A Sampling Technique for Variance Estimation of Long-Range Dependent Traffic

Javier Aracil, *Member, IEEE*

***Abstract*—Due to the long-range dependence of Internet traffic, the sampling distribution of the variance is very hard to obtain and, as a result, confidence intervals cannot be provided. Nevertheless, we show that the r-decimated variance sampling distribution can be approximated by a χ^2 distribution. This sampling technique can be used to provide a confidence interval for the variance, with significant benefits for many applications in Internet dimensioning, traffic forecasting and control.**

***Index Terms*—Self-similarity, variance estimation.**

I. Introduction and Problem Statement

Long-range dependent traffic models constitute the foundation for traffic forecasting algorithms [1] and also for the analysis of buffer dynamics under long range dependent traffic (Fractional Gaussian Noise -FGN-) [2]. However, all of the above algorithms for network dimensioning and control demand *a priori* knowledge of the traffic correlation and marginal distribution parameters (moments). Let the time be slotted, with the slot duration being equal to δ, and let $\{X_i, i \geq 1\}$ denote the number of bytes per slot. Let us consider a traffic sample $(X_1, \ldots, X_n)$ and the well-known variance estimator

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{1}$$

where $\bar{X} = (1/n)\sum_{i=1}^{n} X_i$. Due to the long range dependence, the correlation function can be approximated by $\rho(k) \sim k^{2H-2}$, being H the Hurst parameter and k the correlation lag. Thus, the covariance terms in (1) are not null and the variance estimator is biased. Actually, the bias term $\Delta_n = -(n-1)^{-1}2\sum_{k=1}^{n-1}(1-(k/n))\rho(k)$ converges to zero and s^2 is asymptotically unbiased. Furthermore, if the correlation structure of the traffic can be estimated, then the bias can be removed with the estimated correction factor $(1-\hat{\Delta}_n)^{-1}$ (see [3, p. 156]), where $\hat{\Delta}_n$ is an estimator for Δ_n.

On the other hand, the distributional properties of the estimator are complicated (see [3, Sec. 8.4]) for a more detailed description) and, as a result, confidence intervals cannot be provided in a closed analytical form. Needless to say, confidence intervals are necessary to provide a reliable variance estimator. Alternatively, variance can be estimated with maximum-likelihood estimators (MLEs). Such MLEs achieve the same rate of convergence as under short range dependence. In fact, it can be shown that the estimates are $\sqrt{n}$ consistent and asymptotically normal [3, Th. 5.1].

In this letter, an alternate way to estimate the variance of long-range dependent traffic is presented. In case of traffic with independent increments, and by Fisher's Theorem, it turns out that the sampling distribution of $(n-1)s^2/\sigma^2$ is χ^2_{n-1}, where σ^2 is the marginal distribution variance. Thus, the objective of this letter is to analyze to which extent the use of a sampling technique makes the distribution of (1) become χ^2. The intuition behind is that correlation decreases as the lag between samples increases, thus resembling the independent increments case. In presence of strongly correlated traffic ($H = 0.78$), it will be shown that the use of sampling at a rate $1/4$ or lower provides a distribution for the estimator (1) which can be modeled as χ^2, with significance level 5%. Extremely simple, but also extremely useful, the proposed sampling method provides a reliable variance estimator for a wide range of applications in traffic engineering.

II. Analysis and Results

The estimator is defined as follows:

$$s'^2 = \frac{1}{n/r-1}\sum_{i=1}^{n/r}(X_{ri} - \overline{X'(n,r)})^2 \tag{2}$$

where r denotes the sampling period, for a sample of size n, and $\overline{X'(n,r)} = (1/(n/r))\sum_{i=1}^{n/r} X_{ri}$. Concerning the r-decimated sample mean $\overline{X'(n,r)}$ it has been shown [4] that the relative efficiency $\mathrm{var}(\overline{X'(n,1)})/\mathrm{var}(\overline{X'(n,r)})$ tends to 1 as $r \to \infty$ for all r. However, the deficiency d of $\overline{X'(n,r)}$ relative to $\overline{X'(n,1)} = \bar{X}$ such that $\mathrm{var}(\overline{X'(n+d,r)}) \leq \mathrm{var}(\overline{X'(n,1)}) \leq \mathrm{var}(\overline{X'(n+d-1,r)})$ tends to infinity as $r \to \infty$. Thus, if one is interested in estimating the mean in a given time frame with a finite number of samples, this can be achieved by decimation at a moderate decrease in efficiency [4, p. 16]. In any case, the results of this letter have also been obtained using estimation of the sample mean without decimation, with nearly no difference at all with respect to using the r-decimated mean.

