Alpha Stable Distribution Based FARIMA Modeling and Forecasting for Network Traffic Data

Hu Sheng1 *, Qiuting Yan1, and Kaixin Li1
1School of Electrical and Information Engineering, Dalian Jiaotong University, Dalian, China

*Corresponding author e-mail: hu.sheng@djtu.edu.cn

Abstract. Through research on network traffic in recent years, it is found that actual network traffic not only has long correlation but also has heavy-tailed characteristics. The Alpha stable distribution provides a very useful theoretical tool for such processes. According to the nature of the Alpha stable distribution, we need to improve the relevant parameter value method of the Fractional Autoregressive Integrated Moving Average model and improve the prediction accuracy of the model. In this paper, the Alpha stable distribution based Fractional Autoregressive Integrated Moving Average modelling method is introduced, and the validation of the method is analyzed using Fractional Gaussian Noise time series and impulse noise corrupted Fractional Gaussian Noise time series. Finally, the Alpha stable distribution based Fractional Autoregressive Integrated Moving Average model is applied to BC-p Aug89.TL data provided by Bell Labs. The analysis shows that the improved Fractional Autoregressive Integrated Moving Average model has a better prediction effect on actual network traffic and is a more representative research method.

Keywords: Alpha stable distribution, Fractional Autoregressive Integrated Moving Average model, Network traffic modeling and prediction

1. Introduction
In the field of network analysis and biomedical area, most random signals or noises have non-stationary heavy-tail characteristics [1], with local mutations. Then when the Fractional Autoregressive Integrated Moving Average (FARIMA) model is established, the results of the original FARIMA model's parameter estimation method cannot describe the local mutation changes of these signals, and these mutations may also affect the accuracy of the prediction results [2]. Therefore, it is necessary to improve the parameter estimation method of FARIMA model. Therefore, based on the introduction of a heavy-tailed distribution-Alpha stable distribution[3], this paper fits the distribution of actual network traffic, uses the relevant properties of the Alpha stable distribution to establish the values of the relevant parameters of the FARIMA model, and applies Into real network traffic. The analysis shows that the improved FARIMA model has better prediction effect on network traffic and is more conducive to the development of the network.
2. Parameter Estimation of FARIMA Model Based on Alpha Stable Distribution

2.1. Fractional difference operator
Implementing the fractal difference operator \( d \) is a key to the FARIMA model for modeling network traffic. At this time, in order to improve the prediction accuracy of the FARIMA model, the value of the fractional difference parameter \( d \) in the FARIMA model is changed as shown in equation (1) under the known factors of heavy tails in real network traffic \(^4\).

\[
d = H - 1/\alpha
\]

In the formula, \( H \) represents the Hurst parameter, and \( \alpha \) represents the characteristic index of Alpha stable distribution.

2.2. Autoregressive term order
The generalized Yule-Walker equation is defined as follows: A zero-mean \( ARS\alpha S \) process \( X(n) \) can be expressed as:

\[
X(n) = a_1X(n-1) + a_2X(n-2) + \ldots + a_pX(p-1) + U(n)
\]

Among them, \( P \) is the model order of \( AR \) and \( U(n) \) is the independent and identical distribution process following the distribution \( S\alpha S \). Its characteristic index is \( \alpha \) and its dispersion coefficient is \( \gamma \).

Taking the conditional expectation for \( X(m) \) on both sides of the above formula, and \( n - p \leq m \leq n - 1 \), we can get:

\[
E[X(n)|X(m)] = a_1E[X(n-1)|X(m)] + \ldots + a_pE[X(p-1)|X(m)] + E[U(n)|X(m)]
\]

Because \( U(n) \) and \( X(n) \) are independent of each other, \( \{X(n)\} \) is a stable and symmetric Alpha stable distribution, so \( E[U(n)X(m)] = 0 \), according to the properties expected by the condition, there are:

\[
E[X(n)|X(m)] = \lambda(n-m)X(m)
\]

