An Application of the EM-algorithm to Approximate Empirical Distributions of Financial Indices with the Gaussian Mixtures

Sergey Tarasenko

Abstract—In this study I briefly illustrate application of the Gaussian mixtures to approximate empirical distributions of financial indices (DAX, Dow Jones, Nikkei, RTSI, S&P 500). The resulting distributions illustrate very high quality of approximation as evaluated by Kolmogorov-Smirnov test. This implies further study of application of the Gaussian mixtures to approximate empirical distributions of financial indices.

Keywords—financial indices, Gaussian distribution, mixtures of Gaussian distributions, Gaussian mixtures, EM-algorithm

I. INTRODUCTION

APPROXIMATION of empirical distributions of financial indices using mixture of Gaussian distributions (Gaussian mixtures) (eq. (1)) has been recently discussed by Tarasenko and Artukhov [1]. Here I provide detailed explanation of steps and methods.

\[
GM = \sum_{i=1}^{n} p_i \cdot N(\mu_i, \sigma_i)
\]  

(1)

II. EM-ALGORITHM FOR MIXTURE SEPARATION

A. General Theory

The effective procedure for separation of mixtures was proposed by Day [2], [3] and Dempster et al. [4]. This procedure is based on maximization of logarithmic likelihood function under parameters \(p_1, p_2, ..., p_k, \Theta_1, \Theta_2, ..., \Theta_k\), where \(k\) is number of mixture components:

\[
\sum_{i=1}^{n} \ln \left( \sum_{j=1}^{k} p_j \cdot f(x_i; \Theta_j) \right) \rightarrow \max_{p_j, \Theta_j}
\]  

(2)

In general, the algorithms of mixture separations based on [2] are called Estimation and Maximization (EM) algorithms. EM-algorithm consists of two steps: E - expectation and M-maximization. This section is focused scheme how to construct EM-algorithm.

Let \(g_{ij}\) is defined as posterior probability of observations \(x_i\) to belong to \(j\)-th mixture component (class):

\[
g_{ij} = \frac{p_j \cdot f(x_i; \Theta_j)}{\sum_{j=1}^{k} p_j \cdot f(x_i; \Theta_j)}
\]  

(3)

Posterior probability \(g_{ij}\) is equal or greater then 0 and \(\sum_{j=1}^{k} g_{ij}\) for any \(i\).

S. Tarasenko is independent researcher. Email: infra.core@gmail.com

Let \(\Theta\) be a vector of parameters: \(\Theta = (p_1, p_2, ..., p_k, \Theta_1, \Theta_2, ..., \Theta_k)\). Next, we decompose the logarithms likelihood function into three components:

\[
\ln L(\Theta) = \sum_{i=1}^{n} \ln \left( \sum_{j=1}^{k} p_j \cdot f(x_i; \Theta_j) \right) = \sum_{j=1}^{k} \sum_{i=1}^{n} g_{ij} \ln(p_j) + \sum_{j=1}^{k} \sum_{i=1}^{n} g_{ij} f(x_i; \Theta_j) - \sum_{j=1}^{k} \sum_{i=1}^{n} g_{ij} \ln(g_{ij})
\]  

(4)

For this algorithm to work, the initial value \(\hat{\Theta}^0\) is used to calculate initial approximations for posterior probabilities \(g_{ij}^0\). This is Expectation step. Then values of \(g_{ij}^0\) are used to calculate value of \(\hat{\Theta}^1\) during the Maximization step.

Each of components (6) and (7) are maximized independently from each other. This is possible because component (6) depends only on \(p_j\) \((i=1, ..., k)\), and component (7) depends only on \(\Theta_j\) \((j=1, ..., n)\).

As a solution of optimization task (8)

\[
\sum_{j=1}^{k} \sum_{i=1}^{n} g_{ij} \ln(p_j) \rightarrow \max_{p_1, ..., p_k}
\]  

(8)

a value of \(p_j^{(t+1)}\) for the iteration \(t + 1\) is calculated as:

\[
p_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} g_{ij}^{(t)}
\]  

(9)

where \(t\) is iteration number, \(t = 1, 2, ...\)

A solution of optimization task (10)

\[
\sum_{j=1}^{k} \sum_{i=1}^{n} g_{ij} f(x_i; \Theta_j) \rightarrow \max_{\Theta_1, ..., \Theta_k}
\]  

(10)

depends on a particular type of function \(f(\cdot)\).