Without loss of generality it is assumed that n/r is an integer. If r is large enough, then it will be shown that $(n/r-1)s'^2/\sigma^2$ has a sampling distribution which is approximately equal to $\chi^2_{n/r-1}$.

First, we provide an example that shows how small values of r make the sampling distribution of $(n/r-1)s'^2/\sigma^2$ (2) fit a $\chi^2_{n/r-1}$. Recall that a traffic sample is a n-tuple $(X_1, \ldots, X_n)$ that represents the number of bits per time interval. A number of

Manuscript received January 8, 2004. The associate editor coordinating the review of this paper and approving it for publication was Prof. D. Petr. This work was supported by the Ministry of Science and Technology of Spain.

The author is with Universidad Pública de Navarra, Campus Arrosadía, 31006 Pamplona, Spain (e-mail: javier.aracil@unavarra.es).

Digital Object Identifier 10.1109/LCOMM.2004.833822

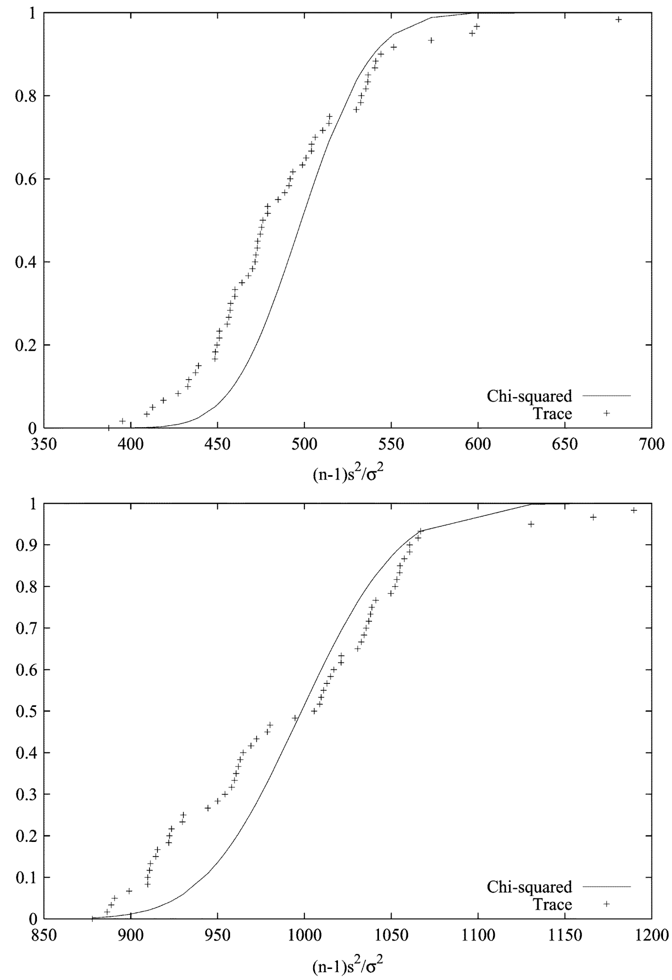

Fig. 1. Comparison between the sampling distribution of ns^2/σ^2 (1) and the χ^2_{n-1} distribution for (top) $n = 500$ (top) and (bottom) $n = 1000$.

