Estimation of the Memory index of transmission rate measurements using an Infinite Source Poisson model

Gilles Faÿ[1], François Roueff[2], and Philippe Soulier[3]

[1] UST de Lille - CNRS UMR8524
U.F.R. de Mathématiques - Bât. M2
59655 Villeneuve d'Ascq Cedex, France
(e-mail: Gilles.Fay@univ-lille1.fr)
[2] GET - Télécom Paris - CNRS LTCI
46 rue Barrault,
75634 Paris Cedex, France
(e-mail: roueff@tsi.enst.fr)
[3] Université de Paris X - Equipe ModalX - UFR SEGMI
200 avenue de la République,
92001 Nanterre Cedex, France
(e-mail: philippe.soulier@u-paris10.fr)

Abstract. We present long memory processes related to some point processes, give their main properties, asymptotic behaviour and discuss some statistical issues with a view on Internet traffic measurements. The Infinite Source Poisson model is a generalisation of the M/G/∞ queue. Arrivals are driven by a homogeneous Poisson process, durations of active periods are independent and identically distributed (iid) and independent of the arrivals. Each active periods (say dowload sessions) is assumed to have a constant transmission rate and the available bandwidth to be unlimited. Theses rates are iid, independent of the arrivals but possibly depending on the durations. In a traffic modelling context, the obtained process $X(t)$ can serve for modelling the bandwith occupation, often called the *workload*. The stability of the model depends on the tail behavior of the duration distribution. Both in the stable and unstable cases, the tail behavior of the durations can be recovered from the dependence structure of $X(t)$. In particular, heavy-tails durations will result in long range dependence (LRD) in $X(t)$ and the corresponding tail and Hurst indices α and H satisfy $H = (3-\alpha)/2$ for all $\alpha \in (0, 2)$. In practical situations, the process $X(t)$ is observed through passive measurements, by counting packets going trough a point of the network, and then by evaluating the instantaneous workload. Such measurements are much simpler than collectiong complete characterizations of flows. However, from a queuing point of view, as mentionned above about the stability, the important parameter is α. The object of this paper is to rely on the relationship between α and H for estimating α from measurements on $X(t)$.
Keywords: Infinite Source Poisson Model, Heavy tails and long range dependence, Traffic modelling.

1 Modelling transmission rates

We consider the Infinite Source Poisson model with random transmission rate defined by

$$X(t) = \sum_{j \in \mathbb{N}} U_j \mathbf{1}_{\{t_j \le t < t_j + \eta_j\}} \, . \tag{1}$$

The transmissions are generated at birth times $\{t_j\}$ which are the points of a unit rate homogeneous Poisson process on the positive half-line and have rates given by $\{U_j\}$. The transmissions have positive durations $\{\eta_j\}$. We assume that the vectors $\{(\eta_j, U_j)\}$ are i.i.d. and independent of the arrivals process. The workload at time t is the sum of rates of all surviving present and past transmission. This model was considered by [Resnick and Rootzén, 2000], [Mikosch *et al.*, 2002] among others.

In the following, we consider that the path of the process is observed along continuous time. From a numerical point of view, since the path of X is piece-wise constant, this means that one observes all the jump times and the workload at these times. In practical situations, the transmission rate is measured by counting the packets going through some point of the network link. From the packet counts, one may compute the overall average rate of transmission over equi-spaced time slots $[k\delta, (k+1)\delta]$ $k \in \mathbb{Z}$. From now on, we take $\delta = 1$ without loss of generality. The process X is not aimed to model the traffic at packets level since the transmission rate at the packets level cannot be assumed to be constant. Nevertheless

$$Y_k = \int_k^{k+1} X(s)\, ds$$

is a reasonable model for the overall transmission rate averaged on $[k, k+1]$ because, by locally averaging the instantaneous rate, one eliminates local variations of it. The estimator we will consider is computed from the wavelet coefficients of X. In the case of Haar wavelet these coefficients can be computed exactely from the discrete sequence $\{Y_k,\, k \in \mathbb{Z}\}$. Otherwise, some adaptations are needed but we will not pursue in this direction here and thus will assume either that the continuous time path of X is observed or that the wavelet ψ used below is the Haar wavelet $\psi = \frac{1}{2}(\mathbf{1}_{[0,1)} - \mathbf{1}_{[1,0)})$.

