MIXTURE-OF-EXPERT ARMA-GARCH MODELS FOR STOCK PRICE PREDICTION

Him Tang, and Lei Xu

Department of Computer Science and Engineering, The Chinese University of Hong Kong
Shatin, New Territories, Hong Kong, P. R. China
Email: {htang,lxu}@cse.cuhk.edu.hk

ABSTRACT

1. INTRODUCTION

In recent years, researchers are interested in modelling nonlinear financial time series. Conventional Autoregressive Moving Average (ARMA) assume that a time series can be modelled using a linear difference equation. However, most time series can be modelled by using more than one time series, i.e. a mixture of linear time series. In the literature, finite mixture of autoregressive (AR) model, finite mixture of autoregressive moving average (ARMA) model [1], and finite mixture of autoregressive generalized autoregressive conditional heteroscedasticity (AR-GARCH) model [2] have been adopted for finance exchange rate prediction. These mixture models have been shown to successfully capture more than one time series information from the data points, and can give a better prediction performance than using single linear time series model. Recently, finite mixture of ARMA-GARCH model has been developed [3] which can model a larger class of time series structures and gives better prediction performance than other finite mixture models.

In this paper, we introduce finite mixture-of-expert ARMA-GARCH model for financial time series forecasting. Mixture-of-expert architecture has been well developed and has been published in many artificial intelligence literature [4, 5, 6]. The mixture-of-expert model is different from finite Gaussian mixture model in the way that data distribution in mixture-of-expert is conditioned on input vector. In the case of ARMA-GARCH model, the input vector corresponds to the lagged time series data. This causes the weights of different experts change along with time, where the weights are constants in the case of Gaussian mixture. The mixture-of-expert ARMA-GARCH model has been derived and we use GEM algorithm to train the mixture model.

Experiments have been conducted for four mixture models: finite Gaussian mixture AR-GARCH and ARMA-GARCH model, finite mixture-of-expert AR-GARCH and ARMA-GARCH model. These models have been used to predict the daily stock prices of HSBC Holding (HSBC HDG) and Cheung Kong Holding (CK HDG). It will be shown that for the second step prediction, mixture-of-expert model outperforms the finite Gaussian mixture model.

2. FINITE MIXTURE-OF-EXPERT ARMA-GARCH MODEL

The mixture-of-expert ARMA-GARCH model is similar to the Gaussian mixture of ARMA-GARCH model proposed in [3]. Specifically, each expert can be denoted as a normal ARMA series

$$x_{t,j} = \sum_{r=1}^{R} b_{rj} x_{t-r} + \sum_{s=1}^{S} a_{sj} \epsilon_{t-s} + \epsilon_{t,j}, \qquad (1)$$

Furthermore, each residual term $\epsilon_{t,j}$ is assumed gaussian white noise with variance denoted by the GARCH model

$$\sigma_{t,j}^2 = \delta_{0j} + \sum_{q=1}^{Q} \delta_{qj} \epsilon_{t-q}^2 + \sum_{p=1}^{P} \beta_{pj} \sigma_{t-p,j}^2, \qquad (2)$$

where $\delta_{qj} > 0$ for $q = 1, \ldots, Q$ and $\beta_{pj} > 0$ for $p = 1, \ldots, P$.

In this paper, we use the alternative mixture-of-expert model proposed in [7]. Mathematically, the finite mixture-of-expert ARMA-GARCH model can be denoted as a K-expert mixture model

$$p(x_t, y_t) = \sum_{j=1}^{K} \alpha_j G(y_t; \mu_j, \Lambda_j) G(x_t; \hat{x}_{t,j}, \sigma_{t,j}^2), \qquad (3)$$

$$\hat{x}_{j,t} = \sum_{r=1}^{R} b_{rj} x_{t-r} + \sum_{s=1}^{S} a_{sj} \epsilon_{t-s}, \qquad (4)$$

The work described in this paper was fully supported by a grant from the Research Grant Council of the Hong Kong SAR (Project No: CUHK 4169/00E).

where $\alpha_j > 0$ and $\sum_{j=1}^{K} \alpha_j = 1$. The gating network is a multivariate Gaussian distribution

$$G(y_t; \mu_j, \Lambda_j) = \frac{1}{(2\pi)^{\frac{n}{2}} \sqrt{|\Lambda_j|}} e^{-\frac{1}{2}(y_t-\mu_j)^T \Lambda_j^{-1}(y_t-\mu_j)}, \tag{5}$$

