# Economic Growth Centre
## Working Paper Series

## A New Direction of Fund Rating Based on the Finite Normal Mixture Model

*by*

**GAO Zhangpeng and Shahidur RAHMAN**

Economic Growth Centre
Division of Economics
School of Humanities and Social Sciences
Nanyang Technological University
Nanyang Avenue
SINGAPORE 639798

Website: http://www.hss.ntu.edu.sg/egc/

**Working Paper No: 2006/03**

Copies of the working papers are available from the World Wide Web at:

Website: http://www.hss.ntu.edu.sg/egc/

# A New Direction of Fund Rating
# Based on the Finite Normal Mixture Model

**By**
**Gao Zhangpeng and Shahidur Rahman**
**Nanyang Technological University, Singapore**

***Abstract***

In this paper we try to develop a theoretical framework for fund rating under the assumption that superior funds could have a higher expected return than that of inferior funds, which could arise from the segmented market information or the differentiated ability of mangers to acquire and analyze the information. Under this setting, the funds are rated based on the cross-sectional distribution of all the funds instead of the preset-percentiles as Morningstar. We use the finite normal mixture for rating fund performance with the number of performance groups determined by likelihood ratio test using parametric bootstrap procedures, and we estimate the model with EM algorithm by treating the group information of funds as missing information.

***JEL Classification***: G0, G1, C1, D4
**Keywords**: Fund Rating, Fund Performance, Finite Normal Mixture, Bootstrap, EM Algorithm

**Address for correspondence**
Dr. Shahidur Rahman,
Associate Professor,
NTU-S3-B2-B-62,
Division of Economics,
Nanyang Technological University,
Singapore, 639798,
Phone: 65+67906404, Fax: 65+67920697
*E-mail: asrahman@ntu.edu.sg*

# A New Direction of Fund Rating
# Based on the Finite Normal Mixture Model

## 1. Introduction

Numerous literatures have been devoted to fund performance study, such as Jensen (1968), Cai, Chan and Yamada (1997), Christopherson, Ferson and Glassman (1998, Grinblatt and Titman (1989), Ippolito (1989) among others. However, there is little study on fund rating till now. There are possibly two reasons. One is that the number of funds is too small before 1990 to have a meaningful rating. Even though the fund history is long enough for fund performance study, where people usually use three-year time series data such as Carhart (1997), Connor and Korajczyk (1991), and Elton, Gruber and Blake (1996), the number of funds at each point of time is small to do a cross-sectional fund rating study from the perspective of its distribution. Second, there is a lack of statistical methodology to specify and estimate the density model. However, since the introduction of the Expectation and Maximization (EM) algorithm into the maximum likelihood estimation with missing data by Dempster, Laird and Rubin (1977), the finite mixture model becomes popular, e.g. in medical and biological study, in 1980s and 1990s. Furthermore, the parametric bootstrap procedure by McLachlan and Basford (1988) overcomes the model specification difficulty to some extent.

With the proliferation of funds in 1990s and well-developed statistical methodology over the last decade, we intend to propose a new direction of fund rating that is based on the finite normal mixture distribution model. This model is more flexible and provides more sensible rating results than current fund rating method by Morningstar, which is commercial fund rating method based on the fixed number of performance groups and preset percentiles.

There are several questions regarding Morningstar's method. First, it is not appropriate to fix the number of performance groups before we investigate the distributions of alphas. It is very possible that we only have one performance group if we find later that the difference of alphas may be just a random effect, caused by the "Luck" of managers. Second, the number of funds in each group shouldn't be fixed before rating. We provide a simple example to illustrate its limitations. Suppose we have 100 funds. 50 funds have alphas around 10% and another 50 funds have alphas around 5%. In this situation, it is obviously not appropriate to say we have five fund performance groups and the top 15 funds are rated as superior funds as implied by Morningstar's method. Instead, it is better to say we only have 2 performance groups and the top 50 funds are superior fund group based on the actual distribution of alphas.

Under the rationale of the example, we propose a method, which is based on the cross-sectional distribution of all of the funds' performances, measured by alphas (Alpha is the measure of fund performance during a period). The method uses a finite normal mixture

model to describe the distribution of funds performances. Doing so, this method partly overcomes the shortcomings of Morningstar's method. First, the number of performance groups is not fixed, but determined by the spectrum of the alphas of all the funds. For example, if the alphas closely cluster around only one value, we may conclude there is only one performance group from estimation; if alphas cluster around three values, then we may conclude there are three performance groups. We exploit the parametric bootstrap procedure to determine the number of groups at that point of time. Second, we can obtain the posterior probability of individual funds after we have specified and estimated the model, so we know the performance group that the fund belongs to by comparing posterior probabilities. Third, after knowing the group of each fund, it is straightforward to know the number of funds in each group. Therefore, the number of performance groups, the number of funds in each performance group, and the performance group that the fund belongs to, are determined by the cross-sectional distribution of alphas, which are not fixed before rating like Morningstar's method.

This paper is arranged this way. In section 2, we present the motivation and justification of the finite normal mixture model. In section 3, we formulate the fund rating issue under the framework of the finite normal mixture model. In section 4, we treat the group information of the fund as missing data, so we can estimate the model under EM framework which is more straightforward and intuitive. In section 5, we show how to determine the number of performance groups by parametric bootstrap procedures. Finally we summarize our fund rating procedures.