Where \( \lambda(n-m) \) represents the covariation coefficient of \( X(n) \) and \( X(m) \), and \( \lambda(0) = 1 \). Substitute the above formula (4) into the formula (3) for simplification, when \( n - p \leq m \leq n - 1 \):

\[
\lambda(n-m) = a_1\lambda(n-m-1) + \ldots + a_p\lambda(n-m-p)
\]

Definition: \( P = [\lambda(1)\lambda(2)\ldots\lambda(p)] \), \( a = [a_0,a_1,a_p] \)

\[
C = \begin{bmatrix}
\lambda(0) & \lambda(1) & \lambda(2) \\
\lambda(1) & \lambda(0) & \lambda(2) \\
\lambda(2) & \lambda(0) & \lambda(2)
\end{bmatrix}
\]

In this way, the coefficient \( p \) of the AR model can be obtained by solving the following linear equation:

\[
Ca = P
\]

Equation (7) is the expression of the generalized Yule-Walker equation, in which the covariance matrix \( C \) uses the Toeplitz matrix.\(^5\) At this time, if the covariance matrix \( C \) is full rank, the above
equation can be solved to obtain the estimated value of the parameter $a$. In practical applications, we need to estimate the covariance matrix $C$ and vector $P$ according to the output observations, that is, we need to estimate and calculate the covariance coefficient $\hat{\lambda}(j)$, here we will use the least squares estimation (LS) to estimate the covariance coefficient $\hat{\lambda}(j)$.

3. Alpha stable distribution based FARIMA modeling and forecasting
(1) First, zero-average the data, that is, preprocessing.

(2) Estimate the Hurst parameter by the residual variance method. The sample feature function method is used to estimate the Alpha stable distribution feature index $\alpha$, and then calculate the estimated value of $d$ according to the relationship $d = H - 1/\alpha$, use the estimated value of $\hat{d}$ Perform fractional differential filtering;

(3) According to the method of ARMA model structure identification, the autocorrelation coefficient and the partial autocorrelation coefficient based on the generalized Yule-walker equation are respectively obtained by calculating the data after the fractional difference. Analyze, preliminarily determine the order of $p,q$, and finally check by Akaike information criterion (AIC) to determine the final order of $p,q$;

(4) Use the established ARMA model for data prediction ARMA;

(5) Integrate the predicted results of the data into $\hat{d}$ parameter score integration.

4. Forecast of FARIMA model to simulate network traffic
In order to verify the effectiveness of the above method for improving the $d$ and $p$ parameters of the FARIMA model for Alpha stable distribution, this study selected two time series with a data length of 10000: randomly generated scores with known Hurst parameters of 0.8 and no noise pollution The order Gaussian noise (FGN) sequence is represented by data 1, and the randomly generated result is shown in Figure 2; the fractional order Gaussian noise (FGN) polluted by Alpha stable distribution noise ($\alpha = 1.5$) is represented by data 2, and the randomly generated result is shown in Figure 3 As shown, they are used to simulate real network traffic data, and model and analyze it[6].
Perform self-similar Hurst parameter estimation on the two time series. Here, the residual variance method is used to estimate the Hurst parameters separately. The feature index $\alpha$ has been given as $\alpha = 1.5$, and the value of the fractional difference parameter $d$ is calculated according to the estimation result. The results of data 2 are shown in Table 1.

According to the obtained $d$ value, the two time series are respectively subjected to fractional differential filtering, and the autocorrelation coefficients and partial autocorrelation coefficients based on the covariance-generalized yule-walker equation are obtained for the filtered time series, and the truncation of the correlation function graph and partial autocorrelation function graph initially determines the order of $p,q$. The autocorrelation function graph and partial correlation function graph of data 1 and data 2 are shown in Figure 4 and Figure 5 respectively.

