A STOCHASTIC CONTEXT-FREE GRAMMAR MODEL FOR TIME SERIES ANALYSIS

W. Wang†, V. Portnoy‡, and I. Pollak†,

† Purdue University
School of ECE
West Lafayette, IN 47907

‡ Jefferies & Company
Quantitative Strategies
New York, NY 10022

ABSTRACT

We propose a stochastic context-free grammar model whose structure can alternatively be viewed as a graphical model, and use it to model time series. We use the inside-outside algorithm to estimate the model parameters. We assume that the time series is a finite-order Markov process generated by our model, and develop an algorithm to forecast the conditional variance of the process. We use this algorithm to forecast the volatility of the S&P 500 index, achieving results that outperform both standard and more recent approaches.

***Index Terms*—** stochastic context-free grammar, volatility forecasting, GARCH, graphical model.

1. INTRODUCTION

Volatility is typically defined as the standard deviation or variance of the return series of a financial instrument. Volatility forecasting is an important task in the analysis of financial markets and has a variety of applications such as the assessment of investment risk and option pricing. Volatility estimates are also used as economic indicators. Overviews of the current and past literature on volatility forecasting as well as examples of its applications can be found in [1, 12].

GARCH model [6, 3] and its variations have been widely used for volatility modeling. However, these models tend to have long memory and sometimes fail to accurately track abrupt changes in market conditions. Hidden Markov models (HMMs) which have been widely used in signal processing for a long time—especially in speech analysis and recognition [13]—have been recently suggested to improve modeling of market changes [5, 14]. Specifically, several HMM-based approaches have been proposed to model the fact that typically high-volatility periods are shorter than low-volatility periods. In [8], a regime-switching volatility model is proposed where different regimes correspond to different volatility levels. Other models have been proposed where the volatility process is assumed to be generated by a hidden finite-state Markov process [5, 14]. In addition, the framework of [14] models the leverage effect, i.e., the empirical observation that negative returns typically lead to higher volatility than positive returns. Variants of GARCH that model the leverage effect are proposed in [7] and [16].

Motivated by the recent literature on HMM-based approaches for volatility forecasting, we propose to use a more general class of models called *stochastic context-free grammars* (SCFGs) [11]. SCFGs have recently been used with success in natural language processing [11], RNA modeling [4], and image processing [15]. In Sections 2–3, we present our results which use a simple SCFG time series model for volatility forecasting. Specifically, we model the return series using a hidden Markov random field whose parameters can be estimated using the inside-outside algorithm [10] which is a generalization of the standard forward-backward algorithm used for estimating the parameters of an HMM. In Section 2 we introduce our SCFG model, show how to estimate its parameters, and propose a forecasting method based on this models. In Section 3, we illustrate our algorithm using the S&P 500 stock index data.

This work was supported in part by the NSF Grant IIS-0329156.

2. STOCHASTIC CONTEXT-FREE GRAMMARS FOR TIME SERIES ANALYSIS

We propose a Markov random field model for volatility prediction. The model can be viewed either as a *graphical model* [9] or, alternatively, as a *stochastic context-free grammar* (SCFG) [11]. We describe one-step-ahead prediction for this model and a parameter estimation procedure using the EM algorithm.

2.1. Model and Volatility Forecasting

We construct a probabilistic model which specifies a probability distribution $f_{\mathbf{y}}$ for a T-dimensional random vector $\mathbf{y} = (y_1, \ldots, y_T)$ where T is a fixed integer. This joint distribution induces a conditional distribution $f_{y_T|\mathbf{y}_{1:T-1}}$ of the random variable y_T given the random vector $\mathbf{y}_{1:T-1} = (y_1, \ldots, y_{T-1})$. We then model the return process r_t for $t = 1, 2, \ldots$ as a Markov process of order $T-1$, with the following joint density of returns up to time t:

$$f_{\mathbf{r}_{1:t}}(\mathbf{R}_{1:t}) \stackrel{\triangle}{=} f_{\mathbf{y}}(\mathbf{R}_{1:T}) \prod_{\tau=T+1}^{t} f_{y_T|\mathbf{y}_{1:T-1}}(R_\tau|\mathbf{R}_{\tau-T+1:\tau-1}), \tag{1}$$

where $\mathbf{R}_{t_1:t_2} = (R_{t_1}, \ldots, R_{t_2})$ denotes the vector of observations of the returns $(r_{t_1}, \ldots, r_{t_2})$ from time $t = t_1$ until time $t = t_2$.

