ESTIMATING RETURN ON INITIAL PUBLIC OFFERING USING MIXTURES OF REGRESSIONS

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Xiyuan Liu

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Statistics

May 2015

Fargo, North Dakota

North Dakota State University

Graduate School

Title

ESTIMATING RETURN ON INITIAL PUBLIC OFFERING USING MIXTURES OF REGRESSIONS

By

Xiyuan Liu

The supervisory committee certifies that this thesis complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Tatjana Miljkovic
Chair

Rhonda Magel
Co-Chair

Yarong Yang

Saleem Shaik

Approved:

06 May 2015
Date

Rhonda Magel
Department Chair

ABSTRACT

Financial advisors working in a stock exchange market are often faced with a situation to convince a client of merits of investing in a company that just entered the market. To predict company's return based on its revenue, a simple linear regression may be used. This thesis finds that a model based on a mixture regressions is superior over a simple linear regression. The error term in each regression component is assumed to follow standard Gaussian distribution.

The data is tested on 116 companies that entered the market as Initial Public Offering (IPO). A 2-component mixture regressions is found to provide the best fit for the data. A simulation study is conducted to verify the performance of this model. Optimum number of components is found using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) as well as the parametric bootstrapping of the likelihood ratio test statistics.

ACKNOWLEDGEMENTS

This thesis was supported by Dr. Tatjana Miljkovic, who provided insight and expertise that greatly assisted the thesis.

I thank Matthew Warner for comments and language support that greatly improved the manuscript.

I would also like to show my gratitude to Dr. Rhonda Magel, Dr. Yarong Yang, and Dr. Saleem Shaik for their time and feedback.

This thesis would not have been done without my family behind me. I am indebted for the mental and physical support they have provided that has spurred me on through tough times during not only my thesis but also my entire graduate study. I am thankful for their love and prayers.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

1. INTRODUCTION

The initial offer of shares of stocks by a company to the public is referred to as Initial Public Offering (IPO). In the financial language, this is also known as "going public." The owner of the company decides to give up a part of this ownership to stockholders. This is usually a good time for the company since the company has grown enough and became successful on the market so that it can continue to grow through expansion by obtaining sufficient cash through the sale of stock. In a stock market a company can raise money by issuing either equity or a debt. If a company has not delivered equity to the public, it is known as an IPO. Two different kind of companies exist on the market: public and private.

Public companies have thousands of shareholders and the market determines the value of the entire company through daily trading. Public companies operate in most sectors of the economy. For example, Williams-Sonoma sells household goods such as sofas and refrigerators, whereas Sky Financial Group, another public company, offers financial products such as commercial and business loans. Some large public companies have very diverse services and products, such as Coca Cola which sells food products and partners with restaurants. All public companies have an appointed board of directors that reports quarterly financial information to Securities and Exchange Commission (SEC) in the United States. Of course, the defining feature of a public company is that anybody can invest in a public company since its stock is traded in the open market. This feature makes investing in a public company attractive for many investors. In contrast to public companies, private companies typically have a smaller number of shareholders and the owners of the company do not need to disclose publicly all information about company activities.

Private companies usually are smaller than public companies, but it is not uncommon to see large private company as well. Such as Koch Industries, a global company owned by Koch family, which sells several types of products including transportation fuels, building, and electronic connectors. Similar to public companies, private companies operate in many sectors of the economy. For example, Mars sells pet foods, and Bechtel, mainly owned by Bechtel family, is a civil engineering company that provides several services such as Financing and Equity Investment, Engineering, and consulting (Nuclear Power, Oil, and Chemicals). Unlike public companies, investing private companies is sometimes difficult or impossible because a small private company is not obligated to accept an offer if the ownership is not interested whereas a public company must accept an offer unless some legal reason exists to deny the potential investors.

When company decide to go public it is an opportunity to open up many financial doors. Trading in an open market provides a prestige, raises cash, provides means for liquidity, and it is easier to do mergers and acquisitions among many other benefits. Employees of these companies can have benefit of having a stock ownership plan which can help attract talented employees and increase the reputation of the company.

Several years ago, only strong private companies would clarify for an IPO and it was difficult to get listed. With the internet boom, this trend has changed so that today even IPO is done by a small startup companies seeking to expend their business.

This thesis uses the IPO data to illustrate how mixture of regression model can be used to estimate a return of a company given the knowledge of its revenue. The market data for IPO statistics available to an investor or a financial advisor may allowed for this type of analysis. A rational behind this approach is based on the idea that a single population model for returns based on revenue does not adequately reflect a market in which companies tend to cluster into subgroups based on some common characteristics within each group such as company size, performance, revenue, etc. Hence, it is more appropriate to model return as a function of company's revenue using a k-component regression model rather than a single regression.

2. LITERATURE REVIEW

The study of financial data has become increasingly dependent upon statistical method. This research is attempting to optimize The Initial Public Offering (IPO) data in a mixture of regression framework. The IPO data is from the book Regression Modeling with Actuarial and Financial Applications [6, Frees 2010].

Based on the facts that companies can be classified into sub-groups based on their sizes, use of a simple linear regression on IPO data may be questionable. In order to analyze IPO data in this situation, mixture model is proposed. Several books such as Finite Mixture Models [10, Peel 2000] and Data Mining [14, Witten 2011] mentioned Expectation-Maximization (EM) algorithm as a common approach for the mixture model. De Veaux [3, 1989] gave a method to approach mixture model using EM algorithm. He developed a mixtures of linear regression and connected the EM algorithm to relevant maximum likelihood equations. De Veaux used this mixture model to explain the experimental data and found out that the data should follow a 2-component linear regression model. In this thesis, a similar method is used to analyze IPO data and also evaluated its performance.

In order to apply EM algorithm on IPO data without giving the initial estimators, an initialization method should be used. Maitra [8, 2009] introduced the initialization method to initialize estimators for EM algorithm and applied it on two different clustering data sets: diurnal microarray gene expressions and industrial releases of mercury. The initialization method and the EM algorithm for both data worked effectively. The result of the first data set identified 21 and 22 as the optimal number of clusters, and the result of the second data indicated 113 optimal clusters. The idea of Maitra was used as a motivation for introducing starting values in the EM algorithm in this thesis. Further, the efficiency of this initialization method was compared to two commercial R packages: Mixtools [2, Benaglia 2009] and Mixreg [13, Turner 2000] on two data sets that are included in these packages: CO2 and Aphids.

In order to test the performance of the EM algorithm which applies the initialization method, two simulation studies are conducted. The simulation study method is provided by Faria and Soromenho [5, 2010]. They conducted simulation study for mixture of regressions in order to compare the performance of the three algorithms: the expectation maximization (EM), the classification version of the EM (CEM), and the stochastic version of the EM (SEM). They found that the performance of these algorithms depends on the parameters of the regression lines and the initialization. In general, the performance of CEM is the best when the true parameter values are used as the starting values in the initialization. In case of parallel regression lines, the SEM's performance is the best under random initialization. When the regression lines are concurrent and the algorithms are randomly initialized, all three algorithms have the same performance.

After testing the performance of the EM algorithm, the EM algorithm is used to analyze the IPO data with a mixture model. However, in order to select the best number of components for a mixture model of IPO data, statistical criteria AIC and BIC are both considered. The formulas for AIC and BIC are provided by Dr. Akaike [1, 1974] and Dr. Schwarz [11, 1978], respectively.

In order to determine the best number of components, bootstrap simulation is employed. Bootstrapping permits through simulation, repeated sampling of a data set which follows a certain distribution. The distribution is necessary for a hypothesis test which provides a way to test the number of components. An example of bootstrapping is mentioned in the study of *Estimating the propagation rate of a viral infection of potato plants via mixtures of regressions.* In this study, Turner [12, 2000] fitted a mixture of two linear regressions on a single predictor variable. He obtained a 95% confidence upper bound for the infection rate by means of maximum likelihood employing the EM algorithm. Optimal number of components is also tested using the bootstrapping of the likelihood ratio test statistic.

