Using Digital Filtration for Hurst Parameter Estimation

Ján PROCHÁSKA[1], *Radoslav VARGIC*[1]

[1] Dept. of Telecommunications, Slovak University of Technology, Ilkovičova 3, 812 19 Bratislava, Slovak Republic

prochaska@ktl.elf.stuba.sk, vargic@ktl.elf.stuba.sk

Abstract. *We present a new method to estimate the Hurst parameter. The method exploits the form of the autocorrelation function for second-order self-similar processes and is based on one-pass digital filtration. We compare the performance and properties of the new method with that of the most common methods.*

Keywords

Self-similarity, Hurst parameter, autocorrelation function, fractional Gaussian noise.

1. Introduction

Although known for decades, self-similarity of random processes (in the distributional sense) gained a great deal of attention in the early 90's, when Leland et al showed in [1] that self-similarity plays a significant role in Local Area Network traffic. Until then, it was thought that network traffic behaves like telephony traffic and follows the traditional Poisson models. Not only has it been proven that this assumption is wrong, other evidence for self-similar behavior in different environments has been gathered and presented [2], [8], [9], [10].

One of the most important things about self-similar behavior of network traffic and self-similar processes is the much higher utilization rate produced in contrary to the Poisson like models [4]. This has an obvious consequence for buffer design and has a big impact on the overall performance of a system. Another important point is whether an underlying process is self-similar or not and, if it is, then how much is it self-similar.

The article is divided as follows: Section 2 gives a short introduction into self-similarity for a discrete random process and discusses the known methods for determining the degree of self-similarity. Section 3 explains how the autocorrelation function (ACF) and digital filtering can be used to estimate the self-similarity parameter. Section 4 provides results and a summary.

2. Self-similarity, Theoretical Background

We call a second-order stationary random process $\mathbf{X}$ exactly second-order self-similar (for details please refer to [3], [4], [5], [6], [7]), when it satisfies

$$Var(\mathbf{X}^{(m)}) = m^{2H-2} Var(\mathbf{X}),\ m \geq 1 \tag{1}$$

$$r(k) = r^{(m)}(k) = \frac{\left[|k-1|^{2H} - 2k^{2H} + (k+1)^{2H}\right]}{2},\ k \geq 0, m \geq 1 \tag{2}$$

where $r^{(m)}(k)$ denotes the ACF of the aggregated process $\mathbf{X}^{(m)}$ defined as

$$\mathbf{X}^{(m)}(k) = \frac{1}{m} \sum_{i=km-(m-1)}^{km} \mathbf{X}(i),\ m \geq 1 \tag{3}$$

and the parameter H is called the Hurst parameter or the self-similarity parameter. A weaker form of self-similarity is known as the asymptotical self-similarity. This happens when (1) and (2) hold for $m \to \infty$.

Because of its practical importance, we stick close to the fractional Gaussian noise (fGn) (for details please refer to [3]) which satisfies both (1) and (2). It's clear that fGn remains the same (in the sense of its statistical properties), over all scales (aggregation levels) and so are the impacts of the self-similar behavior. The range $\frac{1}{2} < H < 1$ is especially of interest, as the process exhibits another interesting property – the long-range dependence. In this case, the correlations decay slowly to zero. The decay is hyperbolic rather than exponential and the variance of time averages tends more slowly to zero as one would expect. From (2) it's clear that the Hurst parameter "determines" the long term behavior of the process. The higher is the value of the Hurst parameter, the more is the process self-similar (for further important implications of self-similarity please refer to [4]).

Besides the autocorrelation structure described by (2), the self-similarity manifests itself in other forms, such as the variances of the aggregated process (1), or in the spectral domain via the power spectral density. These proper

ties form the basis for Hurst parameter estimation. The analysis is done either in the time domain (Variance-time plots, R/S plot) or in the spectral domain (Periodogram, Whittle's estimator, Abry-Veitch estimator using wavelets to analyze the process). Please refer to [4] for further details.