Using this model, a one step ahead volatility forecast at time $t-1$ is computed as the conditional variance of r_t given the observations $\mathbf{r}_{1:t-1} = \mathbf{R}_{1:t-1}$. Since the process is Markov of order $T-1$, this quantity is equal to the conditional variance of r_t given the observations $\mathbf{r}_{t-T+1:t-1} = \mathbf{R}_{t-T+1:t-1}$:

$$\begin{aligned} \mu_t &\stackrel{\triangle}{=} E[r_t|\mathbf{r}_{t-T+1:t-1} = \mathbf{R}_{t-T+1:t-1}] \\ \sigma_t^2 &\stackrel{\triangle}{=} \text{var}[r_t|\mathbf{r}_{t-T+1:t-1} = \mathbf{R}_{t-T+1:t-1}] \end{aligned} \tag{2}$$

We now describe our construction of the distribution $f_{\mathbf{y}}$. Our model is a Markov random field whose graph structure is illustrated in Fig. 1. We select integers $t_0, t_1, \ldots, t_N$ such that $t_0 = 0 < t_1 < \ldots < t_N = T$, and partition the vector $\mathbf{y}$ into N vectors $\mathbf{y}_1, \ldots, \mathbf{y}_N$ such that $\mathbf{y}_n = (y_{t_{n-1}+1}, y_{t_{n-1}+2}, \ldots, y_{t_n})$ for $n = 1, \ldots, N$. For example, in Fig. 1, $N = 5$, and $\mathbf{y}_1 = (y_1, \ldots, y_8)$, $\mathbf{y}_2 = (y_9, \ldots, y_{12})$, $\mathbf{y}_3 = (y_{13}, y_{14})$, $\mathbf{y}_4 = (y_{15})$, and $\mathbf{y}_5 = (y_{16})$.

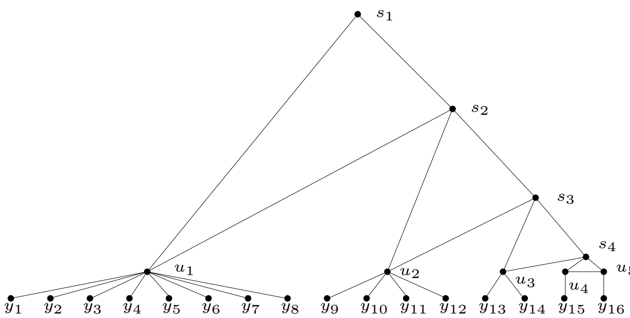

Fig. 1. An example of our model with domain size $L = 16$ and $N = 5$ terminal states.

A hidden state u_n is assigned to each vector $\mathbf{y}_n$. The state u_n is a discrete random variable taking values in the set $\{1, \ldots, M\}$. Given $u_n = i$, the random variables comprising $\mathbf{y}_n$ are conditionally independent and identically distributed according to the following Gaussian density with mean μ and variance σ_i^2: $f_{y_t|u_n}(x|i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x-\mu)^2}{2\sigma_i^2}}$, for $t_{n-1} + 1 \le t \le t_n$.

Note that the conditional mean μ is the same for all observations whereas the variance σ_i depends on the hidden state i. The random variables u_n are called *terminal states*. An additional layer of hidden discrete state variables $s_1, \ldots, s_{N-1}$, called *nonterminal states*, serves to model statistical dependencies among the terminal states and the observations. The nonterminal states take values in the set $\{1, \ldots, M\}$. The nonterminal state s_1 is called the *root state* and has probability p^i to assume the value i, $p^i = \text{Prob}(s_1 = i)$ for $i = 1, \ldots, M$. Given a nonterminal state $s_n = k$ with $1 \le n < N - 2$, the conditional joint probability of the events $\{s_{n+1} = i\}$ and $\{u_n = j\}$ is denoted by p_n^{ijk}, $p_n^{ijk} = \text{Prob}(s_{n+1} = i, u_n = j | s_n = k)$ for $i, j, k = 1, \ldots, M$ and $n = 1, \ldots, N - 2$. Given the last nonterminal state $s_{N-1} = k$, the conditional joint probability of the events $\{u_{N-1} = j\}$ and $\{u_N = i\}$ is denoted by p_{N-1}^{ijk}, $p_{N-1}^{ijk} = \text{Prob}(u_{N-1} = j, u_N = i | s_{N-1} = k)$ for $i, j, k = 1, \ldots, M$. We then define the joint probability distribution of the nonterminal states, terminal states, and the vector $\mathbf{y}$:

$$f_{\mathbf{s},\mathbf{u},\mathbf{y}}(\mathbf{S}, \mathbf{U}, \mathbf{Y}) = p^{S_1} p_{N-1}^{U_N, U_{N-1}, S_{N-1}} \prod_{n=1}^{N-2} p_n^{S_{n+1}, U_n, S_n} \times \prod_{n=1}^{N} \prod_{t=t_{n-1}+1}^{t_n} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(Y_t-\mu)^2}{2\sigma_i^2}}.$$

The probability density of $\mathbf{y}$ is then calculated by marginalizing over all the hidden variables:

$$f_{\mathbf{y}}(\mathbf{Y}) = \sum_{\mathbf{S},\mathbf{U}} f_{\mathbf{s},\mathbf{u},\mathbf{y}}(\mathbf{S}, \mathbf{U}, \mathbf{Y}).$$

We let $\mathcal{G}$ be the set of all parameters. This set consists of the root state probabilities p^i, the transition probabilities p_n^{ijk}, and the mean μ and variances σ_i^2 of the conditional probability distributions of the observations $\mathbf{y}$. In the next subsection we describe our algorithm for estimating the parameters from training data.

In the Appendix, we give a formula for evaluating the conditional probability $\text{Prob}(u_N = i | \mathbf{y}_{1:T-1} = \mathbf{Y}_{1:T-1})$. Using this and Eq. (2), we obtain the following volatility forecast equation:

$$\begin{aligned} &\text{var}[r_t | \mathbf{r}_{t-T+1:t-1} = \mathbf{R}_{t-T+1:t-1}] \\ &= \sum_{i=1}^{M} \text{Prob}(u_N = i | \mathbf{r}_{t-T+1:t-1} = \mathbf{R}_{t-T+1:t-1}) \sigma_i^2. \quad (3) \end{aligned}$$

In order to model the leverage effect, we construct two different parameter sets, $\mathcal{G}^+$ and $\mathcal{G}^-$. We train the model $\mathcal{G}^+$ using only training data sequences in which $Y_{T-1} > 0$. We train the model $\mathcal{G}^-$ using only training data sequences in which $Y_{T-1} \le 0$. The estimate in Eq. (3) then uses model parameters $\mathcal{G}^+$ if $Y_{T-1} > 0$ and $\mathcal{G}^-$ if $Y_{T-1} \le 0$. We call this modified model *thresholded stochastic context-free grammar* (T-SCFG).

2.2. Parameter Estimation

Given training data $R_1, \ldots, R_t$, the maximum likelihood parameter estimation strategy would aim to maximize the log-likelihood function $L(\mathcal{G}) \triangleq \log f_{\mathbf{r}_{1:t}}(\mathbf{R}_{1:t}; \mathcal{G})$ which is the logarithm of the probability distribution defined in Eq. (1), viewed as a function of the model parameters $\mathcal{G}$. In this paper, we estimate the parameters by solving a simpler maximization problem. Specifically, we maximize the following:

$$\tilde{L}(\mathcal{G}) \triangleq \sum_{\tau=T}^{t} \log f_{\mathbf{y}}(\mathbf{R}_{\tau-T+1:\tau}; \mathcal{G}). \quad (4)$$

