Robust estimation of the self-similarity parameter in network traffic using wavelet transform

Haipeng Shen\textsuperscript{a,*}, Zhengyuan Zhu\textsuperscript{a}, Thomas C.M. Lee\textsuperscript{b}

\textsuperscript{a}Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, NC 27599-3260, USA
\textsuperscript{b}Department of Statistics, Colorado State University, USA

Received 9 March 2006; received in revised form 20 December 2006; accepted 20 February 2007
Available online 4 March 2007

Abstract

This article studies the problem of estimating the self-similarity parameter of network traffic traces. A robust wavelet-based procedure is proposed for this estimation task of deriving estimates that are less sensitive to some commonly encountered non-stationary traffic conditions, such as sudden level shifts and breaks. Two main ingredients of the proposed procedure are: (i) the application of a robust regression technique for estimating the parameter from the wavelet coefficients of the traces, and (ii) the proposal of an automatic level shift removal algorithm for removing sudden jumps in the traces. Simulation experiments are conducted to compare the proposed estimator with existing wavelet-based estimators. The proposed estimator is also applied to real traces obtained from the Abilene Backbone Network and a university campus network. Both results from simulated experiments and real trace applications suggest that the proposed estimator is superior.

Keywords: Hurst parameter; Long-range dependence; Internet traffic; Non-stationarity

1. Introduction

Self-similarity and long-range dependence (LRD) are two closely related phenomena observed in many different fields, ranging from biology to econometrics to telecommunication. In particular, the seminal paper [1] discusses the self-similar nature of network traffic. Willinger et al. [2] reveal the relationship between the self-similar property of network traffic and the heavy tail distribution of duration times of Internet connections, which makes self-similar processes a popular tool for modelling Internet traffic flows. Mikosch et al. [3] and Park et al. [4] provide some recent theoretical and empirical insights into the relationship, as well as support for the LRD of network traffic data. Below we briefly discuss the definitions of self-similar and LRD processes; see [5] for more details.

Suppose $X(t)$ is a second-order stationary stochastic process, and denote its spectrum and autocorrelation function by $f_X$ and $\gamma_X$, respectively. The process $X(t)$ is said to be long-range dependent with the LRD parameter $\alpha$, if either, for some constant $c_f > 0$,

$$f_X(v) \sim c_f |v|^{-\alpha} \quad \text{as } |v| \to 0 \quad \text{where } \alpha \in (0, 1),$$

\text{Equation (1)}
or for a different constant $c_\gamma > 0$, $\gamma_X(k) \sim c_\gamma |k|^{-(1-\alpha)}$ as $|k| \to \infty$, where $\alpha \in (0, 1)$. Intuitively, the process is called long-range dependent because $\gamma_X(k)$ goes to zero so slow as $k \to \infty$ that $\sum_{k} \gamma_X(k) = \infty$. A process $Y(t)$ is said to be self-similar with the self-similarity parameter $H$ if and only if $e^{-H} Y(ct) \overset{d}{=} Y(t)$ for all $c > 0$, where $\overset{d}{=} \text{means equality in finite-dimensional distributions}$. Note that $H$ is also known as the Hurst parameter.

There is a close connection between LRD and self-similar processes. Indeed, the increments of any finite variance self-similar process are LRD, as long as $1/2 < H < 1$, with $H$ and $\alpha$ related through $H = (\alpha + 1)/2$. For example, the fractional Brownian motion (fBm) [5] has emerged as a natural model of network traffic in practice, which is a self-similar Gaussian process with stationary increments. The increments are the so-called fractional Gaussian noise (fGn), which are LRD with the parameter $\alpha = 2H - 1$.

The self-similarity parameter $H$, or equivalently the LRD parameter $\alpha$, controls the degree of self-similarity or LRD. Thus, accurate and reliable estimation of $\alpha$ or $H$ is of both theoretic interest and practical importance for modelling network traffic, provisioning service quality and planning capacity [6,7]. The current paper focuses on robust estimation of the Hurst parameter in the context of network traffic.

Many estimators have been proposed in the literature, including the aggregated variance estimator, the periodogram based estimator, and the Whittle estimator; e.g., see [8,9]. Veitch and Abry [10] propose a wavelet-based estimator, which is referred to as the AV estimator in the literature. Abry et al. [11] prefer the AV estimator over the
aggregated variance estimator and the Whittle estimator because of its robustness to polynomial trends. Recently, Soltani et al. [12] propose a modification of the AV estimator, which provides more accurate estimates for simulated fGn. We will refer to this estimator as the SSB estimator. These previous studies indicate that the wavelet-based estimators are, in many cases, the preferred ones for handling stationary LRD processes.

The AV and SSB estimators are designed for stationary processes. However, in reality, network traffic is not stationary, as studied in [13] in connection with issues about scaling properties of TCP flows. One type of non-stationarity involves gradual diurnal trends, and can be handled by the estimators which are robust to polynomial trends [11].

Another common non-stationary phenomenon is the presence of large mean level shifts (Fig. 1(a)). The theory of $a$ and $b$ traffic is proposed in [14], suggesting that large level shifts could be caused by large file transmissions over high bandwidth links. Large level shifts can also be resulted from changes in routing map, IP scanning and other abnormal activities. A third type of non-stationarity can be attributed from the sudden drop of the traffic level to effectively zero (Fig. 2(a)). This is different from level shifts, and sometimes can be considered as a missing value problem. It can be caused by recording device or router failures, or some other problems in the data collection process. The presence of level shifts and missing values can deteriorate the performance of the standard estimators [15,16].

In addition to the non-stationary features, extremely large or small observations are often prominent in network traffic data, which make their empirical distributions to have heavier tails than Gaussian distributions. Therefore, a good Hurst parameter estimator should be robust against such extreme values as well. However, the AV/SSB estimators are least-squares based techniques, which are known to be prone to extreme observations. Thus it is important to seek more robust estimation procedures that can handle these issues.

