New Hybrid Model for Simulation of Investment Return

Akihiro Fukushima, LuoYan Zhou – Mentor: Professor Daniel Egger

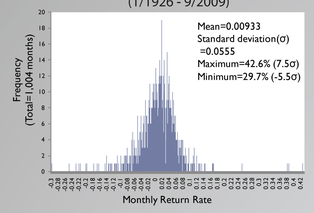
**Data-Mining, Optimization, and Data-Visualization course (2009 Fall)
in the Master of Engineering Management Program**

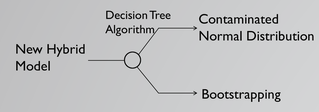
CENTER FOR QUANTITATIVE MODELING

Executive Summary

Future investment return outcomes are often simulated by making random selections with replacement from a normal probability distribution with mean and standard deviation derived from a historical data-set. However, the distribution of investment returns may not be normal; monthly returns of the S&P 500 index, for example, have a significantly taller central "peak" and "fat tails." Our goal was to develop a simple model for simulation purposes that accurately reflects the non-normal features of S&P 500 index data. The resulting hybrid model is a composite of a Contaminated Normal Distribution (CND) model and a Bootstrapping model with a Decision Tree algorithm. CND is a mixture of two different normal distributions. CND can generate a taller peak, but cannot simulate fat tails. On the other hand, Bootstrapping is random sampling with replacement directly from the historical data, without smoothing. Bootstrapping can simulate both the taller peak and the fat tails of the historical data, but it has issues of discrete events because it is based only on observed data. Our hybrid model can generate a taller peak and fat tails, while resolving the issues of discrete events that limit the usefulness of Bootstrapping alone.

Monthly Returns of the S&P 500 index
(1/1926 - 9/2009)

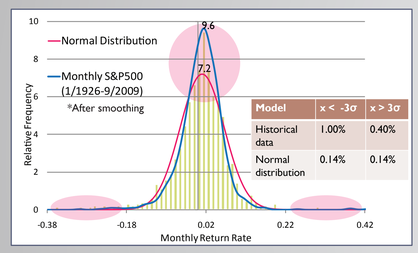




Problem Difinition

Financial service firms have used a Monte Carlo simulation with a normal distribution model in order to simulate a return on investments. However, a normal distribution does not fit the actual distribution of monthly S&P 500 index returns (1/1926 to 9/2009) as the right-hand chart illustrates. Compared to a normal distribution, the historical data has a taller peak and fat tails. With regard to a peak, while the maximum Relative Frequency in the actual distribution is 9.6, that in the normal distribution is 7.2. Moreover, the historical data has 1% of observed data at less than minus three standard deviations (-3σ), and 0.4% of observed data at greater than plus three standard deviations (+3σ), while a normal distribution has only 0.14% <-3σand 0.14% > +3σ.

The objective of this project is to find a model that better fits the actual historical distribution.



Model	x < -3σ	x > 3σ
Historical data	1.00%	0.40%
Normal distribution	0.14%	0.14%

Procedure

Monte Carlo simulation

A Monte Carlo simulation is random sampling from a given probability distribution. As a result of a large number of trials, the mean and the distribution of outcomes are obtained. Financial service firms use Monte Carlo simulations for forecasting the probability of various long-term investment outcomes.

In this project, we conducted different numbers of trials (ranging from 1×10^3 trials up to 3×10^4 trials) in order to evaluate the effect of the number of trials on the accuracy of the simulation. As a Monte Carlo simulation engine, we used RiskSim, distributed by Decision Toolworks (available at www.treeplan.com).

Metrics to evaluate best fit

1. Absolute Difference

Absolute Difference is a metric to evaluate whether or not each model achieves better fit to the actual distribution. As the chart below shows, Absolute Difference is the sum of the absolute values of the differences between the historical distribution and a simulated distribution. This metrics is sensitive to the size of bin used. If the size of bin is small, differences will be exaggerated. Toward achieving an optimal result, we chose 0.01% as the width of bin.

Total historical data
=1,004 months
Bin=0.01% (Return Rate)
H*i*=historical data in *i*th bin
S*i*=simulation data in *i*th bin

Absolute Difference
$=\Sigma|Hi-Si|=314.6$ months

Percentage of Difference
=Absolute Difference
/Total historical data =31.3%

Kernel Density Estimation

In order to smooth histograms obtained from a Monte Carlo simulation, we ran Kernel Density Estimation (KDE). We chose the Matlab program written by Z.Botev (2007) because it can automatically calculate an appropriate bandwidth. Moreover, we applied the parameter of $n=2^9$ in the program (n is defined as "the number of mesh points used in the uniform discretization of the interval"). The bandwidths of the historical data and the new hybrid model (1×10^4 trials) calculated by the program are respectively 0.0110 and 0.0066.

