doi:10.3969/j.issn.1000-436x.2013.06.005

分数阶 Fourier 变换域中网络流量的自相似特性分析

郭通，兰巨龙，黄万伟，张震
(国家数字交换系统工程技术研究中心，河南 郑州 450002)

摘 要：通过分析网络流量数据在 FrFT 域的统计特性发现，实际网络流量在 FrFT 域满足自相似性，进一步地，针对网络流量在 FrFT 域的“时域”和“频域”展开，分别给出了基于改进的整体经验模态分解—去趋势波动分析（MEEMD-DFA）的 Hurst 指数估计法以及基于加权最小二乘回归（WLSR）的 Hurst 指数自适应估计法。实验结果表明，相比于现有估值算法，MEEMD-DFA 法具有较高的估计精度，但计算复杂度高；而 FrFT 自适应估计法则具有更优的估计顽健性，且计算复杂度较低，可作为一种实时在线估计真实网络数据 Hurst 指数的方法。
关键词：自相似特性；分数阶 Fourier 变换；Hurst 指数；整体经验模态分解；去趋势波动分析；加权最小二乘；自适应
中图分类号：TP393.0 **文献标识码**：A **文章编号**：1000-436X(2013)06-0038-11

Analysis the self-similarity of network traffic in fractional Fourier transform domain

GUO Tong, LAN Ju-long, HUANG Wan-wei, ZHANG Zhen
(National Digital Switching System Engineering & Technological Research Center, Zhengzhou 450002, China)

Abstract: Statistical characteristics of network traffic data in FrFT domain were analyzed, which indicates the self-similarity feature. Further, Hurst parameter estimation methods based on modified ensemble empirical mode decomposition-detrended fluctuation analysis (MEEMD-DFA) and adaptive estimator with weighted least square regression (WLSR) were presented, which aimed at displaying network traffic in “time” or “frequency” domain of FrFT domain separately. Experimental results demonstrate that the MEEMD-DFA method has more accurate estimate precision but higher computational complexity than existing common methods. The overall robustness of adaptive estimator is more satisfactory than that of the other methods in simulation, while it has lower computational complexity. Thus, it can be used as a real-time online Hurst parameter estimator for traffic data.
Key words: self-similarity; fractional Fourier transform; Hurst parameter; ensemble empirical mode decomposition; detrended fluctuation analysis; weighted least square regression; adaptive

1 引言

自从 Leland 等人[1]提出自相似的以太网流量模型以来，对计算机网络的自相似性研究就受到了众多研究者的青睐。Pierre 等人[2]通过分析 2001 年~2008 年贯穿太平洋主干链路（MAWI 数据集）上每天的流量数据发现：实际链路中包含有许多异常和拥塞流量，采用草图（sketch）法滤除这些异常流量后得到的数据显示出尺度不变的突发性或自相似性，Himanshu 等人[3]对该主干链路 2001 年~2009 年小时间尺度（1~100 ms）流量行为及特性的纵向分析和研究进一步印证了这一发现。Ciflikli 等人[4,5]基于现有流量特性的研究几乎都是针对 IPv4 相关协议展开这一现状，对连接至日本 WIDE-6Bone 骨干链路的 IPv6 线路上的流量数据进行了深入的分析，研究表明：相比于 IPv4，IPv6 分组到达时间间

收稿日期：2012-05-30；**修回日期**：2013-02-20
基金项目：国家重点基础研究发展计划（“973”计划）基金资助项目（2012CB315900）；国家高技术研究发展计划（“863”计划）基金资助项目（2011AA01A103）
Foundation Items: The National Basic Research Program of China (973 Program) (2012CB315900); The National High Technology Research and Development Program of China (863 Program) (2011AA01A103)

隔与分组长度的统计特性表现出更强的自相似性。这些都说明，即使随着实际链路带宽和负载的增加，随着新型网络应用的快速部署，真实网络流量仍普遍保持着强烈、持续而稳定的自相似（长相关）特性。