Table 1. Test results of two time series by residual variance method.

|       | H estimated value | $\alpha$ estimated value | Formula     | $d$ value |
|-------|------------------|--------------------------|-------------|-----------|
| Data1 | 0.8108           |                          | $d=H-0.5$  | 0.3108    |
| Data2 | 0.8137           | 1.5                      | $d=H-1/\alpha$ | 0.1470   |

Figure 4. FGN time series self / partial correlation coefficient analysis chart.
According to the prediction process in Section 3, the network traffic forecast of the improved FARIMA model is performed on the data of Bell Labs. Select data samples with a length of 10000, and the simulation results are shown in Figure 8.
Self-similar Hurst parameter estimation and feature index $\alpha$ estimation are performed on the pre-processed data, and the results of the estimated values are shown in the following table 3:

### Table 3. Self-similar estimation results.

| Sample | $H$  | $\alpha$ | $d$   |
|--------|------|----------|-------|
| 10000  | 0.7951 | 1.9986  | 0.2947 |

Then perform fractional differential filtering on the real network traffic with $d = 0.2947$, and use the results of the fractional differential to stimulate the ARMA model, namely FARIMA (0, 0.2947, 0) to stimulate the ARMA (p, q) model. Obtain the autocorrelation coefficient (ACF) and partial autocorrelation coefficient (PACF) of the sample data at this time.

![Figure 9. Real network traffic 100 - step prediction chart](image)

**Figure 9.** Real network traffic 100 - step prediction chart

Real network is censored at $q = 2$, so the value of $q$ should be determined to be 2 at this time; from Figure 5.3 the partial autocorrelation coefficient of the sample data of the real network (PACF) there is obvious tailing at $p = 0$, and the value of $p$ is 0 at this time. Use the AIC guidelines for inspection. When the order is 2, the value of the AIC information criterion is the smallest, 14214.3115, that is, the final model parameter is ARMA (0, 2). By calculating the white noise sequence, the autocorrelation coefficient of Gaussian white noise $\{\varepsilon_t\}$ is -0.0649, which is close to zero, and the model is applicable. At this point, the establishment of the ARMA (0, 2) model has been completed. Next, the FARIMA (0, 0.2947, 2) model is used to predict real network traffic data with a length of 10,000. The prediction results are shown in Figure 11. The average absolute error of the prediction is calculated, and the result is 0.0444. It can be seen that the predicted result is very close to the real traffic data, and it is concluded that the improved FARIMA model based on the Alpha stable distribution has a better prediction effect on real network traffic.

### 6. Conclusion

Through the analysis of the FARIMA model, since the FARIMA model also takes into account the long-term and short-term correlation, the FARIMA model is selected for network traffic services. The simulation results show that the improved FARIMA model has smaller prediction errors and the best
prediction effect. Therefore, the improved FARIMA model is used to predict network services. Zero-mean pre-processing of the real Bell Labs data, modeling through the improved FARIMA model, and traffic prediction for the real network data, the prediction result error is closer to 0, indicating that the Alpha distribution based The FARIMA model has good prediction accuracy and can be used as an important research method in network traffic prediction.

Acknowledgments
The study was supported by “Natural science research project of Liaoning Provincial Department of Education, China (Grant No. JDL2019014)”

References
[1] M.Crovella, Bestavros A. Self-similarity in World Wide Web traffic: evidence and possible causes. IEEE/ACM Transaction on Networking. 1996 5(6) 835-846.
[2] M.E. Souse, A. Suarez, J.C. Lopez, Application of the whittle estimator to the modeling of traffic based on the M/G/∞ process, IEEE Communications Letters. 2007 11(10) 817-819.
[3] Y.M. Sun. A high-precision approach for effective fractal-based similarity search of stochastic non-stationary time series, Proceedings of the Seventh International Conference on Machine Learning and Cybernetics. 2008 136-141.
[4] J.P. Montillet, Modeling Geodetic Processes with Levy alpha-Stable Distribution and FARIMA, Mathematical geosciences. 2014 12(12).
[5] Fang daoyuan, wei mingjun, mathematical modeling: method guidance and case analysis, zhejiang university press, 2011.
[6] Zhang Qi, han baoming. Modeling and simulation of passenger interaction in transportation hubs [J]. Journal of system simulation, 2008, 20 (20) : 5442-5445.