The goal of this article is to propose a robust wavelet-based procedure for estimating the Hurst parameter $H$, which we term the robust wavelet (RW) estimator. We show that our estimator provides accurate and reliable estimates of $H$ with the presence of mean level shifts, missing values and extreme observations in the data. Similar to [14], we consider Internet traffic as a mixture of $a$ and $b$ traffic. The $a$ traffic consists of large file transmissions at high speed, while the $b$ traffic can be modelled as an LRD process with a Hurst parameter $H$. Our approach is to first remove the $a$ traffic component using a novel level shift removing procedure, then apply a robust regression technique to the empirical wavelet coefficients of the remaining $b$ traffic component to estimate $H$.

The level shift removing procedure is formulated as a rigorous model selection problem, and the model dimension is selected using the Bayesian information criterion (BIC) [17], which is asymptotically consistent as a model selection criterion [18]. The followup robust regression step then reduces...
the influence of missing values and extreme observations.

Rincon et al. [7] studied a similar problem of segmenting LRD traffic, and proposed an alignment procedure to detect change points in the variance structure of the network traffic. They also used BIC as the model selection criterion, which they referred to as Schwarz information criterion (SIC). However, there are fundamental differences between the current work and [7]; e.g., the adopted traffic models are different.

Numerical studies using simulated fGn with level shifts and missing values show that our RW estimator dramatically reduces the bias in estimating $H$ while having similar standard error as the traditional wavelet estimators. The RW estimator is also applied to four real traces collected from the Abilene Backbone Network and the campus network of the University of North Carolina at Chapel Hill (UNC). For some traces, it is possible for the classical wavelet methods to yield estimates that are greater than 1, which are outside the permissible range of $H$ for stationary LRD time series (0.5–1). On the other hand, our RW estimator results in estimates that lie inside the above permissible range. A graphical tool, robust logscale diagram (RLD), is also proposed to accompany our RW estimator, as an extension of the regular LD [19] for the AV estimator.

The severe effects of non-stationarity on the traditional Hurst parameter estimators are also recognized in [20], where the authors propose to estimate the parameter locally over a set of time windows in order to reduce the non-stationarity effect. A median-based wavelet estimator is also developed to obtain robust estimates. This paper presents an alternative global approach via the viewpoint of $\alpha/\beta$ traffic models. Our approach deals with non-stationarity and missing values in a unified framework, and simulation studies are carried out to show the improvement of our estimator over the traditional estimators. Several robust regression procedures are also compared.

We end this section by pointing out an unfortunate fact that the Greek letter $\alpha$ has been unavoidably used to denote two different entities: the LRD parameter in (1), and the notion of $\alpha$ traffic introduced by [14]. However, confusion should not arise, as it should be clear from the context that which entity a particular $\alpha$ is referring to.

The rest of this paper is organized as follows. Two motivating examples are presented in Section 2. Section 3 introduces the standard wavelet estimators. Section 4 describes the proposed robust wavelet (RW) estimation procedure and the RLD. Section 5 provides the results of our simulation studies, and the RW estimator is applied to real network traffic traces in Section 6. We conclude in Section 7.

2. Motivating examples

The AV estimator is argued to be robust against level shifts in the mean and/or the variance in [21], except for very abrupt jumps of large magnitude. While the argument is valid for moderately large level shifts, a visual inspection of recent Internet traffic data [22] reveals that large and abrupt level shifts do exist, and indeed can adversely affect the estimates of $H$. The effect of a level shift has been shown analytically to be present on all wavelet scales, and becomes more pronounced on large scales used to estimate the LRD parameter [16]. Therefore level shifts may be “confounded” with LRD.

Fig. 1(a) plots a time series of packet counts in 100ms intervals from a trace of the eastbound traffic on the Abilene Backbone Network between Kansas City and Indianapolis. See http://abilene.Internet2.edu/ for detailed information about the trace and the network. The trace \{X(t); t = 1,\ldots,65536\} is clearly non-stationary, showing oscillating level shifts. Fig. 1(b) displays the trace’s LD (see Section 3.1). Intuitively, the AV estimator of $H$ is obtained by fitting a weighted least squares regression to the upper linear region of the diagram. However, the non-stationarity actually biases the diagram upwards, and causes the AV method to estimate $H$ as 1.19, which is larger than 1.

Our RW procedure first locates and removes the level shifts (Fig. 1(c)), and estimates $H$ using robust regression techniques on the wavelet coefficients of the level-shifts-removed trace (Fig. 1(d)). Our method estimates $H$ to be 0.82 (Table 3). Thus, our approach complements [21] by proposing an estimator which is robust against large mean level shifts.

To illustrate the phenomenon of missing values, Fig. 2(a) plots the middle portion of a two-hour trace of packet counts of the incoming traffic into UNC between 1 and 3 PM on April 11, 2002 (UNC02-APR-11). As indicated on the plot, during a 10s period in the middle, very few packets were transferred through the campus network. The LD (Fig. 2(b)) shows two bumps at octaves 7 and 9,
which are statistically significant. The bumps are caused by the short period of missing values, and have a serious impact on the AV estimator [16]. Our proposed RLD shows no apparent bumps. See Sections 4.4 and 6.1 for more discussion about LD and RLD.

3. Standard wavelet estimators

This section reviews two wavelet-based techniques for estimating $H$. We shall restrict ourselves to a brief description of the main features of the wavelet analysis as employed for our purposes here. Further details on wavelets can be found, for example, in [23].

3.1. The AV estimator [10]

Consider the family of wavelet basis functions \( \{\psi_k(t) = 2^{-j/2}\psi_0(2^{-j}t - k), j = 1, \ldots, J, k \in \mathbb{Z}\} \). Suppose \( X(t) \) is a second-order stationary stochastic process with a spectrum given by (1). Let \( dX(j, k) = \langle X(t), \psi_k(t) \rangle \) denote the coefficients of the discrete wavelet transform (DWT) of the process \( X(t) \) at octave \( j \) as shown in [10], if \( X(t) \) is LRD and as \( j \to \infty \), the second moments of \( dX(j, k) \) can be well approximated by

\[
\mathbb{E}d_X(j, \cdot) \approx 2^jc^jC,
\]

(2)

where

\[ C = \int |v|^{-2} |\Psi_0(v)|^2 \, dv \]

and

\[ \Psi_0(v) = \int \psi_0(t)e^{-2\pi ivt} \, dt. \]

Eq. (2) suggests that

\[ \log_2(\mathbb{E}d_X(j, \cdot)) \approx jz + \log_2(c^jC). \]