We now introduce the assumption on the joint distribution of the transmissions rates and durations.

Assumption 1 *The random vectors* $\{(\eta, U),\ (\eta_n, U_n), n = 0, \pm 1, \pm 2, \dots\}$ *are i.i.d. with distribution* ν *on* $\mathbb{R}_+ \times \mathbb{R}$ *and independent of the homogeneous Poisson Point Process on the real line with points* $\{t_j\}_{j \in \mathbb{Z}}$*; there exist a real number* $\alpha \in (0, 2)$ *and a positive integer* k^* *such that* $\mathbb{E}[|U|^{k^*}] < \infty$ *and for each integer* $k = 0, 1, \dots, k^*$

$$\mathbb{E}\left[U^k \mathbf{1}_{\{\eta > t\}}\right] = L_k(t) t^{-\alpha} \, . \tag{2}$$

where L_k *are slowly varying as* $t \to +\infty$.

Defining, for each $k \leq k^*$, the signed measure on $\mathbb{R}_+$

$$\nu_k(\mathrm{d}v) := \int u^k \nu(\mathrm{d}v, \mathrm{d}u),$$

and the function

$$H_k(t) = \nu_k(t, \infty) = \mathbb{E}\left[U^k \mathbf{1}_{\{\eta > t\}}\right], \quad t \geq 0,$$

Condition (2) is equivalent to saying that H_k, $k = 0, 1, \ldots, k^*$, are regularly varying with index α.

Assumption 1 implies in particular that the tails of the distribution of η is regularly varying with index α. This in turns implies Assumption 1 if U and η are independent, in which case the functions L_k differ by a multiplicative constant. A more realistic situation for network traffic modelling is the case where the transmission rate U is independent of the amount of transmitted data during the download session (which is equal to $W := U\eta$), given that the rate is above some threshold. Below this threshold, the accessible amount of data is supposed to have light tails, and above this threshold, W is supposed to have heavy tails. In practice this threshold separate high rate connections (say, xDSL/LAN/Cable connection) from low rate connections (say, RTC connection), which are not suitable for downlaoding large data. In this case, it can be shown that the measure ν_k inherits the heavy tails of W for all k such that $\mathbb{E}|U|^k < \infty$.

2 Stationary version and asymptotic behavior

If $\mathbb{E}[\eta] < \infty$, a stationary version of this process is defined by

$$X_S(t) = \sum_{j \in \mathbb{Z}} U_j \mathbf{1}_{\{t_j \leq t < t_j + \eta_j\}} \; t \in \mathbb{R}, \tag{3}$$

where, in the sequel, $\{t_j\}$ are the points of a unit rate homogeneous Poisson process on the line such that $t_k < t_{k+1}$ for all k and $t_{-1} < 0 \leq t_0$.

By Karamata's Theorem, for all such k, we easily obtained the asymptotic equivalences of standard tail behaviors of ν_k. For instance, if $\alpha > 1$,

$$\mathbb{E}\left[U^k \{\eta - t\}_+\right] \mathrm{sim} \frac{1}{\alpha - 1} L_k(t) t^{1-\alpha} \; . \tag{4}$$

