_i as captured

[52]For recent finance applications of multiple hypothesis testing in asset pricing, see Barras, Scaillet, and Wermers (2010), Fama and French (2010), Ferson and Chen (2015), Harvey and Liu (2017b), and Harvey, Liu, and Zhu (2016).

by $var(\alpha_i)$, which depends on the parameters given in Equations (19), (20), and (21) of the updated GMD (see Appendix A). Since, as discussed previously, the updated GMD takes both time-series and cross-sectional information into account, $var(\alpha_i)$ also incorporates information about the cross-sectional dispersion of the alphas.

These two effects implied by our model make intuitive sense as they allow us to learn from both the mean and the variance of the alpha population. Additionally, the learning effect is more pronounced in small samples and will go away when we have a long enough time series of returns. This can be easily seen from the formulas of our algorithm. When T goes to infinity and based on Equations (19)-(21), the alpha distribution collapses to the point mass at $\bar{a}_i$, which is the estimate based on time-series information only. This implies that $m(\alpha_i) = \bar{a}_i$ and $var(\alpha_i) = 0$. As a result, our MLEs of β_i and σ_i converge to their OLS estimates. The fact that our method implies differential adjustment to the alpha estimate between small and large samples makes it an attractive method for performance evaluation when a large fraction of funds have short time series.

For the update of θ, we seek the parameter vector θ of a GMD that best describes the alpha distribution. The optimization problem we are solving is:

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{N} \frac{1}{M} \sum_{m=1}^{M} \log f(\alpha_i^m | \theta), \tag{25}$$

where $\{\alpha_i^m\}_{m=1}^M$ are randomly generated samples from the conditional distribution of α_i given $\mathcal{R}$ and $\mathcal{G}^{(k)}$. If there were just one fund in the cross-section, then $\hat{\theta}$ will approximately equal the parameters that govern the GMD for a single fund that are given in Equations (19)-(21). With multiple funds in the cross-section, we have multiple GMDs, each governing the alpha distribution of a particular fund. Our method tries to find the best θ that describes the cross-section of GMDs, which can be viewed as a mixture distribution that chooses a fund with equal probability from the cross-section of funds and, conditional on a fund being chosen, draws an alpha from the fund's GMD. Notice that this mixture distribution in our model is very different than the alpha distribution in the equation-by-equation OLS model, where it is simply the cross-section of fitted alphas. Our method allows us to capture the estimation risk of each fund's alpha and leads to a more informed estimate of the alpha distribution at the population level.

E Further Comparisons with CCZ and JS

E.1 Comparison with CCZ

Theoretically, CCZ ignores the updated OLS parameters (i.e., $\mathcal{B}$ and $\mathbf{\Sigma}$), which is part of *Step III* in our algorithm. In particular, as we show in Appendix A, Section 2,

fund-specific factor loadings and residual standard deviations should be reestimated through the MLE once we have updated information on individual fund alphas. This information is different from the equation-by-equation OLS densities for individual fund alphas as it combines individual funds' time-series information with information about the alpha population, which is given in *Step II*. The update of OLS parameters is important not only because it generates better estimates for fund-specific OLS parameters, but also because these parameters are later fed into the next iteration of the algorithm, leading to a better estimate of the cross-sectional alpha distribution. In our online Appendix IA, we provide evidence on the improvement of our model over CCZ in estimating fund-specific OLS parameters. We also experiment with modified versions of CCZ by partially updating either the factor loadings or the residual standard deviations. We show that the updating of both sets of parameters is an important reason that our model outperforms CCZ.

To better illustrate the difference between the two methods, we look at an example. To simplify, we assume that factor loadings are known for sure and we strip them out to focus on realized alphas (i.e., true alpha plus noise). Suppose a fund has an excess return sequence of [2%, 3%, 3%, 2%] over four years. In the CCZ model, the fund is generating an OLS alpha of 2.5% and residual standard deviation is estimated to be 0.5%.[53] So the fund is significantly outperforming and, by feeding the OLS t-statistic for the fund to the CCZ model, the authors are forcing the cross-sectional alpha distribution to explain the extremely good performance of the fund.

In our model, the story is different. Suppose we have a good knowledge of the cross-sectional alpha distribution and we believe that it is unlikely that any fund has an alpha of 2.5%. Due to shrinkage, we will reduce the alpha estimate for this particular fund to, say, 1% (i.e., we shrink its alpha by 1.5%). This new alpha estimate forces us to reestimate the residual standard deviation of the fund as the noise sequence now becomes [1%, 2%, 2%, 1%] (=$[2\% - 1\%, 3\% - 1\%, 3\% - 1\%, 2\% - 1\%]$), so the residual standard deviation estimate becomes 1.58%, which is much higher than CCZ's estimate.

Notice that in our framework, individual fund return residuals do not have to sum up to zero, which leads to a different decomposition of fund returns into skill (i.e., true alpha) and luck (i.e., residuals) than what people usually do by following the equation-by-equation estimation approach. In our example, if we strongly believe that the fund should have an alpha of 1%, we will conclude that the fund has experienced four years of luck, which is very likely given the short return history of the fund. However, in the traditional equation-by-equation estimation approach, such a conclusion is not possible as the luck components have to sum up to zero (i.e., the luck is reflected in the risk-adjusted returns). We believe that our approach is more intuitive than the traditional approach, especially when many funds have a short return history.

[53] We use the MLE. The OLS estimate, due to the adjustment for finite sample bias of the MLE, is 0.58%. Their small difference in inconsequential for our interpretation of the example.

Viewing from the perspective of decomposing returns into luck and skill, the CCZ framework is inconsistent. In their first-stage estimation, they run equation-by-equation OLS and therefore force the luck components for each fund to sum up to zero. In the second-stage estimation, they fit the cross-sectional alpha distribution, through which they can update the individual alpha estimate for each fund. However, under this updated alpha estimate, the luck components for each fund no longer sum up to zero. This contradiction is inherent in their two-stage estimation approach and has a significant bearing on the interpretation of the estimation outcome. Our model, by simultaneously solving for the parameters that govern the alpha population and fund-specific OLS parameters (i.e., $\mathcal{B}$ and $\mathbf{\Sigma}$), provides a coherent framework to think about alpha estimation, as well as decomposing fund performance into skill and luck.

From a methodological perspective, especially in comparison with the literature on multiple shrinkage that we mentioned before, our model shows that when we shrink the set of parameters of interest (i.e., parameters that govern the alpha population in our application), we indirectly introduce interactions among auxiliary parameters (i.e., fund-specific factor loadings and residual standard deviations) that are not of primary interest to us. For example, the inference on a particular fund's factor loadings impacts our inference on the alpha population, which in turn affects the inference on factor loadings for another fund. Our model captures these interactions through the joint estimation of the model MLE. Alternatively, one could directly introduce shrinkage to auxiliary parameters, which is straightforward to achieve in our framework (see Harvey and Liu 2017c). However, such an approach has the risk of misspecifying the distribution for auxiliary parameters, which may lead to biased inference on variables that are of primary interest to us. Hence, we focus on the shrinkage of fund alphas and do not directly model shrinkage on other OLS parameters in this paper.

E.2 Comparison with JS

Can a Bayesian framework that also features learning across funds, such as Jones and Shanken (JS; 2005), arrive at similar estimates to our framework? The answer is no, and we prove this by comparing JS with our model through simulations (detailed results are in the online Appendix IB).

For our initial comparison with JS, we implement our framework assuming a single normal distribution characterizes the alpha population.[54] In particular, we assume that the alpha distribution is characterized by the second component of the GMD in Table 1—that is, a normal distribution with a mean of -0.69% and a standard deviation of 0.59%. We follow JS to estimate the model and present the results in Table IB.1 of online Appendix IB. We find that JS overestimates the standard deviation of the alpha population by 15% (=(0.672-0.586)/0.586) under a diffuse

[54]We use the normal distribution for an "apples-to-apples" comparison with JS. Later we will compare the mixture of normals vs. JS.

prior (which is the more relevant prior from the perspective of model comparison). In contrast, NRA overestimates the standard deviation by 4% (=(0.610-0.586)/0.586). NRA also implies better estimates of different percentiles of the alpha population and fund alphas at the individual fund level. In our online Appendix IB, we provide further analysis and discussion on the relative performance between our model and JS.

When the alpha population is described by a mixture distribution that corresponds to the parameter configuration in Table 1, the difference between JS and the NRA model is more dramatic, as shown in Table 2 and 4. JS, by modeling alphas as random draws from a normal distribution, provides biased estimates of summary statistics for the alpha population.

NRA differs from JS by allowing shrinkage toward the mean of a particular group of funds, as opposed to the overall population mean as in JS. Its improvement over JS is akin to the improvement of multiple shrinkage (see, e.g., Karolyi 1993) over simple shrinkage (see, e.g., Vasicek 1973). Importantly, while multiple shrinkage uses pre-specified instrument variables to partition the data into different groups, our method relies on the likelihood function that only involves the return data to classify funds into different performance groups.

One may argue that a mixture distribution for the alpha population is not a fair assumption in terms of model comparison since JS assumes that alphas are drawn from a normal distribution. However, the main point of our paper, consistent with recent empirical findings on performance evaluation (Barras, Scaillet, and Wermers 2010, Ferson and Chen 2015, Kosowski et al. 2006), is to provide a flexible and parsimonious framework to model the alpha population, through which we can make better estimates of individual fund alphas. When the alpha population features salient departures from normality (as we tested to be the case in the next section), it is important to take these departures into account to sharpen our inference on fund alphas.

It is not our purpose to critique the Bayesian approach, as it does provide a powerful framework to achieve shrinkage. It also provides probabilistic analysis that is solely based on the given data. Our goal is to present a frequentist approach that allows a rich modeling of the alpha population, from which we can make better inference on fund alphas.

F Other Issues

F.1 Misspecification of the Factor Model

Inference on fund alphas both at the population and at the individual fund levels is contingent upon the benchmark model being used. For instance, for mutual funds

performance evaluation, suppose the true benchmark model is a five-factor model that includes the Fama and French (1993) and Carhart (1997) four factors. Then misspecifying the benchmark model as the four-factor model will likely lead to biased alpha estimates, both for the alpha population and for the individual funds.

The concern about model risk is to some extent alleviated by considering the noise-reduced alpha model. Using the aforementioned five-factor model example, suppose the fifth factor — the factor that is missing from the four-factor model — applies only to a small fraction of funds. By using a misspecified four-factor model, the equation-by-equation OLS will imply biased alpha estimates for this small fraction of funds. Under the noise-reduced alpha model, we are able to learn from the entire cross-section of funds, including those that are not exposed to the fifth factor. As a result, the bias in the alpha estimates for the small fraction of funds that are exposed to the fifth factor is likely to be lower under the noise-reduced alpha model than under the OLS model.