## 2. Motivation of Finite Normal Mixture Model

Utilization of the finite normal mixture model is motivated due to the multimodal shape of the distribution of alphas and formal normality tests. The multimodal shape is a strong indication of finite mixture distribution model. In addition, the results from normality tests contradict what we generally believe on the distribution of alphas. The alpha is the management effect, always interpreted as the manager's ability to deliver abnormal return over passive portfolios. It is affected by many factors, like stock selection, the idiosyncratic news shocks to stocks, asset allocation and rotation, and irregular liquidation caused by redemption from investors, so based on the Linderberg-Levy central limit theorem (CLT) it is generally assumed that the fund performance follows a normal distribution. And the statistical inference of alphas in all the traditional measures relies on the assumption that the alpha is normally distributed.

However, if fund managers' decisions are based on different information sets, we have a group structure in the distribution. Therefore, we can not assume the distribution of alphas as a univariate normal distribution, because the expected performance and investment risk will be different for managers who have different information sources. The more information the managers have, the better investment decision they will make. The managers who can make better decisions are expected to deliver higher performance. In this case, the distribution of alphas of all the fund mangers, which are from different information sets, will be a finite normal mixture distribution, i.e. an addition of several normal distributions. This model is a convenient way to model the group structure of the

distribution, as shown by McLachlan and Basford (1988), and it is now widely applied in medicine study, for example, Tao, et al. (2004). In the model each component is a normal distribution which is derived from the corresponding information set. The number of components in the finite normal mixture model is actually interpreted as the number of information sets in the fund market. For example, if managers have private information about firms, they are from superior information set which has higher expected performance, so are expected to deliver higher alphas compared to mangers that don't have the private information.

There are several sources that may result in different information sets of fund managers. First, there is managers' ability at acquiring private information. The managers who have private information from insiders are in the private information set, while other managers are in the public information set. The managers from the private information set are expected to have higher performance. Second, there is managers' information collection ability. There is huge amount of information today. Colleting all the information is both time-consuming and expensive. The mangers that can efficiently collect the relevant information are expected to deliver higher abnormal returns (alphas), since they possess more useful information for investment decision than other managers. Third, there is managers' ability to analyze the information on hand. Only well analyzed and interpreted information can produce higher abnormal returns. Those who correctly analyze the information are actually in a superior information set. Therefore, depending on the information they have and the ability to collect and analyze the information, we may have several information sets, for example, badly-analyzed public information set, well-analyzed public information set, and private information set. The differentiated ability to acquire private information and analyze public information may lead to more information sets in an inefficient market, where information is not well transmitted and absorbed.

Based on the information set the managers are from, the expected performance and the investment risk in that information set will be different. Assuming that alphas from the same information set follow a normal distribution, alphas of all the managers will thus form a finite normal mixture distribution. This is the possible reason that we observe multimodal shape and non-normal characteristics of the distribution of all the alphas when the alpha is theoretically expected to be normally distributed.

In our model, we still assume that the alphas are normally distributed but arise from different normal distributions that are corresponding to different information sets. Under this assumption we may observe both a multimodal shape and fat tail in the distribution of alphas. As long as we can identify the number of information sets in the distribution, we know the number of performance groups. That is the way that we determine the number of performance groups.

There are some advantages of the funds rating method that we propose. First, the number of performance groups is not arbitrarily fixed. It is estimated from the empirical cross-sectional distribution of alphas. The number of performance groups is interpreted as the number of information sets in the fund market. Second, the performance group of the fund is determined by the posterior probability of the fund in the estimated distribution.

So it is not fixed by the preset cut-off percentiles. Third, the number of funds in each performance group is not fixed, which may change from period to period, depending on the distribution of all the alphas in the study.

**3. Parametric Formulation of Finite Normal Mixture Distribution Model**

We assume that the non-normal features are caused by the group-structure of the data. In our fund performance study, we suspect that there are more than one performance group arising from the distinct information sets. In section 2 we have justified that when market information is segmented or exploited at different levels, there are possibly more than one performance group. Thus there exists a group structure in the distribution of alphas. When there is a group structure in the data, finite normal mixture model is a natural way to model the unknown distribution. In this model, it is expected that the more information that the manager has, the higher the alpha.

In this section, we will formulate the finite mixture distribution model based on McLachlan and Basford (1988). Let $Y_1,...,Y_n$ denote a random sample of size n. $Y=(Y_1,...,Y_n)^T$ is a column vector representing the entire random sample. $Y_j$ is the random variable corresponding to the alpha of fund j. And its probability density function is $f(y_j)$ (j=1,…,n). A realization of the random sample is denoted by $y=(y_1,...,y_n)^T$. $y_j$ is the alpha of fund j that we observed, which is estimated by the RBSA measure.

In finite mixture model, the density function $f(y_j)$ is a summation of finite component densities, $f_i(y_j)$. It is written in the form,

$$f(y_j)=\sum_{i=1}^{g}\pi_i f_i(y_j) \tag{1}$$

where $\pi_i\,(i=1,...,g)$ is the mixing proportion or can be called component weight. They are nonnegative and sum to one, i.e.

$$\begin{aligned} &0\le\pi_i\le 1,(i=1,...,g)\\ &\sum_{i=1}^{g}\pi_i=1. \end{aligned} \tag{2}$$

Here g is the number of components, and $f_i(y_j)$ is the component density. $f(y_j)$ is a g-component mixture density, and the corresponding distribution function is denoted by $F(y_i)$.