3. METHODOLOGY

3.1. EM algorithm

First, consider a multivariate linear regression model:

$$Y = X\beta + \epsilon \tag{3.1}$$

where Y is a $n \times 1$ vector, X is a $n \times p$ design matrix and β is a $p \times 1$ parameter vector. The error matrix follows a multi-normal distribution of $N(0, I\sigma^2)$. Now, suppose Y given X $(Y|X)$ is following a multi-normal distribution where mean vector is $X\beta$ and covariance matrix is $I\sigma^2$, since Y is following a PDF

$$f(Y, \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}[(Y-\mu)^T\Sigma(Y-\mu)]} \tag{3.2}$$

the PDF of Y given X appear as below

$$f(Y|X, \beta, \sigma^2) = \frac{1}{(2\pi)^{\frac{n}{2}}|I\sigma^2|^{\frac{1}{2}}} e^{-\frac{1}{2}[(Y-X\beta)^T(I\sigma^2)(Y-X\beta)]}. \tag{3.3}$$

When a log-likelihood is taken for $f(Y|X, \beta, \sigma^2)$, the formula above becomes

$$L = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}\sum\limits_{i=1}^{n}(y_i - x_i\beta)^2} \tag{3.4}$$

where y_i is an element in Y vector, and x_i is a $1 \times p$ vector from the matrix X. After taking a log function to both side of the Equation (3.4), the log-likelihood is developed as

$$log(L) = l(Y|X,\beta,\sigma^2) = -\frac{n}{2}log(2\pi) - \frac{n}{2}log(\sigma^2) - \frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta) \tag{3.5}$$

Now, consider a mixture of two simple linear regressions model:

$$g(Y|X) = \pi\phi(Y|X,\beta_1,\sigma_1^2) + (1-\pi)\phi(Y|X,\beta_2,\sigma_2^2) \tag{3.6}$$

where β_k $(k = 1, 2)$ is a 2×1 vector, Y is a $n \times 1$ vector and X is a $n \times 2$ matrix and ϕ is a PDF of $N(X\beta_k, \sigma_k^2)$ $(k = 1, 2)$. In this case, the complete likelihood function is defined as below.

$$L_c = \prod_{i=1}^{n}\prod_{k=1}^{2}[\pi_k\phi(y_i|x_i,\beta_k,\sigma_k^2)]^{I[z_i=k]} \tag{3.7}$$

where $k = 2$ is the number of components, x_i is a $n \times 1$ vector from matrix X and z_i represents the origin of x_i, that is, which component x_i belongs with (component 1 or component 2). I is an indicator function:

$$I(A) = \begin{cases} 0 & \text{if A is true} \\ 1 & \text{if A is false .} \end{cases}$$

The complete log-likelihood function, based on Equation (3.7), is developed as

$$log(L_c) = l_c = \sum_{i=1}^{n}\sum_{k=1}^{2} I[z_i = k](log(\pi_k) + log(\phi(y_i|x_i,\beta_k,\sigma_k^2))). \tag{3.8}$$

In the next step, parameters of the model are estimated. The conditional expectation given the observed data and the parameter estimates is defined as Q-function

$$Q = E[l_c|\text{obs.data}] = \sum_{i=1}^{n}\sum_{k=1}^{2} \pi_{ik}[log(\pi_k) + log(\phi(y_i|x_i, \beta_k, \sigma_k^2))] \ . \tag{3.9}$$

In order to estimate parameters $(\hat{\pi}_{ik},\ \hat{\beta}_k,\ \hat{\sigma}_k^2)$ of the model, the EM algorithm is used. This algorithm contains two steps: E-step and M-step. In the E-step, the posterior probability (π_{ik}) of a single point (x_i, y_i) belonging to the k^{th} component is computed as

$$\begin{aligned}\pi_{ik} = E[I(z_i = k)|X_i = x_i] = P[z_i = k|X_i = x_i] &= \frac{P(z_i = k \cap X_i = x_i)}{P(X_i = x_i)} \\ &= \frac{P(z_i = k)P(X_i = x_i|z_i = k)}{\sum\limits_{k'=1}^{2} P(z_i = k')P(X_i = x_i|z_i = k')}.\end{aligned} \tag{3.10}$$

Since $P(z_i = k)$ represents π_k, and $P(X_i = x_i|z_i = k)$ represents $\phi(y_i|x_i, \beta_k, \sigma_k^2)$, the following π_{ik} is defined as

$$\pi_{ik} = \frac{\pi_k \phi(y_i|x_i, \beta_k, \sigma_k^2)}{\sum\limits_{k'=1}^{2} \pi_{k'}\phi(y_i|x_i, \beta_{k'}, \sigma_{k'}^2)} \ . \tag{3.11}$$

In the M-step, the Q-function (3.9) should be maximized by taking three partial derivatives of Equation (3.9) with respect to π_k, β_k, and σ_k^2 and setting them to zero. As a result, the following maximum likelihood estimators are derived:

The estimator of π_k is

$$\hat{\pi}_k = \frac{\sum\limits_{i=1}^{n} \pi_{ik}}{n} \ . \tag{3.12}$$

And the estimator of β_1 is

$$\hat{\beta}_1 = (X^T w_1 X)^{-1}(X^T w_1 Y) \ . \tag{3.13}$$

Similarly, the estimator of β_2 is

$$\hat{\beta}_2 = (X^T w_2 X)^{-1}(X^T w_2 Y) \ . \tag{3.14}$$

The estimator of σ_1^2 is

$$\hat{\sigma}_1^2 = \frac{\sum\limits_{i=1}^{n} \pi_{i1}(y_i - x_i\beta_1)^T(y_i - x_i\beta_1)}{\sum\limits_{i=1}^{n} \pi_{i1}} = \frac{||\sqrt{w_1}(Y - X^T\beta_1)||}{trace(w_1)} \ . \tag{3.15}$$

Similarly, the estimator of σ_2^2 can be written as

$$\hat{\sigma}_2^2 = \frac{||\sqrt{w_2}(Y - X^T\beta_1)||}{trace(w_2)} \tag{3.16}$$

where $w_i = diag(\pi_{1i},\ \pi_{2i},\ \pi_{3i}, \ldots,\ \pi_{ni})$.

The derivations in M-step provide the maximum likelihood estimators (MLEs) for π_k, β_k, and σ_k. These MLEs are used to estimate the mixing probabilities for each point on each iteration of the E-step. However, the derivations do not provide an initial value for the E-step. Therefore, initialization is necessary, which the next sub-section will address.

3.2. Initialization method

In situations when we have a single population, the initial value is easy to obtain, because there is only one local maxima. However, one possibility is that the IPO data contains more than one population. If this is the case, the likelihood for the model will contain more than one local maxima. According to Melnykov and Melnykov [9, 2012], even when more than one local maxima exists, the EM algorithm can only generate one solution for one local maxima, because the EM algorithm can only have one starting point. If the local maxima is not the best local maxima, the EM algorithm will loop infinitely, because the likelihood function for this model is unbounded. Initial values can be obtained by initializing the algorithm with a set of initial parameters, or by using an initialization method. Since there is no way to obtain the parameters of the model by visually checking the IPO data set or by manually examining the data set, the problem persists because the initial values remain unknown. A solution for this problem is to use an initialization strategy before beginning the EM algorithm in order to obtain the best initial value.

The initialization methods for the EM algorithm is classified as deterministic or stochastic. Deterministic methods usually obtain the initial value using cluster analysis, which classify the data set. The most popular cluster analysis in this case is called hierarchical clustering, which classify the data set based on the distance between groups. An example for a deterministic method is called random initialization method, which is employed in mixtools [2, Benaglia 2009]. This method is classified as deterministic because it provides the initial random values immediately. This is the major difference between deterministic and stochastic who provides the best initial values by comparing several results that are generated through repeating process. The basic steps of random initialization method are shown below.