When the benchmark model is missing a factor that applies to the majority of funds, it is unlikely that any performance evaluation model will do well. One therefore needs to be cautious when trying to interpret the results of our paper. Our inference relies on a prespecified benchmark model for performance evaluation and could be sensitive to this choice. In our online Appendix IC, we report our model estimates under alternative specifications for the benchmark factor model.

Another possible misspecification of the factor model assumes a constant beta while the true beta is time-varying (see, e.g., Christopherson, Ferson, and Turner 1999). If fund-level characteristics and macroeconomic variables can be used as instruments to model time-varying betas, then the static factor model considered in our current paper would be missing factors that interact these instruments with the benchmark factors. On the other hand, data snooping bias and spurious regressions make it difficult to choose instruments that help enhance the inference on alphas, as shown in Ferson, Sarkissian, and Simin (2008). Hence, we focus on unconditional regressions in this paper. If strong and pre-determined instruments were available, our approach can easily be applied to allow for dynamic betas.

F.2 Sample Selection Bias

As with all approaches to performance evaluation, sample selection may bias our results. On the one hand, studies that condition on fund survival overestimate fund performance, see Brown et al. (1992), Carhart et al. (2002), Elton, Gruber, and Blake (1996). On the other hand, reverse-survivorship may understate fund performance, as shown in Linnainmaa (2013). In particular, Linnainmaa (2013) models fund survival as a function of past performance and estimate the underlying alpha distribution, which is modeled as a normal distribution, through simulated method of moments. We differ from Linnainmaa (2013) by using the GMD to explicitly model the tail

behavior of the alpha distribution and relying on the likelihood function to provide exact and efficient inference.

We believe bias will likely be smaller in our framework compared with the standard equation-by-equation OLS. For example, when there is reverse-survivorship bias, a skilled fund may drop out of sample after having a bad (unlucky) shock. This makes its in-sample alpha an understatement of its true population value. Hence, using the equation-by-equation OLS, if we take the average of the cross-section of fitted alphas, this average will underestimate the overall population mean if there is reverse-survivorship bias. Funds that have a shorter history and a higher level of idiosyncratic volatility are more likely to drop out after experiencing a bad shock. In our framework, the importance of these funds is downwardly weighted. We know their alpha estimates are more noisy, so we put less weight on them in terms of learning about the alpha population.

F.3 Noise-Reduced Alpha vs. Multiple Hypothesis Testing

By treating the alpha of an investment fund as random, our model takes into account the cross-sectional uncertainty in alpha from a population perspective and helps deflate the fund alpha and its t-statistic, thereby imposing a more conservative inference on the fund alpha. This is consistent with the idea of multiple testing that has been applied to performance evaluation (see, Barras, Scaillet, and Wermers 2010; Fama and French 2010; and Ferson and Chen 2015) and to asset pricing in general (see Harvey and Liu 2017b; and Harvey, Liu, and Zhu 2016). What is the connection between the two methods?

Suppose a researcher wants to test the effectiveness of a drug for all patients. The researcher divides the sample into a female group and a male group and separately tests the effectiveness of the drug. Since two tests have been tried, the chance of finding a significant result is higher than the case with a one-shot test. The researcher can apply a multiple testing adjustment to these two tests so that the overall error rate, however defined, is controlled at a prespecified level. However, it does not make sense to use the model in our paper since there are a limited number of gender types in the population (i.e., we do not have hundreds of gender types). It is not appropriate to view the means of the two groups — male and female — as coming from an underlying distribution as there are only two samples from this distribution.

The NRA model applies when it is plausible to view the objects in the cross-section as coming from a certain underlying population. For fund alphas, it makes sense to think that the alphas for different funds are not independent of each other

since there are limited investment opportunities in the financial market and funds compete with each other to generate alphas.[55]

Despite their similarities in discounting fund alphas and their t-statistics, the two models are fundamentally different. The multiple testing approach, and hypothesis testing in general, treats the fund alpha as a dichotomous variable (that is, zero vs. nonzero). Its objective function is also about controlling the probability or the fraction of false discoveries, that is, a zero alpha fund being incorrectly classified as a nonzero fund. On the other hand, the NRA model preserves the continuity of the alpha distribution. Its objective function is the goodness-of-fit of a parametric model to the data. While the fund-by-fund hypothesis testing framework is useful to roughly classify investment managers into different groups, the NRA model is designed to provide inference on the alpha population as well as refining inference about a particular fund.

Another advantage to the NRA approach is that OLS t-statistics are no longer sufficient statistics to rank the cross-section of funds. In contrast, the multiple testing approach always preserves the ranking of funds based on OLS t-statistics. For example, suppose Fund A has a more extreme positive OLS alpha estimate and at the same time a higher OLS t-statistic than Fund B. Then Fund A will always be regarded as more attractive than Fund B under multiple testing, regardless of the multiple testing methods we use. In our framework, due to learning across funds, the more extreme OLS alpha estimate of Fund A is pulled toward the population mean more than the less extreme OLS alpha estimate of Fund B. As a result, the overall relative attractiveness between Fund A and B might be reversed.

Our results suggest that there are more funds with positive alphas than what the literature on fund-by-fund hypothesis testing suggests. We think that this is also related to the difference in loss functions between the two methods. To estimate the fraction of positive-alpha funds, the usual approach follows a two-stage procedure. In the first stage, we adjust for multiple testing, controlling the false discovery rate at a prespecified level (e.g., 5%). In the second stage, we calculate the fraction of funds that survive the multiple testing p-value cutoff. Notice that the first-stage multiple testing cutoff, which is found by meeting the size of the test, will have a big impact on the second-stage estimation. Typically, multiple testing sets a tough cutoff in the first stage, which tends to lead to an underestimation of the fraction of positive-alpha funds in the second stage. In other words, in the usual approach, the stringent multiple testing cutoff in the first stage conflicts with the goal of obtaining an unbiased estimate of the fraction of positive-alpha funds in the second stage. In our framework, our goal is to directly estimate the density of the underlying alpha population. As we show in our simulation study, we are able to provide unbiased estimates of population statistics such as the fraction of positive-alpha funds.

[55]See French (2008) for a similar argument on the competitiveness of the investment funds industry. Also see the argument in Jones and Shanken (2005) for the population perspective on fund alphas.

F.4 Time-Varying Alphas

While our paper focuses on unconditional alphas, we can use fund-level characteristics as instruments to study conditional alphas. Jones and Mo (2016) show that a number of firm characteristics help forecast the cross-section of fund alphas. They also find that the performance of many of these characteristics in explaining fund alphas deteriorates through time. Our model can be easily extended to take into account the predictability and the variation in predictability of fund returns by using fund characteristics. Our framework allows one to make inference by drawing information from the entire cross-section, which can potentially improve the out-of-sample predictability of fund alphas. This is further explored in Harvey and Liu (2017c).

F.5 Modeling Information Ratios

Instead of alphas, we can modify our framework to model the distribution of other performance metrics such as information ratios.[56]

We use the information ratio as an example. In Step II of our algorithm, notice that the residual standard deviations are assumed to be known. Hence, by replacing alphas with information ratios, one can also obtain a GMD representation for the updated distribution for each fund's information ratio. In Step III, when we sample from the updated information ratio distribution for each fund, we can numerically solve the MLE for the regression coefficients and the residual standard deviation for each fund.[57] The other part of Step III for which we update the parameters for the population of information ratios still applies. Taken together, our framework applies fairly straightforwardly to the case of information ratios. If it were the case that information ratios rather than alphas are better represented as a GMD, then we may be able to make better inference on manager skill by directly modeling information ratios. We leave this question to future research.

[56] We thank the anonymous referee for leading us to think about these extensions of our model.

[57] Notice that, unlike for the case with alphas, analytical expressions for the regression coefficients and the residual standard deviation are not available since the idea of a concentrated likelihood function, which is used to derive analytical expressions for the regression parameters in the standard OLS, no longer applies. This is because, by sampling information ratios, there is an interaction term between the regression coefficients and the residual standard deviation that does not allow us to estimate the regression coefficients independently of the residual standard deviation. We have to solve the MLE numerically.

Detecting Repeatable Performance Online Appendix

Campbell R. Harvey

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

Yan Liu*

Texas A&M University, College Station, TX 77843 USA

Current version: October 13, 2017

* Current Version: October 13, 2017. First posted on SSRN: November 19, 2015. Send correspondence to: Campbell R. Harvey, Fuqua School of Business, Duke University, Durham, NC 27708. Phone: +1 919.660.7768, E-mail: cam.harvey@duke.edu . The full version of the paper is available at http://ssrn.com/abstract=2691658.

1 Summary

The on-line appendix is organized as follows.

Appendix IA provides further details on the simulation study. Appendix IA.1 describes the construction of our residual dependence model that mimics the cross-sectional dependence structure for the actual data. Appendix IA.2 reports additional results on the comparison among alternative models when residuals are correlated. Appendix IA.3 reports results on estimation risk for individual funds' factor loadings and residual standard deviations. Appendix IA.4 reports simulation results under a different parameter configuration.

Appendix IB compares our Noise Reduced Alpha (NRA) model with Jones and Shanken (2005).

Appendix IC shows estimation results of our model under alternative benchmark factor models.

Appendix ID provides details of our estimation procedure.

Appendix IE provides two motivating examples that highlight the intuition behind our approach.

Appendix IF details the difference between the usual EM model and our implementation with fund heterogeneity.

Appendix IG places our empirical findings in the context of the literature.

Appendix IH provides answers to some frequently asked questions.

IA Model Simulations

IA.1 Residual Correlations

To allow a more realistic assessment of our model and to provide robust standard errors, we propose a parametric model that to a large extent captures salient features of the residual correlation matrix of the actual data.

Ideally, a bootstrap procedure might be the most effective approach to control for cross-sectional dependency, as in Fama and French (2010). However, many funds have a limited number of overlapping months, reducing the usefulness of the bootstrap approach that jointly samples the cross-section. For our simulation study, since our goal is to compare model performances, we would like to fix the number of funds in the cross-section so that the only factor that causes the difference in model performance is the estimation framework itself.

We propose a simple parametric model to capture residual dependence. Suppose a balanced panel of residuals is a $T \times N$ matrix, where T is the number of time periods and N is the number of funds in the cross-section. At each point in time, let $resid_t$ be a row vector of residuals for N funds. We model $resid_t$ as:

$$\underbrace{resid_t}_{1\times N} = \underbrace{c_t}_{scalar}\underbrace{L_c}_{1\times N} + \underbrace{D_t}_{1\times N} \circ \underbrace{L_d}_{1\times N},$$

where c_t is a scalar common shock, L_c is the cross-sectional loadings on the common shock, D_t is cross-section of idiosyncratic shocks, L_d is the cross-sectional loadings on the idiosyncratic shocks, and '$\circ$' denotes Hadamand product. Both c_t and elements in D_t follow a standard normal distribution, and they are independent contemporaneously and across time. Loadings on the common shock (i.e., L_c) follow a normal distribution with mean μ_c and standard deviation σ_c. Loadings on idiosyncratic shocks (i.e., L_d) follow a Gamma distribution parameterized by γ_{d1} and γ_{d2}.