Hurst指数作为表征网络流量突发特性的重要参数，它反映了网络数据的自相似程度及二阶统计特性，通常，突发网络流量的 Hurst 指数在 0.5~1 之间，表示网络具有正的相关结构，Hurst 指数越大说明网络的自相似(长相关）程度越高，突发性也越强。而当发生网络攻击时，自相似性会降低，当攻击使得网络几乎完全阻塞时，Hurst 指数将趋向于 0.5，因此，快速而准确地估计 Hurst 指数对于网络整体性能的定量分析和网络异常检测具有重要意义。

Hurst 指数估计方法分为时域和频域 2 类。常用的时域法包括方差−时间(V-T, variance-time)法、聚合序列绝对值(Abs, absolute values of the aggregated series)法、回归残差(Res, residuals of regression)法和 R/S (rescaled adjusted range)法[6~8]。频域法有周期图(periodogram)法、Whittle 估计法和小波（wavelet）估计法[6~8]。

近年来，分数阶 Fourier 变换[9]以其具有的时频旋转特性在信号处理、数学和通信等领域得到了广泛的应用。文献[10]分析了网络流量在 FrFT 域中取不同变换阶数时 Hurst 指数估值的变化情况，认为网络流量的自相似特性能够通过 FrFT 表示，但其研究对象仅限于正常网络流量且未给出 FrFT 与 Hurst 指数两者间的直接对应关系。文献[11,12]通过对 FrFT 与小波变换进行比较分析，发现在满足某些特定条件的情况下，FrFT 能够转换成小波变换的形式，并引入局部自相似性分析（LASS, local analysis of self-similarity）方法[13]对 FrFT 能量谱进行 Hurst 指数估计，结果表明，相较于现有的常用方法，FrFT LASS 法的估值更为精确，但该方法要求序列长度至少为 2 000 且估值结果易受 LASS 工具设置的时窗大小影响。

本文通过分析不同流量在 FrFT 域的统计特性，发现实际网络流量在 FrFT 域满足自相似性，在此基础上，针对网络流量数据在 FrFT 域上的“时域”和“频域”展开，分别给出了相应的 2 种 Hurst 指数估计方法，实验结果表明，与现有估计方法相比，FrFT 域上的“时域”估计法具有更高的估计精度，但其计算复杂度高；而 FrFT 域上的“频域”估计法则具有更优的估计顽健性和较低的计算复杂度，此外，该算法要求序列长度较短且不易受时间尺度变化的影响，使其可用于真实网络数据 Hurst 指数的实时在线估计。

2　网络流量在 FrFT 域的自相似特性

2.1　分数阶 Fourier 变换的相关定义

分数阶 Fourier 变换是对经典 Fourier 变换的推广，其变换的数学形式如下。

定义 1[9]　信号 $x(t)$的 a 阶分数阶 Fourier 变换定义为

$$X_a(u)=F_a(u)=\int_{-\infty}^{+\infty}x(t)K_a(t,u)\mathrm{d}t \tag{1}$$

FrFT 的变换核为

$$K_a(t,u)=\begin{cases}\sqrt{(1-\mathrm{i}\cot\alpha)}\mathrm{e}^{\mathrm{i}\pi(t^2\cot\alpha-2ut\csc\alpha+u^2\cot\alpha)}, & \alpha\neq n\pi\\ \delta(t-u), & \alpha=2n\pi\\ \delta(t+u), & \alpha=(2n\pm1)\pi\end{cases} \tag{2}$$

其中，n 为整数，即 $n\in\boldsymbol{Z}$。$\alpha=a\pi/2$，a 为分数阶 Fourier 变换的阶数，FrFT 简记为 $\boldsymbol{F}_a$。

2.2　自相似的相关定义

设 $X=\{X_t:t=0,1,2,\cdots\}$为一协方差平稳随机过程，即 X 具有恒定均值 $\mu=\mathrm{E}[X_t]$和有限方差 $\sigma^2=\mathrm{E}[(X_t-\mu)^2]$，其自相关函数 $r(k)=\mathrm{E}[(X_t-\mu)(X_{t+k}-\mu)]/\sigma^2$ $(k=0,1,2,\cdots)$仅与 k 有关。令 $X_k^{(m)}=(X_{km-m+1}+\cdots+X_{km})/m(k=1,2,3,\cdots)$ 为$\{X_t\}$的 m 阶聚集过程，并记时间序列 $X^{(m)}=(X_1^{(m)},X_2^{(m)},\cdots)$ 的自相关函数为 $r^{(m)}(k)(m=1,2,3,\cdots)$。