1. Rank the data set via independent variable ($x_j,\ j = 1 \ldots n$)

2. Randomly separate the data set into k groups ($k =$ number of components)

3. Estimate the mixing probabilities ($\hat{\pi}_i$) based on the ratio of the size of each group and the size of the data set respectively

4. Use maximum log-likelihood to estimate the parameters ($\hat{\beta}_i$, $\hat{\sigma}_i$) for these groups separately

5. Randomly select a value from the normal distribution with means equals $\hat{\beta}_i$ and variance equals $\hat{\sigma}_i^2(X^T X)^{-1}$, where X is a $n \times 2$ matrix. Assign the value as an initial estimator for β_i

However, there is a huge disadvantage in deterministic methods. The methods can only provide one set of initial parameters, and if the set of initial parameters is not appropriate for the data set, either the EM algorithm reaches an incorrect solution or continues infinitely. However, the stochastic method does not have this disadvantage. The stochastic method determines the best set of initial values by repeatedly calculating the likelihood function using multiple sets of initial values then select the best likelihood function. Stochastic methods solve the problem in deterministic methods by repeated calculations. But this may result in stochastic methods consuming more processing time than deterministic methods. This thesis is motivated by a stochastic method called Initializing Partition-Optimization Algorithm [8, Maitra 2009], which is mentioned in the literature review. This algorithm use the Euclidean distance as the criterion for finding clusters in the data set. The basic steps of this initialization are summarized below.

1. Randomly select k points as initial points (k=number of components)
2. Group other points base on the smallest Euclidean distance of those initial points
3. Develop a linear regression for each group
4. Iterate the process m times
5. Select the best model as the initial parameters for EM algorithm (The best model will be the model with the greatest log-likelihood)

There are mainly two differences between the Initializing Partition-Optimization Algorithm [8, Maitra 2009] and the random initialization method [2, Benaglia 2009]. The Initializing Partition-Optimization Algorithm groups the data set based on the Euclidean distance of initial points; while the random initialization method groups the data set randomly. Also, the Initializing Partition-Optimization Algorithm selects the best initial value by determining the largest likelihood after repeating the process m times and recording the likelihood each time; while the random initialization method provides the initial value immediately.

Two examples are considered below to show the efficiency of the initialization method compared to the initialization methods implemented in two different commercial packages (mixtools and mixreg).

3.2.1. Example 1

The data set in this example comes from the R package called mixreg. Mixreg was developed to support the article about aphids and infected potato plants that was already mentioned in literature review section. The author developed a two components mixture regression model in order to analyze the relationship between the number of aphids (a certain insect that can infect potato plants) and the number of infected potato plants.

The data set includes two variables: independent variable represents the number of aphids, and dependent variable represents the number of infected potato plants. According to the article [12, Turner 2000], the optimal number of components for the aphids data set is when $k = 2$.

Table 3.1 shows the performance of three algorithms when the number of components $k = 2$ in the aphids data.

Table 3.1. Aphids data with k=2

k=2										
Algorithm	iteration	LogL	β_{11}	β_{21}	β_{12}	β_{22}	σ_1	σ_2	π_1	π_2
EM	6	-132.065	3.466	0.055	0.857	0.002	3.119	1.124	0.502	0.498
mixtools	25	-132.065	3.474	0.055	0.859	0.002	3.115	1.125	0.502	0.498
mixreg	25	-132.065	3.474	0.055	0.859	0.002	3.115	1.125	0.502	0.498

Although the log-likelihood and estimators in three algorithms are almost the same, the number of iterations is significantly different. EM provides the smallest number of iterations while mixreg and mixtools have the same number of iterations. This is because the initialization method in EM provides the best local maxima by repeating the process then selects the best likelihood. As a result, the EM algorithm in EM only needs to iterate itself to find the best model using the best local maxima. Meanwhile, other two initialization methods in two packages only provide a local maxima, which may not be the best local maxima. As a result, the EM algorithm in these two package have to find the best model and the best local maxima at the same time.

Table 3.2 shows the performance of three algorithms when k=3 in the aphids data which is provided in mixreg.

Table 3.2. Aphids data with k=3

k=3														
Algorithm	iteration	LogL	β_{11}	β_{21}	β_{12}	β_{22}	β_{13}	β_{23}	σ_1	σ_2	σ_3	π_1	π_2	π_3
EM	22	-127.700	3.407	0.056	0.867	0.002	4.892	0.033	3.315	1.124	0.052	0.429	0.504	0.068
mixtools	55	-127.207	2.563	0.067	0.883	0.002	5.122	0.040	3.117	1.138	0.197	0.388	0.504	0.108
mixreg	69	-127.253	3.545	0.055	0.615	-0.002	4.411	-0.010	3.225	0.507	0.860	0.450	0.350	0.200

The estimators in EM and other two packages are still almost the same although the estimators for intercepts in mixtools are slightly different from other two's. Notice that the optimal number of components in aphids data is proved to be two. The result, which is using 3-component model, draws a conclusion: Both the EM and other two packages provide almost the same results even the number of components in the model is not optimal.

3.2.2. Example 2

The "CO2" data set in this example is coming from the commercial R package called mixtools. This data set is represented as an demonstration of how the mixture regression function works in mixtools package. This data set also contains two variables: Gross national product (GNP) per capita is an independent variable and carbon dioxide (CO2) is a dependent variable. The author believes that there is a strong correlation between "CO2" and "GNP," and "CO2" can be predicted by "GNP" using mixture model. The optimal number of component for this data set is equal to two.

Table 3.3 shows the performance of three algorithms when $k = 2$ in the CO2 data.

Table 3.3. CO2 data with k=2

k=2										
Algorithm	iteration	LogL	β_{11}	β_{21}	β_{12}	β_{22}	σ_1	σ_2	π_1	π_2
EM	2	-70.173	9.912	0.317	6.000	0.073	1.317	2.026	0.212	0.788
mixtools	29	-70.173	9.914	0.317	5.998	0.073	1.316	2.025	0.212	0.788
mixreg	31	-70.173	9.914	0.317	5.998	0.073	1.316	2.025	0.212	0.788

Although the number of iterations in three algorithms is different, the log-likelihood and estimators in three algorithms are the same. It implies that the EM algorithm can reach to the same level through different initialization methods when the number of components is correct.

Table 3.4 shows the performance of three algorithms when k=3 in the "CO2" data set that is provided in mixtools.

The EM algorithm seems to run faster than mixtools and mixreg. Comparing with the estimators, mixreg provides a very different estimators against other two. In addition, mixreg contains the largest number of iterations. This result implies that the initialization method may affect the performance of the EM algorithm by providing bad initial values that may increase the

Table 3.4. CO2 data with k=3

k=3														
Algorithm	iteration	LogL	β_{11}	β_{21}	β_{12}	β_{22}	β_{13}	β_{23}	σ_1	σ_2	σ_3	π_1	π_2	π_3
EM	37	-62.388	7.254	0.106	5.047	0.067	9.434	0.337	0.143	1.599	1.362	0.236	0.524	0.240
mixtools	41	-62.388	7.254	0.106	5.035	0.067	9.425	0.337	0.143	1.594	1.362	0.236	0.523	0.241
mixreg	272	-62.614	2.177	0.263	9.917	-0.073	1.927	0.654	0.252	1.782	0.804	0.196	0.571	0.233

number of iterations in EM algorithm.

In order to select the best mixture regression model, the optimal number of components should be determined. Methods to determine the best number of components is addressed in next sub-section.