μ_j is the mean and Λ_j is the covariance matrix of the Gaussian. y_t is the input vector, which has the form

$$\begin{aligned} y_t &= (x_{t-1}, \ldots, x_{t-R}, \epsilon_{t-1}, \ldots, \epsilon_{t-S}, \\ &\quad \sigma^2_{t-1,1}, \ldots, \sigma^2_{t-P,1}, \ldots, \sigma^2_{t-1,K}, \ldots, \sigma^2_{t-P,K})'. \end{aligned}$$

Once the model has been learned, one-step ahead prediction can be done via taking expectation of x_t

$$\begin{aligned} E(x_t) &= \hat{x}_t = \frac{\alpha_1 G(y_t; \mu_1, \Lambda_1)}{\sum_{i=1}^{K} \alpha_i G(y_t; \mu_i, \Lambda_i)} \hat{x}_{t,1} + \ldots + \\ &\quad \frac{\alpha_K G(y_t; \mu_K, \Lambda_K)}{\sum_{i=1}^{K} \alpha_i G(y_t; \mu_i, \Lambda_i)} \hat{x}_{t,K}, \end{aligned} \tag{6}$$

and so

$$\epsilon_t = x_t - \hat{x}_t. \tag{7}$$

For the second step prediction, it is less straight forward. We need to calculate the second step distribution. However, exact calculation of the second step distribution is often intractable. So we will use the Monte Carlo method. Firstly, we introduce the conditional *density* function of the mixture-of-expert model

$$\begin{aligned} & F(x_t | X_{t-1}), \\ = & \int_{-\infty}^{+\infty} p(x_t, y_t) \partial x_t \end{aligned} \tag{8}$$

where $X_{t-1} = \{x_{t-1}, x_{t-2}, \ldots\}$, and $p(x_t, y_t)$ is same as (3). So (8) represents the one step forwards conditional density function of the mixture-of-expert model at x_t, which is conditioned on the past data X_{t-1}. We called this the first step distribution. We randomly generate many first step values x_t^i using the first step distribution, $F(x_t|X_{t-1})$. For each generated first step value, we treat it as the true first step, and use it to calculate the second step distribution. Mathematically, the second step distribution can be represented as

$$F(x_{t+1}|X_{t-1}) = \frac{1}{N} \sum_{i=1}^{N} F(x_{t+1}|X_{t-1}, x_t^i), \tag{9}$$

where $X_{t-1} = \{x_{t-1}, x_{t-2}, \ldots\}$, $F(x_{t+1}|\cdot)$ is a conditional second step density function of the mixture model. In equation (9), N number of x_t^i have been randomly generated using the first step distribution $F(x_t|X_{t-1})$.

3. DERIVATION OF THE GENERALIZED EXPECTATION-MAXIMIZATION (GEM) ALGORITHM FOR IMPLEMENTATION

3.1. The E step

The joint probability of the finite mixture model is

$$p(X, Y) = \prod_{t=1}^{T} \sum_{j=1}^{K} \alpha_j G(y_t; \mu_j, \Lambda_j) G(x_t; \hat{x}_{t,j}, \sigma^2_{t,j}), \tag{10}$$

and we want to maximize the log-likelihood of (10)

$$\hat{\Theta} = \arg\max_{\Theta} \ln p(X, Y) \tag{11}$$

where $\Theta = \{\{a_{sj}\}_{s=1}^{S}, \{b_{rj}\}_{r=1}^{R}, \{\beta_{pj}\}_{p=1}^{P}, \{\delta_{qj}\}_{q=1}^{Q}, \alpha_j, \mu_j, \Lambda_j\}_{j=1}^{K}$, $X = \{x_t\}_{t=1}^{T}$, $Y = \{y_t\}_{t=1}^{T}$.

We use the EM algorithm [8] to maximize the log-likelihood. Let the unobserved variables are $Z = \{z_t\}_{t=1}^{T}$. We define that if x_t is produced by the uth component, then $z_t = u$. And so

$$p(x_t, y_t | z_t = u) = G(y_t; \mu_u, \Lambda_u) G(x_t; \hat{x}_{t,u}, \sigma^2_{t,u}), \tag{12}$$