定义 2[1]　过程 X 被称为严格二阶自相似的且具有自相似指数 $H=1-\beta/2$，如果其 m 阶聚集过程 $X^{(m)}$ 具有与原过程 X 同样的相关函数，即 $r^{(m)}(k)=r(k)$ 对所有$(m=1,2,\cdots,k=1,2,\cdots)$成立。

定义 3[14]　一个自相似过程 $X(t)$，如果其 δ 增量函数 $Y(\delta,t)$存在并满足

$$\begin{aligned}\{Y(\delta,t)&:=Y_\delta(t)=X(t)-X(t-\delta),t\in R\}\\&\overset{\mathrm{d}}{=}X(\delta)-X(0)\},\forall\delta\end{aligned} \tag{3}$$

那么称这个过程为具有平稳增量 $Y(\delta,t)$的自相似过程，也称为 H-sssi 过程，其中，$0<H<1$。

通常在时间间隔 0~t 内的累积分组到达过程 $X(t)$即为一个 Hurst 指数为 H 的自相似过程，在时间单元 δ 内的分组到达时间序列 $X(t)$满足 H-sssi 过程的定义条件。

2.3　FrFT 域中网络流量的统计自相似性

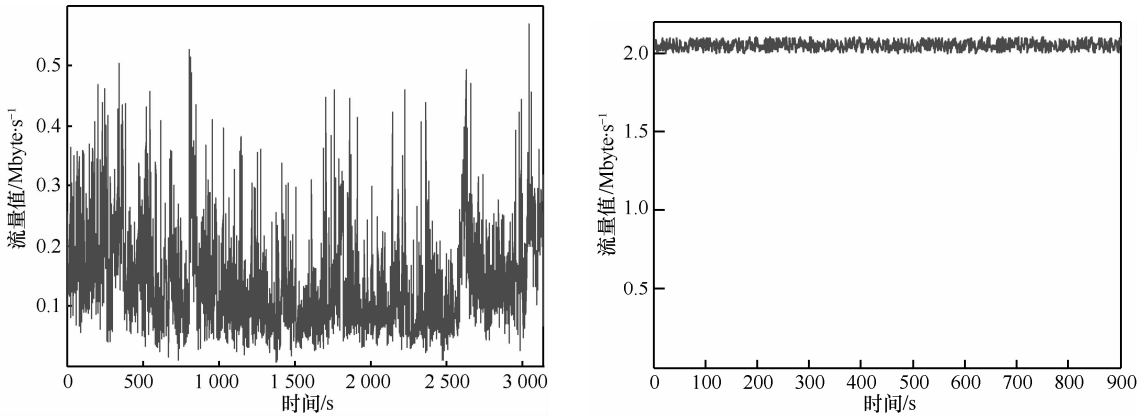
文献[15]研究发现，混合有噪声的信号 $x(t)$经分数阶 Fourier 变换后，能够实现信号与噪声的部分分离，尤其是当 FrFT 的时频“正交”（即变换阶数 a=0.5）时，分离效果最好。笔者将这一发现引入至网络流量分析领域，并提出如下假设。

假设 1　实际网络流量实质也是一种信号与噪声的混合体（异常流量即为其中的噪声），经由 a=0.5 的 FrFT 后，流量中异常部分将实现最大程度的滤除，变换后得到数据的自相似特性依然保持。

为了验证这一假设，本文采用 Bellcore 在 LAN 上采集的流量数据 BC_pAug89 [16]以及 MAWI 数据集上 2003 年 6 月 3 日的异常流量数据[17]作为原始数据，对它们在 FrFT 域上的统计特征分别进行分析。