Veitch and Abry [10] suggest estimating the variance \( \mathbb{E}d_X^2(j, \cdot) \) using the sample variance at octave \( j \), \( \mu_j = 1/n_j \sum_k d_X^2(j, k) \), where \( n_j \) is the number of wavelet coefficients at octave \( j \), and further assume that the \( d_X(j, \cdot) \)'s are i.i.d. Gaussian for a fixed \( j \), and that \( d_X(j, \cdot) \) and \( d_X(j', \cdot) \) are independent when \( j \neq j' \). As a result,

\[ \log_2 \mu_j \approx jz + \log_2(c^jC) - \log_2(n_j) + \ln X_{n_j}/\ln 2, \]

where \( \approx \) means approximate equality in distributions, and \( X_{n_j} \) is a Chi-squared random variable with \( n_j \) degrees of freedom. Furthermore, they show that

\[ \mathbb{E}(\log_2 \mu_j) \approx jz + \log_2(c^jC) + g_j \]

and

\[ \text{var}(\log_2 \mu_j) \approx \zeta(2, n_j/2)/\ln 2^2, \]

where \( g_j = \psi(n_j/2)/\ln 2 - \log_2(n_j/2), \) \( \psi(z) = \Gamma'(z)/\Gamma(z) \) is the Psi function and \( \zeta(z, v) \) is a generalized Riemann Zeta function.

Let \( y_j = \log_2 \mu_j - g_j \), the bias-corrected response, which are asymptotically unbiased estimators of the quantities \( \log_2(\mathbb{E}d_X^2(j, \cdot)) \). Then the AV estimator of \( z \) is the weighted least squares (WLS) estimate obtained from the following linear regression,

\[ y_j = jz + \log_2(c^jC) + \varepsilon_j, \]

(3)

where \( \varepsilon_j \) has mean 0 and variance \( \zeta(2, n_j/2)/\ln 2^2. \) Consequently, the Hurst parameter \( H \) can be estimated as \( \hat{H} = \langle \hat{z} + 1/2 \rangle. \) Under mild regularity conditions, the consistency and asymptotic normality of the AV estimator is established by [24].

\[ \text{Logscale Diagram:} \] The plot of \( y_j \) against \( j \) is referred as the Logscale Diagram (LD), along with Gaussian confidence intervals corresponding to the variability of the \( y_j \) [19]. The LD is often used to show the variability of stochastic processes at different octaves (or scales). For LRD processes, the upper part of the LD forms a straight line of slope \( z \) (or \( 2H - 1 \)). Pictorially speaking, the AV estimator of \( z \) is obtained by fitting a WLS line to a range of octaves on the LD. Thus, the LD also provides a good diagnostic tool for the existence of LRD. The choice of the range is a subtle problem though, which is partly addressed in [25].

3.2. The SSB estimator [12]

As one can see, the errors \( \varepsilon_j \) (3) have the same distribution as \( \log_2(X_{n_j}) \), which is different from a Gaussian distribution. Soltani et al. [12] propose an alternative approach to estimate \( z \). The approach is developed for fBm, and can be easily generalized to fGn and other stationary LRD processes. Below we will illustrate the approach for an arbitrary second-order stationary LRD process.

Let \( D_{jk} = (d_X^2(j, k) + d_X^2(j, k + n_j/2))/2. \) According to [10] and Eq. (2), \( d_X(j, \cdot) \) is approximately a Gaussian random variable with variance \( 2^jc^jC \), which leads to

\[ D_{jk} \equiv \log_2 D_{jk} \approx jz + \log_2(c^jC) - 1 + \ln X_{n_j}/\ln 2, \]

(4)

where \( X_{n_j} \) is a Chi-squared random variable with two degrees of freedom.

Let \( Z = -\ln X_{n_j}/\ln 2 \), then \( Z \) has a Gumbel distribution with parameters \( \gamma = -1 \) and \( \beta = 1/\ln 2 \). Its mean and variance are \( -1 + \delta/\ln 2 \) and \( \pi^2/(6\ln^2 2) \), respectively, with \( \delta \) being the Euler’s constant. As a result, \( D_{jk} \) has a negative
transformation. The AV method first averages the wavelet coefficients, rather than taking the log’s of the wavelet coefficients, and then applying a robust regression procedure to the empirical wavelet coefficients of the residuals. We first outline the details of the level shift removing method and the robust regression procedure, then provide a summary of the RW procedure. Analogous to the LD, we also propose an RLD as a graphical diagnostic tool for the RW estimator.

4.1. Level shift removing method

Suppose we observe the process \( X(t) \) at a set of \( n \) discrete time points. The following model can be used to model the mean level shifts,

\[
X(t) = \alpha(t) + \beta(t), \quad \alpha(t) = \sum_{i} \mu_i 1_{[t_i, t_{i+1})}(t),
\]

where \( \beta(t) \) is a stationary LRD process with a Hurst parameter \( H \). Let \( T = \{t_1, \ldots, t_{m-1}\} \) denote the collection of the shift locations.

To fit the above model, or equivalently, to remove the mean shifts \( \alpha(t) \) from \( X(t) \) to obtain \( \beta(t) \), two issues need to be addressed. The first issue is to define a “best” fitting model while the second issue is the need for a practical algorithm to locate such a defined “best” fitting model. These two issues are discussed in the next two sections. Quite often a first good step prior to level shift removal is to remove low-frequency trends hidden in real traces (e.g., [8, Ch. 9]). Nevertheless, a close examination of the traces considered in our paper reveals that such low frequency trends do not appear to exist; hence we will not consider this possible step of removing low-frequency trends in later analysis.