$$p(z_t = u) = \alpha_u. \tag{13}$$

And the Q function of the EM algorithm will be

$$\begin{aligned} & Q(\Theta, \Theta^*) \\ &= E_{Z|X,Y\Theta^*} \left\{ \sum_{t=1}^{T} \ln[p(x_t, y_t|z_t) p(z_t)] \right\} \\ &= \sum_{z_1=1}^{K} \cdots \sum_{z_T=1}^{K} \\ &\quad \left\{ \sum_{t=1}^{T} \ln[p(x_t, y_t|z_t) p(z_t)] \prod_{l=1}^{T} p(z_l|x_l, y_l, \Theta^*) \right\} \\ &= \sum_{z_t=1}^{K} \sum_{t=1}^{T} \ln[p(x_t, y_t|z_t) p(z_t)] p(z_t|x_t, y_t, \Theta^*) \\ &= \sum_{t=1}^{T} \sum_{j=1}^{K} \ln[p(x_t, y_t|z_t) p(z_t = j)] p(z_t = j|x_t, y_t, \Theta^*) \\ &= \sum_{t=1}^{T} \sum_{j=1}^{K} [p(z_t = j|x_t, y_t, \Theta^*) \\ &\quad \ln \alpha_j G(y_t; \mu_j, \Lambda_j) G(x_t, y_t; \hat{y}_{t,j}, \sigma^2_{t,j}).] \end{aligned} \tag{14}$$

The probability $p(z_t = j|x_t, y_t, \Theta^*)$ (denoted as $h_j(t)$) can be obtained as follows,

$$\begin{aligned} & p(z_t = j|x_t, y_t, \Theta^*) \\ = \; & h_j(t) \\ = \; & \frac{\alpha_j G(y_t; \mu_j, \Lambda_j) G(x_t; \hat{x}_{t,j}, \sigma^2_{t,j}, y_t)}{\sum_{u=1}^{K} \alpha_u G(y_t; \mu_u, \Lambda_u) G(x_t; \hat{x}_{t,u}, \sigma^2_{t,u}, y_t)} \end{aligned} \tag{15}$$

This is the E step of the EM algorithm.

3.2. The M step

In the M step we maximize (14). Since β, δ and α must bigger than zero, so we replace them by:

$$\delta_{0j} = e^{\gamma_{0j}}, \tag{16}$$
$$\delta_{qj} = e^{\gamma_{qj}}, \quad \text{where } q = 1, \ldots, Q \tag{17}$$
$$\beta_{pj} = e^{\rho_{pj}}, \quad \text{where } p = 1, \ldots, P \tag{18}$$

To ensure $\{\alpha_j\}_{j=1}^K$ sum to unity and each α_j is larger than zero, we use

$$\alpha_j = \frac{e^{m_j}}{\sum_{i=1}^K e^{m_i}}. \tag{19}$$

The parameters that we will need to adjust are $\{m_j, \mu_j, \Lambda_j\}_{j=1}^K$ and $\Omega = \{a_{1j}, \ldots, a_{Sj}, b_{1j}, \ldots, b_{Rj}, \rho_{1j}, \ldots, \rho_{Pj}, \gamma_{0j}, \ldots, \gamma_{Qj}\}_{j=1}^K$.

The first derivatives of (14) with respect to m_j, μ_j, and Λ_j are

$$\frac{\partial Q(\Theta, \Theta^*)}{\partial m_j} = h_j(t) - \alpha_j. \tag{20}$$
$$\frac{\partial Q(\Theta, \Theta^*)}{\partial \mu_j} = h_j(t)\frac{1}{2}(\Lambda_j^{-1} + \Lambda_j^{-T})(y_t - \mu_j) \tag{21}$$
$$\frac{\partial Q(\Theta, \Theta^*)}{\partial \Lambda_j} = h_j(t)\frac{1}{2}\left[\Lambda_j^{-T}(y_t - \mu_j)(y_t - \mu_j)^T - I\right]\Lambda_j^{-T} \tag{22}$$

For all other parameters, the first derivative is

$$\frac{\partial Q(\Theta, \Theta^*)}{\partial \omega} = h_j(t)\left[\frac{1}{2\sigma_{t,j}^2}\left(\frac{\epsilon_{t,j}^2}{\sigma_{t,j}^2} - 1\right)\frac{\partial \sigma_{t,j}^2}{\partial \omega} - \frac{\epsilon_{t,j}}{\sigma_{t,j}^2}\frac{\partial \epsilon_{t,j}}{\partial \omega}\right] \tag{23}$$