There is also a component label variable $Z_j$, which is a vector with g elements. The *ith* element of $Z_j$ is $Z_{ij}(i=1,...,g;\ j=1,...,n)$, which is an indicator variable being one or zero. In our fund performance study, we assume that there are g performance groups. If the fund j is from performance group i, then $Z_{ij}$ is one, otherwise $Z_{ij}$ is zero. In the model, $\pi_i$ is interpreted as the proportion of funds that belong to performance group i

($i = 1,...,g$). It is straightforward that $\pi_i$ is also the probability that fund j is generated from performance group i if we don't know the group information. Therefore, $Z_j$ follows a multinomial distribution,

$$pr\{Z_j = z_j\} = \pi_1^{z_{1j}} \pi_2^{z_{2j}} ... \pi_g^{z_{gj}} \tag{3}$$

where $z_j$ is a realization of $Z_j$. We look at $f(y_j)$ and $f_i(y_j)$ again with the component label variable $Z_j$. Given the group information that fund j is from group i, so $Z_{ij} = 1$, $f_i(y_j)$ can be viewed as the conditional probability density of $Y_j$. And $f(y_j)$ can be viewed as the unconditional density without group information.

The finite mixture model in (1) can be viewed as a semi-parametric model between the fully parametric model as represented by a single parametric family (g=1) and a nonparametric kernel model (g=n). But the single parametric model is usually inadequate to describe the actual distribution. In the finite mixture model, $f_i(y_j)$ is from a parametric family and specified by $f_i(y_j;\theta_i)$, where $\theta_i$ is a set of unknown parameters in the component density. The finite mixture model thus can be written as,

$$f(y_j;\Psi) = \sum_{i=1}^{g} \pi_i f_i(y_j;\theta_i) \tag{4}$$

where $\Psi = (\pi_1,...,\pi_{g-1},\xi^T)^T$ contains all the unknown parameters in the model, and $\xi$ is a vector containing all the parameters in component densities, from $\theta_1$ to $\theta_g$. Since the summation of $\pi_i$ ($i = 1,...,g$) is one, we only need to estimate g-1 mixing proportions. We arbitrarily leave out the $g^{th}$ mixing proportion, $\pi_g$. The parameter space of $\Psi$ is denoted by $\Omega$, the parameter space of $\theta_i$ is denoted by $\Theta$.

In our study we assume that there are g information sets in the market, arising from the differentiated ability of acquiring and analyzing both public and private information. The different information sets lead to heterogeneous performance groups. We further assume that the alphas of each group follow a normal distribution, denoted as $N(\mu_i,\sigma_i^2)$. The finite mixture model views the alphas of funds as having been generated from one of the g performance groups with mean and variance as,

$$\alpha_j = \mu_i + \varepsilon_{ij}, Var(\varepsilon_{ij}) = \sigma_i^2, (i = 1,...g;\ j = 1,...n)$$

where $\alpha_j$ ($j = 1,...,n$) is the performance of fund j. $\mu_i$ ($i = 1,...,g$) is interpreted as the expected performance in this performance group. $\sigma_i^2$ ($i = 1,...,g$) is interpreted as the investment risk of a fund in performance group i. The higher the $\sigma_i^2$, the higher the risk to

invest in this kind of funds. $\varepsilon_{ij}$ $(i=1,...,g;\ j=1,...,n)$ follows a normal distribution with mean zero and variance $\sigma_i^2$.

With the above assumption, the finite mixture model can be written as,

$$f(y_j;\Psi)=\sum_{i=1}^{g}\pi_i f_i(y_j;\theta_i)$$

where

$$f_i(y_j;\theta_i)=\phi(y_j;\mu_i,\sigma_i^2)=(2\pi\sigma_i^2)^{-\frac{1}{2}}\exp\{-\frac{1}{2}(y_j-\mu_i)^2/\sigma_i^2\}\ . \qquad (5)$$

The vector $\Psi$ is $(\pi_1,...,\pi_{g-1},\mu_1,...\mu_g,\sigma_1^2,...\sigma_g^2)^T$, $(i=1,...,g;\ j=1,...,n)$, containing 3g-1 parameters.

**4. EM Algorithm for Finite Normal Mixture Model**

To obtain parameter estimates in the finite normal mixture model, Redner and Walker (1984) recommended the application of the expectation-maximization (EM) algorithm, synthesized in the celebrated paper by Dempster et al (1977). As pointed out by Woodward and Sain, (2003), the EM algorithm is an effective tool to deal with various missing data problems. In the fund performance study, there exists missing information, i.e. the component label vector $Z_i$, so we can formulate the maximum likelihood estimation problem under EM framework.

To estimate the maximum likelihood estimator (MLE) of $\Psi$ with observed data, the likelihood function is written as,

$$L(\Psi)=\prod_{j=1}^{n}\sum_{i=1}^{g}\pi_i\phi_i(y_j;\mu_i,\sigma_i^2) \qquad (6)$$

and its log likelihood is given by,

$$logL(\Psi)=\sum_{j=1}^{n}\log\{\sum_{i=1}^{g}\pi_i\phi_i(y_j;\mu_i,\sigma_i^2)\} \qquad (7)$$

To find MLE, we take first-order derivatives of the log likelihood function,

$$\frac{\partial logL(\Psi)}{\partial\Psi}=0 \qquad (8)$$

Since it is in a summation form of the component density function, it poses a computational difficulty. However, it is straightforward to find MLE, under EM framework.

We introduce the component label vector $z_1,...,z_n$ to the observed data. Formulating the finite normal mixture model under the EM framework, the observed data vector $y=(y_1,...,y_n)$ is viewed as incomplete data because the component label vectors, $z=(z_1,...,z_n)$, are not available. In our study, each $y_j$ of fund j is regarded as being from one of the performance groups, corresponding to one of components in the finite normal mixture model. $z_j$ is a g dimensional indicator vector for fund j. $z_{ij}=1$ means that the fund j is from performance group i. Zero means that the fund is not from this group. Therefore the complete data vector is

$$y_c=(y,z) \tag{9}$$

where the component vector $z=(z_1,...,z_n)$ is the realization of the random vector $Z_1,...,Z_n$. As reported in (3) the vector $Z_j$ follows a multinomial distribution.