3.3. Estimating optimal number of components

After introducing the initialization method for the IPO data set and presenting two comparing examples, our focus is on selecting the optimal number of components for the data set. As the number of components increases, the log-likelihood will also increase. However, as the number of components increases, the model efficiency decreases, therefore, a small number of components is preferable. There are many ways to estimate the optimal number of components: Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and bootstrapping of Likelihood Ratio Test Statistic (LRTS).

3.3.1. Akaike information criterion (AIC)

The Akaike Information Criterion (AIC) provides a criterion for model selection with different numbers of components. The AIC selects the smallest number of components by balancing the error of the fitting model (log-likelihood function) against the number of free parameters (includes the number of components) using the following formula

$$AIC = -2 \times log(L) + 2 \times M \tag{3.17}$$

where M is the number of free parameters in the model.

The criterion AIC selects the best model by balancing the error and number of components. However, there is one more factor that AIC does not included. BIC, on the other hand, includes this factor.

3.3.2. Bayesian information criterion (BIC)

Although AIC includes the number of components for model selection, it does not account the number of observations in the data set. The BIC not only considers the log-likelihood function and the number of free parameters, it also includes the number of observations in the data set as follows

$$BIC = -2 \times log(L) + M \times ln(n) \tag{3.18}$$

where $log(L)$ is log-likelihood function, M is the number of free parameters in the model, and n is the number of observations in the data set.

Although both AIC and BIC can determine the best model hence to determine the optimal number of components, the determination is not sufficient. In order to ensure the optimal number of components, a hypothesis test should be employed. Therefore, bootstrapping is involved.

3.3.3. Bootstrapping

Another method for selecting a number of optimal components is based on bootstrapping of the likelihood ratio test statistic. In this method, a hypothesis test is considered. The null hypothesis and alternative hypothesis in this test are

H_o: number of components $= k$

H_α: number of components $= k + 1$

In order to test the hypothesis, the log-likelihood ratio test (LRT) is used. It is expected that the likelihood ratio test statistic, LRTS, follows a Chi-square distribution with a degree of freedom equal to the difference between the number of parameters under H_o and the number of parameters under H_α. Unfortunately, in the mixture model case, because the mixing probabilities under H_o are unknown, the number of parameters is unknown; hence the LRTS does not follow the Chi-square distribution. This problem is solved by bootstrapping.

Bootstrapping is a useful resampling method that deals with a data set whose distribution is undetermined. It can estimates nearly any statistics for the data set such as the distribution, the mean, and the variance. The main idea for bootstrapping is to obtain the estimators through repeated resampling when the size of the data set is insufficient to generate model estimators.

By resampling the data set, the data set appears to show some obvious characteristics so that the estimators can be obtained easily. The bootstrapping in this case allows LRTS to reach an acceptable level of accuracy through repeated resampling (in this case, 500 times) of a data set based on the MLEs of H_o and computing the LRTS each time. As a result, the true value of the LRTS can be determined to an acceptable level using the p-value. Bootstrapping of the procedure used below is mentioned by T. Rolf Turner [12, 2000]:

1. Use EM algorithm to fit the data assuming the number of components equals to K and $K+1$

2. Calculate the log-likelihood ratio Q

3. Simulate data from the model using K as the number of components

4. Use both K and $K+1$ component models to fit the simulated data, and calculate the corresponding log-likelihood ratio statistic $Q*$

5. Compute the p-value for Q using

$$\frac{1}{N-1}\sum_{i=1}^{N} I(Q \geq Q*) \tag{3.19}$$

   where $I(\cdot)$ is an indicator function

After the optimal number of components is selected using AIC, BIC, and bootstrapping of the likelihood ratio, the EM algorithm is completed. In order to perform the EM algorithm in an appropriate situation, the performance of the EM algorithm should be tested. To test the performance of the EM algorithm, two sets of simulation studies are conducted.

4. SIMULATION STUDIES

Two simulation studies were conducted in order to test the performance of the initialization method. The first study uses small σs and the second study uses large σs. Both simulation studies contain sample size equal to 100 and include two different scenarios: (1) two parallel lines, (2) two concurrent lines.

Consider the following simulation model for a mixture of two simple linear regressions:

$$g(Y|X) = \pi_1 \phi(Y|X\ \beta_1, \sigma_1^2) + \pi_2 \phi(Y|X\ \beta_2, \sigma_2^2) \tag{4.1}$$

where $\phi(Y|X\ \beta_i, \sigma_i^2)$ is a Gaussian density with mean $X^T\beta$ and variance σ^2; Y is a response vector of size $n \times 1$, X is a $n \times 2$ design matrix having first column filled with "1"s, β_i is a 2×1 parameter vector, and $\epsilon_i \sim N(0, \sigma_i^2)$.

- Scenario-1. In this scenario, two parallel lines were chosen with the following parameters

$$\beta_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \text{ and } \beta_2 = \begin{pmatrix} 11 \\ 1 \end{pmatrix}$$

- Scenario-2. In this scenario, two concurrent lines were chosen with the following parameter values

$$\beta_1 = \begin{pmatrix} 52 \\ 1 \end{pmatrix} \text{ and } \beta_2 = \begin{pmatrix} 0 \\ 3 \end{pmatrix}$$

For both scenarios two sets of mixing probabilities are considered,

$$\pi_1 = 0.5, \pi_2 = 0.5 \text{ and } \pi_1 = 0.2, \pi_2 = 0.8$$

Independent variable (x) and dependent variable (y) need to be generated when the true variables are set up. Two different sets of x and y are needed because of the fact that there are two different sets of mixing probabilities.

When $\pi = \left(\begin{smallmatrix} 0.5 & 0.5 \end{smallmatrix}\right)$, X is a combination of two of the same sub-sequences; each sub-sequence is from 1 to 50.

When $\pi = \left(\begin{smallmatrix} 0.2 & 0.8 \end{smallmatrix}\right)$, X is a combination of two different sub-sequences; one contains 20 points that from 1 to 80 by 4, and the other is from 1 to 80.

Dependent variable Y is generated according to the formula

$$Y = X\beta_i + \epsilon_i. \tag{4.2}$$

y_i, which are the elements of Y, are following a normal distribution with mean equal to $X\beta_i$ and variance σ_i^2.

This study considers $N = 5000$ simulation runs. Data were produced under a two component mixture linear model with sample size equal to 100. Maximum likelihood estimators $(\hat{\beta}, \hat{\pi}, \hat{\sigma})$, log-likelihood and the number of iterations are also recorded during each simulation runs.

The observed measures were $BIAS(\hat{\theta}_j)$ and $MSE(\hat{\theta}_j)$ for $j = 1, ..., N$ are defined as follow

$$BIAS(\hat{\theta}_j) = \frac{1}{N} \sum_{m=1}^{N} \hat{\theta}_j^{(m)} - \theta_j, \tag{4.3}$$

$$MSE(\hat{\theta}_j) = \frac{1}{N} \sum_{m=1}^{N} (\hat{\theta}_j^{(m)} - \theta_j)^2 \tag{4.4}$$

where $N = 5000$, $\theta_j = (\beta_j,\ \sigma_j^2,\ \pi_j)$, and $\theta_j^{(m)} = (\beta_j^{(m)},\ \sigma_j^{2(m)},\ \pi_j^{(m)})$.

The criterion BIAS calculates the difference between the average of estimators, after running the EM algorithm 5000 times, and the true parameters. Hence, the result with the smallest BIAS generates the most accurate estimator. The criterion MSE calculates the mean square error of 5000 estimators, and the result with the smallest MSE has a more stable performance. After deriving and discussing the criteria to determine the performance of EM-algorithm, we conducted two simulation studies for two scenarios. The results are shown below.

4.1. Results of simulations in Scenario-1

The first set of simulation in Scenario-1 will set $\sigma_1 = 1$ and $\sigma_2 = 0.2$, while the second set of simulation in Scenario-1 will set $\sigma_1 = 1$ and $\sigma_2 = 2$.