Although a number of methods working in the time domain exist, only few are known to directly analyze the ACF. The Correlogram [3] is useful for an initial heuristic analysis of the data as the decay of the correlations is hyperbolic, but it is not suitable for accurate estimation. In [11] a moment estimator has been introduced by Kettani working with lag 1 of the ACF. The concept has been revisited in [12], where an iterative approach has been used to analyze the ACF for one arbitrary chosen lag. In the next section we will show how the ACF can be analyzed using digital filtration.

3. Analysis of the Autocorrelation Function

A straightforward estimation of H involving the ACF would be to find the best-fit $r(k)$ (and thus H) for a sample ACF $\hat{r}(k)$. It may be obtained as

$$\hat{r}(k) = \frac{\sum_{i=1}^{N-|k|} \left(X(i) - \hat{\mu}\right)\left(X(i+|k|) - \hat{\mu}\right)}{N\hat{\sigma}^2} \tag{4}$$

where N is the sample length, $\hat{\mu}$ and $\hat{\sigma}^2$ are the sample mean and sample variance, respectively. Using the method of least squares (MLS), the equation to solve is

$$\frac{\partial}{\partial H} \sum_k \left(\hat{r}(k) - r(k)\right)^2 = 0 . \tag{5}$$

However an analytic solution of (5) seems not to exist and the corresponding H has to be found iteratively. To avoid this, a suitable transform domain can be used.

We propose a Z–transform approach. The idea behind is similar to the approach above, but the estimator is based on one-pass digital filtration of the ACF. Moreover, the result is expected to have a constant function form, so no complicated regression is needed. We see that the basis for the $r(k)$ is the term k^{2H} (from (2)). Therefore, we rewrite the ACF as

$$r(k) = \frac{1}{2}\delta(k) + \frac{1}{2}\left[x(k-1) - 2x(k) + x(k+1)\right],\ k \geq 0 \tag{6}$$

where $\delta(k)$ is the Kronecker delta function and $x(k)$ is a function with form

$$x(k) = \begin{cases} k^{2H} & ;k \geq 0 \\ 0 & ;k < 0 \end{cases} \tag{7}$$

Using (two-sided) Z-transform, (6) can be transformed into Z-domain

$$R(z) = \frac{1}{2} + \frac{1}{2}X(z)\left[z^{-1} - 2 + z\right] = \frac{1}{2} + \frac{1}{2}X(z)\left[\frac{(z-1)^2}{z}\right] \tag{8}$$

where $X(z) = Z\{x(k)\}$. Filtering both sides of (8) with a digital filter defined by the transfer function $F(z)$ or equivalently by the impulse response $f(k)$

$$F(z) = 2\frac{z}{(z-1)^2} \leftrightarrow f(k) = \begin{cases} 2k & ;k \geq 0 \\ 0 & ;k < 0 \end{cases} \tag{9}$$

leads to

$$X(z) = F(z)R(z) - \frac{z}{(z-1)^2} \tag{10}$$

or equivalently in the time domain we get

$$x(k) = f(k) * r(k) - k \tag{11}$$

where * denotes the linear convolution. The computation of (11) can be done with complexity of O(n) by using following formula

$$x(k) = 2[x(k-1) + r(k-1)] - x(k-2),\ k \geq 2 \tag{12}$$

where $x(0) = 0$ and $x(1) = 1$ (from (11)). Let's now define

$$H(k) = \frac{1}{2}\frac{\log x(k)}{\log k},\ k \geq 2 . \tag{13}$$

Using (7) we see, that $H(k) = H$. Analogically to (11), for a self-similar data set, the $\hat{x}(k)$ can be obtained from $\hat{r}(k)$. Then the function

$$\hat{H}(k) = \frac{1}{2}\frac{\log \hat{x}(k)}{\log k},\ k \geq 2 \tag{14}$$

is expected to be a constant function with the value of the Hurst parameter as we expect $\hat{r}(k)$ to follow $r(k)$ and thus $\hat{x}(k)$ to follow $x(k)$. Since the ACF of a real process never exactly fits to (2), the $\hat{H}(k)$ will likely not be exactly constant for all $k \geq 2$. In this case, the estimation of H from (14) can be made, for example, by choosing the sample mean of $\hat{H}(k)$. We will refer to this estimate as $\hat{H}$.