4.1.1. Defining a best fitting model

For the current problem, defining a “best” fitting model for \( \alpha(t) \) is the same as defining the best combination of \( m \) and \( T \). Note that once these quantities are specified, estimates of \( \mu_i \) can be obtained by

\[
\hat{\mu}_i = \frac{1}{t_{i+1} - t_i} \sum_{j=i}^{t_{i+1}-1} X(t).
\]

This “best” model identification problem can be formulated as a statistical model selection problem, in which different candidate models may have different model dimensions (in our case related to different values for \( m \)). The BIC [17] seems to serve our purpose very well (Section 5), and outperforms other model selection criteria such as Akaiake information criterion in our simulation studies. BIC has also been used successfully for related problems [7]. Specifically, we define our “best”
fitting model as the minimizer of the following BIC function:

$$\text{BIC}(m, T|H_0) = -2\{l(m, T|H_0)\} + (2m - 1) \ln n,$$

where \((2m - 1)\) is the number of parameters in the model being fitted \((m \text{ mean parameters } \mu_i \text{ and } m - 1 \text{ location parameters } t_i)\); and \(l(m, T|H_0)\) is the conditional log likelihood value for the fitted model under the assumption that \(\beta(t)\) is fGn with a Hurst parameter \(H_0\) and the variance \(\sigma^2\) estimated by maximum likelihood,

$$l(m, T|H_0) = \frac{n}{2} \ln n - \frac{n}{2} (1 + \ln(2\pi))$$

$$- \frac{n}{2} \ln \sum_{i,j} (X(t) - \tilde{z}(t))(X(s) - \tilde{z}(s)) W_{ts},$$

where \(W_{ts}\) is the \((t, s)\)th element of the inverse of the correlation matrix, whose \((t, s)\)th entry is given by the autocorrelation \(\gamma(t - s)\) as defined in (6) of Section 5.1. \(\sigma^2\) are not counted in \(\text{BIC}(m, T|H_0)\), which do not affect the minimization.

For later use we also define the corresponding residual sum of squares \(\text{RSS}_m\) of the model being fitted: \(\text{RSS}_m = \sum_t (X(t) - \tilde{z}(t))^2\), where \(\tilde{z}(t) = \sum_i \hat{H}_i \tilde{z}(t|\hat{H}_i)\). For simplicity the dependence on \(T\) is suppressed in \(\text{RSS}_m\).

An astute reader will have noticed that, in order to use \(\text{BIC}(m, T|H_0)\) to remove the level shifts to obtain a more accurate estimate \(\hat{H}\), we must first supply an initial value \(H_0\) for \(l(m, T|H_0)\). To solve this problem, we consider a set of initial candidate values for \(H_0\). For each candidate value, we derive the “best” model using the algorithm defined in Section 4.1.2. We then compare the \(\text{BIC}\) values among this set of “best” models, and select \(H_0\) as the one that gives the smallest \(\text{BIC}\) value. The corresponding best combination of \(m\) and \(T\) under the chosen \(H_0\) determines the final level shifts. From numerical experiments we observed that the final estimate depends, to a certain extent, on the choice of the initial candidate values. In practice, we advocate choosing this initial set as \(\{0.5, 0.55, 0.6, \ldots, 0.95\}\). Our extensive numerical experience suggests this initial set works very well for the estimation of \(H\). Practical implementation of this strategy is given in Section 4.3.

4.1.2. Locating the defined best model

Now the second issue is the practical minimization of \(\text{BIC}(m, T|H_0)\) with respect to \(m\) and \(T\). This is certainly not a trivial optimization task, as \(n\) is typically huge and \(m\) varies amongst different candidate models. Here we propose using the following greedy merging algorithm to (approximately) minimize \(\text{BIC}(m, T|H_0)\).

The algorithm begins with an over-fitting model whose shift locations \(t_i\)’s are believed to be a superset of the shift locations of the “best” fitting model. In our implementation, the initial over-fitting model is chosen to have mean shift locations at every other time points; i.e., there are \(n/2\) initial segments, each of length 2. Then, at each time step, two adjacent segments are merged to form a new and bigger segment. These two segments are selected in such a way that, when they are merged, it produces the smallest increase in the value of \(\text{RSS}_m\). This merging process continues until there is only one segment left, which corresponds to a model with no mean shift. At the end of this merging process, a nested sequence of fitted models are obtained and the one that gives the smallest \(\text{BIC}\) value is taken as our “best” fitting model.

The following updating formula for \(\text{RSS}_m\) is applied to further speed up the merging process. Suppose at a particular time step the algorithm computes the increase in \(\text{RSS}_m\) if the \(i\)th and \((i + 1)\)th segments are to be merged. Denote the \(\text{RSS}_m\) values before and after the merge by \(\text{RSS}_b\) and \(\text{RSS}_a\), respectively. Then the following relationship exists:

$$\text{RSS}_a = \text{RSS}_b + (t_i - t_{i-1})^2 \hat{\mu}_i^2 + (t_{i+1} - t_i)^2 \hat{\mu}_{i+1}^2$$

$$- (t_i - t_{i-1}) \hat{\mu}_i + (t_{i+1} - t_i) \hat{\mu}_{i+1} / (t_{i+1} - t_i).$$

4.2. Robust regression method

Consider the following linear regression model, \(y_j = x_j^T \beta + \epsilon_j\), where \(y_j\) is the \(j\)th response, \(x_j\) represents the corresponding \(p\)-dimensional covariate vector, and \(\beta\) is the regression coefficient. The classical least-squares (LS) regression estimates \(\beta\) by

$$\hat{\beta}_{\text{LS}} = \text{argmin} \sum_j (y_j - x_j^T \beta)^2.$$

The AV and SSB estimators \([10,12]\) use LS to obtain estimates of \(H\). When the errors \(\epsilon_j\) follow a Gaussian distribution, LS procedures enjoy excellent properties such as unbiasedness and minimum variance. However, LS procedures are very sensitive to extreme observations. If the error distribution has a heavier tail than a Gaussian distribution, then LS procedures can be very inefficient and unstable.
When estimating the Hurst parameter, even if the random processes under consideration are exact fBm, $D_{jk}^L$ (4) have a Gumbel distribution instead of a Gaussian distribution. The SSB method uses the average of $D_{jk}^L$ as the regression dependent variable. At a small octave $j$, this average is taken over a large number of wavelet coefficients, and the Central Limit Theorem implies that the average $A_j$ (5) has an approximate Gaussian distribution. This limits us to using only $A_j$s at small octaves. However, in practice, only $A_j$s at large octaves are useful for estimating $H$. For real Internet traffic data, in almost all cases, the empirical distribution of $D_{jk}^L$ is heavier than Gaussian. Furthermore, if there are some abnormalities in the data, such as abrupt level shifts to record data for a short period of time, or missing values, even more extreme observations will be created among the $D_{jk}^L$’s. Our simulation results (Section 5) suggest that the SSB estimator performs well for fGn, but gives erroneous results if the fGn is contaminated by mean level shifts or missing values.