Table 4.1 below provides the BIAS and the MSE of estimators, the average number of iterations and maximum log-likelihood of 5000 replicates for Scenario-1 with two parallel lines.

Table 4.1. Simulation with parallel lines

parallel													
Algorithm	σ	π	iteration	MaxLogL	Criteria	β_{11}	β_{21}	β_{12}	β_{22}	σ_1	σ_2	π_1	π_2
EM	$\sigma = 1, 0.2$	$\pi = 0.5, 0.5$	2.00	-102.00	BIAS	-0.01	0.00	0.00	0.00	-0.03	-0.01	0.00	0.00
					MSE	0.09	0.00	0.00	0.00	0.01	0.00	0.00	0.00
		$\pi = 0.2, 0.8$	2.52	-42.96	BIAS	0.01	0.00	0.00	0.00	-0.06	0.00	0.00	0.00
					MSE	0.07	0.00	0.00	0.00	0.02	0.00	0.00	0.00
EM	$\sigma = 1, 2$	$\pi = 0.5, 0.5$	2.36	-216.12	BIAS	0.00	0.00	-0.01	0.00	-0.03	-0.06	0.00	0.00
					MSE	0.07	0.00	0.03	0.00	0.01	0.05	0.00	0.00
		$\pi = 0.2, 0.8$	4.38	-226.63	BIAS	0.05	0.00	0.03	0.00	-0.06	-0.02	0.00	0.00
					MSE	0.09	0.00	0.02	0.00	0.03	0.03	0.00	0.00

The results in Table 5 imply that the EM algorithm performs very well. The MLEs for these simulation settings are unbiased. The EM algorithm coverages quickly in all four cases as we see with low number of iterations reported. An increase in σ seems not to affect the performance of the algorithm for these settings. These results were compared to those generated by mixtools and mixreg R packages and it was found that higher BIAS and MSE were reported by both of the packages for all parameters. However, no differences were found in log-likelihood.

Figure 4.1 shows both sets of simulations in Scenario-1(parallel lines).

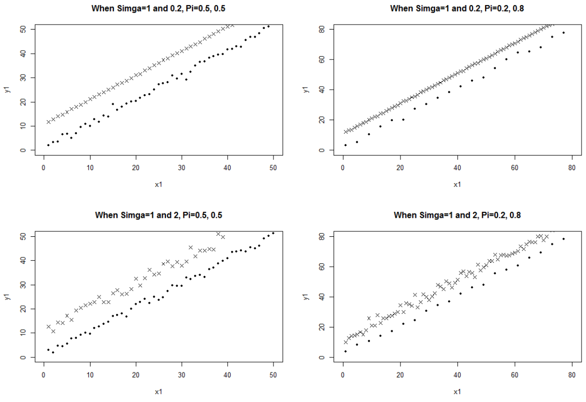

Figure 4.1. Scenario-1 with parallel lines

The figure above shows that when σ is small ($\sigma = 1$ and 0.2), it is easier to observe two distinguish parallel lines. The simulation results for the study on two concurrent lines are presented in the following subsection.

4.2. Results of simulations in Scenario-2

The first set of simulation in Scenario-2 (concurrent) will set $\sigma_1 = 1$ and $\sigma_2 = 2$, while the second set of simulation in Scenario-2 will set $\sigma_1 = 4$ and $\sigma_2 = 5$.

Table 4.2 provides the BIAS and the MSE for all MLEs, the average of numbers of iterations and maximum log-likelihood of 5000 replicates for scenario-2.

Table 4.2. Simulation with concurrent lines

concurrent													
Algorithm	σ	π	iteration	MaxLogL	Criteria	β_{11}	β_{21}	β_{12}	β_{22}	σ_1	σ_2	π_1	π_2
EM	$\sigma = 1, 2$	$\pi = 0.5, 0.5$	5.55	-212.00	BIAS	0.00	0.00	0.00	0.00	-0.02	-0.02	0.00	0.00
					MSE	0.00	0.00	0.00	0.00	0.01	0.04	0.00	0.00
		$\pi = 0.2, 0.8$	5.15	-165.00	BIAS	0.00	0.00	0.03	0.00	-0.02	-0.13	0.00	0.00
					MSE	0.00	0.00	0.00	0.00	0.01	0.12	0.00	0.00
EM	$\sigma = 4, 5$	$\pi = 0.5, 0.5$	8.32	-321.00	BIAS	-0.02	0.01	0.02	-0.01	-0.01	-0.06	0.00	0.00
					MSE	0.03	0.01	0.01	0.01	0.04	0.51	0.00	0.00
		$\pi = 0.2, 0.8$	4.94	-307.18	BIAS	-0.05	0.00	-0.02	0.00	-0.09	-0.33	0.00	0.00
					MSE	0.01	0.00	0.05	0.00	0.01	0.79	0.00	0.00

From Table 4.2, we observe that all MLEs are unbiased and the performance of the EM algorithm does not change with an increase in σ or changes in π. For these settings, optimal solutions were found with a very few iterations. The results were compared to those obtained by mixreg and mixtools R packages and it was found that the R packages produce larger BIAS and MSE compared to EM algorithm.

Figure 4.2 shows both sets of simulations in Scenario-2(concurrent lines).

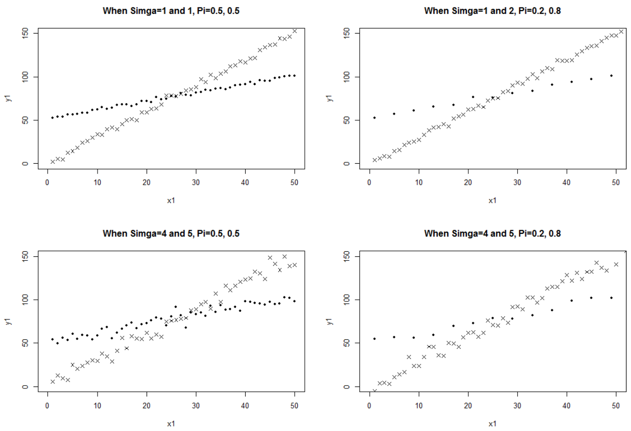

Figure 4.2. Scenario-2 with concurrent lines

Figure 4.2 shows that when σ increases, it becomes harder to observe two concurrent lines. Conclusions for both scenarios are presented in the following subsection.

4.3. Conclusion

Two simulation studies are conducted to test the performance of the EM algorithm. Two independent conclusions can be drawn by comparing the results within the scenarios and the results between two scenarios. The first conclusion is: Neither the set of σs nor the set of πs can affect the stability and accuracy of the EM algorithm. The second conclusion is based on the fact that almost the same results are produced from Scenario-1 and Scenario-2: Neither the parallel lines nor the concurrent lines can affect the performance of the EM algorithm, although the parallel lines seem to have a larger likelihood. The performance of two R packages, mixtools and mixreg, is also tested using these two simulation studies. Even though EM, mixtools, and mixreg share the same likelihood, the performance of EM is always better than these two R packages because it provides a smaller BIAS and MSE. After the performance of the EM algorithm is tested and the conclusions are made, the EM algorithm can be applied to the IPO data set.

5. ANALYSIS

5.1. Mixture approach in IPO data set

Initial Public Offering (IPO) is a type of investment offering that a privately owned company issues to the public. An IPO usually is issued when a company wants to become publicly traded or to raise expansion capital. In an IPO, an underwriting firm helps the company or the issuer to determine what type of security to issue and the best offering price.

IPO data for this project is provided in the book *Regression Modeling with Actuarial and Financial Applications* [6, Frees 2010]. The data is obtained from 116 companies that priced during January 1, 1998 through June 1, 1998. It contains 6 variables: Company name, Ticker symbol, Return (in percent), REV (in millions of dollars) which stands for the company's revenues, LnREV that denotes logarithm of revenues, and Price IPO. For the purpose of this analysis, LnREV and Return were only used.