Our model succinctly captures cross-sectional dependency through the common shock c_t. The average cross-sectional loading on the common shock, captured by μ_c, controls the sign and magnitude of the average pairwise correlation among the residuals. Parameters in the Gamma distribution give us flexibility to capture the heterogeneity in the relative weight of the variance of the common shock and the variance of the idiosyncratic shock in the cross-section.

We calibrate our model to match key statistics of the collection of pairwise residual correlations in the actual data. Table IA.1.1 shows the results. Overall, our model provides a reasonably good approximation to the actual correlation structure in the data, especially for the left tail of the distribution of residual correlations. On the right tail, our model implies a somewhat higher level of correlation than the actual

data. To the extent that positive residual correlations result in larger standard errors for the model parameter estimates (as shown in Table 2 in the main paper), our calibrated model provides robust standard errors that control for a higher level of residual correlation than what the data suggests.

Using our calibrated model, for each simulation run, we generate a $T \times N$ ($= 336 \times 3619$) matrix of residuals. We next scale the residual time-series for each fund by its estimated residual standard deviation. This makes sure that the residual standard deviations for the simulated data are consistent with what we see from the actual data. At the same time, scaling does not affect the correlation structure. Finally, for each fund in our sample, we assign the corresponding simulated residuals to the time periods during which the fund exists. This makes sure that the months during which a fund exists in our simulated sample are the same as in the actual data.

Table IA.1.1: **Residual Correlations: Realized vs. Model Implied**

Summary statistics of pairwise correlations for factor model residuals. "Realized corr." look at the realized residual correlations for the data. "Model implied corr." look at the residual correlations corresponding to our parametric model that is parameterized by $(\mu_c, \sigma_c, \gamma_{d1}, \gamma_{d2})' = (0.9, 1.4, 1.5, 1.8)'$. "$p(i)$" reports the i-th percentile of the collection of pairwise correlations.

	Realized corr.	Model implied corr.
$p(10)$	-0.286	-0.265
$p(25)$	-0.087	-0.078
$p(50)$	0.060	0.056
$p(75)$	0.208	0.253
$p(90)$	0.352	0.519

IA.2 Model Comparisons When Residuals Are Correlated

We report the simulation results for different models when there is residual correlation in the cross-section. In particular, Table IA.2.1 shows the population statistics when the pairwise correlation is set at $\rho = 0.2$. Table IA.2.2 shows the population statistics when residual dependence mimics the dependence structure in the actual data, as modeled in on-line appendix IA.1. Table IA.2.3 shows the results on individual funds when the pairwise correlation is set at $\rho = 0.2$. Table IA.2.4 shows the results on individual funds when residual dependence mimics the dependence structure in the actual data, as modeled in on-line appendix IA.1.

Table IA.2.1: **A Simulation Study: Population Statistics,** $\rho = 0.2$

Population statistics based on the model estimates in a simulation study. We fix the model parameters at $\mathcal{G}^*$ (Table 1 of the main paper) and generate D sets of data sample. For each set of data sample, we estimate our model using the proposed noise reduced alpha model ("NRA"), the standard equation-by-equation OLS (OLS), and the model in Chen, Cliff and Zhao (CCZ). We then calculate summary statistics for the alpha population for both models based on the estimated model parameters. "Mean" is the mean of the alpha distribution. "Stdev." is the standard deviation of the alpha distribution. "Iqr." is the inter-quartile range of the alpha distribution. "*p10*" is the *10*th percentile of the alpha distribution. The other percentiles are similarly defined. "True" reports the population statistics based on the true model. "Estimate" reports the averaged estimate of the population statistics across the D sets of simulations. "RMSE" reports the square root of the mean squared estimation error, that is, $\sqrt{\sum_{d=1}^{D}(s_d - s)^2/D}$, where s is the true statistic and s_d is the estimated statistic based on the d-th simulated sample. Pairwise residual correlation is set at 0.2.

		NRA	OLS	CCZ
Mean(%)	Estimate	−1.121	−1.063	−1.089
(True = −1.136)	RMSE	0.428	0.552	0.440
Stdev.(%)	Estimate	1.125	3.877	1.415
(True = 1.187)	RMSE	0.099	2.901	0.252
Iqr.(%)	Estimate	1.122	3.254	1.312
(True = 1.144)	RMSE	0.179	2.123	0.261
p5(%)	Estimate	−3.580	−5.305	−3.941
(True = −3.700)	RMSE	0.562	1.701	0.577
p10(%)	Estimate	−2.701	−4.249	−2.937
(True = −2.832)	RMSE	0.666	1.516	0.666
p50(%)	Estimate	−0.851	−1.064	−0.845
(True = −0.860)	RMSE	0.373	0.552	0.354
p90(%)	Estimate	0.028	2.099	0.241
(True = 0.008)	RMSE	0.392	2.174	0.487
p95(%)	Estimate	0.279	3.117	0.656
(True = 0.284)	RMSE	0.426	2.899	0.660

Table IA.2.2: **A Simulation Study: Population Statistics, Data Dependence**

Population statistics based on the model estimates in a simulation study. We fix the model parameters at $\mathcal{G}^*$ (Table 1 of the main paper) and generate D sets of data sample. For each set of data sample, we estimate our model using the proposed noise reduced alpha model ("NRA"), the standard equation-by-equation OLS (OLS), and the model in Chen, Cliff and Zhao (CCZ). We then calculate summary statistics for the alpha population for both models based on the estimated model parameters. "Mean" is the mean of the alpha distribution. "Stdev." is the standard deviation of the alpha distribution. "Iqr." is the inter-quartile range of the alpha distribution. "*p10*" is the *10*th percentile of the alpha distribution. The other percentiles are similarly defined. "True" reports the population statistics based on the true model. "Estimate" reports the averaged estimate of the population statistics across the D sets of simulations. "RMSE" reports the square root of the mean squared estimation error, that is, $\sqrt{\sum_{d=1}^{D}(s_d - s)^2/D}$, where s is the true statistic and s_d is the estimated statistic based on the d-th simulated sample. Residual dependence mimics the dependence structure in the actual data and is modeled in on-line Appendix IA.1.

		NRA	OLS	CCZ
Mean(%)	Estimate	−1.115	−1.109	−1.116
(True = −1.136)	RMSE	0.310	0.375	0.315
Stdev.(%)	Estimate	1.163	3.951	1.450
(True = 1.187)	RMSE	0.190	2.932	0.341
Iqr.(%)	Estimate	1.167	3.311	1.348
(True = 1.144)	RMSE	0.337	2.192	0.390
p5(%)	Estimate	−3.543	−5.379	−4.073
(True = −3.700)	RMSE	0.482	1.761	0.633
p10(%)	Estimate	−2.702	−4.337	−3.050
(True = −2.832)	RMSE	0.529	1.588	0.591
p50(%)	Estimate	−0.872	−1.067	−0.865
(True = −0.860)	RMSE	0.283	0.437	0.268
p90(%)	Estimate	0.048	2.157	0.264
(True = 0.008)	RMSE	0.346	2.198	0.446
p95(%)	Estimate	0.312	3.155	0.677
(True = 0.284)	RMSE	0.398	2.915	0.599

Table IA.2.3: **A Simulation Study: Individual Funds,** $\rho = 0.2$

Summary statistics on model performance at the individual fund level. We fix the model parameters at $\mathcal{G}^*$ (Table 1 of the main paper) and generate D sets of data sample. For each set of data sample, we estimate our model using the proposed noise reduced alpha model (NRA), the standard equation-by-equation OLS (OLS), and the model in Chen, Cliff, and Zhao (CCZ, 2015). For NRA and CCZ, given the parameter estimates, we use equations (12)-(14) in the main paper to first construct the density forecast for each individual fund, and then obtain the point estimate and the confidence interval. For OLS, its point estimate is the estimate for the intercept, and its confidence interval is constructed using the point estimate and the standard error for the intercept. "Mean absolute deviation" is the averaged (across simulations) mean absolute distance between the estimated alpha and the true alpha for the cross-section of funds. "Stdev. of mean absolute deviation" is the averaged (across simulations) standard deviation of the absolute distance between the estimated alpha and the true alpha for the cross-section of funds. "Length, p" reports the averaged (across simulations) p-th percentile of the length of the 90% (or 95%) confidence intervals for the cross-section of funds. "Coverage probability" reports the averaged (across simulations) probability for the 90% (or 95%) confidence intervals to cover the true alpha values for the cross-section of funds. Other variables are similarly defined. Pairwise residual correlation is set at 0.2.

		NRA	OLS	CCZ
	Mean absolute deviation(%)	0.728	1.841	0.829
	Stdev. of mean absolute deviation(%)	0.690	3.198	0.746
90% confidence interval	Length, *p10*(%)	1.886	3.295	2.095
	Length, *p50*(%)	2.699	6.146	3.122
	Length, *p90*(%)	3.867	12.445	4.604
	Coverage probability	0.873	0.894	0.870
95% confidence interval	Length, *p10*(%)	2.328	3.926	2.593
	Length, *p50*(%)	3.401	7.323	3.951
	Length, *p90*(%)	4.574	14.827	5.480
	Coverage probability	0.931	0.945	0.930

Table IA.2.4: **A Simulation Study: Individual Funds, Data Dependence**

Summary statistics on model performance at the individual fund level. We fix the model parameters at $\mathcal{G}^*$ (Table 1 of the main paper) and generate D sets of data sample. For each set of data sample, we estimate our model using the proposed noise reduced alpha model (NRA), the standard equation-by-equation OLS (OLS), and the model in Chen, Cliff, and Zhao (CCZ, 2015). For NRA and CCZ, given the parameter estimates, we use equations (12)-(14) in the main paper to first construct the density forecast for each individual fund, and then obtain the point estimate and the confidence interval. For OLS, its point estimate is the estimate for the intercept, and its confidence interval is constructed using the point estimate and the standard error for the intercept. "Mean absolute deviation" is the averaged (across simulations) mean absolute distance between the estimated alpha and the true alpha for the cross-section of funds. "Stdev. of mean absolute deviation" is the averaged (across simulations) standard deviation of the absolute distance between the estimated alpha and the true alpha for the cross-section of funds. "Length, p" reports the averaged (across simulations) p-th percentile of the length of the 90% (or 95%) confidence intervals for the cross-section of funds. "Coverage probability" reports the averaged (across simulations) probability for the 90% (or 95%) confidence intervals to cover the true alpha values for the cross-section of funds. Other variables are similarly defined. Residual dependence mimics the dependence structure in the actual data and is modeled in on-line Appendix IA.1.