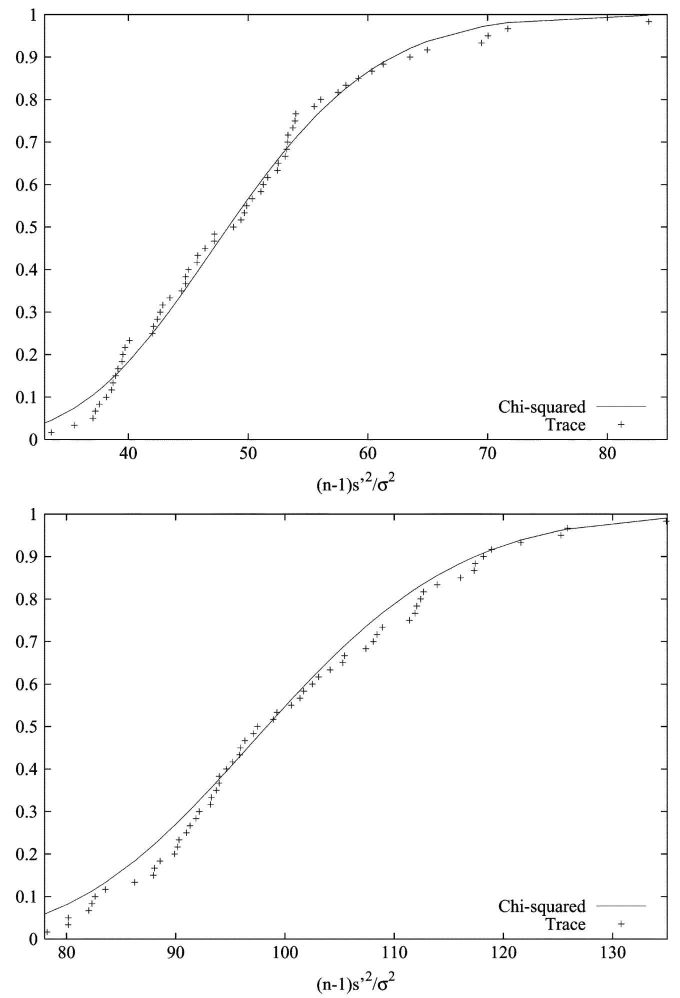

Fig. 2. Comparison between the sampling distribution of $(n/r - 1)s'^2/\sigma^2$ (2) and the $\chi^2_{n/r-1}$ distribution $(r = 10)$ for (top) $n = 500$ and (bottom) $n = 1000$.

60 independent FGN traffic samples are generated using the fast and approximate method proposed in [5] (DTFT-based). The FGN parameters are set to those inferred from the Ethernet *Bellcore traces*-pOct.TL-(coefficient of variation $\sigma^2/\mu^2 = 0.1, \mu =$ 2200 kb/s, Hurst parameter $H = 0.78$) [2]. In order to visually illustrate the effect of correlation on variance estimation, Fig. 1. shows the sampling distribution of $(n-1)s^2/\sigma^2$ (1), together with the χ^2_{n-1} distribution, for a sample size $n = \{500, 1000\}$.

On the other hand, Fig. 2 shows the sampling distribution of $(n/r - 1)s'^2/\sigma^2$ (2) for $n = \{500, 1000\}$ and $r = 10$ with the same parameters (number of samples, sample size, FGN mean, variance and Hurst parameter). Note that the χ^2_{49} and χ^2_{99} distributions fit the sampling distribution much better, in comparison to the results obtained in Fig. 1.

More formally, the Pearson statistic was used to test goodness of fit of the sampling distribution of $(n/r - 1)s'^2/\sigma^2$ (2) to a $\chi^2_{n/r-1}$ distribution. Traces are generated using the DTFT-based method [5] and the random midpoint displacement method [6]. Furthermore, and in order to provide a model as close as possible to real Internet traffic, traces have also been obtained by superposition of Poisson arriving Pareto-distributed bursts [7]. The results are shown in Tables I $(n = 500)$ and II $(n = 1000)$. The first column is the lag r for the estimator (2), the second column is the Pearson statistic and the third column is the reject threshold for the null hypothesis H_0 of goodness of fit to a χ^2 distribution, for a significance level of 5%. In order to obtain the Pearson statistic, the number of bins were selected so that more than five samples per bin were always available. Thus, the goodness of fit threshold varies depending on the value of r. Regarding the variance estimator (1) the Pearson statistic takes on values well above the null hypothesis threshold. For example, for the DTFT-based generator, the Pearson statistic is equal to 58.033 77 for $n = 500$ (goodness of fit threshold $5\% = 13.276\,704$) and 44.686 554 for $n = 1000$ (goodness of fit threshold $5\% = 16.811\,894$). Values of the Pearson statistic well above the goodness of fit threshold are also obtained with the RMD and $M/Par/\infty$ generators. Thus, the null hypothesis H_0 cannot be accepted.

The following conclusions can be obtained from Tables I $(n = 500)$ and II $(n = 1000)$. First, the null hypothesis of goodness of fit cannot be rejected for nearly all r values with $r \geq 4$. Furthermore, the Pearson statistic takes on low values in comparison with the H_0 reject threshold. Second, since a sampling distribution can be identified, the proposed estimator (2)