Note that this would be the correct log-likelihood function if the sequences $\mathbf{R}_{1:T}, \mathbf{R}_{2:T+1}, \ldots, \mathbf{R}_{t-T+1:t}$ were independent samples from the density $f_{\mathbf{y}}$. To simplify notation, we define $D = t - T + 1$ and let $\mathbf{Y}^{(d)} = \mathbf{R}_{d:d+T-1}$, for $d = 1, \ldots, D$. Eq. (4) can then be rewritten as follows:

$$\tilde{L}(\mathcal{G}) = \sum_{d=1}^{D} \log f_{\mathbf{y}}(\mathbf{Y}^{(d)}; \mathcal{G}).$$

We use the EM algorithm [2] to perform the maximization. It can be shown that this maximization problem is a special case of the log-likelihood maximization problems for standard stochastic context-free grammars [11] and spatial random trees [15]. Exact algorithms for the EM update for these two problems, called the inside-outside algorithm [10] and the center-surround algorithm [15] respectively, can therefore be adapted to yield an exact EM update for our problem. The formulas for the EM updates are described in the Appendix. The derivation of the algorithm can be found in [15].

3. EXAMPLES

We now apply our model to volatility prediction based on daily returns of S&P 500 composite stock index. Observations of S&P 500 composite index, p_t, are taken from April 3, 1995 to December 31, 1999, resulting in a total of 1200 log-returns r_t, defined as follows: $R_t = \log(p_t/p_{t-1})$. The first 800 observations are used as training data for parameter estimation, while the remaining 400 data points are used as test data, as shown in Fig. 2. We take a constant volatility model of sample variance in the training set as a benchmark. We compare the forecasting performance among four models Gaussian GARCH(1,1), GJR(1,1), SCFG and T-SCFG. When we evaluate the performance, squared log-returns are used as a proxy for volatility.

For the SCFG model, we specify initial parameter estimates $\mathcal{G}$ as follows: similar to [14] we only allow transitions between adjacent

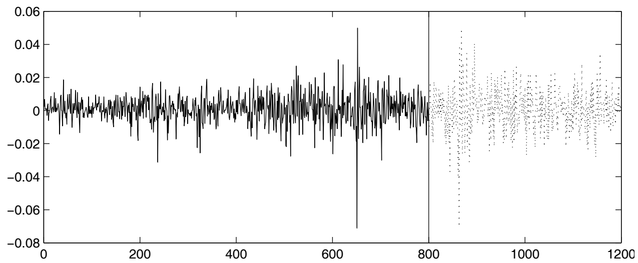

Fig. 2. Log-returns of S&P 500 composite index, ranging from Apri 3, 1995 to December 31, 1999. The vertical line separates the training set and the test set.

	MSE$\times 10^{-7}$	% improvement
Sample Variance	1.340	0
GARCH(1,1)	1.198	10.6
SCFG	1.204	10.1
GJR(1,1)	1.160	13.4
T-SCFG	1.141	14.9

Table 1. Mean-square errors for various algorithms, as defined in Eq. (5), and percentage improvement over the sample variance.

states, i.e. $p_n^{ijk} = 0$ for $|i-k| + |j-k| > 1$. Values of $\sigma_i, i = 1, 2, \ldots, M = 8$ are initialized so that $\sigma_1 < \sigma_2 < \ldots < \sigma_8$. Initial root probabilities are $p^i = 1/8$ for all states i. We do not estimate μ and set $\mu = 0$. We train our SCFG model to find a local maximum of the log-likelihood by using the inside-outside algorithm. We moreover build our T-SCFG model in a similar strategy. After the $\mathcal{G}^+$ and $\mathcal{G}^-$ are initialized, the models $\mathcal{G}^+$ and $\mathcal{G}^-$ are then trained only using the training sequences $\mathbf{Y}^{(d)} = (R_d, \ldots, R_{d+T-1})$ for which R_{d+T-2} is, respectively, positive and nonpositive. The volatility at time t is then estimated using $\mathcal{G}^+$ if $R_{t-1} > 0$ and using $\mathcal{G}^-$ if $R_{t-1} \leq 0$.