		NRA	OLS	CCZ
	Mean absolute deviation(%)	0.691	1.849	0.809
	Stdev. of mean absolute deviation(%)	0.687	3.257	0.744
90% confidence interval	Length, *p10*(%)	1.922	3.313	2.139
	Length, *p50*(%)	2.735	6.161	3.155
	Length, *p90*(%)	3.991	12.444	4.678
	Coverage probability	0.876	0.893	0.889
95% confidence interval	Length, *p10*(%)	2.360	3.948	2.642
	Length, *p50*(%)	3.456	7.341	3.998
	Length, *p90*(%)	4.707	14.827	5.566
	Coverage probability	0.937	0.944	0.942

IA.3 Estimation Risk

We provide evidence on the improvement of our model over CCZ/OLS in estimating fund specific OLS parameters. Notice that CCZ and OLS generate the same parameter estimates for factor loadings and residual standard deviations as CCZ rely on the equation-by-equation OLS in the first stage of their estimation procedure to estimate these parameters. Table IA.3.1 reports the results.

Across different correlation specifications, the estimation error (as measured by the mean absolute deviation) under NRA is uniformly smaller than that under OLS/CCZ. For certain parameters, the reduction is rather substantial. For example, for the loadings on *umd* (i.e., the momentum factor), the mean absolute deviation under NRA is 0.099, which is on average 11% (=(0.109 − 0.099)/0.089) lower than that under OLS/CCZ relative to the mean absolute loading on *umd* for the underlying true model. For some parameters, the reduction is relatively mild. This is not surprising since we know that OLS is the best linear unbiased estimator (BLUE) of the regression coefficients under the usual OLS assumptions. In our context, we break one of the OLS assumptions by assuming that the regression intercept is drawn from a certain underlying distribution. Our NRA framework recognizes this new assumption and yields more efficient parameter estimates.

Table IA.3.1: **Estimation Risk: NRA vs. OLS/CCZ**

Estimation risk on fund specific OLS parameters. We fix the model parameters at $\mathcal{G}^*$ (Table 1 of the main paper) and generate D sets of data sample. For each set of data sample, we estimate our model using the proposed noise reduced alpha model (NRA), the standard equation-by-equation OLS (OLS), and the model in Chen, Cliff, and Zhao (CCZ, 2015). OLS and CCZ generate the same estimates for factor loadings and residual standard deviations. "Average absolute (true)" reports the averaged cross-sectional absolute values for factor loadings and residual standard deviations for the underlying true model. "MAD" reports the averaged (i.e., across D sets of simulations) cross-sectional mean of the absolute deviation between the true and the estimated parameter value. We calculate the mean absolute deviation (MAD) for fund specific OLS parameter estimates across different models. "MAD (NRA)" reports the MAD for NRA and "MAD (OLS&CCZ)" reports the MAD for OLS, which generates the same parameter estimates and MAD as CCZ.

		$\beta_{i,mkt}$	$\beta_{i,smb}$	$\beta_{i,hml}$	$\beta_{i,umd}$	$\sigma_i(\%)$
	Average absolute (true)	0.996	0.293	0.234	0.089	1.531
$\rho = 0$	MAD (NRA)	0.103	0.142	0.166	0.101	0.242
	MAD (OLS/CCZ)	0.112	0.146	0.170	0.105	0.271
$\rho = 0.2$	MAD (NRA)	0.100	0.147	0.151	0.097	0.255
	MAD (OLS/CCZ)	0.103	0.158	0.157	0.105	0.288
Data dependence	MAD (NRA)	0.102	0.146	0.141	0.099	0.226
	MAD (OLS/CCZ)	0.106	0.152	0.147	0.109	0.256

Table IA.3.2: **Parameter Estimates: NRA vs. Modified CCZ**

Model estimates in a simulation study. We fix the model parameters at $\mathcal{G}^*$ (Table 1 of the main paper) and generate D sets of data sample. For each set of data sample, we estimate our model using the proposed noise reduced alpha model (NRA), the standard equation-by-equation OLS (OLS), and the model in Chen, Cliff and Zhao (CCZ, 2015), and two modified versions of CCZ. One version (CCZ, β only) updates the estimates for factor loadings (based on the updated alpha population in NRA) but does not update the estimates for residual standard deviations. The other version (CCZ, σ only) updates the estimates for residual standard deviations (based on the updated alpha population in NRA) but does not update the estimates for factor loadings. ρ is the assumed level of pairwise correlation for the correlation model that assumes an equal correlation for each pair of residual series. "Data Depen." corresponds to the correlation model that resembles the realized correlations for the actual data. 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 $\mathcal{G}^*$. "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, that is, $\sqrt{\sum_{d=1}^{D}(\gamma_d - \gamma)^2/D}$. "*p(10)*" reports the 10th percentile of the parameter estimates and "*p(90)*" reports the 90th percentile of the parameter estimates. μ_l and σ_l are the (annualized) mean and the (annualized) standard deviation for the l-th component normal distribution, and π_l is the probability for drawing from the l-th component, $l = 1, 2$.

		NRA	OLS	CCZ	CCZ (β only)	CCZ (σ only)
$\mu_1(\%)$	Bias	0.160	0.455	0.367	0.206	0.207
(True = −2.277)	RMSE	0.187	2.592	0.394	0.228	0.259
	p(10)	−2.233	−4.457	−2.097	−2.185	−2.252
	p(90)	−2.007	0.481	−1.749	−1.964	−1.844
$\sigma_1(\%)$	Bias	0.046	13.255	0.564	0.163	0.256
(True = 1.513)	RMSE	0.081	16.501	0.598	0.191	0.274
	p(10)	1.486	2.594	1.898	1.583	1.647
	p(90)	1.631	27.107	2.237	1.776	1.900
$\pi_1(\%)$	Bias	0.023	−0.128	0.084	0.036	0.062
(True = 0.283)	RMSE	0.029	0.324	0.089	0.039	0.067
	p(10)	0.284	0.016	0.341	0.302	0.315
	p(90)	0.324	0.979	0.398	0.333	0.376
$\mu_2(\%)$	Bias	−0.012	0.066	−0.015	−0.013	0.016
(True = −0.685)	RMSE	0.027	1.549	0.051	0.030	0.050
	p(10)	−0.729	−1.180	−0.756	−0.733	−0.728
	p(90)	−0.675	1.870	−0.647	−0.671	−0.624
$\sigma_2(\%)$	Bias	0.009	5.752	0.055	0.018	0.041
(True = 0.586)	RMSE	0.018	13.429	0.059	0.022	0.047
	p(10)	0.579	2.204	0.618	0.590	0.601
	p(90)	0.608	23.765	0.669	0.617	0.655
$\pi_2(\%)$	Bias	−0.023	0.128	−0.084	−0.036	−0.062
(True = 0.717)	RMSE	0.029	0.324	0.089	0.039	0.067
	p(10)	0.676	0.022	0.602	0.667	0.624
	p(90)	0.719	0.984	0.659	0.698	0.685

Table IA.3.3: **Population Statistics: NRA vs. Modified CCZ**

Population statistics based on the model estimates in a simulation study. We fix the model parameters at $\mathcal{G}^*$ (Table 1 of the main paper) and generate D sets of data sample. For each set of data sample, we estimate our model using the proposed noise reduced alpha model ("NRA"), the standard equation-by-equation OLS (OLS), and the model in Chen, Cliff and Zhao (CCZ), and two modified versions of CCZ. One version (CCZ, β only) updates the estimates for factor loadings (based on the updated alpha population in NRA) but does not update the estimates for residual standard deviations. The other version (CCZ, σ only) updates the estimates for residual standard deviations (based on the updated alpha population in NRA) but does not update the estimates for factor loadings. We then calculate summary statistics for the alpha population for all models based on the estimated model parameters. "Mean" is the mean of the alpha distribution. "Stdev." is the standard deviation of the alpha distribution. "Iqr." is the inter-quartile range of the alpha distribution. "*p10*" is the *10*th percentile of the alpha distribution. The other percentiles are similarly defined. "True" reports the population statistics based on the true model. "Estimate" reports the averaged estimate of the population statistics across the D sets of simulations. "RMSE" reports the square root of the mean squared estimation error, that is, $\sqrt{\sum_{d=1}^{D}(s_d - s)^2/D}$, where s is the true statistic and s_d is the estimated statistic based on the d-th simulated sample. Residual correlation is set at zero.

		NRA	OLS	CCZ	CCZ (β only)	CCZ (σ only)
Mean(%)	Estimate	−1.133	−1.140	−1.142	−1.135	−1.150
(True = −1.136)	RMSE	0.034	0.061	0.039	0.030	0.046
Stdev.(%)	Estimate	1.191	3.908	1.577	1.247	1.337
(True = 1.187)	RMSE	0.031	2.886	0.294	0.073	0.157
Iqr.(%)	Estimate	1.163	3.385	1.362	1.192	1.286
(True = 1.144)	RMSE	0.050	2.250	0.198	0.063	0.158
p5(%)	Estimate	−3.642	−5.499	−4.270	−3.757	−3.941
(True = −3.700)	RMSE	0.108	1.806	0.481	0.131	0.275
p10(%)	Estimate	−2.815	−4.420	−3.150	−2.883	−3.043
(True = −2.832)	RMSE	0.091	1.595	0.341	0.109	0.243
p50(%)	Estimate	−0.892	−1.120	−0.882	−0.885	−0.879
(True = −0.860)	RMSE	0.044	0.285	0.043	0.037	0.041
p90(%)	Estimate	0.041	2.148	0.244	0.077	0.148
(True = 0.008)	RMSE	0.051	2.145	0.243	0.076	0.153
p95(%)	Estimate	0.307	3.174	0.675	0.369	0.469
(True = 0.284)	RMSE	0.054	2.893	0.407	0.096	0.206

Table IA.3.4: **Individual Funds: NRA vs. Modified CCZ**

Summary statistics on model performance at the individual fund level. We fix the model parameters at $\mathcal{G}^*$ (Table 1 of the main paper) and generate D sets of data sample. For each set of data sample, we estimate our model using the proposed noise reduced alpha model (NRA), the standard equation-by-equation OLS (OLS), and the model in Chen, Cliff, and Zhao (CCZ, 2015), and two modified versions of CCZ. One version (CCZ, β only) updates the estimates for factor loadings (based on the updated alpha population in NRA) but does not update the estimates for residual standard deviations. The other version (CCZ, σ only) updates the estimates for residual standard deviations (based on the updated alpha population in NRA) but does not update the estimates for factor loadings. For NRA and CCZ (and modified versions of CCZ), given the parameter estimates, we use equations (12)-(14) in the main paper to first construct the density forecast for each individual fund, and then obtain the point estimate and the confidence interval. For OLS, its point estimate is the estimate for the intercept, and its confidence interval is constructed using the point estimate and the standard error for the intercept. "Mean absolute deviation" is the averaged (across simulations) mean absolute distance between the estimated alpha and the true alpha for the cross-section of funds. "Stdev. of mean absolute deviation" is the averaged (across simulations) standard deviation of the absolute distance between the estimated alpha and the true alpha for the cross-section of funds. "Length, p" reports the averaged (across simulations) p-th percentile of the length of the 90% (or 95%) confidence intervals for the cross-section of funds. "Coverage probability" reports the averaged (across simulations) probability for the 90% (or 95%) confidence intervals to cover the true alpha values for the cross-section of funds. Other variables are similarly defined. Residual correlation is set at zero.