2. Relative Frequency

Relative Frequency is a measure of how tall a peak is in each distribution. Relative Frequency is one of the outputs from the Botev's KDE program, which is defined as "density" in the program.

3. Probabilities of tails

Probabilities of tails is a measure for the evaluation of the fat tails problem. After calculating probabilities of less than minus three sigma and more than three sigma, we compared the probabilities of a simulated distribution with those of the historical distribution (1% for x<-3σ, 0.4% for x>3σ).

Tested Models

A. Contaminated Normal Distribution model

A Contaminated Normal Distribution (CND) model can change the kurtosis (Skew Normal Distribution can change skewness). This project adopts the CND which is a composite of two normal distribution having different variances. The equation below illustrates the parameters of CND.

$$CND = F_{ND} * \sigma1*(1-P) + F_{ND} *\sigma2*P + \mu$$
$$(\sigma1 < \sigma s < \sigma2,\ 0<P<1)$$

F_{ND}=Standardized Normal Distribution
μ=Mean of monthly S&P 500 index
σs=Standard Deviation of monthly S&P 500 index

We estimated the appropriate parameters of σ1 and σ2 and P with Chi-square goodness-of-fit test. Specifically, after creating different CNDs by changing σ1 and σ2 and P, the CNDs were divided into 80 areas. In this case, the expected number of elements in each area is 12.55 (1,004 months divided by 80 areas). By comparing the expected number and the numbers of elements in 80 areas in each CND, we found the CND with appropriate parameters (σ1=0.04, σ2=0.11, P=0.13). The strength of the CND model is that it generates a taller peak than is possible with a normal distribution alone.

B. Bootstrapping model

A Bootstrapping model randomly samples a return rate with replacement from a collection of observed data. For example, in the case of 3×10^3 trials, a Bootstrapping model chooses a return rate from 1,004 samples of monthly S&P 500 index (1/1926 to 9/2009) 3×10^3 times. A Bootstrapping model uses the same monthly data any number of times. So, if the total number of observed data is small, a Bootstrapping model tends to choose the same return rate. The strength of the Bootstrapping model is to generate fat tails (simulate outlier events) which a normal distribution model and CND cannot do.

Historical Data

Date	S&P500
Jan-1995	0.02593
Feb-1995	**0.03898**
Mar-1995	0.02951
Apr-1995	0.02945
May-1995	0.03997
Jun-1995	**0.02323**
Jul-1995	0.03316
Aug-1995	0.00251

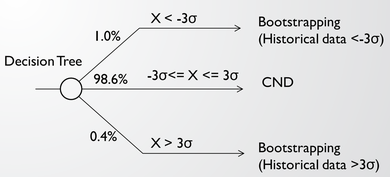
Random Selection with replacement

Simulated data

Trial	Return Rate
1	**0.02323**
2	**0.03898**
3	**0.02323**
4	
5	
….	
2999	
3000	

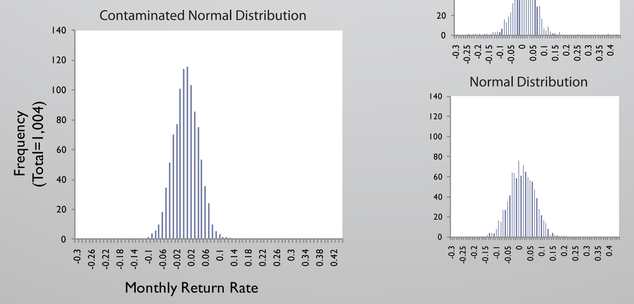
C. New hybrid model

Our new hybrid model is a composite of the CND model and the Bootstrapping model with a Decision Tree algorithm. With the probabilities of 1.0% and 0.4%, the hybrid model chooses the Bootstrapping model. Otherwise, the hybrid model chooses the CND model. By implementing this decision tree algorithm, the hybrid model achieves the benefits of Both CND and Bootstrapping.



A. Result of Contaminated Normal Distribution

As expected, the CND model generated a taller peak. The histograms are the results of a Monte Carlo simulation with 3×10^3 trials. While a normal distribution model has a lower peak, the CND model has a taller peak. However, the CND model cannot solve the fat tails problem.



B. Result of Bootstrapping model

The Bootstrapping model can solve the fat tails problem. As the table below shows, Bootstrapping (3×10^4 trials) contains only 2.4% of Absolute Difference. In addition, Bootstrapping (2×10^4 trials) has tails similar to the historical data. However, Bootstrapping may exaggerate the influence of discrete observations when using a limited data-set (here, 1,004 months). In other words, Bootstrapping is sampling with replacement from a specific collection of observed data rather than from a smooth probability density function.