We compare the forecasting performance of different algorithms in Table 1 by calculating the mean-square error (MSE) loss function for each algorithm, as follows:

$$\text{MSE} = \frac{1}{400} \sum_{t=801}^{1200} (R_t^2 - \sigma_t^2)^2, \qquad (5)$$

where σ_t^2 is the volatility estimate. Note that the performance of SCFG model is close to GARCH(1,1), and is significantly better than sample variance estimates; Also note that the T-SCFG model has the lowest MSE. In addition, the HMM-based method of [14] results in a larger MSE than GARCH on a similar data set, as reported in [14]. Therefore this indirect comparison suggests that our method outperforms the method of [14]. The data used in [14] is the S&P 500 daily log-return series for January 3, 1995 through December 31, 1999; however, a direct comparison to [14] is difficult since the data in [14] appears to be scaled and interpolated, and the specifics of these modifications are not reported. In Fig. 3, we show forecasting results for the GARCH and T-SCFG models: the solid line is the square root of the volatility forecast. The GARCH model is more "persistent", resulting in smoother volatility forecasts. On the other hand, high persistence of GARCH model causes overestimates of the volatility during highly volatile periods. The structure of our T-SCFG model with discrete hidden states allows for the parameters to adapt to different data patterns, as long as these patterns are well represented in the training data. It produces more noisy estimates which adapt to rapid changes more quickly than the GARCH estimates.

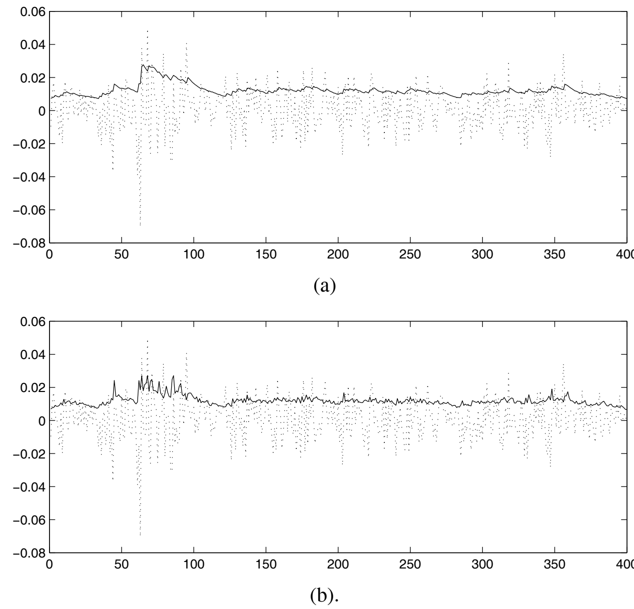

Fig. 3. S&P 500 index log-returns (dotted lines) and square root of the volatility forecasts (solid lines) using (a) GARCH model and (b) T-SCFG model.

	$\hat{b}_0 \times 10^{-4}$	$\hat{b}_1$	ρ
GARCH(1,1)	0.53 ($\pm$0.58)	0.71 ($\pm$0.29)	0.232
SCFG	0.35 ($\pm$0.70)	0.90 ($\pm$0.41)	0.207
GJR(1,1)	0.58 ($\pm$0.47)	0.71 ($\pm$0.22)	0.302
T-SCFG	0.05 ($\pm$0.61)	1.10 ($\pm$0.34)	0.303

Table 2. Regression parameters $\hat{b}_0$ and $\hat{b}_1$ with 95% confidence intervals, and the correlation coefficient between each set of estimates and the data.

To further evaluate the performance, we calculate the best linear fit of the estimates to the test data. In other words, we calculate constants b_0 and b_1 that minimize the following quantity:

$$\sum_{t=801}^{1200} [R_t^2 - (b_0 + b_1 \hat{\sigma}_t^2)]^2$$

If σ_t^2 is an unbiased estimate of r_t^2 then we would have $b_0 = 0$ and $b_1 = 1$. Table 2 lists the coefficients b_0 and b_1 for the four estimators, together with their 95% confidence intervals. As evident from the table, the T-SCFG model produces estimates that are the least biased among the four models—i.e., the values of b_0 and b_1 for T-SCFG are the closest to 0 and 1, respectively. In the same table, we also provide the correlation coefficient ρ between the data and its estimates. The correlation coefficient of 1 would mean perfect correlation; the correlation coefficient of zero means uncorrelatedness. Note that the T-SCFG model achieves the highest correlation coefficient.