Therefore, viewing the component label vector as part of completer data, we can rewrite the likelihood function as,

$$L_c(\Psi)=\prod_{j=1}^{n}\prod_{i=1}^{g}\{\pi_i\phi_i(y_j;\mu_i,\sigma_i^2)\}^{z_{ij}}. \tag{10}$$

The log likelihood is written as,

$$logL_c(\Psi)=\sum_{j=1}^{n}\sum_{i=1}^{g}z_{ij}\{\log\pi_i+\log\{\phi_i(y_j;\mu_i,\sigma_i^2)\}\} \tag{11}$$

where $z_{ij}$ is linear in the log likelihood function. Now it is much easier to calculate it iteratively in (11). $z_{ij}$ is treated as missing data when we apply the EM algorithm to the problem. There are two steps: E for expectation step, and M for maximization step. We use E step to deal with the additional missing data $z_{ij}$. Given the observed data y, we take conditional expectation of the complete data log likelihood $logL_c(\Psi)$ using $\Psi^{(k)}$, which is MLE of $\Psi$ in the k$^{th}$ iteration. $\Psi^{(0)}$ is the initial value that we specified in the initial step.

On the first iteration we need to calculate the expectation of complete data log likelihood given y and $\Psi^{(0)}$. It is expressed as,

$$Q(\Psi;\Psi^{(0)})=E_{\Psi^{(0)}}\{\log L_c(\Psi)\,|\,y\}. \tag{12}$$

The subscript under the expectation operator E means that the expectation is also depending on $\Psi^{(0)}$, which changes over time in iterations. After k$^{th}$ iteration, it is written as,

$$Q(\Psi;\Psi^{(k)}) = E_{\Psi^{(k)}}\{\log L_c(\Psi)\big|y\} \tag{13}$$

To calculate the conditional expectation of the complete data log likelihood, we only need to calculate the conditional expectation of $Z_{ij}$ given the observed data y, because $Z_{ij}$ is linear in the log like likelihood function (13). Here, $Z_{ij}$ is the random variable corresponding to the realized value $z_{ij}$.

$$E_{\Psi^{(k)}}(Z_{ij}\big|y) = pr_{\Psi^{(k)}}\{Z_{ij}=1\big|y\} = \tau_i(y_j;\Psi^{(k)}) \tag{14}$$

where $\tau_i(y_j;\Psi^{(k)})$ is the posterior probability that fund j belongs to group i given observed data y. Because the expected probability of $Z_{ij}=1$ is just $\pi_i$, according to the Bayesian theorem, it is straightforward to find $\tau_i(y_j;\Psi^{(k)})$, which is given by

$$\begin{aligned}\tau_i(y_j;\Psi^{(k)}) &= \pi_i^{(k)}\phi_i(y_j;\mu_i^{(k)},\sigma^{2^{(k)}}_i)/\ f(y_j;\Psi^{(k)})\\ &= \pi_i^{(k)}\phi_i(y_j;\mu_i^{(k)},\sigma^{2^{(k)}}_i)/\sum_{c=1}^{g}\pi_c^{(k)}\phi_c(y_j;\mu_c^{(k)},\sigma^{2^{(k)}}_c).\end{aligned} \tag{15}$$

Since $\phi(y_j;\mu_i,\sigma_i^2) = (2\pi\sigma_i^2)^{-\frac{1}{2}}\exp\{-\frac{1}{2}(y_j-\mu_i)^2/\sigma_i^2\}$, after substituting it into the posterior probability in (15), we obtain,

$$\tau_i(y_j;\Psi^{(k)}) = \frac{\pi_i^{(k)}\sigma_i^{-1^{(k)}}\exp\{(y_j-\mu_i^{(k)})^2/\sigma_i^{2^{(k)}}\}}{\sum_{c=1}^{g}\pi_c^{(k)}\sigma_c^{-1^{(k)}}\exp\{(y_j-\mu_c^{(k)})^2/\sigma_c^{2^{(k)}}\}}, (k=0,1,....; i=1,...,g; j=1,...,n). \tag{16}$$

Therefore, the conditional expectation of complete data log likelihood is given by

$$Q(\Psi;\Psi^{(k)}) = \sum_{j=1}^{n}\sum_{i=1}^{g}\tau_i(y_j;\Psi^{(k)})\{\log\pi_i + \log\{\phi_i(y_j;\mu_i,\sigma_i^2)\}\}\ . \tag{17}$$

With the observed data y and parameters, which are estimated from k$^{th}$ maximization $\Psi^{(k)}$, we take conditional expectation of complete data log likelihood. This is E step. Then in the M step, we maximize $Q(\Psi;\Psi^{(k)})$ with respect to $\Psi$ over the parameter space $\Omega$ to get the updated $\Psi^{(k+1)}$. The calculation of updated mixing proportions, $\pi_i^{(k+1)}$, of $\pi_i$ are independent of the calculation of updated parameter , $\xi^{(k+1)}$, of $\xi$ , containing the parameters in component densities. If the missing information, $z_{ij}$ is known, the complete data MLE of $\pi_i$ is simply,

$$\hat{\pi}_i = \sum_{j=1}^{n} z_{ij} / n, (i = 1, ..., g). \tag{18}$$

Since $z_{ij}$ is not known, we use, $\tau_i(y_j; \Psi^{(k)})$ to replace $z_{ij}$ in the above estimation, which is the conditional expectation of $z_{ij}$ in complete data log likelihood. The updated mixing proportion is,

$$\pi_i^{(k+1)} = \sum_{j=1}^{n} \tau_i(y_j; \Psi^{(k)}) / n, (i = 1, ..., g). \tag{19}$$

When updating mixing proportions $\pi_i$ on the $(k+1)^{th}$ iteration, we sum up all the posterior probabilities that the fund belongs to performance group i. Each $y_j$ contributes to the update.