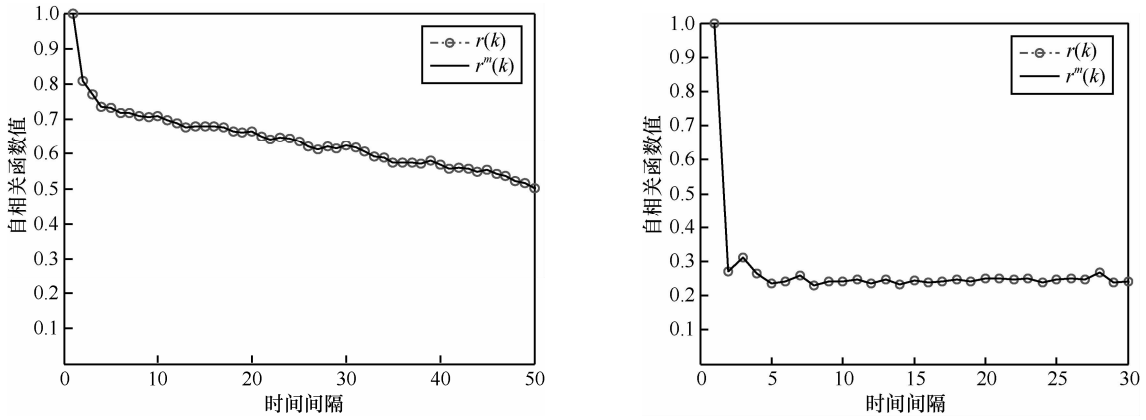
图 1(a)给出了 BC_pAug89 在 1 s 级时间间隔内的流量累积字节数据，其原始数据长度为 1×10^6，经累积后最终形成了 3 143 个数据点，将该段数据命名为 BC_pAug89(1)。图 1(b)给出了 2003 年 6 月 3 日 14:00~14:15 在 MAWI 主干链路上测得的 1 s 时间尺度上的字节计数时间序列，序列长度为 900，将其命名为 MAWI_20030603(1)。

计算上述两段数据在 FrFT 域（a=0.5）的统计均值及方差，如表 1 所示，可以看出，它们的均值基本不随时间的推移而变化，即均值恒定，且其方差均为有限值。图 2(a)和图 2(b)分别给出了这两段数据及其聚集过程 $X^{(m)}$在 FrFT 域的自相关函数 $r(k)$



(a) BC_pAug89(1)　　(b) MAWI_20030603(1)

图 1　原始数据在 1 s 级时间间隔内的流量累积字节数据



(a) BC_pOct89(1)及其聚集序列在 FrFT 域的自相关函数对比(k=50, m=60)

(b) MAWI_20030603(1) 及其聚集序列在 FrFT 域的自相关函数对比(k=30, m=30)

图 2　原始数据在 FrFT 域的自相关函数

表 1　　**不同数据集在 FrFT 域上的统计均值及方差**

原始数据	均值(byte·s^{-1})/μ_1	方差/σ_1^2	延迟 5 均值	延迟 10 均值	延迟 15 均值
BC_pAug89(1)	$6.439\,8\times10^5$	$1.303\,1\times10^{12}$	$6.439\,6\times10^5$	$6.439\,5\times10^5$	$6.439\,2\times10^5$
MAWI_20030603(1)	$2.125\,8\times10^6$	$1.479\,1\times10^{12}$	$2.125\,9\times10^6$	$2.125\,6\times10^6$	$2.125\,4\times10^6$

与 $r^{(m)}(k)$的对比，其对应的时间间隔 k 分别为 50 和 30，聚集阶数 m 为 60 和 30。从图 2 中可以看出，$r^{(m)}(k)= r(k)$。选取不同的聚集阶数 m 和时间间隔 k，并计算其对应的自相关函数，得到与图 2 类似的结果，在此不再赘述。

进一步地，为了定量地衡量数据 BC_pAug89(1)与 MAWI_20030603(1)的自相似特性，采用时域法中的绝对值法与 R/S 法、频域法中的周期图法与小波估计法以及 FrFT 域的 LASS 法[14]分别对这两段数据进行 Hurst 指数估计，估计结果如表 2 所示。从表 2 中可以看出，对于数据 BC_pAug89(1)，时域与频域法的 H 值估计结果在 0.834~0.912 之间波动，说明该段数据为正常网络流量，而 FrFT 域的 H 估值为 0.918，则说明正常网络流量经分数阶 Fourier 变换后其自相似特性保持不变；对于数据 MAWI_20030603(1)，时域与频域法的 H 估值结果均小于 0.5，可以判定 MAWI 主干链路在 2003 年 6 月 3 日 14:00~14:15 时间段发生异常，将该段数据按 a=0.5 变换至 FrFT 域后，得到的 H 估计值为 0.796，这与文献[2]中采用 Sketch 法得到 H 值为 0.8 的结果基本一致，说明分数阶 Fourier 变换具备与 Sketch 法同样的异常流量过滤功能。