The effect of extreme observations and non-Gaussian errors can be reduced by techniques such as robust regressions. Below we briefly discuss three robust regression techniques.

One of the first few attempts at getting more robust regression estimators came from [27], who proposed minimizing the sum of absolute values of the residuals ($L_1$ regression) instead of the sum of squares to obtain the $L_1$ estimate as

$$\hat{\beta}_{L_1} = \text{arg min} \sum_{j} |y_j - x_j^T \beta|.$$ 

The second technique is the $M$-estimator (Maximum Likelihood type estimator) [28], which estimates the parameters by minimizing the sum of some function $\rho(\cdot)$ of the residuals,

$$\hat{\beta}_M = \text{arg min} \sum_{j} \rho(y_j - x_j^T \beta).$$

A good choice of $\rho(\cdot)$ can reduce the influence of extreme observations by down-weighting them, while maintain a high statistical efficiency. Here we employ an iteratively re-weighted least squares (IRLS) algorithm to approximate the $M$-estimators [29], which uses the Huber and Bisquare weight functions, defined, respectively, as

$$f(u) = \begin{cases} 1, & |u| \leq c, \\ \frac{c}{|u|}, & |u| > c. \end{cases}$$

The weight function $f(u)$ is related to $\rho(u)$ by $f(u) = \rho(u)/u$. Typically one takes the tuning parameter $c = 1.345$ for the Huber function and $c = 4.685$ for the Bisquare weight function to make the technique 95% efficient if the error does have a Gaussian distribution.

The third technique is the least trimmed squares (LTS) estimator [30], which minimizes the sum of the first $h$ ordered squared residuals:

$$\hat{\beta}_{LTS} = \text{argmin} \sum_{j=1}^{h} r^2_{(j)},$$

where $r_{(j)}$ is the $j$th order statistic of the residual $r_j = y_j - x_j^T \beta$, and $h \in (n/2, n]$ determines the breakdown point of the procedure. LTS can have breakdown points as high as 50%, but is numerically more difficult to obtain.

To reduce the effect of extreme observations, we propose to use $D_{jk}^L$ as the response instead of $A_j$, and apply some robust regression procedure to $D_{jk}^L$. The three aforementioned robust procedures ($L_1$, IRLS, and LTS) have similar performance for simulated fGn as shown in Section 5.2. We recommend to use the IRLS procedure because of its computational advantage.

### 4.3. Robust wavelet estimation of $H$

Our robust wavelet (RW) procedure for estimating $H$ can be summarized as follows:

1. For each $H_k \in \{0.5, 0.55, 0.6, \ldots, 0.95\}$,
   a. Use the level shift removing method in Section 4.1 to determine the corresponding number and locations of mean level shifts $(m_k, T_k)$ in the data trace $X(t)$.
   b. Choose the combination $(m_k, T_k, H_k)$ that gives the smallest BIC.
2. Let $\hat{\beta}(t)$ be the estimated level shifts from the fitted model with the chosen $H_k$. Remove $\hat{\beta}(t)$ from $X(t)$ and obtain the estimated $\beta$ traffic $\hat{\beta}(t) = X(t) - \hat{\beta}(t)$.
3. Apply IRLS to the wavelet coefficients of $\beta(t)$ to obtain a final robust estimate $\hat{H}$. 

In our simulation studies (Section 5), this robust procedure is compared with the SSB
method [12], and appears to provide better estimates of $H$ with smaller bias and standard error. Our estimator also provides more stable $H$ estimates for real data traces when certain abnormalities make the other estimators inappropriate (Section 6).

4.4. Robust logscale diagram

The LD [19] provides a good diagnostic tool for the AV estimator. Correspondingly, we define the RLD as the plot of $y_j' = \arg \min \sum k f(y - D'_j,k)(y - D'_j,k)^2$ against octave $j$, where $f(\cdot)$ is the weight function employed in the IRLS procedure. For an appropriately defined $f(\cdot)$, $y_j'$ can be regarded as a robust estimator of the center of $D'_j,k$. Roughly speaking, our robust method estimates $x$ (and eventually $H$) by fitting a line to the upper part of the RLD using the IRLS regression technique. Thus, the RLD plays a similar role as a diagnostic tool for our robust method as the LD does for the AV method. It can be viewed as a robust version of the LD. A comparison of the RLD with the LD is provided in Fig. 3(d).

5. Simulation results

In this section we empirically compare our proposed robust wavelet (RW) estimator with the SSB estimator [12]. The AV estimator [10] is not compared, as the SSB estimator is shown to be better than the AV estimator [12,26]. Three types of traces are simulated in the studies: $f$Gn, $f$Gn with missing values, and $f$Gn with mean level shifts. These different setups are employed to mimic some scenarios observed in real Internet traffic traces.

![Fig. 3. Sample fGn with level shifts.](image-url)
5.1. Fractional Gaussian noise and its simulation

Fractional Gaussian noise (fGn) \( \epsilon_t \) is the increment of fBm [5]. It is a stationary Gaussian time series with mean zero, variance \( \sigma^2 \) and autocorrelation function
\[
\gamma(k) = \langle (k+1)^{2H} + (k-1)^{2H} - 2|k|^{2H} \rangle / 2, \quad k \geq 0.
\]

For \( H = 1/2 \), \( \gamma(k) = 0 \) for \( k \neq 0 \) and \( \epsilon_t \) becomes white noise. For \( H \neq 1/2 \), as \( k \to \infty \), \( \gamma(k) \sim H(2H-1)|k|^{2H-2} \) and \( \epsilon_t \) has LRD if \( 0.5 < H < 1 \). The fGn is one of the simplest examples of LRD time series. A simple model for the time series of bin counts \( X_t \) is \( X_t = \mu + \epsilon_t \), which only has three parameters, \( \mu, \sigma^2 \) and the Hurst parameter \( H \). We employ the circulant embedding method described in [31] to simulate fGn. This is the only theoretically exact method that has computational complexity of \( n \log n \), where \( n \) is the length of the fGn.