		NRA	OLS	CCZ	CCZ (β only)	CCZ (σ only)
	Mean absolute deviation(%)	0.611	1.853	0.803	0.721	0.736
	Stdev. of mean absolute deviation(%)	0.659	3.187	0.724	0.666	0.672
90% confidence interval	Length, *p10*(%)	1.975	3.297	2.169	2.016	2.097
	Length, *p50*(%)	2.812	6.163	3.232	2.867	3.025
	Length, *p90*(%)	4.095	12.487	4.770	4.217	4.486
	Coverage probability	0.890	0.893	0.907	0.896	0.906
95% confidence interval	Length, *p10*(%)	2.433	3.928	2.680	2.487	2.584
	Length, *p50*(%)	3.545	7.343	4.083	3.636	3.821
	Length, *p90*(%)	4.822	14.878	5.677	4.985	5.285
	Coverage probability	0.943	0.943	0.952	0.947	0.952

IA.4 Model Performance under Alternative Parameterizations

To make sure that our model performs well in various situations, we evaluate our model under an alternative parameter configuration.

Our choice of the factor loadings and residual standard deviations are the same as before, which correspond to the equation-by-equation OLS estimates. For parameters in θ that governs the cross-sectional distribution of alphas, instead of using the optimal estimates for our application to mutual funds, we simply set them at the parameter estimates that correspond to the equation-by-equation OLS estimates. In particular, we run equation-by-equation OLS to obtain the fitted alphas. We then fit a two-component GMD on these alphas and obtain the estimate for θ. Table IA.4.1 reports the estimation results. Notice that the parameter estimates are drastically different from the parameter estimates by the NRA model in Table 1 of the main paper. This shows that the equation-by-equation OLS is unable to uncover the true underlying alpha distribution. This new set of parameters also helps evaluate our model's performance under a parameter configuration that is far away from our choice in the main paper.

Table IA.4.2 reports the simulation results on parameter estimates. Table IA.4.3 reports the simulation results on population statistics. Table IA.4.4 reports the simulation results on the alpha estimates for individual funds. Overall, the results are qualitatively similar to the results under the parameterization given in Table 1 of the main paper. The NRA model dominates both OLS and CCZ in making inference on both the alpha population and individual funds.

Table IA.4.1: **Alternative Parameter Vector (θ^*) for the Simulated Model**

Parameter vector (θ^*) for the simulated model. We run equation-by-equation OLS for a cross-section of 3,619 mutual funds that at least have eight months of return observations for the 1983-2011 period. We obtain the cross-section of fitted alphas. We then fit a two-component GMD on these alphas. μ_l and σ_l are the (annualized) mean and the (annualized) standard deviation for the l-th component normal distribution, and π_l is the probability for drawing from the l-th component, $l = 1, 2$.

	First component ($l = 1$)	Second component ($l = 2$)
$\mu_l(\%)$	-6.443	-1.273
$\sigma_l(\%)$	16.951	2.606
π_l	0.040	0.960

Table IA.4.2: **A Simulation Study with An Alternative Parameter Configuration: Parameter Estimates for the Alpha Population**

Model estimates in a simulation study. We fix the model parameters at $\mathcal{G}^*$ (Table IA.4.1 of the on-line appendix) and generate D sets of data sample. For each set of data sample, we estimate our model using the proposed noise reduced alpha model (NRA), the standard equation-by-equation OLS (OLS), and the model in Chen, Cliff and Zhao (CCZ, 2015). ρ is the assumed level of pairwise correlation for the correlation model that assumes an equal correlation for each pair of residual series. "Data Depen." corresponds to the correlation model that resembles the realized correlations for the actual data. 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 $\mathcal{G}^*$. "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, that is, $\sqrt{\sum_{d=1}^{D}(\gamma_d - \gamma)^2/D}$. "*p(10)*" reports the 10th percentile of the parameter estimates and "*p(90)*" reports the 90th percentile of the parameter estimates. μ_l and σ_l are the (annualized) mean and the (annualized) standard deviation for the l-th component normal distribution, and π_l is the probability for drawing from the l-th component, $l = 1, 2$.

		$\rho = 0$			$\rho = 0.2$			Data Depen.		
		NRA	OLS	CCZ	NRA	OLS	CCZ	NRA	OLS	CCZ
μ_1(%)	Bias	0.126	1.568	0.548	−0.160	0.993	0.215	−0.083	1.394	0.384
(True = −6.443)	RMSE	1.659	2.220	1.615	1.760	2.341	1.732	1.626	2.422	1.542
	p(10)	−8.667	−6.888	−8.186	−8.705	−7.732	−7.928	−8.712	−7.978	−8.009
	p(90)	−4.200	−3.034	−4.032	−4.443	−2.640	−4.032	−4.508	−2.703	−4.194
σ_1(%)	Bias	−0.040	2.930	−0.416	0.005	3.324	−0.363	0.088	3.562	−0.313
(True = 16.951)	RMSE	1.234	6.261	1.275	1.252	6.130	1.301	1.296	7.318	1.272
	p(10)	15.485	15.236	15.167	15.469	15.328	15.241	15.615	14.912	15.385
	p(90)	18.596	26.295	18.327	18.759	27.298	18.464	19.059	29.081	18.651
π_1(%)	Bias	0.001	0.015	0.005	0.000	0.013	0.004	0.000	0.014	0.004
(True = 0.040)	RMSE	0.005	0.018	0.008	0.005	0.017	0.007	0.005	0.018	0.007
	p(10)	0.035	0.043	0.039	0.032	0.040	0.035	0.032	0.037	0.037
	p(90)	0.047	0.068	0.052	0.046	0.069	0.051	0.046	0.070	0.051
μ_2(%)	Bias	−0.006	−0.007	−0.004	−0.066	−0.071	−0.064	0.029	0.030	0.032
(True = −1.273)	RMSE	0.054	0.062	0.054	0.549	0.595	0.554	0.399	0.430	0.404
	p(10)	−1.351	−1.353	−1.348	−2.003	−2.047	−1.991	−1.739	−1.784	−1.739
	p(90)	−1.213	−1.199	−1.204	−0.546	−0.477	−0.534	−0.738	−0.696	−0.729
σ_2(%)	Bias	−0.007	0.847	0.184	−0.067	0.793	0.133	0.002	0.859	0.169
(True = 2.606)	RMSE	0.054	0.850	0.154	0.094	0.796	0.104	0.145	0.874	0.213
	p(10)	2.533	2.204	2.685	2.461	3.290	2.613	2.455	3.293	2.612
	p(90)	2.669	3.544	2.827	2.622	3.497	2.774	2.800	3.711	2.946
π_2(%)	Bias	−0.001	−0.015	−0.005	0.000	−0.013	−0.004	0.000	−0.014	−0.004
(True = 0.960)	RMSE	0.005	0.018	0.008	0.005	0.017	0.007	0.005	0.018	0.007
	p(10)	0.953	0.932	0.948	0.954	0.932	0.949	0.954	0.930	0.949
	p(90)	0.965	0.957	0.961	0.968	0.960	0.965	0.968	0.963	0.964

Table IA.4.3: **A Simulation Study with An Alternative Parameter Configuration: Population Statistics**

Population statistics based on the model estimates in a simulation study. We fix the model parameters at $\mathcal{G}^*$ (Table IA.4.1 of the online appendix) and generate D sets of data sample. For each set of data sample, we estimate our model using the proposed noise reduced alpha model ("NRA"), the standard equation-by-equation OLS (OLS), and the model in Chen, Cliff and Zhao (CCZ). We then calculate summary statistics for the alpha population for both models based on the estimated model parameters. "Mean" is the mean of the alpha distribution. "Stdev." is the standard deviation of the alpha distribution. "Iqr." is the inter-quartile range of the alpha distribution. "*p10*" is the *10*th percentile of the alpha distribution. The other percentiles are similarly defined. "True" reports the population statistics based on the true model. "Estimate" reports the averaged estimate of the population statistics across the D sets of simulations. "RMSE" reports the square root of the mean squared estimation error, that is, $\sqrt{\sum_{d=1}^{D}(s_d - s)^2/D}$, where s is the true statistic and s_d is the estimated statistic based on the d-th simulated sample. Residual correlation is set at zero.

		NRA	OLS	CCZ
Mean(%)	Estimate	−1.481	−1.475	−1.484
(True = −1.477)	RMSE	0.083	0.090	0.082
Stdev.(%)	Estimate	4.361	5.723	4.679
(True = 4.350)	RMSE	0.198	1.529	0.270
Iqr.(%)	Estimate	3.649	4.926	3.881
(True = 3.511)	RMSE	0.210	1.422	0.391
p5(%)	Estimate	−6.143	−7.895	−6.483
(True = −6.223)	RMSE	0.182	1.684	0.302
p10(%)	Estimate	−4.909	−6.162	−5.207
(True = −4.946)	RMSE	0.130	1.225	0.242
p50(%)	Estimate	−1.333	−1.319	−1.327
(True = −1.435)	RMSE	0.153	0.166	0.149
p90(%)	Estimate	2.183	3.432	2.430
(True = 2.077)	RMSE	0.168	1.361	0.347
p95(%)	Estimate	3.269	4.984	3.602
(True = 3.353)	RMSE	0.180	1.638	0.276

Table IA.4.4: **A Simulation Study with An Alternative Parameter Configuration: Individual Funds**

Summary statistics on model performance at the individual fund level. We fix the model parameters at $\mathcal{G}^*$ (Table IA.4.1 of the on-line appendix) and generate D sets of data sample. For each set of data sample, we estimate our model using the proposed noise reduced alpha model (NRA), the standard equation-by-equation OLS (OLS), and the model in Chen, Cliff, and Zhao (CCZ, 2015). For the random alpha model, given the parameter estimates, we use equations (12)-(14) in the main paper to first construct the density forecast for each individual fund, and then obtain the point estimate and the confidence interval. For OLS, its point estimate is the estimate for the intercept, and its confidence interval is constructed using the point estimate and the standard error for the intercept. "Mean absolute deviation" is the averaged (across simulations) mean absolute distance between the estimated alpha and the true alpha for the cross-section of funds. "Stdev. of mean absolute deviation" is the averaged (across simulations) standard deviation of the absolute distance between the estimated alpha and the true alpha for the cross-section of funds. "Length, p" reports the averaged (across simulations) p-th percentile of the length of the 90% (or 95%) confidence intervals for the cross-section of funds. "Coverage probability" reports the averaged (across simulations) probability for the 90% (or 95%) confidence intervals to cover the true alpha values for the cross-section of funds. Other variables are similarly defined. Residual correlation is set at zero.