Figure 5.1 below shows the scatter plot of IPO data.

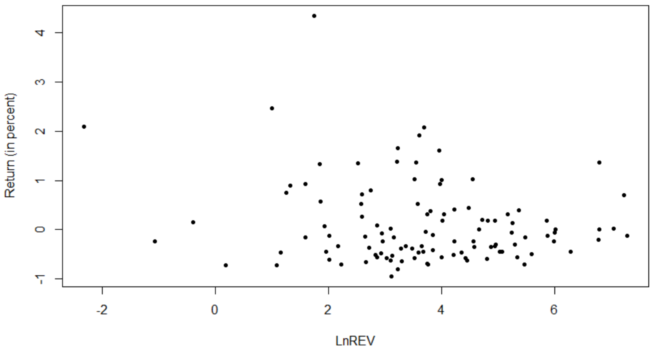

Figure 5.1. Scatter plot of IPO data

In the beginning of the analysis, a simple linear regression model is considered to analyze the data set. This model corresponds to a mixture of linear regression when $k = 1$. The best fitted line for this model is

$$\hat{y} = 0.438 - 0.090x \tag{5.1}$$

and it is shown in Figure 5.2.

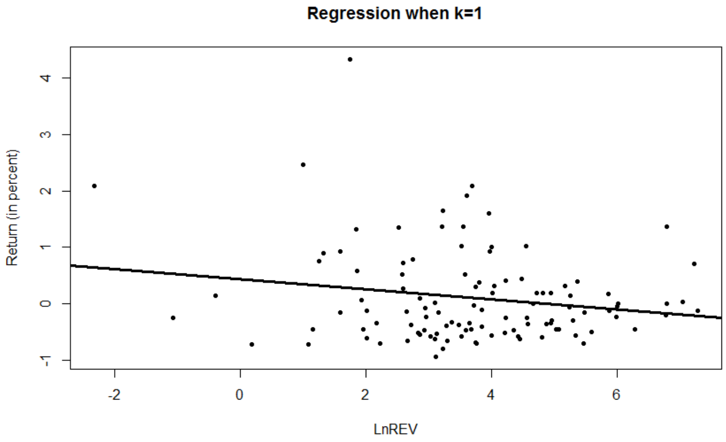

Figure 5.2. Simple linear regression

Notice that the coefficient of x equals to 0.090, which is close to zero. The p-value for the coefficient of x is 0.052, close to 0.05, which suggests that the H_o may not be rejected. Under H_o, a claim is made that the population parameter for the slope coefficient is equal to zero. Therefore, the simple linear regression may not be suitable for the IPO data set.

In the next step, a k-component mixture model is tested. The summary of results for $k = 1, 2, 3, 4$ is provided in Table 5.1.

Table 5.1. Results for IPO data by k-components

k	logL	β	$\beta_{,1}$	$\beta_{,2}$	$\beta_{,3}$	$\beta_{,4}$	σ_1	σ_2	σ_3	σ_4	π_1	π_2	π_3	π_4	BIC	AIC
1	-132.388	$\beta_{1,}$ $\beta_{2,}$	0.438 -0.090				0.814				1.000				278.877	270.775
2	-102.034	$\beta_{1,}$ $\beta_{2,}$	1.478 -0.194	-0.501 0.047			0.855	0.279			0.378	0.622			236.972	218.068
3	-96.282	$\beta_{1,}$ $\beta_{2,}$	1.663 -0.213	-0.765 0.075	-0.100 0.022		0.881	0.154	0.312		0.306	0.374	0.320		244.270	214.564
4	-85.333	$\beta_{1,}$ $\beta_{2,}$	1.398 -0.168	-0.114 0.021	6.408 -1.178	-0.754 0.070	0.622	0.253	0.017	0.158	0.299	0.271	0.029	0.401	241.173	200.666

Table 5.1 shows the number of components ($k = 1, 2, 3, 4$), log-likelihood, MLEs ($\hat{\beta}_k$, $\hat{\sigma}_k$, $\hat{\pi}_k$), AIC and BIC. The log-likelihood of simple linear regression in Table 5.1 is -132.388, whereas the value of log-likelihood is largely increased when $k = 2$ (-102.034), this indicates that mixture model is more appropriate to analyze IPO data. Traditionally, the model with the largest value of log-likelihood is considered as the best model, however, from the table above, since the log-likelihood became greater when the number of components increases, criteria AIC and BIC are also considered. The BIC criterion indicated that the best model has $k = 2$. The criterion AIC is not used because the AIC continues to decrease when the number of components increases. It is a well known practice in mixture modeling to use BIC rather than AIC due to the fact that BIC produces more conservative results [10, Peel 2000].

Figure 5.3 illustrates the two-component regression model.

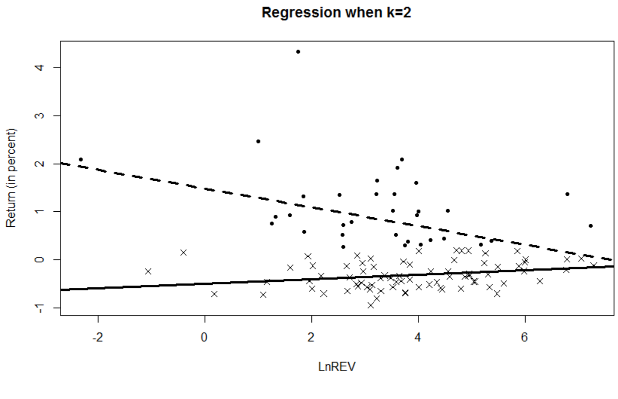

Figure 5.3. Two-component regression model

From Figure 5.3, we may observe that there are two different trends in the model, which means the return is affected by presence of subgroups with exhibit different patterns of correlation between revenue and return. The dashed line indicates a downward trend with weak negative correlation between revenue and return. The solid line in the plot indicates an upward trend with weak positive correlation between revenue and return. Additionally, from Figure 5.3, a larger return is more likely to be classified into the dashed line, while a small return is more likely to be classified into the solid line. In order to ensure that the return is affected by two trends, which means the number of component is 2, a bootstrapping of the likelihood ratio test statistic is conducted.

5.2. Bootstrapping approach in IPO

To confirm that the best number of components is $k = 2$, bootstrapping is performed twice. The hypotheses for the first bootstrapping are:

H_o: number of components = 1

H_α: number of components = 2

while the hypotheses for the second bootstrapping are:

H_o: number of components = 2

H_α: number of components = 3 .

The number of iterations in both hypothesis tests is equal to 500. Figure 5.4 shows the histograms generated based on the bootstrapping results.

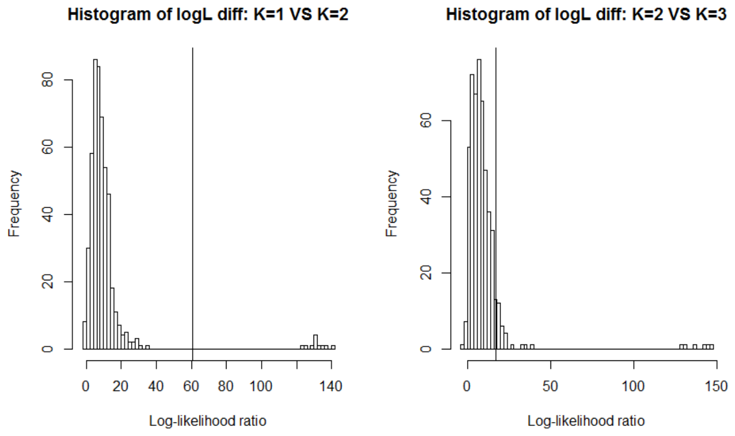

Figure 5.4. Histograms of bootstrapping results

The vertical line in Figure 5.4 represents the observed value of the log-likelihood Ratio Tests (LRTS). This is the true value based on H_o. If H_o is true then the histogram of the bootstrapping LRTS results should capture the true LRTS, which means the histogram and the vertical line in the plot will overlap. Similarly, when the H_o is rejected, the histogram of the bootstrapping LRTS results will not capture the true LRTS, thus the vertical line and the histogram will be apart. The p-value of 0.022 is reported providing sufficient evidence to reject H_o when $k = 1$ vs $k = 2$. The p-value of 0.076 is reported providing insufficient evidence to reject H_a, therefore H_o is not rejected when $k = 2$ vs $k = 3$. In conclusion, the optimal number of components, $k = 2$, seems to fit the model best.