Where $\omega \in \Omega$. To calculate (23), we also need the following first derivatives:

$$\frac{\partial \sigma_{t,j}^2}{\partial a_{sj}} = 2\sum_{c=1}^{Q} \epsilon_{t-c,j} e^{\gamma_{cj}} \frac{\partial \epsilon_{t-c,j}}{\partial a_{sj}} + \sum_{d=1}^{P} e^{\rho_{dj}} \frac{\partial \sigma_{t-d,j}^2}{\partial a_{sj}},$$
$$\frac{\partial \epsilon_{t,j}}{\partial a_{sj}} = -\epsilon_{t-s,j} - \sum_{c=1}^{S} a_{cj} \frac{\partial \epsilon_{t-c,j}}{\partial a_{sj}},$$
$$\frac{\partial \sigma_{t,j}^2}{\partial b_{rj}} = 2\sum_{c=1}^{Q} \epsilon_{t-c,j} e^{\gamma_{cj}} \frac{\partial \epsilon_{t-c,j}}{\partial b_{rj}} + \sum_{d=1}^{P} e^{\rho_{dj}} \frac{\partial \sigma_{t-d,j}^2}{\partial b_{rj}},$$
$$\frac{\partial \epsilon_{t,j}}{\partial b_{rj}} = -y_{t-r,j} - \sum_{c=1}^{S} a_{cj} \frac{\partial \epsilon_{t-c,j}}{\partial b_{rj}},$$
$$\frac{\partial \sigma_{t,j}^2}{\partial \gamma_{0,j}} = e^{\gamma_{j0}} + \sum_{c=1}^{P} e^{\rho_{cj}} \frac{\partial \sigma_{t-c,j}^2}{\partial \gamma_{0j}},$$
$$\frac{\partial \sigma_{t,j}^2}{\partial \gamma_{qj}} = \epsilon_{t-q,j}^2 e^{\gamma_{qj}} + \sum_{c=1}^{P} e^{\rho_{cj}} \frac{\partial \sigma_{t-c,j}^2}{\partial \gamma_{qj}},$$
$$\frac{\partial \sigma_{t,j}^2}{\partial \rho_{pj}} = \sigma_{t-p,j}^2 e^{\rho_{pj}} + \sum_{c=1}^{P} e^{\rho_{cj}} \frac{\partial \sigma_{t-c,j}^2}{\partial \rho_{pj}},$$
$$\frac{\partial \epsilon_{t,j}}{\partial \gamma_{0j}} = \frac{\partial \epsilon_{t,j}}{\partial \gamma_{qj}} = \frac{\partial \epsilon_{t,j}}{\partial \rho_{pj}} = 0.$$

Where $r = 1, \ldots, R$, $s = 1, \ldots, S$, $q = 1, \ldots, Q$, $p = 1, \ldots, P$.

Since we use GEM algorithm, we only adjust each parameter more towards the maximum in each M step. Let θ be one of the parameters. To update θ we use:

$$\theta^{(n+1)} = \theta^{(n)} + \lambda_n \frac{\partial E(l_t)}{\partial \theta}, \tag{24}$$

where $\theta \in \Theta$.

To ensure the estimated model will be stationary, the initial characteristic equations of every ARMA model

$$1 - b_{1,j}z - b_{2,j}z^2 - \cdots - b_{R,j}z^R = 0, \tag{25}$$

and every GARCH model

$$1 - \beta_{1,j}z - \beta_{2,j}z^2 - \cdots - \beta_{P,j}z^P = 0, \tag{26}$$

must have all their roots lie outside the unit circle [9]. During our experiments, if the initial characteristic equations have all their roots lie outside the unit circle, the estimated results will also be stationary.

4. EXPERIMENTS: FIRST STEP PREDICTION

The period that we will be used to train the model is from September 15, 1997 to July 11, 1998, which consists of 300 data. Two stocks are under investigation,

Cheung Kong Holding (CK HDG) and HSBC Holding (HSBC HDG). The model ARMA(1,1)-GARCH(1,1) is being used. We then predict the following 300 data points. Figures 1 to 6 show the first step prediction for the two stocks using three different approaches: conventional ARMA-GARCH, Gaussian mixture ARMA-GARCH and mixture-of-expert ARMA-GARCH. Solid lines are the actual stock prices, dashed lines are the predicted prices.

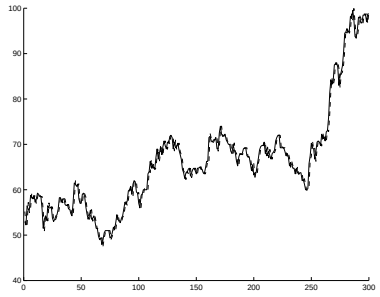

Figure 1: First step prediction of CK prices with conventional ARMA-GARCH.