5.2. Simulation I: fGn

In the first simulation study, we compare the performances of the SSB estimator with three robust regression estimators, where IRLS, LTS, and \( L_1 \) regression techniques are applied to \( D_{jk} \) of the simulated traces. Here no level shift removing is used.

We first simulate 50 fGn of length \( n = 2^{14} \) for \( H = \{0.5, 0.75, 0.9\} \) separately with \( \mu = \sigma = 20 \), and estimate \( H \) using the four estimators. Since our aimed application is for the Internet traffic data, we are more interested in scenarios with \( H \in (0.5, 1) \), which corresponds to LRD time series. The case \( H = 0.5 \) is included for comparison purpose only. From the simulated traces, we observe that a fGn trace becomes more variable and more non-stationary-like as \( H \) gets closer to 1. When \( H = 0.9 \), there are a lot of natural oscillations, which are not caused by non-stationarity but rather the strong autocorrelation.

The bias, standard error (SE) and root mean square errors (RMSE) of the 50 estimated \( H \)'s are calculated. Let \( H_0 \) be the true value of \( H \), \( \bar{H}_i \) be the estimated \( H \) for the \( i \)th trace, and \( \bar{H} \) be the average of the 50 \( \bar{H}_i \)'s. Then these quantities are defined as:
\[
\text{bias} = \frac{1}{50} \sum_{i=1}^{50} \bar{H}_i - H_0, \quad \text{SE} = \sqrt{\frac{1}{49} \sum_{i=1}^{50} (\bar{H}_i - \bar{H})^2},
\]
\[
\text{RMSE} = \sqrt{\text{bias}^2 + \text{SE}^2}.
\]

The analysis results [22] suggest that all four methods yield very accurate estimates of \( H \) when the processes are exact fGn. IRLS gives marginally better estimates than LTS and \( L_1 \). It is also better than the SSB method when \( H \) is large. In sequel we will only use IRLS in the following simulation studies, as all three robust procedures are expected to have similar performances, and IRLS is computationally faster.

We also applied the level shifts removing method to the simulated traces, which correctly found no or very few level shifts. In all cases, the estimated \( H \)'s are almost identical to those of the original traces. This indicates that our level shifts removing method works well in practice when no artificial level shifts are introduced in the simulated traces. Simulation results in Section 5.4 show that it also works well when artificial level shifts are present.

5.3. Simulation II: fGn with missing values

To mimic traffic breaks encountered in real traces such as UNC02-APR-11, we simulate fGn traces with missing values. To be more specific, we take the traces simulated in Section 5.2, and manually set portions of them to be 0 according to the following two scenarios: (i) \( X(t) = 0 \) for \( t = 1, \ldots, 200 \), and (ii) \( X(t) = 0 \) for \( t = 2^{13} + 1, \ldots, 2^{13} + 200 \).

We compare the performances of the SSB, IRLS and RW estimators, and the results are summarized in Table 1. The last two estimators are different in the following way. The RW estimator employs the mean level shifts removing algorithm first before applying the IRLS regression on the wavelet coefficients of the residual trace; while the IRLS estimator fits the IRLS regression on the wavelet coefficients of the original trace.

We can see that for large \( H \), the IRLS estimator is on average about 15% better than the SSB method in terms of RMSE, and the improvement is mostly in the reduction of the bias. A two-sample \( t \)-test shows that the bias of the IRLS estimator is significantly smaller than that of the SSB method (\( p \)-value < 0.002), and an \( F \)-test of equal variance suggests no difference in the variances of the two estimators (\( p \)-value = 0.59). For \( H = 0.5 \), IRLS is significantly better when missing in the middle (\( p \)-value = 10\(^{-6} \) for bias reduction). This
comparison shows the usefulness of using robust regression technique when missing values are present, even with a tiny percentage (\( \approx 1\% \)). A comparison between the RW and IRLS estimators shows that the former in general has a smaller bias, although the difference is statistically significant only for small \( H \). The biggest difference occurs when \( H = 0.5 \), and the difference is more dramatic for missing in the middle. This is caused by boundary effects. The data near boundaries are discarded when estimating \( H \) using wavelet methods.

We close this section with the following comment regarding the unbiasedness and efficiency of the estimators considered in Table 1. Under the assumption of fGn with no missing values or level shifts, we performed a test of unbiasedness and a test that compares SE with the Cramer–Rao lower bound (CRLB). The test results show that all the estimators are biased, and cannot reach the CRLB, including the AV/SSB estimator. In principle, one can use maximum likelihood to derive an estimator that can reach the CRLB asymptotically. However, the computational burden would override the practical feasibility.

5.4. Simulation III: fGn with level shifts

In this simulation study, we take the simulated fGn traces in Section 5.2 and manually introduce some mean level shifts to the traces. For a particular simulated fGn trace, the unit for the level shifts is the grand mean of that trace. The magnitude of level shifts being added varies for different \( H \), which is 1 unit of the grand mean for \( H = 0.5 \) and 0.75 while 4 units of the grand mean for \( H = 0.9 \). For illustration purpose, a typical fGn trace is plotted in Fig. 3 for each \( H \) with the corresponding level shifts superimposed.

Our level shift removing method using BIC is able to select the right number and locations of level shifts to remove in most cases. In particular, the method finds exactly four shifts 98% of the time for \( H = 0.5 \), 84% of the time for \( H = 0.75 \), and 68% of the time for \( H = 0.9 \). When \( H \) is large, it becomes increasingly difficult to distinguish the level shifts from the natural variation caused by strong autocorrelations (Fig. 3).