Regarding the update of $\xi$ on the M step in $(k+1)^{th}$ iteration, we take the first order derivative of the conditional expectation log likelihood with respect to parameters, and then solve the equations to find out MLE of $\xi^{(k+1)}$ in the $(k+1)^{th}$ iteration:

$$\frac{\partial Q(\Psi; \Psi^{(k)})}{\partial \xi} = 0, \tag{20}$$

gives,

$$\frac{\partial Q(\Psi; \Psi^{(k)})}{\partial \xi} = \sum_{j=1}^{n} \sum_{i=1}^{g} \tau_i(y_j; \Psi^{(k)}) \frac{\partial \log\{\phi_i(y_j; \mu_i, \sigma_i^2)\}}{\partial \xi} = 0. \tag{21}$$

$\xi$ contains the parameters $(\mu_i, \sigma_i^2), (i = 1, ..., g)$.

Since $\phi(y_j; \mu_i, \sigma_i^2) = (2\pi\sigma_i^2)^{-\frac{1}{2}} \exp\{-\frac{1}{2}(y_j - \mu_i)^2 / \sigma_i^2\}$, we have,

$$\begin{aligned} &\frac{\partial Q(\Psi; \Psi^{(k)})}{\partial \mu_i} = \sum_{j=1}^{n} \tau_i(y_j; \Psi^{(k)})(y_j - \mu_i) = 0, (i = 1, ..., g), \\ &\frac{\partial Q(\Psi; \Psi^{(k)})}{\partial \sigma_i^2} = \sum_{j=1}^{n} \tau_i(y_j; \Psi^{(k)})(\frac{1}{\sigma_i^2} - \frac{(y_j - \mu_i)^2}{\sigma_i^4}) = 0, (i = 1, ..., g). \end{aligned} \tag{22}$$

Then, MLE of $(\mu_i, \sigma_i^2)$, $(i = 1, ..., g)$ are obtained as,

$$\mu_i = \frac{\sum_{j=1}^{n} \tau_i(y_j; \Psi^{(k)}) y_j}{\sum_{j=1}^{n} \tau_i(y_j; \Psi^{(k)})}, (i = 1, ..., g), \tag{23}$$

$$\sigma_i^2 = \frac{\sum_{j=1}^{n} \tau_i(y_j;\Psi^{(k)})(y_j-\mu_i)^2}{\sum_{j=1}^{n} \tau_i(y_j;\Psi^{(k)})}, (i=1,...,g).$$

Since we know that $\pi_i^{(k+1)} = \sum_{j=1}^{n} \tau_i(y_j;\Psi^{(k)})/n, (i=1,...,g)$ in (19), we can simplify the above two equations as,

$$\mu_i^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_i(y_j;\Psi^{(k)})y_j}{n\pi_i^{(k+1)}}, (i=1,...,g), \quad (24)$$

$$\sigma_i^{2^{(k+1)}} = \frac{\sum_{j=1}^{n} \tau_i(y_j;\Psi^{(k)})(y_j-\mu_i)^2}{n\pi_i^{(k+1)}}, (i=1,...,g).$$

Note that k+1 denotes the updated parameters for the k+1$^{th}$ iteration. We repeat the E step and M step alternatively until the estimates of parameters in Ω converge. A desirable feature of the EM algorithm is that the solutions are in closed form for the finite normal mixture model.

***Procedural steps of EM algorithm are summarized below:***

1. Choose initial values of $(\pi_i^{(0)}, \mu_i^{(0)}, \sigma_i^{2^{(0)}}), (i=1,...,g)$, given the g-component finite normal mixture model.
2. Estimate posterior probability that fund j $(j=1,...,n)$ belongs to performance group i $(i=1,...,g)$ given the observed data y and $\Psi^{(k)}$ that are estimated parameters in kth iteration.

$$\tau_i(y_j;\Psi^{(k)}) = \frac{\pi_i^{(k)}\sigma_i^{-1^{(k)}} \exp\{(y_j-\mu_i^{(k)})^2/\sigma_i^{2^{(k)}}\}}{\sum_{c=1}^{g} \pi_c^{(k)}\sigma_c^{-1^{(k)}} \exp\{(y_j-\mu_c^{(k)})^2/\sigma_c^{2^{(k)}}\}},$$

$(k=0,1,....; i=1,...,g; j=1,...,n).$

3. Update $(\pi_i^{(k)}, \mu_i^{(k)}, \sigma_i^{(k)}), (i=1,...,g)$ by the following equations in sequence,

$$\pi_i^{(k+1)} = \sum_{j=1}^{n} \tau_i(y_j;\Psi^{(k)})/n, (i=1,...,g),$$

$$\mu_i^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_i(y_j;\Psi^{(k)})y_j}{n\pi_i^{(k+1)}}, i=(1,...,g),$$

$$\sigma_i^{2^{(k+1)}} = \frac{\sum_{j=1}^{n} \tau_i(y_j; \Psi^{(k)})(y_j - \mu_i)^2}{n\pi_i^{(k+1)}}, i = (1,...,g) .$$

4. Repeat step 3 and step 4 until the difference between $(\pi_i^{(k+1)}, \mu_i^{(k+1)}, \sigma_i^{(k+1)}), (i = 1,...,g)$ and $(\pi_i^{(k)}, \mu_i^{(k)}, \sigma_i^{(k)}), (i = 1,...,g)$ is smaller than the preset tolerance level.