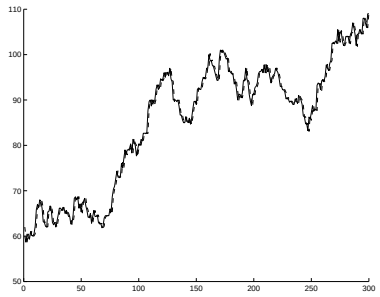

Figure 2: First step prediction of HSBC prices with conventional ARMA-GARCH.

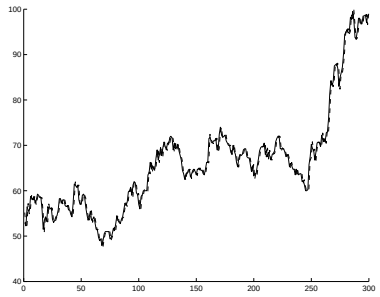

Figure 3: First step prediction of CK prices with Gaussian mixture ARMA-GARCH.

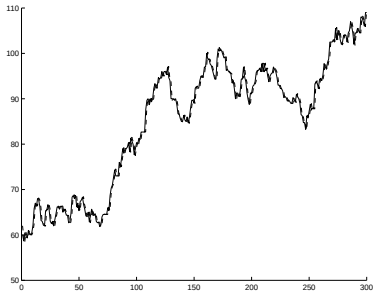

Figure 4: First step prediction of HSBC prices with Gaussian mixture ARMA-GARCH.

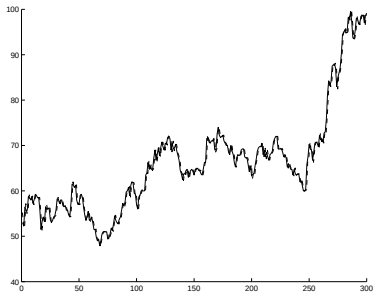

Figure 5: First step prediction of CK prices with mixture-of-expert ARMA-GARCH.

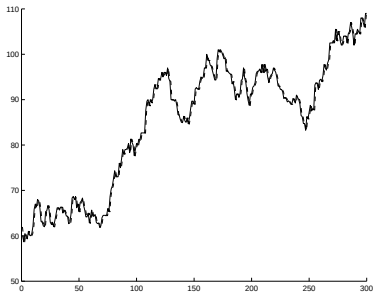

Figure 6: First step prediction of HSBC prices with mixture-of-expert ARMA-GARCH.

	CK HDG	HSBC HDG
Conventional	2.3799	2.1651
Gaussian Mixture	2.1101	1.9899
Mixture-of-Expert	2.0030	1.9225

Table 1: Mean square errors of first step prediction using different ARMA-GARCH models.

	CK HDG	HSBC HDG
Conventional	4.9627	4.2001
Gaussian Mixture	4.8216	3.9935
Mixture-of-Expert	4.4147	3.7120

Table 2: Mean square errors of second step prediction using different ARMA-GARCH models.

Table 1 shows the mean square errors of the two stocks. From the table, we can see that using mixture-of-expert is only slightly better than using Gaussian mixture. It seems that it is worthless to apply the mixture-of-expert model. However, next section will show the real power of using mixture-of-expert.

5. EXPERIMENTS: SECOND STEP PREDICTION

We use the method introduced in section 2 to conduct the second step prediction for the two stocks. We use 1500 random samples (i.e. $N = 1500$) for the Monte Carlo approximation. Figures 7 to 12 show the results of the second step prediction for different approaches.

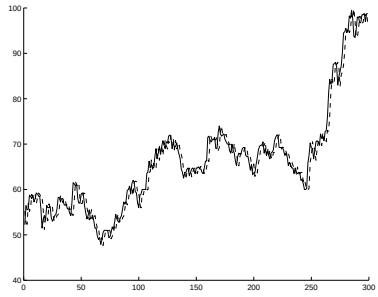

Figure 7: Second step prediction of CK prices with conventional ARMA-GARCH.

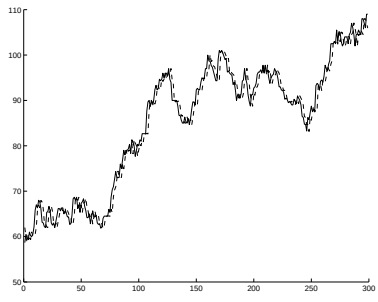

Figure 8: Second step prediction of HSBC prices with conventional ARMA-GARCH.