		NRA	OLS	CCZ
	Mean absolute deviation(%)	1.213	1.851	1.332
	Stdev. of mean absolute deviation(%)	1.218	3.110	1.332
90% confidence interval	Length, *p10*(%)	2.931	3.300	2.935
	Length, *p50*(%)	4.793	6.165	4.844
	Length, *p90*(%)	6.942	12.483	7.183
	Coverage probability	0.881	0.892	0.869
95% confidence interval	Length, *p10*(%)	3.487	3.932	3.498
	Length, *p50*(%)	5.719	7.345	5.777
	Length, *p90*(%)	8.331	14.873	8.623
	Coverage probability	0.937	0.944	0.926

IB Comparison with Jones and Shanken (2005)

We compare our model with Jones and Shanken (JS, 2005). In particular, we follow JS to implement their Bayesian estimates on simulated data. There are three prior specifications in JS: a diffuse prior and two informative priors that are calibrated based on the sample of mutual funds in JS. We focus on the diffuse prior as: 1. The informative priors in JS are not very useful in our context since the sample period in JS is different from ours;[1] and 2. A diffuse prior allows a fairer comparison between the Bayesian approach and the frequentist approach. Nonetheless, one can show that informative priors usually lead to a deterioration in model performance compared to a diffuse prior in our simulation study.[2]

We implement JS with a diffuse prior based on three samples of funds. The first sample is the full sample. The second sample is the first half of funds in our sample (i.e., if we arrange the CRSP mutual fund data into a $T \times N$ matrix, where T is the number of time periods and N is the number of funds, then we look at funds that belong to column one through column $N/2$), and the third sample is the second half of funds. One notable difference between the second and the third sample is the median number of monthly observations across funds. It is 116 for the second sample and 54 for the third sample, which is due to the entries of many funds with a short return history for the third sample. We use the three samples to take a deeper look into the performance of JS with a diffuse prior.

Table IB.1 compares the NRA model and JS with a diffuse prior by calculating the model implied estimates for population statistics as well as the estimation error for alpha at the individual fund level. It also assumes that the underlying alpha population follows a normal distribution that is the same as the second component distribution of the GMD in Table 1 of the main paper, which implies that JS is correctly specified. We also present results on JS when it misspecifies the distribution of the alpha population in the main paper.

From Table IB.1, we see that the NRA model performs better than JS for the full sample, especially for the estimation of the dispersion parameter (i.e., "Stdev." in Table B.1). While the NRA model on average overestimates the dispersion parameter by 4.1% (= (0.610-0.586)/0.586), JS overestimates it by 14.7% (=(0.672-0.586)/0.586). Considering the difference in the likelihood function between the NRA model and JS, we believe that this difference in performance comes from the difference in the specification of the likelihood function for each individual fund's time-series. While noninformative priors for OLS regressions usually imply Bayesian estimates that are close to the MLE when the same size is large, the differences between the Bayesian

[1] For the sample of JS, for which most funds are performing well, informative priors will have the effect of shrinking a fund's alpha towards zero. In our sample, since many funds are not performing well, informative priors would have the opposite effect. It is thus difficult to interpret the informative priors in JS for our sample. We thank the anonymous referee for pointing this out.

[2] Our results are available upon request.

estimates and the MLE may not be negligible for funds with a short return history. To further investigate this issue, we look at the two sub-sample estimates for JS. Our results in Table IB.1 show that the degree of bias for JS for the estimation of the dispersion parameter for the second half of our sample is even higher than that for the full sample. This is consistent with the fact that funds in the second half of our sample on average have fewer observations than in the full sample.

Table IB.1: **NRA vs. Jones and Shanken (2005): A Single Normal Distribution**

Simulation results for NRA and Jones and Shanken (2005). The underlying alpha population is assumed to follow a normal distribution that is the same as the second component distribution of the GMD in Table 1 of the main paper. For each set of the full data sample, we estimate our model using the proposed noised reduced alpha model ("NRA") and Jones and Shanken (2005) under a diffuse prior. We also split the full data sample into two parts, "first half" including the first half of the funds and "second half" including the second half. For each part, we separately estimate Jones and Shanken (2005) under a diffuse prior. We calculate several summary statistics for the alpha population for both models. "Mean" is the mean of the alpha distribution. "Stdev." is the standard deviation of the alpha distribution. "Iqr." is the interquartile range of the alpha distribution. "*p10*" is the *10*th percentile of the alpha distribution. The other percentiles are similarly defined. "True" reports the population statistics based on the true model. "Estimate" reports the averaged estimate of the population statistics across the D sets of simulations. "RMSE" reports the square root of the mean squared estimation error, that is, $\sqrt{\sum_{d=1}^{D}(s_d - s)^2/D}$, where s is the true statistic and s_d is the estimated statistic based on the d-th simulated sample. We also calculate the estimation error for alpha at the individual fund level. For the NRA model, given the parameter estimates, we use equations (12)-(14) in the main paper to first construct the density forecast for each individual fund and then obtain the point estimate. For Bayesian estimates, we use the posterior mean, estimated through the MCMC sampling procedure in Jones and Shanken (2005). "Mean absolute deviation" is the averaged (across simulations) mean absolute distance between the estimated alpha and the true alpha for the cross-section of funds. "Stdev. of mean absolute deviation" is the averaged (across simulations) standard deviation of the absolute distance between the estimated alpha and the true alpha for the cross-section of funds. Residual correlation is set at zero.

		NRA	JS full sample	JS first half	JS second half
Mean(%)	Estimate	−0.668	−0.677	−0.681	−0.685
(True = −0.683)	RMSE	0.052	0.028	0.043	0.046
Stdev.(%)	Estimate	0.610	0.672	0.632	0.702
(True = 0.586)	RMSE	0.025	0.123	0.072	0.140
p5(%)	Estimate	−1.639	−1.789	−1.720	−1.840
(True = −1.647)	RMSE	0.024	0.165	0.125	0.237
p10(%)	Estimate	−1.417	−1.545	−1.491	−1.585
(True = −1.434)	RMSE	0.027	0.130	0.100	0.187
p50(%)	Estimate	−0.636	−0.677	−0.681	−0.685
(True = −0.683)	RMSE	0.052	0.028	0.043	0.046
p90(%)	Estimate	0.145	0.178	0.130	0.215
(True = 0.068)	RMSE	0.082	0.126	0.103	0.184
p95(%)	Estimate	0.367	0.423	0.359	0.470
(True = 0.281)	RMSE	0.091	0.161	0.128	0.233
Mean absolute deviation(%)		0.433	0.444	0.439	0.451
Stdev. of mean absolute deviation(%)		0.334	0.337	0.333	0.343

IC Model Estimates under Alternative Benchmark Models

Table IC.1 shows our model estimates under alternative specifications of the benchmark factor model. We explore two specifications: CAPM and the Fama-French three-factor model (Fama and French, 1993). Overall, the structure of the estimated alpha population under these specifications is consistent with our estimate under the four-factor model in the main paper, that is, around 20% of alphas are drawn from a "bad" group with a large negative alpha and the rest are drawn from an "OK" group with a small negative alpha. This demonstrates the stability of the estimated alpha population, regardless of the choice of the benchmark model. We also see that, not surprisingly, funds tend to perform better under CAPM than under the two alternative specifications, especially for funds whose alphas are drawn from the "bad" group.

Table IC.1: **The Alpha Population: Mutual Funds under Alternative Benchmark Models**

Model estimates and population statistics for mutual funds under alternative specifications for the benchmark factor model. For a cross-section of 3,619 mutual funds covering the 1983–2011 period, we estimate our model, which is based on a two-component GMD specification for the alpha population. "CAPM" refers to the benchmark model specification that only includes the market excess return as the risk factor. "FF 3-factor" refers to the benchmark model specification that includes the Fama and French (1993) three factors as risk factors. Panel A reports the parameter estimates for the model. μ_l and σ_l are the (annualized) mean and the (annualized) standard deviation for the l-th component normal distribution, and π_l is the probability for drawing from the l-th component, $l = 1, 2$. Panel B reports the estimated population statistics for the alpha distribution. "Mean" is the mean of the alpha distribution. "Standard deviation" is the standard deviation of the alpha distribution. "Interquartile range" is the inter-quartile range of the alpha distribution. "*10*th percentile" is the *10*th percentile of the alpha distribution. The other percentiles are similarly defined. For both Panel A and B, "*p(5)*" and "*p(95)*" report the *5*th and *95*th percentiles of the variable of interest across simulations, respectively.

Panel A: Parameter Estimates for the Alpha Population

	CAPM	FF 3-factor
$\mu_1(\%)$	−0.900	−2.362
$\sigma_1(\%)$	2.783	1.767
π_1	0.217	0.214
$\mu_2(\%)$	−0.750	−0.798
$\sigma_2(\%)$	0.661	0.659
π_2	0.783	0.786

Panel B: Population Statistics for the Alpha Population

	CAPM	FF 3-factor
Mean(%)	−0.783	−1.133
Standard deviation(%)	1.424	1.193
Interquartile range(%)	1.101	1.105
*5*th percentile(%)	−2.961	−3.641
*10*th percentile(%)	−1.966	−2.579
*50*th percentile(%)	−0.760	−0.944
*90*th percentile(%)	0.393	0.029
*95*th percentile(%)	1.232	0.294
Fraction of positive alphas	0.180	0.107

ID Estimation Details

In this appendix, we detail the implementation of the estimation method that is described in Section 3 of the main paper.

In *Step I*, we choose a set of starting values to initialize our estimation. We have a large number of parameters that are given by $\mathcal{G} = [\theta', \mathcal{B}', \mathbf{\Sigma}']'$. However, the time-series information of each fund helps us estimate each fund's risk loadings (i.e., β) and residual variance, providing a reasonable set of starting values. Therefore, for $\mathcal{B}$ and $\mathbf{\Sigma}$, we start with their equation-by-equation OLS estimates, that is:

$$\mathcal{B}^0 = \mathcal{B}^{OLS}, \ \mathbf{\Sigma}^0 = \mathbf{\Sigma}^{OLS}, \tag{.1}$$

where a superscript of zero denotes the starting values.

For parameters that govern the GMD (i.e., θ), we randomly generate multiple sets of starting values to avoid local optimums. In particular, for a L-component GMD and for the L parameters that govern the means of the component distributions, we randomly choose L numbers that are uniformly distributed over the interval of $[-20\%, 20\%]$ (per annum). The boundary of 20% reflects our knowledge of the mutual fund data. Our prior is that it is unlikely to have a population of funds that are concentrated around a mean that resides outside of the $[-20\%, 20\%]$ interval. Our estimation results confirm this prior. We never obtain optimal mean estimates for the component distributions that are close to the boundaries. After randomly generating the L mean parameters, we rank them in an ascending order for model identification.