5.3. Analysis of results for IPO data

In this section an analysis of the results is considered in the context of the overall problem. We test the model for small, medium, and large size companies. The data set consists of twenty small companies, twenty medium-size companies, and twenty large companies, selected based on the revenue from IPO data. The range of revenue for small companies is $(0.099 - 9.283)$, the range of revenue for medium-size companies is $(23.64 - 42.72)$, and the range of revenue for large companies is $(121.5 - 398.2)$, all in millions of dollars.

Two different models are generated base on the IPO data: simple linear regression model and 2-components mixture regression model. The estimated coefficients of simple linear regression model are

$$\beta = \begin{pmatrix} 0.438 \\ -0.090 \end{pmatrix}.$$

The value of slope in simple linear regression for the IPO data equals -0.090, which implies that the revenue (LnREV) is inversely proportional to the return. That means when the LnREV increases, the return will decrease. Based on this phenomena, a conjecture was made about the relationship between the return and the size of the company. As the size of the company increases, the return will decrease. Fama and French [4, 1993] proposed an explanation for this relationship: Large companies tend to have larger revenues, but those companies will be exposed to more risks than small companies, so investors are likely to hesitate investing in these large companies. Hence large companies seeking an IPO may not receive the expected amount of investment. The expenses of large companies are greater than small companies, which may lower the net income resulting in a lower return. As a result, the return of large companies decreases.

However, notice that the coefficient of x is very small (-0.090), the linear correlation between revenue and returns is questioned. The p-value from F-test of this model is 0.052, which is larger than the significant level ($\alpha = 0.05$). Although the p-value is very close to the reject region, we still conclude that the simple linear regression model cannot interpret IPO data set accurately. Compared with simple linear regression model, the mixture regression of 2-components provides a better result. The estimated regression coefficients of 2-component model are

$$\beta_1 = \begin{pmatrix} 1.478 \\ -0.194 \end{pmatrix}, \ \beta_2 = \begin{pmatrix} -0.501 \\ 0.047 \end{pmatrix},$$

$$\pi_1 = 0.378 \ , \pi_2 = 0.622 \ .$$

From both coefficients above and Figure 5.3, we observed that there are two different trends in the model. The slope coefficient for the first component (β_1) indicates a downward trend, and the slope coefficient of the second component(β_2) indicates an upward trend. The model implies that return is determined by these two different regression lines. Based on the set of mixing probabilities ($\pi_1 = 0.378$ and $\pi_2 = 0.622$), the second line (β_2) seems to have more effects on the mixture model.

The fitted returns in mixture model are generated using LnREV, mixing probabilities, and both trends. Fitted values of returns based on 2-component mixture model are calculated as

$$\hat{y}_i = \pi_1 f_1(x_i) + \pi_2 f_2(x_i), \tag{5.2}$$

where $f_k(x)$ for $k = 1, \ 2$ represents the model for each component.

Fitted returns compared with the true value of returns for both $k = 1$ and $k = 2$ are plotted in Figure 5.5.

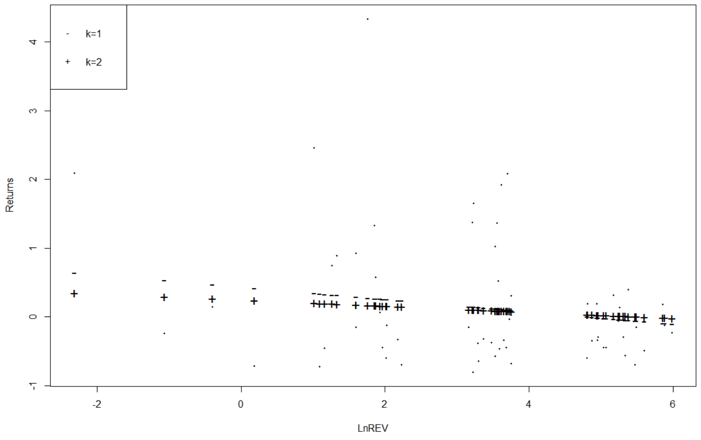

Figure 5.5. True vs. fitted value of the return

Fitted returns from simple linear regression model in Figure 5.5 are labeled as (”+”), and fitted returns from 2-component regression model are labeled as (”-”). When the simple linear regression model and 2-component model are generated, the following observations can be drawn: (1) The return will become lower when the size of company becomes larger. (2) The second line (β_2) that indicates an upward trend has more effects than the first line (β_1) that indicates a downward trend. The result generated from simple linear regression model indicates a downward trend, which supports (1). However, the result from 2-component regression model does not support (2). The fitted returns that are generated from 2-component model also indicate a downward trend, which implies that the model largely depends on the first line (β_1). The difference between this conclusion and (2) is reasonable: Even though the mixing probability for β_2 is larger (0.622), the coefficient of the log revenue in the second line is relatively small (0.047). Therefore, the coefficient of the log revenue that is used to calculate fitted return turns into a negative number (-0.044).

Although both simple linear regression model and 2-component regression model indicate a negative relation between returns and revenues, Figure 5.5 indicates that when the value of revenue is small ($0.099 - 9.283$), the fitted returns from simple linear regression model are aways overestimated. On the other hand, the fitted returns from 2-component model are more close to the true returns. Furthermore, the difference between simple linear regression model and 2-component model become smaller when the revenue increases, which implies that when the revenue increases, simple linear regression model and 2-component model generate the same fitted returns.

Two residuals plot are also created to test normality assumptions for both regression models. Figure 5.6 provides the residual plot of the simple linear regression.

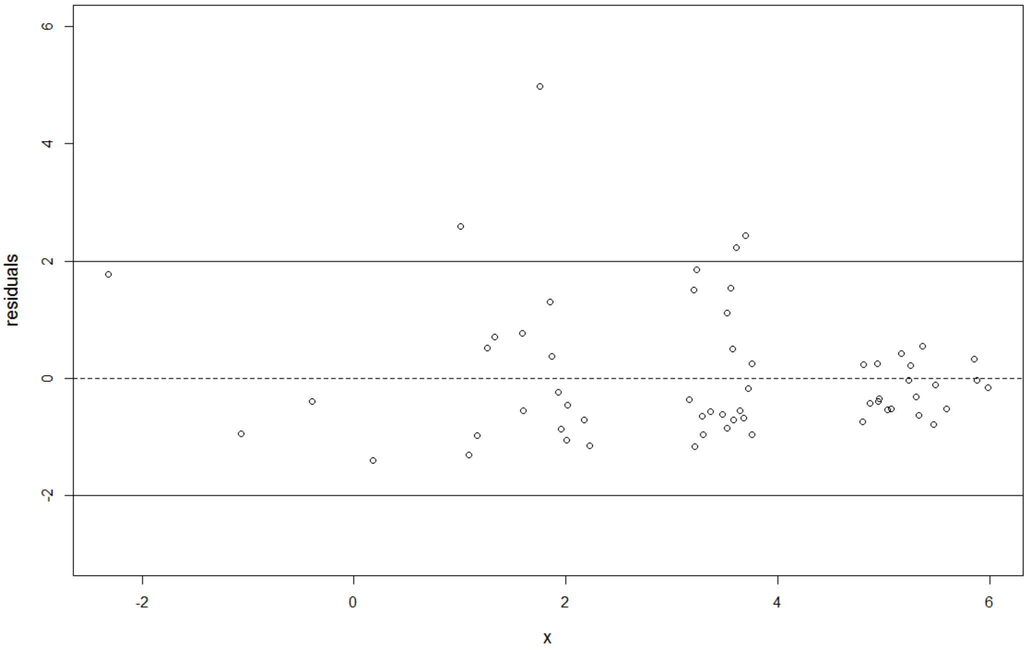

Figure 5.6. Residual plot for simple regression model

The horizontal lines in Figure 5.6 are reference lines: $h = \pm 2$. The residual plot indicates that there are four outliers: Inktomi Corp., IBS Interative, MIPS Technologies, and Broadcom Corp.