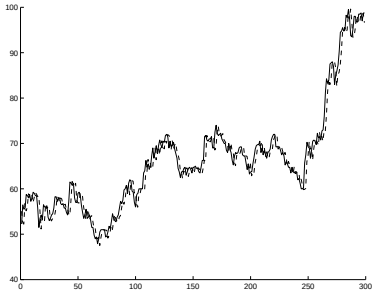

Figure 9: Second step prediction of CK prices with Gaussian mixture ARMA-GARCH.

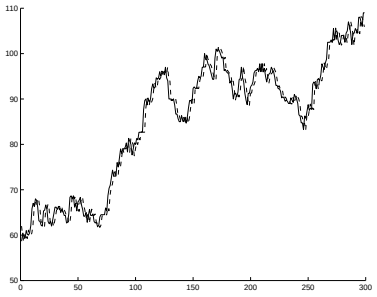

Figure 10: Second step prediction of HSBC prices with Gaussian mixture ARMA-GARCH.

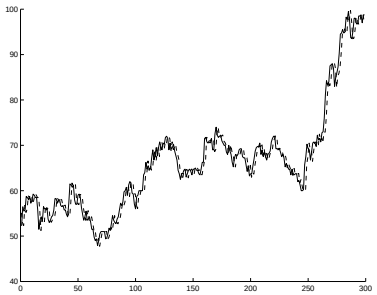

Figure 11: Second step prediction of CK prices with mixture-of-expert ARMA-GARCH.

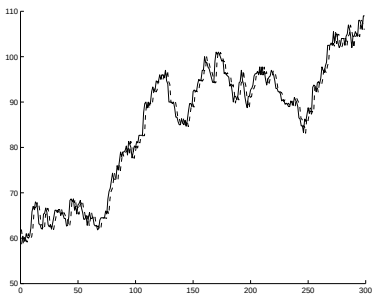

Figure 12: Second step prediction of HSBC prices with mixture-of-expert ARMA-GARCH.

Table 2 shows the mean square errors for second step prediction. As can be seen from the table, mixture-of-expert can yield a much better second step prediction than Gaussian mixture, which is due to the addition of the gating network in the model.

6. CONCLUSION

In this paper, we derive a GEM algorithm for the finite mixture-of-expert ARMA-GARCH model. Its relative empirical performance in stock price prediction against the conventional ARMA-GARCH and Gaussian mixture ARMA-GARCH model is investigated. Results reveal that both mixture models outperform the conventional ARMA-GARCH model. First step prediction using Gaussian mixture is as good as using mixture-of-expert. However, the mixture-of-expert do much better second step prediction than Gaussian mixture. This results show that the forecasting power of mixture-of-expert extends beyond a single step prediction. This is due to the additional gating network exists in mixture-of-expert, which can decouple the weight in one expert from other experts [4]. We conclude that whenever we do first step prediction, Gaussian mixture should be used since it required much less training time and resources. For second step prediction, mixture-of-expert is preferred.

7. REFERENCES

[1] H. Y. Kwok, C. M. Chen, and L. Xu, "Comparison between mixture of arma and mixture of ar model with application to time series forecasting," in *Proceedings of Fifth International Conference on Neural Information Processing*, 1998, pp. 1049–1052.

[2] W. C. Wong, F. Yip, and L. Xu, "Financial prediction by finite mixture garch model," in *Proceedings of Fifth International Conference on Neural Information Processing*, 1998, pp. 1351–1354.

[3] Chiu K. C. Tang, H. and Lei. Xu, "Finite mixture of arma-garch model for stock price prediction," in *The Third International Workshop on Computational Intelligence in Economics and Finance*, 2003.

[4] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Computation*, vol. 3, pp. 79–87, 1991.

[5] M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the em algorithm," *Neural Computation*, vol. 6, pp. 181–214, 1994.

[6] M. I. Jordan and L. Xu, "Convergence resutls for the em approach to mixtures of experts architectures," *Neural Networks*, vol. 8, no. 9, pp. 1409–1431, 1995.

[7] L. Xu and Hinton G. E. Jordan, M. I., "An alternative model for mixtures of experts," in *Advances in Neural Information Processing Systems*, 1995, vol. 7, pp. 633–640.

[8] A. P. Depmster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of The Royal Statistical Society Series B – Statistical Methodology*, vol. 39, no. 1, pp. 1–38, 1977.

[9] W. H. Greene, *Econometric Analysis*, Prentice Hall, New Jersey, fourth edition, 2000.