Proposition 1 *Let Assumption 1 hold. The process X_S is well defined and strictly stationary if and only if $\mathbb{E}[\eta] < \infty$. If moreover $k^* \geq 2$, then X_S is weakly stationary with expectation and autocovariance function given by*

$$\mathbb{E}[X_S(t)] = \mathbb{E}[U\eta] \; ,$$
$$\mathrm{cov}(X_S(0), X_S(t)) = \mathbb{E}[U^2(\eta - t)_+] \mathrm{sim} \frac{1}{1-\alpha} L_2(t) t^{1-\alpha} \text{ if } \alpha > 1 \; ,$$

where the equivalence holds as $t \to +\infty$.
The process X *is nonstationary with expectation* $\mathbb{E}[X(t)] = \mathbb{E}[U(\eta \wedge t)]$ *and autocovariance function given, for* $s \leq t$ *by*

$$\mathrm{cov}(X(s), X(t)) = \mathbb{E}[U^2\{s - (t-\eta)_+\}_+] = \int_{t-s}^{t} H_2(v)\mathrm{d}v \ .$$

If $\alpha \in (0,1)$ and if t and s tend to infinity at the same rate, the following asymptotic equivalent of $\mathrm{cov}(X(s), X(t))$ holds. For all $t, s > 0$, as $T \to \infty$,

$$\mathrm{cov}(X(Ts), X(Tt)) \mathrm{sim} \frac{1}{1-\alpha} L_2(T) T^{1-\alpha} \{(s \vee t)^{1-\alpha} - |t-s|^{1-\alpha}\} \ . \qquad (5)$$

The proof of Proposition 1 is a straightforward application of well known properties of Poisson point processes.

If $\mathbb{E}[\eta] < \infty$, the non-stationary process X converges to X_S. By definition, the difference between X and X_S is given by

$$X_S(t) - X(t) = \sum_{k<0} U_k \mathbf{1}_{\{t_k \leq t < t_k + \eta_k\}}, \quad t \geq 0 \ .$$

Since $\mathbb{E}[\eta] < \infty$ and since the η_k are i.i.d and independent of the birth times t_k, a Borel-Cantelli argument yields that this sum has almost surely a finite number of terms, which is at most the number of indices $k < 0$ such that $t_k + \eta_k \geq 0$. Hence, almost surely, $\lim_{t\to\infty}\{X_S(t) - X(t)\} = 0$. This limit also holds in the mean $\mathbb{E}[|X_S(t) - X(t)|] \leq \mathbb{E}[|U|(\eta - t)_+] \to 0$. The asymptotic behavior of the cumulative workload is now investigated.

If we are not in the stable case, that is, for $\mathbb{E}[\eta] = \infty$, the process X_S is not defined (see Proposition 1). We may still consider the weak limit of the cumulative workload but this limit will be very different in the two cases as shown by the next proposition.

For $\alpha < 1$ (implying $\mathbb{E}[\eta] = \infty$), the next proposition gives a straightforward extension of the results of [Resnick and Rootzén, 2000] to the case of random transmission rate U_j. In the case $\alpha > 1$ (implying $\mathbb{E}[\eta] < \infty$), it has been proved under slightly different assumptions by [Mikosch *et al.*, 2002], [Maulik *et al.*, 2002] or [Mikosch and Resnick, 2004].

Proposition 2 *Denote* $H = (3-\alpha)/2$. *If* $0 < \alpha < 1$, i.e. $1 < H < 3/2$, *and if Assumption 1 holds with* $k^* = \infty$, *then the sequence of processes* $\{L_2^{-1/2}(T) T^{-H} \int_0^{Tt} (X(s) - \mathbb{E}[X(s)])\,\mathrm{d}s, t \geq 0\}$ *converges weakly to the Gaussian process* W *with autocovariance function*

$$\mathrm{cov}(W(s), W(t)) = \frac{1}{1-\alpha} \int_0^t \int_0^s \{(u \vee v)^{1-\alpha} - |u-v|^{1-\alpha}\}\,\mathrm{d}u\,\mathrm{d}v \ .$$

If $1 < \alpha < 2$, i.e. $1/2 < H < 1$, *then* $T^{-H} \int_0^{Tt} X(s)\mathrm{d}s$ *converges in probability to 0, and the sequence* $\{T^{-1/\alpha} \int_0^{Tt} (X(s) - \mathbb{E}[X(s)])\,\mathrm{d}s, t \geq 0\}$ *converges weakly to an* α*-stable Levy process.*

This proposition illustrates a change of behavior between the stationary and non-stationary cases.