We follow a similar procedure to choose the starting values for the standard deviations of the component distributions. In particular, we randomly choose L numbers that are uniformly distributed over the interval of $[0.1\%, 20\%]$ (per annum). Again, the choices of the boundaries reflect our priors about the standard deviations of the component distributions. Our estimation results confirm that these boundaries are never violated for the optimized estimates of the standard deviations of the component distributions.

For the drawing probabilities, the selection of the starting values is more complicated than the selection of the previous two sets of parameters as we now have the parameter constraint that the sum of the L drawing probabilities should be one. We therefore follow a sequential procedure to choose the staring values. We first draw a number (i.e., p_1) that is randomly distributed over the unit interval. After drawing the first number, we draw a second number that is uniformly distributed over $[0, 1 - p_1]$. We continue in this way to draw the rest of the probabilities. In particular, after choosing the first l probabilities (i.e., $\{p_i\}_{i=1}^{l}$), we choose the $(l+1)$-th probability by drawing a number that is uniformly distributed over $[0, 1 - \sum_{i=1}^{l} p_i]$. Lastly, after choosing the $(L-1)$-th probability, the last probability is simply set as $1 - \sum_{i=1}^{L-1} p_i$.

After following the above steps, we now have a randomly generated set of initial parameter values $\mathcal{G}^0 = [(\theta^0)', (\mathcal{B}^0)', (\mathbf{\Sigma}^0)']'$, where θ^0 contains the parameters that govern the GMD. Taking this set of parameter values as input for our algorithm, our estimation becomes automatic. In particular, staring from $\mathcal{G}^0$ and following *Step II-IV*, we arrive at a new set of parameters $\mathcal{G}^1$. Next, starting at $\mathcal{G}^1$, we follow our algorithm and arrive at $\mathcal{G}^2$. We continue in this way and obtain a sequence of parameter estimates $\{\mathcal{G}^k\}_{k=0}^K$. This sequence of parameter estimates are converging as K gets larger. The speed of convergence for our algorithm seems high in that the variations in parameter values become very small after ten to fifteen iterations. To terminate the program, we set a tough threshold for the distance of the parameter estimates between adjacent iterations. In particular, we stop the program at the K-th iteration if the L_1 distance between θ^{K-1} and θ^K is within d^{lim}. To prevent the program from running too many iterations, another criterion we impose is that if the program does not stop until the K^{lim}-th iteration, we stop it at K^{lim}. The choices of d^{lim} and K^{lim} depend on whether the estimation is the intermediate step or the final step, as we shall explain next.

We have explained how our estimation works for one set of starting values. We need to try multiple sets of starting values to avoid local optimums. In particular, following the aforementioned generating procedure for starting values, we randomly generate 100 sets of starting values. For each set, we run our algorithm by setting $d^{lim} = 10^{-1}$ and $K^{lim} = 30$ and obtain 100 sets of parameter estimates. This an intermediate optimization step in which we try to save the computational time by setting d^{lim} and K^{lim} at lenient thresholds and obtain 100 sets of rough estimates. Next, we rank the 100 sets of parameter estimates by the corresponding values of the optimized likelihood function. We choose the top 20 sets and rerun our program by starting at the estimated parameter values. This time, we set $d^{lim} = 10^{-2}$ and $K^{lim} = 50$. We again rank the resulting 20 sets of parameter estimates by the corresponding values of the likelihood function. We choose the top five sets and rerun our program by starting at the estimated parameter values obtained from the previous step. This is the final step estimate and we set $d^{lim} = 10^{-3}$ and $K^{lim} = 100$. We choose the best one (in terms of the value of the likelihood function) among the five sets of estimates as our final estimate. We often see that five sets of parameter estimates in the final step are very close to each other. This assures us that the local optima have been thrown out during the intermediate steps.

IE Motivating Examples

At the core of our method is the idea of extracting information from the cross-section of funds. This information can be used both to make inference on the alpha population and to refine our inference on a particular fund. To motivate the idea, we use two examples throughout our paper. The first example is what we call a *one-cluster* example. Suppose all the funds in the cross-section generate an alpha of approximately 2% per annum and the standard error for the alpha estimate is about 4%. Since the t-statistics are all approximately 0.5 (=2%/4%), which is not even high enough to surpass the single test t-statistic cutoff of 2.0, let alone the multiple testing adjusted cutoffs, we would declare all the funds to be zero-alpha funds. Using our method, the estimate of the mean of the alpha population would be around 2%. In this case, we think our approach provides a better description of the alpha population than the usual hypothesis testing approach. Declaring all the funds to be zero-alpha funds ignores information in the cross-section.

While the one-cluster example illustrates the basic mechanism of our approach, it is too special. Indeed, a simple regression that groups all the funds into an index and tests the alpha of the fund index will also generate a positive and significant estimate for the mean of the alpha population. This motivates the second example, which we call the *two-cluster* example. For the two-cluster example, suppose we have half of the funds having an alpha estimate of approximately 2% per annum and the standard error for the alpha estimate is about 4%. The other half have an alpha estimate of approximately -2% per annum and also have a standard error of about 4%. Similar to the one-cluster example, no fund is statistically significant individually. However, we throw information away if we declare all the funds to be zero-alpha funds. Different from the one-cluster example, if we group all the funds into an index and estimate the alpha for the index fund, we will have an alpha estimate close to zero. In this case, the index regression approach does not work as it fails to recognize the two-cluster structure of the cross-section of fund alphas. Our approach allows us to take this cluster structure into account and make better inference on the alpha population.

The one-cluster and two-cluster examples are special cases of the alpha distributions that our framework can take into account. They correspond to essentially a point mass distribution at 2% and a discrete distribution that has a mass of 0.5 at -0.2% and 0.5 at 0.2%, respectively. Our general framework uses the GMD to model the alpha distribution and seeks to find the best fitting GMD under a penalty for model parsimony. It therefore extracts information from the entire cross-section of alphas.

After we estimate the distribution for the cross-section of alphas, we can use this distribution to refine the estimate of each individual fund's alpha. For instance, for the one-cluster example, knowing that most alphas cluster around 2.0% will pull our estimate of an individual fund's alpha towards 2.0% and away from zero. Similarly, for the two-cluster example, knowing that the alphas cluster at -2.0% and 2.0% with

equal probabilities will pull our estimate of a negative alpha towards -2.0% and a positive alpha towards 2.0%, and both away from zero. In our general framework, after we identify the GMD that models the alpha cross-section, we use it to update the density estimate of each fund's alpha, thereby using cross-sectional information to refine the alpha estimate of each individual fund.

IF The Literature on the EM Algorithm

From a methodological perspective, our framework contributes to the literature on EM algorithm by allowing heterogeneous funds in the cross-section and simultaneously estimating fund specific parameters and other structural parameters.[3] In particular, we allow both factor loadings and residual standard deviations to be fund specific and update the entire cross-section of fund-specific variables along with other structural parameters in the maximization step of the EM algorithm. This is an important and necessary extension for the purpose of our application as we know there is estimation uncertainty as well as a large amount of heterogeneity in the risk-taking behavior of mutual funds. Failing to take either the heterogeneity or the estimation uncertainty into account may bias our estimate of the alpha population. On the other hand, allowing fund heterogeneity does not compromise the simplicity and the intuitive appeal of the standard EM algorithm. We show that our new algorithm simply embeds a constrained OLS estimate for fund specific parameters (i.e., factor loadings and residual standard deviations) into an otherwise standard EM algorithm. This greatly reduces the computational burden of our model. We provide a comprehensive simulation study to demonstrate the performance of our estimation procedure.

[3]See Dempster, Laird, and Rubin (1977) for the original paper that proposes the EM algorithm. See McLachlan and Krishnan (2007) for a more detailed discussion of the algorithm and its extensions. Different from these papers on the EM algorithm, our method allows for heterogeneous factor loadings and residual standard deviations in the cross-section. Chen et al. (2015) use a modified EM algorithm to group funds into different categories. They employ a two-step estimation procedure to first estimate the equation-by-equation OLS and then use the t-statistics of alphas to classify funds. We put fund-specific variables on an equal footing with other structural parameters and simultaneously estimate the model parameters. This allows us to take into account the estimation uncertainty for fund-specific variables using information from the entire fund cross-section. In contrast, CCZ only take into account the estimation uncertainty using fund-specific information.

IG Relating Our Empirical Findings to the Literature

Linking to the existing literature, three approaches are proposed to evaluate mutual fund performance. The first method uses the extreme test statistics and tries to evaluate the significance of the best/worst funds, while controlling for test multiplicity (see, for example, Kosowski et al. 2006, Fama and French, 2010, Harvey and Liu, 2017b). It is based on fund-by-fund hypothesis testing and its null hypothesis is that each fund has a zero alpha. It is designed to answer the question of whether there exists any funds that significantly outperform/underperform and cannot further classify funds into different performance groups. Using this approach, Kosowski et al. (2006) find that there exist managers that significantly outperform. Refining the method in Kosowski et al. (2006) to control for cross-sectional dependency, Fama and French (2010) find no outperforming funds.

The second approach tries to classify funds into broad categories. Papers that follow this procedure include Barras et al. (2010) and Ferson and Chen (2015). The assumption of this approach is less stringent than the assumption under the previous method in that not all funds need to have a zero alpha. Certain funds can have nonzero alphas and this approach tries to control the false discovery rate at 5%. Using this approach, Barras et al. (2010) find that about 75% of funds are zero-alpha funds. Ferson and Chen (2015) refine this method by allowing a non-zero probability for true alphas to disguise themselves as zero, and find that 50% or fewer have zero alphas. Neither paper finds evidence of funds that significantly outperform.

From an methodological perspective, there are several important differences between our approach and the false classification (FC) method in Barras et al. (2010) and Ferson and Chen (2015). The FC approach, being essentially a variant of the traditional hypothesis testing framework, postulates that fund alphas can only take a few particular values, each value for a certain performance group. While this offers a simplification of the inference problem, there is no particular reason to think that fund alphas can only take a few values. As a result, if a fund has a true alpha that is very different from these assumed values, the estimation error by assigning this fund to any particular performance group might be large. Our approach allows us to flexibly model the alpha population as following a continuous distribution, thereby reducing the estimation error in the FC approach where fund alphas are forced to take a few values.