4. Results and Conclusions

To show the performance of the new method we have synthesized several fGn traces with the R statistical environment [15] using the Paxson's algorithm [13], [14]. For each value of $H = \{0.6, 0.7, 0.8, 0.9\}$, 100 different traces have been analyzed, each trace with sample length of 16384. That means, that for each data set (with known H) we have obtained 100 estimates of the Hurst parameter $\hat{H}_n$ (n=1,2…100). The sample mean of $\hat{H}_n$ has been calculated and is given in Tab. 1. Besides that, other reference data

have been analyzed as well: the Nile River levels data listed in [3] and two Bellcore (BC) data sets that were analyzed in [1]. Our results are compared to other most common methods of estimation.

Method / Trace	FACF	MLS	RS	VT	PG	WH	AV
fGn (H=0.6)	0.5991	0.5993	0.6479	0.5948	0.6049	0.6002	0.6023
fGn (H=0.7)	0.6963	0.6952	0.7154	0.6895	0.7060	0.6998	0.6774
fGn (H=0.8)	0.7901	0.7864	0.7767	0.7753	0.8080	0.8000	0.8056
fGn (H=0.9)	0.8726	0.8658	0.8299	0.8519	0.9113	0.9000	0.9057
Nile	0.8353	0.8291	0.8395	0.8467	0.9927	0.8374	0.9031
BC (Aug89)	0.8332	0.8322	0.5581	0.8196	0.8570	0.8282	0.8049
BC (Oct89Ext)	0.9021	0.9001	0.8224	0.8906	0.9581	0.8943	0.9769

Tab. 1. Estimators comparison. FACF – filtration of the ACF, MLS – method of least squares, RS – rescaled statistics, VT – aggregated variance-time plot, PG – periodogram, WH – Whittle's estimator, AV – Abry/Veitch estimator.

We see that for the synthesized data, the estimator performs well overall and gives accurate estimations. For $H = 0.9$ the algorithm seems to underestimate the Hurst parameter. In this case the spectral density fits the theoretical form (an assumption under which the spectral-domain estimators are derived), but the dependence between the samples in the time-domain is no longer correct and the ACF does not follow (2) as expected. This issue is addressed in [16] and is also present in [13], [14]. For longer sequences, this is not a problem and both time and spectral characteristics of the generated fGn traces are correct. We consider this not to be a problem of the estimator, but rather a synthesis issue, where the method provides approximation of the fGn only. For Nile data, the estimator gives estimates very close to those suggested by Beran in his analysis. Even though this data set is small, it is Gaussian and the assumption (2) holds. The Bellcore data follows the model very closely and the estimates are $H = 0.83$ and $H = 0.9$ respectively.

Tab. 2. lists the empirical 95% Confidence Intervals (CI) analysis $\hat{H}_n$. For CI calculations [17], we assumed $\hat{H}_n$ to have normal distribution, which has been confirmed by χ^2 tests. The width of CIs is rather narrow compared e.g. to [11]. This may be due to the fact that our method uses several lags of the ACF when estimating the Hurst parameter.

	fGn (H=0.6)	fGn (H=0.7)	fGn (H=0.8)	fGn (H=0.9)
Mean of $\hat{H}_n$	0.5991	0.6963	0.7901	0.8726
95% CI	[0.5978,0.6004]	[0.6947,0.6980]	[0.7883,0.7919]	[0.8705,0.8747]

Tab. 2. Confidence intervals for $\hat{H}_n$.

Our simulations have shown that the estimator is already accurate for small number of lags. It is therefore not necessary to calculate the whole ACF and the corresponding $\hat{H}(k)$. Using very large number of lags may lead to the estimation error. This is caused by data versus model inaccuracies for very high lags and use of finite signal length. The estimated values of the Hurst parameter are correct so long as the assumption (2) holds and the data approximately follows the theoretical model. In this case, the number of lags is not an issue and the differences noted when using a different number of lags are negligible. The results presented in the tables have been obtained using the first 64 lags. The results also show that our presented method provides estimations very close to the MLS, but without the need for an iteration as in [12]. It also exploits the whole form of the ACF and not just the asymptotic behavior or a particular lag as in [11]. Due to (12) we consider this to be a fast estimator, which we think could be used as real-time estimator e.g. in a network traffic classification system [18]. An obvious extension of this work will be to use this approach for analysis of the asymptotical self-similar processes.