4. CONCLUSIONS

We have introduced an SCFG model for time series, and developed a volatility forecasting algorithm based on this model. Our experiments indicate that our algorithm outperforms both the standard GARCH-based forecasting and more recent methods.

5. ACKNOWLEDGMENT

We thank Branko Stojkovic for helpful discussions.

6. APPENDIX: PARAMETER ESTIMATION ALGORITHM

We define the *inside variables* $p_{in}^{(d)}(n,i)$ and $q_{in}^{(d)}(n,i)$ and the *outside variables* $p_{out}^{(d)}(n,i)$ and $p_{in}^{(d)}(n,i)$. Even though the inside and outside variables depend on the model parameters, we suppress this dependence in the remainder of this section for notational convenience. The inside variable $q_{in}^{(d)}(n,i)$ is the conditional probability density of the observed realization $\mathbf{Y}_n$ of the random vector $\mathbf{y}_n$, conditioned on the event $\{u_n = i\}$. The outside variable $q_{out}^{(d)}(n,i)$ is the joint probability distribution of the remaining data and the event $\{u_n = i\}$. The inside variable $p_{in}^{(d)}(n,i)$ is the conditional probability density of all the observed values that are children of $u_n, \ldots, u_N$, given the event $s_n = i$. We also define $p_{in}(N,i) \triangleq q_{in}(N,i)$ for notational convenience. The outside variable $p_{out}^{(d)}(n,i)$ is the joint probability distribution of all the observed values that are children of $u_1, \ldots, u_{n-1}$ and the event $s_n = i$. These variables can be calculated recursively from the observed data, using the following formulas.

$$q_{in}^{(d)}(n,i) = \prod_{\tau=t_{n-1}+1}^{t_n} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(Y_\tau^{(d)}-\mu)^2}{2\sigma_i^2}}, \ \ 1 \le n \le N$$

$$p_{in}^{(d)}(n,k) = \begin{cases} q_{in}^{(d)}(N,k) & n = N, \\ \sum_{i=1}^{M}\sum_{j=1}^{M} p_n^{ijk} p_{in}^{(d)}(n+1,i) q_{in}^{(d)}(n,j) & \text{otherwise,} \end{cases}$$

$$p_{out}^{(d)}(n,i) = \begin{cases} p^i & n = 1, \\ \sum_{j=1}^{M}\sum_{k=1}^{M} p_n^{ijk} p_{out}^{(d)}(n-1,k) q_{in}^{(d)}(n-1,j) & \text{otherwise,} \end{cases}$$

$$q_{out}^{(d)}(n,j) = \begin{cases} \sum_{i=1}^{M}\sum_{k=1}^{M} p_{N-1}^{ijk} p_{out}^{(d)}(N-1,k) q_{in}^{(d)}(N-1,i) & n = N, \\ \sum_{i=1}^{M}\sum_{k=1}^{M} p_n^{ijk} p_{out}^{(d)}(n,k) p_{in}^{(d)}(n+1,i) & \text{otherwise.} \end{cases}$$

Following [15], EM updates can be written in terms of the inside and outside variables. The resulting formulas are as follows:

$$\hat{p}^i = \frac{\sum_{d=1}^{D} \frac{p^i p_{in}^{(d)}(1,i)}{f_{\mathbf{y}}(\mathbf{Y}^{(d)})}}{D - T + 1},$$

$$\hat{p}_n^{ijk} = \frac{\sum_{d=1}^{D} \frac{p_n^{ijk} p_{out}^{(d)}(n,k) p_{in}^{(d)}(n+1,i) q_{in}^{(d)}(n,j)}{f_{\mathbf{y}}(\mathbf{Y}^{(d)})}}{\sum_{d=1}^{D} \frac{p_{out}^{(d)}(n,k) p_{in}^{(d)}(n,k)}{f_{\mathbf{y}}(\mathbf{Y}^{(d)})}},$$

$$\hat{\sigma}_i = \frac{\sum_{d=1}^{D}\sum_{n=1}^{N}\sum_{\tau=t_{n-1}+1}^{t_n} \frac{q_{out}^{(d)}(i) q_{in}^{(d)}(i) (Y_\tau^{(d)} - \mu)^2}{f_{\mathbf{y}}(\mathbf{Y}^{(d)})}}{\sum_{d=1}^{D}\sum_{n=1}^{N} \frac{q_{out}^{(d)}(i) q_{in}^{(d)}(i)(t_n - t_{n-1})}{f_{\mathbf{y}}(\mathbf{Y}^{(d)})}}.$$

where $I_A(\cdot)$ is the indicator function of the set A, and $f_{\mathbf{y}}(\mathbf{Y}^{(d)}) = \sum_{i=1}^{M} p^i p_{in}^{(d)}(1,i)$.