综上，根据 2.2 节中的定义，可以近似地认为 FrFT 域网络流量数据为平稳随机过程，具有统计自相似性，这说明网络流量经分数阶 Fourier 变换后其自相似性并没有发生改变，可以对 FrFT 域的网络流量数据直接进行 Hurst 指数估计。同时，这也从网络流量的统计特性方面验证了假设 1 的正确性。

3　网络流量在 FrFT 域的 Hurst 指数估计方法

Hurst 指数的顽健估计算法是定量刻画网络流量自相似特性的关键，本文针对网络流量在 FrFT 域的“时域”展开，给出了基于改进的整体经验模态分解−去趋势波动分析（MEEMD-DFA）的 Hurst 指数估计方法；针对其“频域”展开，给出了基于加权最小二乘回归（WLSR）的 Hurst 指数自适应估计方法。

3.1　FrFT 域基于 MEEMD-DFA 的 Hurst 指数估计方法

将网络流量数据变换至 FrFT 域，即按式(1)进行变换，得到复数序列 $F_a(t)$(t 取整数)，计算该复数序列大小｜$F_a(t)$｜，笔者将其看作是网络流量在 FrFT 域的“时域”展开，并对序列的 Hurst 指数进行估计。

近年来，作为一种简单而有效的方法，DFA 法已被广泛地应用于分析时间序列的长相关特性中[18,19]。文献[20]提出将经验模态分解（EMD）作为一种去趋势工具来修正 DFA 法，实验结果表明，基于 EMD 的 DFA 方法在 Hurst 估值时比经典的 DFA 法具有更高的精度。然而，在 EMD 分解过程中由于信号极值点的不均匀分布经常会出现模态混叠问题。为消除模态混叠，Wu 等[21]提出了一种整体经验模态分解（EEMD）法以替代 EMD 方法，通过在分析信号中加入一定白噪声，EEMD 方法可以自动消除存在的模态混叠问题。但是，如何选择合适的白噪声幅度并确定整体实验的次数仍然是一个需要进一步研究的问题。对此，文献[22]提出了一种改进的 EEMD 方法，实例测试说明该方法在消除模态混叠和计算代价方面具有比 EEMD 法更优的效果。本文在借鉴文献[20,22]思想的基础上，提出了一种 MEEMD-DFA 方法对网络流量在 FrFT 域的“时域”展开进行 Hurst 指数估计。

3.1.1　DFA 算法

对于长度为 N 的时间序列$\{X_k\}$，k=1,2,…,N，DFA 过程包括以下 5 个步骤[18]。

Step1　构造序列 X_k 的距平累加值 $Y(i)$为

$$Y(i)=\sum_{k=1}^{N}[X_k-\langle X\rangle] \tag{4}$$

其中，〈X〉为原序列$\{X_k\}$的均值。

Step2　将新序列 $Y(i)$划分为长度为 m 的不重叠等长度子区间，长度为 N 的序列共被分为 N_m=[N/m]。将每个子区间标记为 u_v，对于 1≤i≤m，有 $u_v(i)$=$u(l+i)$，其中，l=(v−1)m。

表 2　不同方法对 2 个累积流量字节序列的 Hurst 指数估计

原始数据	时域法		频域法		FrFT 域
	Abs	R/S	Periodogram	Wavelet	FrFT LASS
BC_pAug89(1)	0.848 8	0.834 1	0.843 0	0.912	0.918
MAWI_20030603(1)	0.413 2	0.488 3	0.419 4	0.416	0.796

Step3 对每个子区间 u_v 的数据进行多项式回归拟合，得到局部趋势函数 $\tilde{u}_v(i)$，计算消除趋势序列为

$$\varepsilon_v(i)=u_v(i)-\tilde{u}_v(i) \tag{5}$$

Step4 每个子区间 u_v 上的消除趋势序列的平均值为

$$F^2(v,m)=\frac{1}{m}\sum_{i=1}^{m}[\varepsilon_v(i)]^2 \tag{6}$$

计算整个原始序列共 N_m 个子区间的 $F^2(v,m)$的均值平方根为

$$F(m)=\left[\frac{1}{N_m}\sum_{v=1}^{N_m}F^2(v,m)\right]^{1/2}\propto m^{\mathrm{H}} \tag{7}$$