TABLE I
PEARSON STATISTIC VERSUS r FOR $n = 500$

	DTFT-based [5]		*RMD [6]*		$M/Par/\infty$ *[7]*	
r	Pearson stat.	H_0 thresh. (5%)	Pearson stat.	H_0 thresh. (5%)	Pearson stat.	H_0 thresh. (5%)
2	10.905331	9.487729	51.966728	13.276704	13.084941	11.070498
3	4.422827	9.487729	51.966728	13.276704	3.974490	9.487729
4	6.514215	9.487729	29.947941	15.086272	6.094170	12.591587
5	2.452535	11.070498	15.506250	15.086272	4.301482	11.070498
6	4.891207	9.487729	6.442152	16.811894	3.273594	11.070498
7	11.997296	12.591587	13.050696	16.811894	4.683629	12.591587
8	2.885880	12.591587	17.400188	16.811894	12.212222	14.067140
9	6.009133	12.591587	5.490216	11.344867	18.085748	14.067140
10	8.875018	11.070498	3.587152	13.276704	0.795887	12.591587
11	13.564603	11.070498	4.238611	16.811894	7.868512	11.070498
12	5.107333	12.591587	5.463474	16.811894	7.913223	9.487729
13	2.569502	11.070498	1.738125	15.086272	5.036819	12.591587
14	2.662108	12.591587	6.212236	15.086272	6.212778	11.070498
15	5.130196	11.070498	2.528135	15.086272	5.791853	12.591587

TABLE II
PEARSON STATISTIC VERSUS r FOR $n = 1000$

	DTFT-based [5]		*RMD [6]*		$M/Par/\infty$ *[7]*	
r	Pearson stat.	H_0 thresh. (5%)	Pearson stat.	H_0 thresh. (5%)	Pearson stat.	H_0 thresh. (5%)
2	8.122409	12.591587	81.438702	15.086272	16.183464	11.070498
3	14.641594	9.487729	24.512192	15.086272	4.928069	11.070498
4	5.519479	9.487729	34.980645	15.086272	4.374393	9.487729
5	4.839687	9.487729	9.555719	15.086272	2.705874	11.070498
6	9.216731	11.070498	7.659267	15.086272	3.821124	9.487729
7	6.109357	12.591587	6.298723	16.811894	1.727433	11.070498
8	1.749234	11.070498	18.414237	15.086272	8.561873	12.591587
9	5.548344	11.070498	13.697659	15.086272	13.443318	11.070498
10	4.739715	11.070498	7.338337	15.086272	7.188905	11.070498
11	3.226673	11.070498	13.758732	16.811894	2.812539	9.487729
12	6.038950	11.070498	9.181775	16.811894	7.916704	12.591587
13	1.601226	12.591587	3.920196	16.811894	14.876435	9.487729
14	6.563014	11.070498	13.831221	16.811894	8.938706	12.591587
15	4.116972	12.591587	12.439699	16.811894	3.156371	12.591587

serves to provide a confidence interval for the variance. Such confidence interval, for a significance level α, is equal to

$$\left(\frac{(n/r-1)s'^2}{\chi^2_{n/r-1;\alpha/2}}, \frac{(n/r-1)s'^2}{\chi^2_{n/r-1;1-\alpha/2}}\right) \tag{3}$$

being $\chi^2_{n/r-1;\alpha}$ the $1-\alpha$ percentile of a $\chi^2_{n/r-1}$ distribution.

Finally, similar results have also been obtained with different values of H. As the value of H increases the sampling interval also increases, due to the stronger long-range dependence. Such results are not shown in the letter for brevity.

III. CONCLUSION

In this letter, a simple estimator for the marginal distribution variance of long-range dependent traffic has been presented, that provides a confidence interval for reliable estimation of the traffic variance.

REFERENCES

[1] A. Sang and S.-q. Li, "A predictability analysis of network traffic," in *Proc. INFOCOM'00*, 2000, pp. 342–351.

[2] I. Norros, "On the use of fractional Brownian motion in the theory of connectionless networks," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 953–962, Aug. 1995.

[3] J. Beran, *Statistics for Long-Memory Processes*. London, U.K.: Chapman&Hall, 1994.

[4] D. B. Percival, "On the sample mean and variance of a long-memory process," Dep. Statistics, University of Washington, Seattle, WA, Technical Report 69, 1985.

[5] V. Paxson, "Fast, approximate synthesis of fractional Gaussian noise for generating self-similar network traffic," *Computer Commun. Rev.*, vol. 27, no. 1, pp. 5–18, 1997.

[6] W. Lau, A. Erramilli, J. L. Wang, and W. Willinger, "Self-similar traffic generation: The random midpoint displacement algorithm and its properties," in *Proc. ICC'95*, 1995, pp. 466–472.

[7] A. Erramilli, P. Pruthi, and W. Willinger, "Fast and physically-based generation of self-similar network traffic with applications to ATM performance evaluation," in *Proc. Winter Simulation Conf.*, 1997, pp. 997–1004.