The residuals plot of 2-component mixture model is displayed in Figure 5.7.

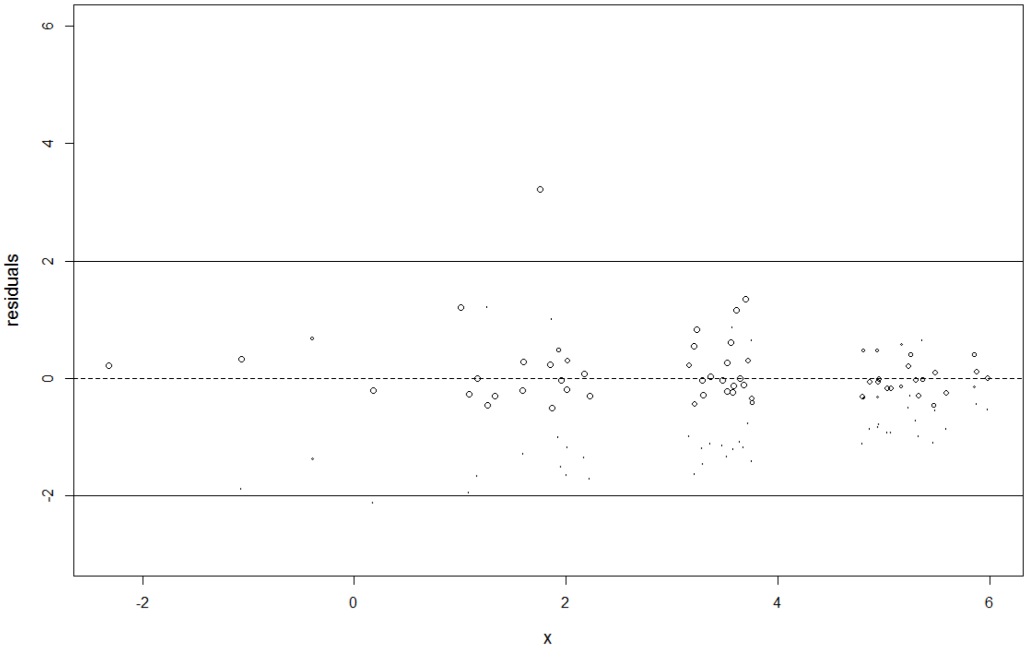

Figure 5.7. Residual plot for two-component regression model

This residual plot is generated using the method provided by Dr. Turner [12, 2000]. This method plots residuals on both lines first, then re-sizes the symbol of residuals according to its probability. When the symbol is smaller, the fitted value of return ($\hat{y}$) is less likely to be used, which means in order to read the residual plot correctly, small symbols should be ignored. There is only one outlier founded from the plot, which is less than the number of outliers in simple linear regression. The outlier is a company called Inktomi Corporation, a company that has a small revenue (5.785) contains the largest return (4.33) in whole IPO data set.

After eliminating four outliers that are shown in the simple linear regression model, we conducted both the simple linear regression model and the mixture model on the data set. This time, both the simple linear regression and the 2-components model were acceptable. The p-value from F-test in the simple linear regression is less than 0.05 (0.0443), which means there is a linear relation between return and revenue. Although the simple linear regression model performs well in this data set, the 2-component model generates a better result: the likelihood of the 2-component model is -38.58, while the likelihood of the simple linear regression model is -54.34. Additionally, both BIC and AIC select the 2-component model as the best model: AIC for the simple linear model equals 114.7, while AIC for the 2-component model equals 91.2; BIC for the simple linear model equals 121, while BIC for the 2-component model equals 105.

6. CONCLUSION

The main contribution of this thesis is fitting mixture regression model into a financial data and explain the relationship between revenue and return, which are two important variables in financial analysis. Normally, the relation between revenue and return is linear when the financial data has only one population. According to the IPO data however, there is more than one population in the data, thus, mixture model is proposed.

To determine the number of components in mixture model or how many sub-populations are there in IPO data, the statistical criteria BIC, AIC, and bootstrapping of likelihood ratio test statistic are used. The 2-component linear model is selected based on the smallest BIC and the results of the bootstrapping test. In order to estimate the parameters in 2-component linear model, the EM algorithm is employed.

To apply the EM algorithm without providing the initial values, an effective initialization method based on Euclidean distance is used. Compared with two commercial packages, this initialization method produces the most stable result.

At this point, two simulation studies are conducted to test the EM algorithm. The performance of the EM algorithm with the initialization method cannot be affected by either σs, πs or different scenarios. In addition, the performance of EM is better than other two R packages: mixtools and mixreg.

Finally, the EM algorithm is used to model return based on revenue of 116 IPO companies. A 2-component mixture of regressions is selected as the best model. The result indicates that a greater revenue implies a smaller return, which proves the assumption that summarized by Dr. Hand [7, 2007]: When the size of a company is increasing, the return of this company decreases. The results also indicate that when the company has a large return, the return of this company is more likely to decrease when the revenue of this company is increasing; meanwhile, when the company has a small return, the return of this company is more likely to increase when the revenue of this company is increasing. By comparing the 2-component mixture model with the simple linear regression, we conclude that when the revenue is increasing, the difference between the fitted return that produced by the simple linear regression model and the fitted return that produced by the 2-component mixture model decreases. Additionally, according to the log-likelihood, AIC, and BIC; although the performance of the simple linear regression model is acceptable after 4 outliers are removed, the 2-component model still provides the best results.

REFERENCES

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic control*, 19:716–723, 1974.

[2] T. Benaglia, D. R. Hunter, D. Chauveau, and D. S. Young. *Mixtools*. CRAN, 2009.

[3] R. D. De Veaux. Mixtures of linear regressions. *Computational Statistics & Data Analysis*, (8):227–245, 1989.

[4] E. Fama and K. French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 43:3–56, 1993.

[5] S. Faria and G. Soromenho. Fitting mixtures of linear regressions. *Journal of Statistical Computation and Simulation*, 80(2):201–225, 2010.

[6] E. W. Frees. *Regression modeling with actuarial and financial applications*. Cambridge University Press, 2010.

[7] J. R. M. Hand. Determinants of the round-to-round returns to pre-ipo venture capital investments in u.s. biotechnology companies. *Journal of Business Venturing*, 22(1):1–28, 2007.

[8] R. Maitra. Initializing partition-optimization algorithms. *IEEEACM Transactions on computational biology and bioinformatics*, 6(1):114–157, 2009.

[9] V. Melnykov and I. Melnykov. Initializing the em algorithm in gaussian mixture models with an unknown number of components. *Computational Statistics and Data Analysis*, 56:1381–1395, 2012.

[10] D. Peel and G. McLachlan. *Finite mixture models*. John & Sons, 2000.

[11] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[12] R. Turner. Estimating the propagation rate of a viral infection of potato plants via mixture of regressions. *Applied Statistics*, 49:371–384, 2000.

[13] R. Turner. *Mixreg*. CRAN, 2000.

[14] I. H. Witten, E. Frank, and M. A. (2011) Hall. *Data mining*. Morgan Kaufmann, 2011.