3 Estimation

3.1 Terminology

The most important parameter for this process is thus the parameter α. In accordance with the notation in use in the context of long memory processes, we define the Hurst index of the process X as $H = (3-\alpha)/2$, because the variance of partial sums scales as T^{2H}. We can also define $d = H - 1/2 = 1 - \alpha/2$, in relation to fractionally integrated processes, such as ARFIMA processes, but this would be quite arbitrary in this context where no fractional integration is involved.

3.2 Methods

The parameter α is a tail index, so traditional methods to estimate a tail index could be used. But it is well known that these methods are not very efficient in the case of dependent data (cf. [Resnick and Stărică, 1995] for instance). Moreover, in the model under consideration here, α is not the tail index of the marginal distribution of the observed process, which has finite variance whereas $\alpha < 2$. Thus it is not at all clear how to use these methods.

But as shown by Proposition 1, the coefficient α is related to the second order properties of the process: the coefficient $H = (3-\alpha)/2$ can be viewed as its Hurst index, *i.e.* H governs the rate of decay of the autocovariance function of the process. Therefore it seems natural to use an estimator of the Hurst index.

3.3 The (wavelet) coefficients

Let ψ be a bounded $\mathbb{R} \to \mathbb{R}$ function with compact support included in $[0, M]$ and such that

$$\int \psi(s)\, \mathrm{d}s = 0\,. \tag{6}$$

For integers $j \geq 0$ and $k \in \mathbb{Z}$, define

$$\psi_{j,k}(s) = 2^{-j/2}\psi(2^{-j}s - k). \tag{7}$$

The wavelet coefficients of the path are defined as

$$d_{j,k} = \int \psi_{j,k}(s)X(s)\, \mathrm{d}s\,, \quad d^S_{j,k} = \int \psi_{j,k}(s)X_S(s)\, \mathrm{d}s\,. \tag{8}$$

Asume that a path is observed between time 0 and T. Since $\psi_{j,k}$ has support in $[k2^j, (k+M)2^j]$, the above coefficients can be computed for all (j,k) such that $T2^{-j} \geq L$ and $k = 0, 1, \ldots, T2^{-j} - M$.

Lemma 1 *Define*

$$\mathcal{L}(z) = z^{\alpha} \int_0^{\infty} \left[\int_{-\infty}^{\infty} \left\{ \int_t^{t+zv} \psi(u)\,\mathrm{d}u \right\}^2 \mathrm{d}t \right] \nu_2(\mathrm{d}v) \ . \tag{9}$$

Then $\mathcal{L}$ is slowly varying at infinity and

$$\mathbb{E}[d_{j,k}^S] = 0 \ , \quad \mathrm{var}(d_{j,k}^S) = \mathcal{L}(2^j)\, 2^{(2-\alpha)j} \ , \tag{10}$$

$$\mathbb{E}[d_{j,k}] = O\left(L_1(k2^j)\, 2^{(3/2-\alpha)j} k^{-\alpha} \right) \ , \tag{11}$$

$$\mathrm{var}(d_{j,k} - d_{j,k}^S) = O\left(L_2(k2^{-j})\, 2^{(2-\alpha)j}\, k^{-\alpha} \right) \ . \tag{12}$$

Remark 31 *The coefficients $d_{j,k}$ are centered in the case where U and η are independent and U is centered, even in the nonstationary case.*