Second, the loss functions in our approach and the FC method are different. FC relies on the multiple hypothesis testing approach and aims to strike a balance between Type I (i.e., false discovery rate) and Type II error rates. Our maximum likelihood-based approach tries to find the best parametric model that fits the data through optimally weighting the likelihood from fitting the panel of return time-series and the likelihood from fitting the cross-section of alphas. Hence, a material advantage

of our framework is that it allows us to take into account the parameter uncertainty in estimating both fund alphas and other OLS parameters (i.e., factor loadings and residual standard deviations) when we try to fit the cross-section of estimated alphas. On the other hand, our structural approach also allows us to address the Type I error concern that is the focus of the FC method. In particular, assuming all funds have a zero alpha, if we estimate the alphas of a thousand funds, on average 25 funds will appear to have a significant positive alpha from a single test perspective. In our framework, these 25 funds will likely not have a significant positive alpha as the posterior distribution of alpha weights the information from the time-series (which is what the single test p-values are based on) by using information from the alpha cross-section. Since our estimate of the mean of the alpha population will likely be zero, learning across funds allows us to downwardly adjust the significance of each individual fund, leading us to correctly declare the 25 funds as insignificant. Equation (**??**) shows the precise formula for how our model adjusts the statistical significance of individual funds when the alpha population has a zero mean.

The third approach, as taken by our paper, is to treat alphas as continuous and try to estimate the underlying distribution for alphas. We deviate from the usual hypothesis testing approach in that we do not think an alpha of zero is any different than an alpha of other value. Another salient feature of our model is that we take various sources of estimation risk into account.

One can think of the three approaches as following an order that tries to obtain a finer and finer understanding of the alpha distribution. The first approach tries to answer the very basic question of whether there exists any fund that has a non-zero alpha. If the answer is yes, we proceed to the second approach to classify funds into broad categories. Finally, viewing alphas as coming from an underlying distribution, we use the third approach to provide a more precise description of this distribution.

Fundamentally, our approach is different from the first two approaches that rely on fund-by-fund hypothesis testing. Viewing fund alphas as coming from an underlying distribution, our model estimates suggest that mutual fund managers are doing better than what people have previously thought. We estimate that a little more than 10% of funds are generating a positive alpha. Our estimate is higher than those reported in the literature and likely due to the fact our structural approach has more power in identifying small but non-negligible alphas. If decreasing return to scale were the underlying economic mechanism that drives alpha dynamics (Berk and Green, 2004), then small but positive alphas are usually associated with large funds.[4] Given that larger funds have a greater impact on the mutual fund industry than smaller funds, it would be a mistake to label these funds as zero alpha funds from an economic perspective.

[4]See Harvey and Liu (2017c) for the evaluation of economies of scale using a similar approach.

IH FAQ

- *In short, what is the most compelling reason to consider the random effects model for performance evaluation?*

  Quoting Searle, Casella, and McCulloch (1992), "Effects are fixed if they are interesting in themselves or random if there is interest in the underlying population." For performance evaluation, we are interested in both the effects themselves (that is, to evaluate which manager outperforms) and the population (that is, the underlying distribution for alphas). A random effects framework with our NRA approach provides a suitable way to think about both.

  NRA provides a suitable framework to think about both.

- *Why not do NRA with a multiple testing adjustment?*

  The mechanisms for the NRA model and the multiple testing framework to discount the significance of fund alphas are different. The random alpha model forces the cross-section of alphas to fit a parametric density. Observations that are too extreme according to the fitted density are adjusted. Multiple testing adjustment invokes the hypothesis testing framework and uses the p-value to measure the distance between the estimated alpha and zero. A smaller p-value indicates a larger distance from zero and we are trying to identify alphas that are sufficiently distant from zero. It is possible to make mistakes by falsely declaring a zero alpha as nonzero. To control for the false discovery rate, we need to adjust the p-values upward. Both the quantities of interest (i.e., raw alpha vs. p-value of alpha) and the objectives (i.e., goodness-of-fit to a density vs. false discovery rate) are different between the two methods. Applying both will likely overkill the significance of fund alphas.

- *Why is MLE better than the moments-based approach for the estimation of a GMD?*

  In general, we need an infinite number of moments — properly weighted — to achieve the estimation efficiency that MLE provides. For example, for a two-component GMD, although it is identified and its five parameters can be estimated using the first five sample moments alone, the sixth moment as well as other higher moments provide additional information for the estimation of the model and should be incorporated into the estimation to improve estimation efficiency.

- *How does model misspecification affect the results of the paper?*

There are different kinds of model misspecifications. A misspecification of the return residuals changes our MLE into a QMLE, which will not bias our estimates. The loss in estimation efficiency is also small, as we show in the simulation study. On the other hand, a misspecification of the factor model (e.g., omitting a true factor) in general will introduce bias for the alpha estimates. Compared to existing models, our model can to some extent alleviate the model misspecification issue, thanks to its ability to use information in the entire cross-section to provide inference. We have a discussion on this towards the end of the paper. However, both our model as well as existing models are sensitive to the issue of model misspecification.

- *Does it make sense to treat all funds equally? It seems that there is more information for a fund with a $1 trillion AUM than a fund with a $10 million AUM.*

  From the perspective of making inference on the alpha population, we think that the alpha for the $10 million fund is just as important as the alpha for the $1 trillion AUM fund. If an investor invests $1 million in either fund, the alpha she gets is simply the alpha for either fund. The alpha for the smaller fund will not be discounted because the fund is smaller. It is likely that the returns for smaller funds are more noisy than returns for larger funds. Our method takes the estimation uncertainty into account.

- *How many funds have a t-statistic over 2.0 under OLS?*

  For our sample, under equation-by-equation OLS, 1.7% of funds have a single test t-statistic above 2.0. However, since we have run thousands of tests, we need to adjust for multiple testing. Applying multiple testing adjustments, few funds are found to be significantly outperforming.

  Our method departs from the hypothesis testing framework and assumes a continuous distribution for fund alphas. In our framework, alphas are almost surely not zero by construction. To see the difference between our framework and hypothesis testing, suppose we have 100 funds, each one having an OLS intercept of 1% (per annum) and a standard error of 2% (per annum). Under hypothesis testing, there is no outperformer, as none of the t-statistics is able to pass the single test t-statistic threshold, let alone the multiple testing t-statistic threshold. Under our model, we estimate the alpha distribution to be, say, normal around a mean of 1%. If we test the significance of each alpha under our model, it might as well be the case that none of the t-statistics is above 2.0, especially if the 2% standard error is high enough at the individual fund level. However, this is not evidence against our model since it is not based on hypothesis testing. In our framework, it is possible that all individual funds have a t-statistic below 2.0 while at the same time the population mean is positive and statistically different from zero.

- *What is the intuition behind the EM algorithm to refine the OLS estimates of alphas?*

  Imagine that the parameters that govern the alpha population (i.e., the normal mixture distribution) are given. In the "expectation" step, we calculate the conditional distribution of alphas by mixing information from the time-series and the cross-section. Essentially, OLS estimates are adjusted for the information in the mixture distribution. More noisy OLS alpha estimates (which are likely due to higher levels of residual standard deviations) are adjusted more aggressively than less noisy OLS alpha estimates. Hence, the new alpha estimates after the "expectation" step are less noisy than the OLS estimates that are based on time-series information alone. However, these new alpha estimates should change our initial guess of the alpha population (i.e., parameters in the normal mixture distribution). As a result, in the "maximization" step, we try to find a new set of parameters that best explain these new alpha estimates. We iterate between the "expectation" step and the "maximization" step to refine our estimates of both the individual alphas and the parameters that govern the alpha population.

- *In the two-cluster example, if a manager realizes a -20% return, why would you want to shrink it towards -4.0%? This would lead investors to not to fire the manager and hence a high Type II error in terms identifying true managers.*

  The two-cluster example is a simplified example, where everything below -4.0% will be pulled towards -4.0%. For the actual method, first of all, -20% will not be pulled all the way to -4.0%. Exactly how much it gets pulled depends on the particular fund's time-series uncertainty vs. cross-sectional uncertainty. On the other hand, if there were a group of managers whose performance concentrate at -20%, our method will allow us to extend the support of the alpha distribution, trying to capture the alpha distribution around -20%.

- *How does our paper relate to* regularization*?*

  Our critique of the traditional approach is consistent with the recent advances in statistics, and in particular in machine learning, that emphasize *regularization*. In general, regularization refers to the process of introducing additional information or constraints to achieve model simplification that helps reduce model overfitting. In the context of our application, we have a complex dataset given the multidimensional nature of the cross-section of investment funds. The standard approach, by treating each fund as a separate entity and running equation-by-equation (that is, fund-by-fund) OLS to obtain a separate t-statistic to summarize its performance, does not reduce the complexity of the dataset. In contrast, our framework imposes a parametric distribution on the cross-section of alphas and thereby substantially reduces the model complexity. It is unlikely to produce a time-series fit that is as good as the equation-by-

equation OLS. However, the better fit by the equation-by-equation estimation may reflect overfitting, which means that the estimated cross-sectional distribution of alphas may be a poor estimate of the future distribution. Our method seeks to avoid overfitting with the goal of getting the best forecast of the future distribution.

- *Does your out-of-sample forecasting exercise capture time-varying alphas and betas?*

  To capture higher frequency variation in funds' risk loadings and alphas, one may want to run the out-of-sample exercise over shorter horizons. However, the cross-sectional distribution of alphas over a short horizon is likely to be significantly affected by contemporaneous macroeconomic and market conditions and therefore unable to reflect the long-run stationary alpha distribution. One can also potentially better capture beta variability if there are pre-determined instruments that help predict funds' factor loadings. However, to the extent that there is uncertainty around the choices of these instruments, an incorrect specification of these instruments may lead to biased inference of the cross-sectional alpha distribution (see our discussion on misspecification of factor models in the next section). We therefore pursue an out-of-sample forecasting exercise that is closest to our model setup, while leaving possible extensions to future research.

- *Can an alternative MCMC sampling scheme (a scheme that is different from Jones and Shanken, 2005) perform better than the NRA?*

  As we show in our paper, NRA seems to perform better than Jones and Shanken (JS) even when JS is correctly specified, that is, the cross-section of alphas follow a normal distribution. The underperformance of JS relative to our model is more likely due to the prior specifications than the particular MCMC sampling scheme that JS use. Hence, we do not think that an alternative MCMC sampling scheme under the JS setup can perform better than our model in general.

- *What is the intuition behind the difference between your 10% estimate of the fraction of funds that outperform and the much lower estimate using hypothesis testing?*

  Our results suggest a different answer to: What proportion of mutual funds outperform? While the existing literature suggests few if any funds are deemed to outperform, our results suggest that over 10% of funds generate positive risk-adjusted performance. Two effects contribute to our estimate. In the usual fund by fund regressions, 0-1% of funds have positive significant alphas. However, due to the high level of estimation uncertainty at the individual fund level, funds with small positive alphas are likely deemed insignificant from the perspective

of the traditional approach. Our framework provides a more powerful procedure to identify these funds by directly modeling the underlying alpha population. On the other hand, we cannot take the fund-by-fund OLS alpha estimate at face value as the cross-sectional learning effect dictates that we should shrink positive alphas towards zero given that the median fund has a negative alpha. Notice that these two effects work against each other. The overall impact is to have a larger estimate for the fraction of outperforming funds to account for funds with small positive alphas, despite the various degrees of shrinkage for these funds.