Table 2 compares the results of the SSB, IRLS, and RW estimators for fGn with level shifts. When \( H = 0.5 \), the amount of level shifts caused a substantial bias of size 0.068 for the SSB method. The bias is positive but smaller when the fGn has a larger \( H \). This pattern is more apparent in another simulation study where we add 16 level shifts of magnitude one unit of the grand mean.

| Table 1 | Comparison of the SSB, IRLS and RW estimators for fGn with missing values |
|---------|-----------------------------|-----------------------------|-----------------------------|
|         | SSB          | IRLS          | RW           |
| Missing at the beginning |               |               |               |
| \( H = 0.5 \) |   |               |               |
| bias     | 0.006 | 0.004 | 0.002 |
| SE       | 0.015 | 0.014 | 0.015 |
| RMSE     | 0.016 | 0.015 | 0.015 |
| \( H = 0.75 \) |   |               |               |
| bias     | 0.011 | 0.007 | 0.005 |
| SE       | 0.016 | 0.015 | 0.015 |
| RMSE     | 0.019 | 0.016 | 0.016 |
| \( H = 0.9 \) |   |               |               |
| bias     | 0.010 | 0.006 | 0.005 |
| SE       | 0.017 | 0.016 | 0.015 |
| RMSE     | 0.020 | 0.017 | 0.016 |

| Missing in the middle |               |               |               |
| \( H = 0.5 \) |   |               |               |
| bias     | 0.028 | 0.020 | -0.001 |
| SE       | 0.014 | 0.014 | 0.015 |
| RMSE     | 0.031 | 0.024 | 0.015 |
| \( H = 0.75 \) |   |               |               |
| bias     | 0.014 | 0.011 | 0.009 |
| SE       | 0.016 | 0.015 | 0.015 |
| RMSE     | 0.021 | 0.018 | 0.018 |
| \( H = 0.9 \) |   |               |               |
| bias     | 0.013 | 0.009 | 0.008 |
| SE       | 0.018 | 0.016 | 0.016 |
| RMSE     | 0.022 | 0.019 | 0.017 |

| Table 2 | Comparison of the SSB, IRLS, and RW estimators for fGn with level shifts |
|---------|-----------------------------|-----------------------------|-----------------------------|
|         | SSB          | IRLS          | RW           |
| \( H = 0.5 \) |   |               |               |
| bias     | 0.068 | 0.053 | 0.001 |
| SE       | 0.013 | 0.015 | 0.015 |
| RMSE     | 0.069 | 0.055 | 0.015 |
| \( H = 0.75 \) |   |               |               |
| bias     | 0.023 | 0.022 | 0.003 |
| SE       | 0.014 | 0.015 | 0.013 |
| RMSE     | 0.027 | 0.026 | 0.014 |
| \( H = 0.9 \) |   |               |               |
| bias     | 0.059 | 0.051 | 0.004 |
| SE       | 0.029 | 0.026 | 0.015 |
| RMSE     | 0.066 | 0.057 | 0.015 |
mean for every $H$. This is not surprising, as a larger $H$ means that $X(t)$ is more positively autocorrelated, and adding level shifts increases the autocorrelation of $X(t)$. Some explanations are offered by [21] on why level shifts have a smaller influence on estimating $H$ when the fGns have a $H$ closer to 1. The IRLS estimator can, to a certain extent, reduce the bias, but the results are still far from satisfactory. On the other hand, the RW estimator dramatically reduces the bias for all $H$ values, especially when comparing to the SSB estimator. The reduction in the bias is highly significant ($p$-value $< 10^{-5}$) for all $H$ according to two-sample $t$-tests.

For the trace shown in Fig. 3(c), Fig. 3(d) plots the LDs of the original fGn (True LD) and the level-shifts-added fGn (LD) along with the RLD. Approximate Gaussian confidence intervals are also provided along the True LD using the variance of $e_j$ (Section 3.2). As one can see, the upper part of the RLD looks very similar to the True LD, and falls well within the confidence intervals, while the upper part of the LD is completely outside the confidence intervals. This illustrates the big bias of the SSB estimator graphically. Note that the difference between the RLD and the True LD over the small octaves is caused by using different estimates for the center of $D_{jk}$.

As for the standard error (SE) of the estimated $H$’s, the RW estimator has a slightly larger SE when $H = 0.5$. However, the RW estimator reduces the SE almost 50% when $H = 0.9$, which is the case we are mostly interested in for real network traffic modelling. After combining the bias and the SE together, the RW estimator has the smallest RMSE for all $H$, which is only 22% to 52% of the RMSE of the SSB estimator.

6. Real data examples

In this section we apply the RW estimator to four real Internet traffic traces, which have mean level shifts and/or missing values. They are the two traces in Figs. 1 and 2, and two more traces collected from the UNC campus network, between 3:00 AM and 5:00 AM on April 9, 2002 (UNC02-APR-09) and between 1:00 PM and 3:00 PM on April 13, 2002 (UNC02-APR-13). The four traces are selected from a much larger collection of traces analyzed in [32] because they are all problematic. We plotted in [22] the real traces, the detected level shifts, as well as the detrended traces from which $H$ is estimated. For UNC02-APR-11 and UNC02-APR-13, zoomed plots are also provided for the problematic portions.

We will only compare our RW estimator with the AV estimator. The reason is that, when estimating $H$ for real traces, one needs to pre-select the range of octaves for which the scaling behavior of LRD exists. An automatic procedure to select such a range for the AV estimator is proposed in [25]. For simulated fGns, one could simply use all possible octaves.

6.1. Comparison of logscale diagrams

Figs. 1(b) and 2(b) show the comparison of the LD and the RLD for Abilene and UNC02-APR-11. The RLDs are for the traces with level shifts removed, and are produced using the Bisquare weight function. The regular LDs appear to be more variable, as $y_j$ is a function of the mean of the squared wavelet coefficients, which is not resistant to the influence of extreme observations. This leads to upward bias or bumps at small time scales, which results in an unstable estimate of $H$ when the original trace has level shifts or missing values. On the other hand, the RLDs are not biased or have no suspicious bumps at small scales, because $y_j$ is a robust estimate of the center of $D_{jk}$. At coarse scales the RLDs are smaller than the LDs due to the removal of level shifts. The variances of the $y_j$’s at coarser scales are much larger than those at finer scales, as indicated by the wider confidence intervals. Thus the large variations of the $y_j$’s at the largest few scales are not completely unexpected, and they have fairly little effect on estimating $H$ since their weights are small in the WLS.

6.2. Comparison of the AV and RW estimators

Table 3 reports the Hurst parameter estimates for the four traces using the two methods. The automatic octave selection procedure developed in [25] is implemented for the AV method and the optimal range of octaves used to estimate $H$ is reported as well.

Park et al. [4] perform an extensive study of the LRD of UNC traces, and for most regular traces, select the octave range to be around 5–13. It is conceivable that LRD appears after a certain scale for all data traces with similar connection duration time distributions. This motivates us to choose octaves from 5 to 13 to estimate $H$ for our RW estimator. The optimal number of level shifts
selected according to BIC is also presented in Table 3. As one can see, the $H$ estimates are very different between the two methods.