Next we consider solution of optimization task (10), when \(f(\cdot)\) is Gaussian distribution.
B. Mixtures of Gaussian Distributions

Here we employ Gaussian distributions:

\[ N(x; \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \]  

(11)

Therefore, a specific formula to compute posterior probabilities in the case of Gaussian mixtures is

\[ g_{ij} = \frac{\exp\left(-\frac{(x-\mu_j)^2}{2\sigma_j^2}\right) + \ln(p_j) - \ln(\sigma_j)}{\sum_{j=1}^{k} \exp\left(-\frac{(x-\mu_j)^2}{2\sigma_j^2}\right) + \ln(p_j) - \ln(\sigma_j)} \]

(12)

According to the EM-algorithm, the task is to find values of parameters \( \Theta_j = (\mu_j, \sigma_j) \) by solving maximization problem

\[ \sum_{j=1}^{k} \ln L_j = \sum_{j=1}^{k} \sum_{i=1}^{n} \ln L_j = \]

(13)

\[ \sum_{j=1}^{k} \sum_{i=1}^{n} \ln\left(\frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(x-\mu_j)^2}{2\sigma_j^2}\right) \right) \rightarrow \max \Theta_1, \Theta_2, ..., \Theta_k \]

(14)

The solution of this maximization problem is given by eqs. (15) and (16):

\[ \hat{\mu}_j = \frac{1}{\sum_{i=1}^{n} g_{ij}} \sum_{i=1}^{n} g_{ij} x_i \]

(15)

\[ \hat{\sigma}_j = \frac{1}{\sum_{i=1}^{n} g_{ij}} \sum_{i=1}^{n} g_{ij} (x_i - \hat{\mu}_j)^2 \]

(16)

Having calculated optimal values for weights \( p_j \) and parameters \( \Theta_j \) \((j=1, ..., k)\) during a single iteration, we apply these optimal values to obtain estimates of posterior probabilities during the Expectation step of the next iteration.

As a stop criterion, we use difference between values of loglikelihood on iteration \( t \) and iteration \( t + 1 \):

\[ \ln L^{(t+1)}(\Theta) - \ln L^{(t)}(\Theta) < \epsilon \]

(17)

where \( \epsilon \) is infinitely small real value.

III. A N APPLICATION OF THE EM-ALGORITHM TO APPROXIMATE EMPIRICAL DISTRIBUTIONS OF FINANCIAL INDICES WITH GAUSSIAN MIXTURES

In this section, I provide several examples of EM-algorithm applications to approximate empirical distributions of financial indices with Gaussian mixtures. I consider the following indices: DAX, Dow Jones Industrial, Nikkei, RTSI, and S&amp;P 500.

In Figs. 1-5 the green line corresponds to the Gaussian distribution, the red line illustrates a Gaussian mixture and the blue lines represent components of the Gaussian mixture.
TABLE III
MIXTURE MODEL FOR NIKKEI

| Component | Weight | Mean    | Standard Deviation |
|-----------|--------|---------|-------------------|
| Component 1 | 0.167  | -0.014  | 0.014             |
| Component 2 | 0.180  | 0.002   | 0.013             |
| Component 3 | 0.367  | 0.011   | 0.008             |
| Component 4 | 0.286  | -0.002  | 0.005             |

Fig. 3. Empirical distribution of Nikkei index values during the period 14 April 2003 - 14 April 2004, p < 0.01, KSSTAT=0.027

TABLE IV
MIXTURE MODEL FOR RTSI

| Component | Weight | Mean    | Standard Deviation |
|-----------|--------|---------|-------------------|
| Component 1 | 0.062  | -0.014  | 0.044             |
| Component 2 | 0.294  | -0.005  | 0.019             |
| Component 3 | 0.303  | 0.005   | 0.014             |
| Component 4 | 0.341  | 0.011   | 0.011             |

Fig. 4. Empirical distribution of RTSI values during the period 14 April 2003 - 14 April 2004, p < 0.01, KSSTAT=0.021

TABLE V
MIXTURE MODEL FOR S&P 500

| Component | Weight | Mean    | Standard Deviation |
|-----------|--------|---------|-------------------|
| Component 1 | 0.014  | 0.011   | 0.027             |
| Component 2 | 0.331  | 0.000   | 0.009             |
| Component 3 | 0.470  | 0.001   | 0.009             |
| Component 4 | 0.186  | 0.000   | 0.001             |

Fig. 5. Empirical distribution of S&P 500 values during the period 14 April 2003 - 14 April 2004, p < 0.01, KSSTAT=0.024

IV. DISCUSSION AND CONCLUSION

The results presented in this study illustrate that EM-algorithm can be effectively used to approximate empirical distributions of log daily differences of financial indices. Throughout the data for five selected indices, EM-algorithm provided very good approximation of empirical distribution with Gaussian mixtures.

The approximations based on Gaussian mixtures can be used to improve application of Value-at-Risk and other methods for financial risk analysis.

This implies further explorations in applying Gaussian mixtures and the EM-algorithm for the purpose of approximation of empirical distributions of financial indices.

REFERENCES

[1] Tarasenko, S., and Artukhov, S. (2004) Stock market pricing models. In the Proceedings of International Conference of Young Scientists Lomonosov 2004, p. 248-250.
[2] Day, N.E. (1969) Divisive cluster analysis and test for multivariate normality. Session of the ISI, London, 1969.
[3] Day, N.E. (1969) Estimating the components of a mixture of normal distributions., Biometrika, 56, N3.
[4] Dempster, A., Laird, G. and Rubin, J. (1977) Maximum likelihood from incomplete data via EM algorithm. Journal of Royal Statistical Society, B, 39.