Acknowledgements

The research described in the paper was financially supported by the Slovak Grant Agency under grant No.1/4084/07.

References

[1] LELAND, W., TAQQU, M. S., WILLINGER, W. On the self-similar nature of Ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, 1994, vol. 2, no. 1.

[2] PAXSON, V. Wide-area traffic: the failure of Poisson modeling. In *Proceedings of the conference on Communications Architectures, Protocols and Applications*, 1994.

[3] BERAN, J. *Statistics for Long-Memory Processes*. New York: Chapman and Hall, 1994.

[4] STALLINGS, W. *High-Speed Networks TCP/IP and ATM Design Principles*. Prentice Hall, 1998.

[5] PARK, K., WILLINGER, W. Self-similar network traffic: An overview. Published in PARK, K., WILLINGER, W. *Self Similar Network Traffic Analysis and Performance Evaluation*, 1999.

[6] WILLINGER, W., PAXSON, V. Long-range dependence and data network traffic. *Long range Dependence : Theory and Applications*, 2001.

[7] WILLINGER, W., PAXSON, V. Self-similarity and heavy tails: structural modeling of network traffic. *A practical Guide to Heavy Tails: Statistical Techniques and Applications*, 1998.

[8] CROVELLA, M. E., BESTAVROS, A. Self-similarity in World Wide Web traffic: evidence and possible causes. *IEEE/ACM Transactions on Networking*, 1997, vol. 5, no. 6.

[9] GARRETT, M., WILLINGER, W. Analysis, modeling and generation of self-similar VBR video traffic. *ACM SIGCOMM Computer Communication Review*, 1994, vol. 5, no. 4.

[10] DUFFY, D. E. Statistical analysis of CCSN/SS7 traffic data from working CCS subnetworks. *IEEE Journal on Selected Areas in Communications*, 1994, vol. 12, no. 3.

[11] KETTANI, H., GUBNER, J. A. A novel approach to the estimation of the long-range dependence parameter. *IEEE Transactions on Circuits and Systems II*, 2006, vol. 53, no. 6.

[12] REZAUL, K. M., GROUT, V. Exploring the reliability and robustness of HEAF(2) for quantifying the intensity of long-range dependent network traffic. *International Journal of Computer Science and Network Security*, February 2007, vol. 7, no. 2.

[13] PAXSON, V. Fast, approximate synthesis of fractional Gaussian noise for generating self-similar network traffic. *ACM SIGCOMM Computer Communication Review*, October 1997, vol. 27, no. 5.

[14] PAXSON, V. Fast approximation of self-similar network traffic. *Technical report, Lawrence Berkeley Laboratory and EECS Division*, University of California, Berkeley, April 1995.

[15] R – PROJECT: Environment for statistical computing and graphics. Available online under www.r-project.org.

[16] AITKEN, G. Long and short-term correlation properties of computer-generated fractional Gaussian noise. *Physica A: Statistical and Theoretical Physics*, 2004.

[17] MONTGOMERY, D. C., RUNGER, G. C. *Applied Statistics and Probability for Engineers*. Third Ed., Wiley, pp.254-256, 2003.

[18] MRAČKA, I., ORAVEC, M. Classification of traffic of communication networks by multilayer perceptron. In *Proc. of International Conference New Information and Multimedia Technologies NIMT–2008*. Brno (Czech Republic), September 2008, pp. 46-49.

About Authors...

Ján PROCHÁSKA was born in Trnava, Slovakia, on November 5, 1979. He received Ing. degree in Electrical Engineering from the Slovak University of Technology in Bratislava, in 2004. Since 2004 he is external PhD student at the Dept. of Telecommunications, Slovak University of Technology in Bratislava.

Radoslav VARGIC was born in Myjava, Slovakia, on September 7, 1972. He received Ing. degree in Electrical Engineering in 1995 and PhD degree in Applied Informatics in 1999, both from the Slovak University of Technology in Bratislava. Now he is university teacher at the Dept. of Telecommunications, Slovak University of Technology in Bratislava.