3.4 The estimator

Lemma 1 provides the rationale for the following minimum contrast estimator of α which is related to the local Whittle estimator, cf. [Künsch, 1987], [Robinson, 1995b]. The obtained estimator has been introduced by Moulines, Roueff and Taqqu (2004) and is called the wavelet Whittle estimator. For positive integers $J_0 < J$, define

$$\Delta = \{(j,k)\ , J_0 < j \leq J \ , \ \ 0 \leq k \leq 2^{J-j} - 1\} \ \text{ and } \ \delta = \frac{1}{\#\Delta} \sum_{(j,k)\in\Delta} j \ .$$

The scale index J is the maximal scale index available from the data while J_0 is a cut-off tuned by the user. The local Whittle estimator of α is then defined as:

$$\hat{\alpha} = \arg \min_{\alpha' \in (0,2)} \log \left(\sum_{(j,k)\in\Delta} \frac{d_{j,k}^2}{2^{(2-\alpha')j}} \right) + \delta \log(2)(2-\alpha') \ .$$

Equivalently, we could have defined $\hat{H} = (3-\hat{\alpha})/2$ or $\hat{d} = 1 - \hat{\alpha}/2$.

Theorem 31 *Let $\alpha \in (1,2)$ and let Assumption 1 hold. Suppose that $\mathcal{L}(x) \to 1$ as $x \to \infty$. If $J_0 \to \infty$, $J \to \infty$ and $J_0 < J/\alpha$ then $\hat{\alpha}$ is a consistent estimator of α.*

A corresponding results hold in the case where $\alpha \in (0,1]$ but some adaptations are needed in the definition of the estimator and a second vanishing moment is needed on ψ.

4 Simulations

We have simulated M/G/∞ processes, which correspond to the process X with $U_k = 1$ for all k's, and estimated α via different classical estimators of long range dependence. The obtain paths are represented in Figure 1 and Figure 2, respectively in non-stable ($\alpha = 0.7 < 1$) and stable ($\alpha = 1.5 > 1$) situation. Monte-carlo simulations provided the boxplots and MSE estimates for the several estimators, also represented on these figures. In those graphs, the X-coordinates $V1$, $V2$, ... $V10$ correspond to the scale cut-off VJ_0.

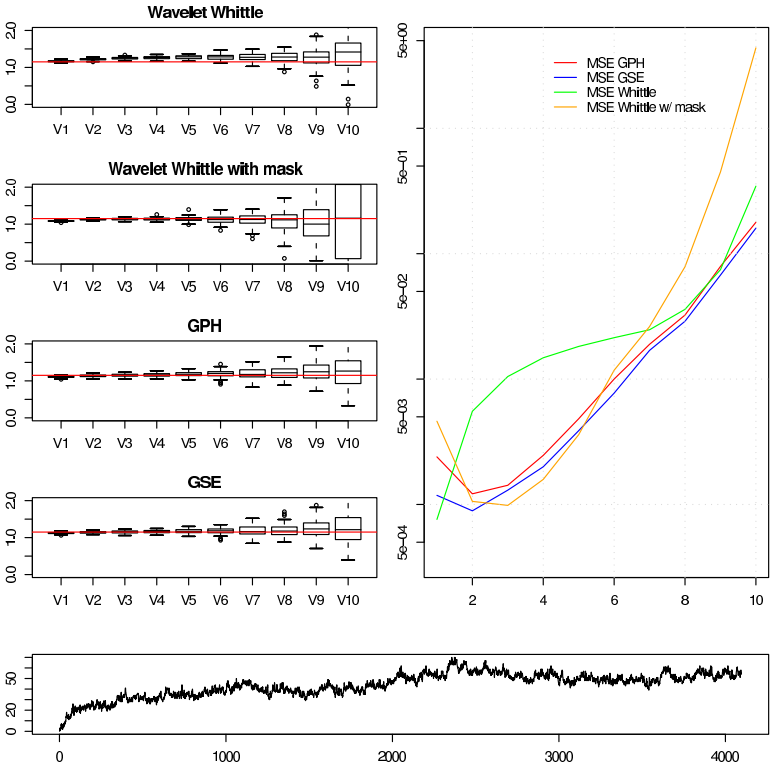

Fig. 1. $\alpha = 0.7$: the process do not converge to a stationary. Its cumulative load is approximately gaussian.