Step5 取 F_m 与 m 的对数的函数关系图，求 Hurst 指数为

$$H=\frac{\mathrm{lb}F(m)}{\mathrm{lb}m}=\frac{\mathrm{lb}\left[\frac{1}{N_m}\sum_{v=1}^{N_m}F^2(v,m)\right]^{1/2}}{\mathrm{lb}m} \tag{8}$$

3.1.2 改进的 EEMD 算法

改进的 EEMD 算法主要由 3 部分组成[22]。

1) 对目标数据 $x(t)$进行 EEMD 分解，得到低频成分 $x_1(t)$和高频弱瞬时成分 $x_2(t)$。选择皮尔逊相关系数（PCC, pearson's correlation coefficient）作为评估在不同的白噪声幅值下 $x_1(t)$与 $x_2(t)$的相关性参数。PCC 定义为

$$r=\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{s_X}\right)\left(\frac{Y_i-\bar{Y}}{s_Y}\right) \tag{9}$$

其中，X 和 Y 分别代表 $x_1(t)$与 $x_2(t)$，$\frac{X_i-\bar{X}}{s_X}$、$\bar{X}$ 和 s_X 分别为标准化变量、样本均值和样本标准差。

2) 确定加入的最佳白噪声幅度，该幅值位于 $x_1(t)$与 $x_2(t)$平均功率的平方根之间。

3) 将合适幅度的白噪声 $w_m(t)$,m=1,2,⋯,N 加入到目标数据 $x(t)$中，再次执行 EEMD 分解。最终得到的分解结果为

$$x(t)=\frac{1}{N}\sum_{i=1}^{L}\sum_{m=1}^{N}c_{m,i}(t)+\frac{1}{N}\sum_{m=1}^{N}r_{m,L}(t) \tag{10}$$

其中，$c_{m,i}(t)$为第 m 次实验时的第 i 个本征模态函数（IMF），$r_{m,L}(t)$为第 m 次实验时的残差序列，L 为由 EMD 方法分解得到的 IMF 个数，N 为 EEMD 方法的整体实验次数。在本文中，N 设置为 100。

3.1.3 基于 MEEMD 的 DFA 算法

将 MEEMD 算法插入至 DFA 算法中，用以对 DFA 算法的 Step3 进行改进，算法的其他步骤保持不变。

Step3′ 对每个子区间 u_v，得到局部趋势函数 $\tilde{u}_v= r_n(i)$，则可以获得残差序列为

$$\varepsilon_v(i)=\tilde{u}_v(i)-r_n(i),1\leqslant i\leqslant m \tag{11}$$

需要注意的是，趋势 $r_n(i)$应由不同时间尺度上的子区间分别确定。

上述给出了基于 MEEMD 的 DFA 估计方法。

3.2 FrFT 域基于 WLSR 的 Hurst 指数自适应估计方法

将网络流量数据变换至 FrFT 域，得到复数序列 $F_a(t)$(t 取整数)，设其对应的能量谱为 E[$g^2(j)$]，将其看作是网络流量在 FrFT 域的“频域”展开。文献[13]得到 E[$g^2(j)$]的对数尺度与 Hurst 指数间关系满足以下条件

$$G(j)\leftrightarrow(2H-1)j+constant \tag{12}$$

其中，j 为二进制尺度系数，*constant* 表示常数。

然而，文献[13]中的 FrFT LASS 法却不能较好地实现尺度区间的自动选择，为此，本文借鉴文献[23,24]的思想，提出了一种基于 WLSR 的自适应参数估计法对网络流量在 FrFT 域的“频域”展开进行 Hurst 指数估计。