5. Determination of the Number of Components
by Parametric Bootstrap Procedures

In the procedures outlined in the previous section, we have to specify the number of components in the first step to initiate the iteration. This is a model specification problem. In some situations, the number of components is given as a priori information. However, in other situations, the number of components has to be inferred from observed data along with other estimates of density function. In the fund performance study, the group information is not known and is of particular interest for us to rate funds. The number of components in the finite normal mixture model indicates the number of performance groups that exist among all the funds. In addition, the number of components directly affects the classification of funds. In an extreme example, suppose that our test shows there is only one performance group, and then it is not necessary to group funds. The abnormally high or low alphas are just the results of "Luck".

There are three approaches to estimate the number of components in the finite normal mixture model. The first method is nonparametric by investigating the number of modes in an estimated kernel density. We know that multimodal shape is a strong implication of mixture model. Roeder (1994) argued that if there is no priori information about the number of components and component densities, it is appropriate to assess the number of modes. Inferential procedures to asses the number of modes include Titterington et al. (1985) and Silverman (1981, 1986), in which Silverman used a kernel method to estimate the density function and develop a technique to assess the number of modes. Other studies using the number of modes include Hartingan and Mohanty (1992), Wong (1985), and Fisher, et al. (1994). But there is an obvious drawback of this approach. If the means of component densities are not sufficiently separate enough, the number of modes are less than the number of components. Therefore it is difficult to identify the right number of components.

The second stream is based on penalized log likelihood, such as AIC and BIC. As the log likelihood increases with the addition of a component to the finite normal fixture model, the log likelihood is penalized by the subtraction of a term that penalizes the model for the number of parameters in it. Using this method, the results are acceptable as discussed in Solka et al. (1998). But the main purpose of this approach is for density estimation, not the identification of the number of terms. In addition, it produces no confidence of results, so we have no idea of Type I error if we reject the null hypothesis.

In our study, we assess the number of components by hypothesis test, using likelihood ratio as the test statistic. The approach focuses on group finding, and most importantly it

provides a p value to assess the confidence about the number of components. In the finite normal mixture model, the likelihood ratio test statistic is,

$$-2\log(\lambda) = 2\{\log L(\hat{\Psi}_1) - \log L(\hat{\Psi}_0)\}, \qquad (25)$$

where $\hat{\Psi}_0$ and $\hat{\Psi}_1$ are MLE of $\Psi$ under $H_0 : g = g_0$ and $H_1 : g = g_1$ respectively Usually we increase the number of components $g_0$ one by one in sequence to see if the increase in log likelihood starts to fade away after some threshold value $g_0$. After adding a new component into finite normal mixture model, if the increase of log likelihood is not significant, then we can conclude that there is no sufficient evidence to reject the hypothesis that there are $g_0$ components in the model. From the above analysis, we know that as long as we know the sampling distribution of the likelihood ratio test statistic, $-2\log(\lambda)$, we can proceed to hypothesis test, and finally identify the number of components.

Unfortunately, in the finite normal mixture model, regularity conditions (Cramer, 1946) do not hold for $-2\log(\lambda)$ to have its usual asymptotic null distribution of Chi-square, where the degrees of freedom are equal to the difference of the number of parameters under null hypothesis and the number of parameters under alternative hypothesis. In the work by Titterington, et al. (1985) and McLachlan and Basford (1988), it is well discussed that conventional asymptotic results for the null distribution of the likelihood ratio test statistic do not hold because the null hypothesis lies on the boundary of the alternative hypothesis (in null hypothesis one mixing proportion is specified as zero).

To rescue it, parametric bootstrap procedures proposed by McLachlan (1992) are used to assess the p value of likelihood ratio test statistic, $-2\log(\lambda)$. Simulation is needed in this occasion. Feng and McColloch (1996) pointed out that this approach leads to valid statistical inference. Wolfe (1971) proposed a modified likelihood ratio test statistic by the rule of thumb, but McLachlan (1987) showed the results may not be applicable in heteroscedastic case where component variances are unequal.

In finite normal mixture model, we test,

$$H_0 : g = g_0 \text{ versus } H_1 : g = g_1.$$

We let $g_1 = g_1 + 1$ in order to find the smallest g that is consistent with the data. Since the null distribution is unknown, we use parametric bootstrap procedures to asses the p value of likelihood ratio test statistic, $-2\log(\lambda)$. Bootstrap samples are generated from the finite normal mixture model with $\Psi$ replace by the MLE, $\hat{\Psi}_0$, which is estimated under null hypothesis by EM algorithm with the observed data. Then we fit the bootstrap sample under null hypothesis and alternative hypothesis respectively by EM, to obtain the bootstrapped likelihood ratio value, $-2\log(\lambda)^{(b)}$, where b means the $b^{th}$ likelihood ratio value from the bth bootstrapped sample. We repeat the sampling for a number of times B,

so we have a sequence of likelihood ratios $\{-2\log(\lambda)^{(b)}\}$. The sequence of values provides an approximation of the unknown null hypothesis distribution. Then we refer the original likelihood ratio, computed from the observed data, to the sequence $\{-2\log(\lambda)^{(b)}\}$. We find the p value of $-2\log(\lambda)$ as,

$$p = 1 - \frac{j}{B+1} \tag{26}$$

where j is the number of replicated likelihood ratio values that are smaller than the original likelihood ratio. If we reject null hypothesis under $g = g_0$, then we can increase the number of components under null hypothesis by one, and move forward to $H_0 : g = g_0 + 1$ versus $H_1 : g = g_0 + 2$ until we don't have sufficient evidence to reject null hypothesis. The threshold g is the number of performance groups in our study.