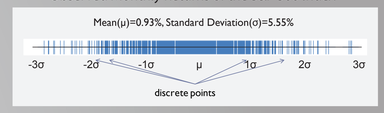
Absolute Difference

Simulation	Absolute difference	Percentage of difference
Normal distribution 3×10^3 trials	314.6	31.3%
CND 3×10^3 trials	223.6	22.3%
CND 1×10^4 trials	187.3	18.7%
Bootstrapping (1×10^3 trials)	128.3	12.8%
Bootstrapping (3×10^3 trials)	63.9	6.4%
Bootstrapping (1×10^4 trials)	45.8	4.6%
Bootstrapping (2×10^4 trials)	34.8	3.5%
Bootstrapping (3×10^4 trials)	24.4	2.4%

Probabilities of tails

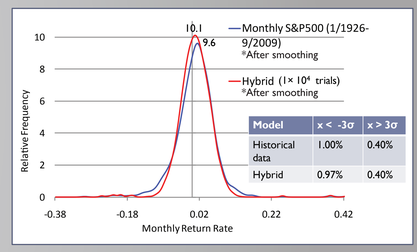
Model	Probability of x < -3σ	Probability of x > 3σ
Historical data (1/1926-9/2009)	1.00%	0.40%
Normal distribution	0.14%	0.14%
Bootstrapping (3×10^3 trials)	0.63%	0.37%
Bootstrapping (1×10^4 trials)	0.88%	0.30%
Bootstrapping (2×10^4 trials)	1.02%	0.40%

Observed Monthly Returns of the S&P 500 Index



C. Result of the new hybrid model

The new hybrid model generates a taller peak and fat tails. The maximum Relative Frequency in the hybrid model (10.1) is much closer to that in the actual distribution (9.6) than a normal distribution model (7.2). Also, the hybrid model can generate fat tails (Probabilities of tails are 0.97% and 0.40%). Moreover, the hybrid model can resolve the issues of discrete events because it does not rely only on observed data. The hybrid model compensates for Bootstrapping's weakness of limited results while it encompasses Bootstrapping's strength of simulating outlier events.



Model	x < -3σ	x > 3σ
Historical data	1.00%	0.40%
Hybrid	0.97%	0.40%

Conclusion

As the comparative chart illustrates, the hybrid model is superior to other three models with regards to generating a taller peak, generating fat tails, and resolving issues of discrete events. However, the hybrid model cannot simulate serial correlation which is the correlation between two consecutive months (e.g. correlation between 10/2009 and 11/2009). As a next step, we could examine the importance of serial correlation and a Vector Autoregression (VAR) model which can simulate serial correlation. Furthermore, we could implement these models into an actual financial simulation and examine how influential the differences between the models are.

	Normal Distribution	CND	Boot strapping	New hybrid model
Generate fat tails?	No	No	Yes	Yes
Generate a taller peak?	No	Yes	Yes	Yes
Resolve issues of discrete events?	Yes	Yes	No	Yes
Implement Cross-sectional correlation?	Yes	Yes	Yes	Yes
Implement Serial correlation?	No	No	No/Yes(long-term data)	No

Acknowledgement and Bibliography

Grateful acknowledgement is given to the following people and journals.

-Professor Daniel Egger for providing crucial ideas such as Absolute Difference and Kernel Density Estimation as well as organizing this project.
-Lois Scheirer and Rajeev Dharmapurikar for sponsoring this project.
-Alex Murguia for offering great articles and suggestions.
-Carlos Coutin for assistances with user scenarios and software.

[1] The Uses And Limits Of Volatility, David Harper, Investopedia (2004)
[2] The Strengths and Weaknesses of Various Financial Simulation Methods , Joseph H. Davis, Nelson W. Wicas and Francis M. Kinniry, The journal of Wealth Management (Spring 2004)
[3] Some Lessons from 250,000 Years of Stock Returns, Truman A. Clark, Dimensional Fund Advisor Inc. (April 2003)
[4] More Risk in Retirement Simulations, Marlena I. Lee, Dimensional Fund Advisor Inc. (January 2009)
[5] When Monte Carlo analysis meets a black swan, Moshe A. Milevsky, Investment News (May 2009)
[6] Modeling the Future , Glenn Kautt and Lynn Hopewell, FPA Journal (2000)
[7] MGP Response to Wall Street Journal Article on Monte Carlo Simulations, PIEtech, Inc. (2009)
[8] Contaminated Normal Distribution – Applied Analysis of Variance in Behavioral Science (1993), Lynne K. Edwards
[9] Chi-square test – Simulation Modeling and Analysis (Third Edition), Averill M. Law and W. David Kelton
[10] Kernel Density Estimation– Kernel Density Estimator, Zdravko Botev (21 Feb 2007, Updated 28 Jun 2009), http://www.mathworks.com/matlabcentral/ fileexchange/14034