3.2.1 基于 WLSR 的自适应参数估计法

对于随机变量 $Y_j,x_j,j=1,\cdots,J$，给定回归模型为 $\tilde{Y}_j=\beta_0+\beta_1\tilde{x}_j+\tilde{\varepsilon}_j$，其中，$\mathrm{E}(\tilde{\varepsilon}_j)=0,\mathrm{Var}(\tilde{\varepsilon}_j)=\sigma^2/N_j$。用矩阵形式表示，该式可以写成

$$\tilde{\boldsymbol{y}}=\tilde{\boldsymbol{X}}\boldsymbol{\beta}+\tilde{\boldsymbol{\varepsilon}} \tag{13}$$

其中，$\tilde{\boldsymbol{X}}$ 是一个 $J\times2$ 矩阵。令 $\boldsymbol{W}=\mathrm{diag}\{w_1,\cdots,w_J\}$，$w_j=N_j$，定义变量：$\boldsymbol{y}=\boldsymbol{W}^{1/2}\tilde{\boldsymbol{y}},\boldsymbol{X}=\boldsymbol{W}^{1/2}\tilde{\boldsymbol{X}},\boldsymbol{\varepsilon}=\boldsymbol{W}^{1/2}\tilde{\boldsymbol{\varepsilon}}$，则有

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon} \tag{14}$$

其中，$\boldsymbol{\varepsilon}\sim(0,\sigma^2\boldsymbol{I})$。因此，式(13)所表示的加权最小二乘估计模型就等同于式(14)的一般最小二乘估计模型。

令 $\boldsymbol{b}$ 为式(13)中 $\boldsymbol{\beta}$ 的最小二乘估计，则 $\boldsymbol{b}$ 的无偏估计 $\hat{\boldsymbol{b}}$ 为

$$\hat{\boldsymbol{b}}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}=\boldsymbol{H}\boldsymbol{y} \tag{15}$$

其中，$\boldsymbol{H}$ 称为“帽子”矩阵，h_{ij} 为其第(i,j)个元素。

指定 e_j 为第 j 项残差，此时，

$$\hat{e}_j = Y_j - \hat{Y}_j = Y_j - x_j^{\mathrm{T}}\,\hat{\boldsymbol{b}} \tag{16}$$

其中，$\hat{Y}_j$ 为 Y_j 的回归值。

由式(12)可知尺度区间 $[j_1, j_2]$ 的选择对拟合结果有很大的影响，本文在借鉴回归诊断[25]和变点分析[26~28]思想的基础上，采用方差分析进行拟合优度评估，分析不同尺度区间对 Hurst 指数估计的拟合程度，从而自适应地选择最优的尺度区间。

定理 1[29]　$\boldsymbol{x}_j$ 表示 X 的第 j 行向量，$\boldsymbol{b}(j)$表示式(14)中第 j 行元素被移除后 $\boldsymbol{\beta}$ 的最小二乘估计。$\boldsymbol{b}$ 和 $\boldsymbol{b}(j)$间的关系可表示为

$$\boldsymbol{b} - \boldsymbol{b}(j) = \frac{\left(X^{\mathrm{T}}X\right)^{-1} x_j e_j}{1 - h_{jj}} \tag{17}$$

该数量关系被用作一种回归诊断工具，用于检测第 j 个观测量是否会对回归模型的估计结果有影响。

定理 2[29]　定义残差平方和：$SSR = \sum_i (Y_i - \hat{Y}_i)^2$；令 s^2 为 σ^2 的一般估计，$s^2(j)$ 为式(14)中第 j 行元素被移除后 σ^2 的估计，定义 $s^2 = \dfrac{SSR}{n-1} = \dfrac{1}{n-1}\cdot\sum_{i=1}^{J}(Y_i - x_i^{\mathrm{T}}b)^2$，类似地，定义 $s^2(j) = \dfrac{1}{(n-1)-1}\cdot\sum_{i\neq j}^{J}\{Y_i - x_i^{\mathrm{T}}b(j)\}^2$，则 s^2 和 $s^2(j)$ 满足

$$(n-2)s^2(j) = (n-1)s^2 - \frac{e_j^2}{1-h_{jj}} \tag{18}$$

定理 1 和定理 2 构成了尺度区间选择时检验统计量的基础。可将选择尺度区间这一问题看作从回归模型中选取一个展示强大线性的子模型。因此，对于每个索引标记为 j 的子模型，该问题可归结为检验假设。

$$H_0(j): EY = constant$$
$$H_1(j): EY = linear \tag{19}$$

令 $SSR_j(\mathrm{old})$ 表示零假设的残差平方和，$SSR_j(\mathrm{new})$ 表示备择假设的残差平方和，利用统计量

$$T(j) = \frac{(SSR_j(\mathrm{old}) - SSR_j(\mathrm{new}))/1}{SSR_j(\mathrm{new})/\left(N(j)-2\right)} = \frac{\sum_{i=j}^{J}(\bar{Y} - \hat{Y}_i)^2}{\sum_{i=j}^{J}(Y_i - \hat{Y}_i)^2/\left(N(j)-2\right)} \sim F(1, N(j)-2)$$