Using Bayes rule, we can evaluate the conditional probability $\text{Prob}(u_N = i | \mathbf{r}_{t-T+1:t-1} = \mathbf{R}_{t-T+1:t-1})$ used in the conditional variance formula of Eq. (3) as follows:

$$\text{Prob}(u_N = i | \mathbf{r}_{t-T+1:t-1} = \mathbf{R}_{t-T+1:t-1}) = \frac{f_{\mathbf{y}_{1:T-1}, u_N}(\mathbf{Y}_{1:T-1}^{(d)}, i)}{f_{\mathbf{y}_{1:T-1}}(\mathbf{Y}_{1:T-1}^{(d)})} = \frac{q_{out}^{(d)}(N,i)}{\sum_{j=1}^{M} q_{out}^{(d)}(N,j)}.$$

7. REFERENCES

[1] T. G. Andersen, T. Bollerslev, P. F. Christoffersen, and F. X. Diebold, "Volatility forecasting," National Bureau of Economic Research, Inc, NBER Working Papers 11188, Mar. 2005.

[2] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *The Annals of Mathematical Statistics*, vol. 41, no. 1, pp. 164–171, 1970.

[3] T. Bollerslev, "Generalized autoregressive conditional heteroskedasticity," *J. Econometrics*, vol. 31, no. 3, pp. 307–327, April 1986.

[4] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis*. Cambridge University Press, 1998.

[5] R. J. Elliott, W. C. Hunter, and B. M. Jamieson, "Drift and volatility estimation in discrete time," *J. Economic Dynamics and Control*, vol. 22, no. 2, pp. 209–218, February 1998.

[6] R. F. Engle, "Autoregressive conditional heteroscedasticity with estimates of variance of United Kingdom inflation," *Econometrica*, vol. 50, no. 4, pp. 987–1008, July 1982.

[7] L. R. Glosten, R. Jagannathan, and D. E. Runkle, "On the relation between the expected value and the volatility of the nominal excess return on stocks," *J. Finance*, vol. 48, no. 5, pp. 1779–1801, 1993.

[8] J. D. Hamilton and R. Susmel, "Autoregressive conditional heteroskedasticity and changes in regime," *J. Econometrics*, vol. 64, no. 1-2, pp. 307–333, 1994.

[9] M. I. Jordan, *Learning in Graphical Models*. MIT Press, Nov. 1998.

[10] K. Lari and S. J. Young, "The estimation of stochastic context-free grammars using the inside-outside algorithm," *Computer Speech and Language*, vol. 4, no. 1, pp. 35–56, 1990.

[11] C. Manning and H.Schütze, *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[12] S.-H. Poon and C. W. J. Granger, "Forecasting volatility in financial markets: A review," *J. Economic Literature*, vol. 41, no. 2, pp. 478–539, 2003.

[13] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in *Proc. of the IEEE*, vol. 77, no. 2, pp. 257–286 1989.

[14] A. Rossia and G. M. Gallo, "Volatility estimation via hidden Markov models," *J. Empirical Finance*, vol. 13, no. 2, pp. 203–230, 2006.

[15] W. Wang, I. Pollak, T.-S. Wong, C. A. Bouman, M. P. Harper, and J. M. Siskind, "Hierarchical stochastic image grammars for classification and segmentation," *IEEE Trans. Im. Proc.*, vol. 15, no. 10, pp. 3033–3052, Oct. 2006.

[16] J.-M. Zakoian, "Threshold heteroskedastic models," *J. Economic Dynamics and Control*, vol. 18, no. 5, pp. 931–955, 1994.