To facilitate programming, we outlined the parametric bootstrap procedures as follows:

1. Given the observed data y, fit the original data under $H_0 : g = g_0$ and $H_1 : g = g_0 + 1$ respectively by the EM algorithm to get estimates $\widehat{\Psi}_0$ and $\widehat{\Psi}_1$.
2. Substitute $\Psi$ in finite normal mixture model with the estimated $\widehat{\Psi}_0$ and $\widehat{\Psi}_1$ to get probability density function under null hypothesis and alternative hypothesis respectively.
3. From the density functions we compute the original likelihood ratio value, $-2\log(\lambda) = 2\{\log L(\hat{\Psi}_1) - \log L(\hat{\Psi}_0)\}$.
4. Take a bootstrap sample from the finite normal mixture model with parameters, $\Psi$, replaced by $\widehat{\Psi}_0$ that we estimate in step 1.
5. Fit bootstrap sample we obtained in step 4 under $H_0 : g = g_0$ and $H_1 : g = g_0 + 1$ respectively by EM algorithm to get estimate of $\widehat{\Psi}_0^{(b)}$ and $\widehat{\Psi}_1^{(b)}$. The superscript b represents the estimate from b$^{th}$ bootstrap sample.
6. Substitute $\Psi$ in finite normal mixture model with the estimated $\widehat{\Psi}_0^{(b)}$ and $\widehat{\Psi}_1^{(b)}$ to get probability density function under null hypothesis and alternative hypothesis respectively.
7. From the density functions we compute the likelihood ratio value, $-2\log(\lambda)^{(b)} = 2\{\log L(\widehat{\Psi}_1^{(b)}) - \log L(\widehat{\Psi}_0^{(b)})\}$.
8. Repeat step 4 through step 7 for B times to get a sequence $\{-2\log(\lambda)^{(b)}\}$ for $b = 1, ..., B$.
9. Order the sequence of likelihood ratios, and then count the number of values that are smaller than $-2\log(\lambda)$, which is the original likelihood ratio, in step 3.
10. Find p value of $-2\log(\lambda)$ as $1 - \frac{j}{B+1}$, where j is the number of counts in step 9.

Usually large B is required to get a precise p value. However, the amount of computation involved is considerable. We choose B as 200 as the number of the bootstrapped samples. The B is sufficient, because our main concern is to see whether we can reject the null hypothesis not to get the precise p values.

## 6. Fund Rating Procedures

We use finite normal mixture model to study the distribution of alphas attempting to find the number of performance groups and assign a rating to each fund. In our research, we provide a new direction of fund rating method that is more flexible and theoretically solid than current fund rating method, like Morningstar's method. The model is implemented in the following steps.

***Step One: Normality Check***

We will check the normality of the distributions first. If they are normal then no further steps are necessary. It implies that there are no superior or inferior funds in the market. The abnormal negative or positive alphas we observed in the last period are just the consequence of "Luck". In other words, the managers happened to have picked the right stocks and correctly timed the market. If the distributions show non-normal features, such as multimodal shape in kernel density, then it is a good indication of group structure in the data. This may be caused by the different information sets that managers are from. We can also test the normality by formal tests, such as the Jarque-Bera test and Lilliefors test. When we find that the distributions can not be described by a univariate normal distribution, the natural way to model it is a finite normal mixture model. The model provides an intuitively appealing interpretation about the number of components and the expected values and the variances of component densities. They are interpreted as the number of performance groups, the expected performance of the fund, and the expected investment risk of the fund respectively. Note that the expected performance here is not expected alpha of all the funds, instead it is the expected alpha of the funds in the performance group that the fund is from. In addition, the model provides the posterior probability that the fund belongs to each group. With this information we can group and rate the funds.

***Step Two: Specification of the Finite Normal Mixture Model***

Before estimating the model parameters, we have to specify the number of components. This is theoretically difficult. A number of approaches are proposed. We use the parametric bootstrap procedures outlined in section 5.5 to identify the number of components. We assess the p value based on the empirical distribution of likelihood ratio. In searching the appropriate number of components we increase g in the null hypothesis gradually one by one until we find the smallest threshold g that is consistent with data.

***Step Three: Estimation of the Finite Normal Mixture Model***

After having fixed the number of components in the model, then we proceed to estimate the model. We use EM algorithm outline in section 4 to solve out the likelihood function. This is not only for straightforward computation of MLE of $\Psi$, but also for the intuitive interpretation of group information. We introduced a component label vector $Z_j$ with value of one or zero, indicating whether the fund was generated for the performance group or not. There are two steps in EM algorithm. In the E step we take the conditional expectation of $Z_j$, given y and current $\Psi^{(k)}$ in finite normal mixture model, to obtain the posterior probability. Then we proceed to the M step to update

$(\pi_1,...,\pi_g;\mu_1,...,\mu_g;\sigma_1^2,...,\sigma_g^2)$ sequentially with posterior probabilities of all the funds. We repeat the E step and the M step until estimates converge.