References

[Künsch, 1987]H. R. Künsch. Statistical aspects of self-similar processes. In Yu.A. Prohorov and V.V. Sazonov (eds), *Proceedings of the first World Congres of the Bernoulli Society*, volume 1, pages 67–74. Utrecht, VNU Science Press, 1987.

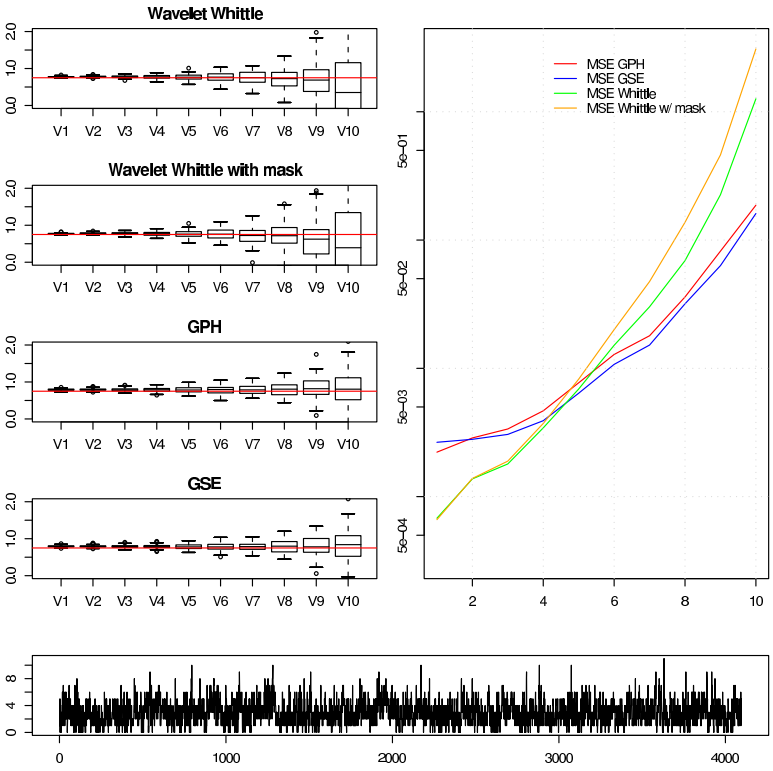

Fig. 2. $\alpha = 1.5$

[Maulik *et al.*, 2002]Krishanu Maulik, Sidney Resnick, and Holger Rootzén. Asymptotic independence and a network traffic model. *Journal of Applied Probability*, 39(4):671–699, 2002.

[Mikosch and Resnick, 2004]Thomas Mikosch and Sidney Resnick. Activity rates with very heavy tails. Technical Report 1411, Cornell University, 2004.

[Mikosch *et al.*, 2002]T. Mikosch, S.I. Resnick, H. Rootzen, and A. Stegeman. Is network traffic approximated by stable Levy motion or fractional Brownian motion? *Annals of Applied Probability*, 12:23–68, 2002.

[Moulines *et al.*,]E. Moulines, F. Roueff, and M. S. Taqqu. A wavelet Whittle estimator. working paper.

[Resnick and Rootzén, 2000]Sidney Resnick and Holger Rootzén. Self-similar communication models and very heavy tails. *The Annals of Applied Probability*, 10(3):753–778, 2000.

[Resnick and Stărică, 1995]Sidney Resnick and Cătălin Stărică. Consistency of Hill's estimator for dependent data. *Journal of Applied Probability*, 32(1):139–167, 1995.

[Robinson, 1995b]P.M. Robinson. Gaussian semiparametric estimation of long range dependence. *Annals of Statistics*, 24(5):1630–1661, 1995b.