For the four traces considered here, because of the abnormal problems discussed earlier, the AV method does not yield reasonable answers. For the Abilene, UNC02-APR-09 and UNC02-APR-13 traces, the AV estimates are greater than 1, which are outside the range of $H$ for stationary LRD time series (0.5–1), indicating possible non-stationarity. For the UNC02-APR-11 trace, the small amount of missing values presented in the middle of the trace causes suspicious bumps in the LD (see Fig. 2(b)), leading to instability in selecting the optimal range of octaves. The selected range only includes the three largest octaves, which is much shorter than the usual range and results in a large uncertainty of the $H$ estimate. On the other hand, all the RW estimates are in the reasonable range (0.5–1).

6.3. Subtrace self-consistency

An alternative way to assess an estimation procedure is to apply the procedure to different subtraces of a trace and compare the corresponding estimates of $H$. If a procedure is robust and reliable, one would expect that similar estimates for $H$ will be obtained amongst the different subtraces. We employ this idea to compare the two methods.

Each trace is divided into four equal-length subtraces and the AV and RW methods are applied to each subtrace. The estimated $H$’s are reported in Table 4. If we treat $\hat{H}$, the estimate for the whole trace, as the truth, then the RMSE of the estimates for the subtraces can be defined as: $\text{RMSE} = \sqrt{\sum_{i=1}^{4}(\hat{H}_{Si} - \hat{H})^2}/4$, where $\hat{H}_{Si}$ is the estimated $H$ for the $i$th subtrace of a trace. Table 4 shows that, for the UNC traces, the RW estimator is much more consistent among the subtraces than the AV estimator. As for the Abilene trace, the RMSEs are similar; however, all the AV estimates are larger than one, indicating non-stationarity.

7. Conclusion

In this paper we developed a robust wavelet estimator for $H$, the Hurst or self-similarity parameter, which is an important index for network traffic data. Our motivation for developing such a robust estimator is based on the observation that real network traffic often exhibits non-stationary behaviors such as abrupt level shifts and missing values. Two key components of our estimator are an automatic level shift removal procedure and the use of robust regression techniques. Our approach of modelling Internet traffic as a mixture of mean level shifts and a stationary component is similar to the notion of $\alpha$ and $\beta$ traffic introduced by [14]. We demonstrated the promising empirical properties of our estimator via both simulation and application to real data.

Our current model assumes only non-stationarity in the mean traffic, while the variance is assumed to be stationary. This is a reasonable assumption for the traces we analyzed partly due to their relatively short durations. However, in practice, network traffic can also exhibit non-stationarity in the variance. Although the wavelet based estimators (AV/SSB/RW) are robust against moderate non-stationarity in variance [21], if the traces exhibit significant evidence against the constant variance assumption, our method could be improved by incorporating this information. It is of our interest to extend the current method by considering models where the variance is allowed to be non-stationary. As discussed earlier, [7] proposed a segmentation algorithm that can locate change points in the

<table>
<thead>
<tr>
<th>Trace</th>
<th>AV Octave</th>
<th>AV $\hat{H}$</th>
<th>RW $m$</th>
<th>RW $\hat{H}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abilene</td>
<td>6–13</td>
<td>1.19</td>
<td>160</td>
<td>0.82</td>
</tr>
<tr>
<td>UNC02-APR-09</td>
<td>7–13</td>
<td>1.28</td>
<td>37</td>
<td>0.87</td>
</tr>
<tr>
<td>UNC02-APR-11</td>
<td>11–13</td>
<td>0.92</td>
<td>110</td>
<td>0.82</td>
</tr>
<tr>
<td>UNC02-APR-13</td>
<td>9–13</td>
<td>1.51</td>
<td>410</td>
<td>0.82</td>
</tr>
</tbody>
</table>

Notes: The RW estimator uses octaves 5–13 in all four traces, and $m$ is the number of level shifts selected by the level shifts removing method described in Section 4.1.

Table 4 Comparison of the AV and RW estimators for the subtraces

<table>
<thead>
<tr>
<th>Trace</th>
<th>Estimator</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abilene</td>
<td>AV</td>
<td>1.29</td>
<td>1.27</td>
<td>1.17</td>
<td>1.19</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>RW</td>
<td>0.80</td>
<td>0.83</td>
<td>0.85</td>
<td>0.92</td>
<td>0.06</td>
</tr>
<tr>
<td>UNC02-APR-09</td>
<td>AV</td>
<td>1.49</td>
<td>0.91</td>
<td>0.92</td>
<td>1.32</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>RW</td>
<td>0.93</td>
<td>0.87</td>
<td>0.85</td>
<td>0.93</td>
<td>0.05</td>
</tr>
<tr>
<td>UNC02-APR-11</td>
<td>AV</td>
<td>0.90</td>
<td>1.37</td>
<td>0.88</td>
<td>0.95</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>RW</td>
<td>0.84</td>
<td>0.79</td>
<td>0.82</td>
<td>0.87</td>
<td>0.04</td>
</tr>
<tr>
<td>UNC02-APR-13</td>
<td>AV</td>
<td>0.92</td>
<td>0.85</td>
<td>0.92</td>
<td>0.75</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>RW</td>
<td>0.86</td>
<td>0.87</td>
<td>0.79</td>
<td>0.85</td>
<td>0.05</td>
</tr>
</tbody>
</table>
variance structure of the traffic. We expect that the segmentation algorithm can be combined with our approach to handle cases where both mean and variance exhibit level shifts. For example, one can envision first segmenting the trace into subtraces with constant variance, then removing mean level shifts in each subtrace and following up with the robust regression.

Acknowledgment

The authors thank the referees whose comments greatly improved the paper. Special thanks are also due to Steve Marron, Felix Hernández-Campos, Don Smith, Patrice Abry and Darryl Veitch. The authors are partially supported by SAMSI, U.S. NSF Grants DMS-0606577, DMS-0605434 and DMS-0203901.

References

[27] F.Y. Edgeworth, On observations relating to several quantities, Hermathena 6 (1887) 279–285.