The results have an intuitively appealing interpretation. $\pi_i$ is interpreted as the proportion of funds in performance group $i, i=1,...g$. $\mu_i$ is interpreted as the expected alpha for performance group $i, i=1,...g$. $\sigma_i^2$ is interpreted as the investment risk of funds in performance group $i, i=1,...g$. The higher the $\sigma_i^2$, the higher the risk. The high $\sigma_i^2$ implies that the performance is volatile in this group. We also have the posterior probability that each fund belong to each group, which provides us with a basis for grouping and rating.

***Step Four: Fund Rating***

We rank $\mu_i(i=1,...,g)$, which is the expected alpha of performance group i. The funds in the group that has the highest ranking are viewed as superior funds, whereas the funds in the group that has the lowest ranking are viewed as inferior funds. The differences of alphas in the performance group are regarded as random effects. Thus we consider the funds in the same performance group have the same expected performance.

**References**


Arnott, A. C., 1996, *Morningstar mutual funds*. December 6.

Cai, J, K. C. Chan, and T. Yamada, 1997, The performance of Japanese mutual funds, *Review of Financial Studies* 10, 237-273.

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

Chan, K. C., Nai-fu Chen, and David A. H., 1985, An explanatory investigation of the firm size effect, *Journal of Financial Economics* 14, 451-471.

Chiang, A. C., 1984, Fundamental methods of mathematical economics.

Conover, W. J., 1980, *Practical Nonparametric Statistics.* Wiley.

Christopherson, J. A., Spring 1995, Equity Style Classifications. *Journal of Portfolio Management*, 32-43.

Cochrane, J., 1996, A cross-sectional test of production-based asset pricing model, *Journal of Political Economy* 104, 572-621.

Cramer, H., 1946. *Mathematical methods of statistics.* (Princeton University Press, Princeton).

Cummisford, R. and Lummer, S., Oct. 1996, Controlling the Limitations of Style Analysis. *Journal of Financial Planning*, 70-76.

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

Epanechnikov, V. K., 1969, Non-parametric estimation of a multivariate probability density, *Theory of Probability and Applications* 14, 153-158.

Fisher, N.I., E. Mammen, and J.S. Marron, 1994, Tesing for multimodality, *Computational Statistics and Data Analysis* 18, 499-512.

Glosten, L., and R. Jagannathan, 1994, A contingent claims approach to performance evaluation, *journal of Empirical Finance* 1, 133-166.

Hartingan, J., and S. Mohanty, 1992, The run test for multimodality, *Journal of Classification* 9.

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

Karmarkar, N., 1984, A new polynomial-time algorithm for linear programming, *Combinatorica* 4, 373-395.

Kosowski, R., A. Timmermann, H. White, and R. Wermers, 2001, Can mutual fund "stars" really pick stocks? evidence from a bootstrap analysis, *Working paper*.

Kosowski, R., A. Timmermann, H. White, and R. Wermers, 2001, Can mutual fund "stars" really pick stocks? new evidence from a bootstrap analysis, *Working paper*.

McLachlan, G.J., 1987, On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture, *Applied Statistics* 36, 318-324.

McLachlan, G.J., 1992, Cluster analysis and related techniques in medical research., *Statistical Method in Medical Research* 1, 27-49.

McLachlan, G.J., and K.E. Basford, 1988. *Mixture models: inference and applications to clustering* (Marcel Dekker, New York).

Nesterov, Y., and A. Nemirovski, 1994, Interior-point polynomial algorithms in convex programming, *SIAM*.

Priebe, C.E., 1994, Adaptive mixture density estimation, *Journal of the American Statistical Association* 89, 796-806.

Redner, R.A., and H.F. Walker, 1984, Mixture densities, maximum likelihood and Em algorithm., *SIAM Review* 26, 195-239.

Roeder, K., 1994, A graphical technique for determining the number of components in a mixture of normals., *Journal of the American Statistical Association* 89, 487-495.

Roll, R., 1977, A critique of the asset pricing theory's tests, *Journal of Financial Economics* 4, 129-176.

Sharpe, W., 1998, Morningstar's risk-adjusted ratings. *Financial Analysts Journal* 54, 21-33

Silverman, B.W., 1981, Using kernel density estimates to investigate multimodality, *Journal of the Royal Statistical Society, Series B* 43, 97-99.

Silverman, B.W., 1986. *Density estimation for statistics and data analysis* (Chapman and Hall, London).

Solka, J.L., E.J. Wegman, C.E. Priebe, W.L. Poston, and W. Rogers, 1998, Mixture structure analysis using Akaike criterion and the bootstrap, *Statistics and Computing* 8, 177-188.

Tao, J., N. Z. Shi, and S. Y. Lee, 2004, Drug risk assessment with determining the number of sub-populations under finite mixture normal models, *Computational Statistics and Data Analysis* 46, 661-676.

Titterington, D.M., A.F.M. Smith, and U.E. Makov, 1985. *Statistical analysis of finite mixture distributions* (Wiley, New York).

Wolfe, J.H., 1971, A Monte Carlo study of sampling distribution of the likelihood ratio for mixtures of multinormal distributions, *Technical Bulletin STB* 72, San Diego: US Naval Personnel and Training Research Laboratory.

Wong, W.A., 1985, A bootstrap testing procedure for investigating the number of subpopulations., *Journal of Statistical Computation and Simulation* 22, 99-112.

Woodward, W.A., and S.R. Sain, 2003, Testing for outliers from a mixture distribution when some data are missing, *Computational Statistics and Data Analysis* 44, 193-210.

Yudin, D.B., and A. Nemirovski, 1976, Information complexity and efficient methods for the solution of convex extremal problems, *Matekon* 13, 3-25.