来进行最优度检验，得到最优尺度区间 $[j_1, j_2]$，其中，$j_1 = \arg\min_{j\geq 1}\{T(j)\}$，$j_2 = J$。

3.2.2　算法的实现过程

下面给出基于 FrFT 的 Hurst 指数自适应估计方法，具体估计步骤如下。

输入参数：原始数据 $X[n]$，分数阶 Fourier 变换阶数 a。

Step1　利用 FFT 实现原始数据的分数阶 Fourier 变换。

Step2　计算实现 FrFT 后原始数据的能量谱 $\mathrm{E}[g^2(j)]$。

Step3　根据 $G(j)$ 与能量谱的对数关系获得 $G(j)$。

Step4　对不同的尺度区间进行方差拟合度检验，得到最优的尺度区间 $[j_1, j_2]$。

Step5　根据最优尺度区间进行参数估计，应用式(12)和式(15)。

Step6　计算 Hurst 指数估计值。

在上述方法中，对 Hurst 指数的估计是无偏的。

4　仿真结果与分析

本文采用 FGN 序列以及 Bellcore 在 LAN 上采集的网络流量数据[16]作为原始数据，估计它们的 Hurst 指数。在仿真试验中，笔者采用 Michigan 大学 Stilian Stoev 教授提供的基于 FFT 的 FGN 序列生成算法[30]产生 H 值在 0.55~0.95（间隔为 0.05）之间，长度 N=2^{16}=65 536 的 FGN 样本序列（每一个 H 值重复实现 100 次）。采用 R/S 分析法、周期图法、小波估计法、EMD-DFA 法[21]、FrFT LASS 法[13]以及本文提出的 FrFT 域的 2 种方法对各 H 值对应的不同 FGN 序列分别进行 Hurst 指数估计，并将其平均值作为各估值算法对应于每个 H 值的最终估计结果。

4.1　顽健性分析

Hurst 指数估计易受网络流量序列中存在的周期成分、采样噪声、趋势成分(表示该过程均值是稳定递增或递减的)[31]等因素的影响，这些因素都有可能导致算法误判。为了有效评估算法的估计顽健性，笔者通过各估值算法分别对以下 4 类数据进行 Hurst 指数估计，并对估计结果进行综合比较与分析。

1) 已知参数的 FGN 序列。

2) FGN 序列与余弦周期信号组成的混合序列。

表 3　　不同方法对 FGN 序列的 Hurst 指数估计结果

H	Hurst 指数估计方法							$[j_1, j_2]$
	R/S	Periodogram	Wavelet	EMD-DFA	FrFT LASS	MEEMD-DFA	FrFT Adaptive	
0.55	0.556	0.523	0.566	0.534	0.584	0.556	0.552	[7,16]
0.60	0.614	0.581	0.607	0.582	0.628	0.586	0.652	[5,16]
0.65	0.659	0.647	0.664	0.669	0.661	0.640	0.652	[5,16]
0.70	0.698	0.701	0.695	0.721	0.682	0.677	0.698	[6,16]
0.75	0.774	0.739	0.756	0.772	0.732	0.741	0.750	[6,16]
0.80	0.799	0.815	0.806	0.824	0.809	0.782	0.805	[5,16]
0.85	0.825	0.849	0.878	0.875	0.835	0.857	0.847	[7,16]
0.90	0.853	0.884	0.942	0.927	0.873	0.909	0.895	[7,16]
0.95	0.907	0.965	0.984	0.978	0.939	0.960	0.953	[5,16]

3) FGN 序列与高斯白噪声组成的混合序列。

4) FGN 序列与趋势成分组成的混合序列。

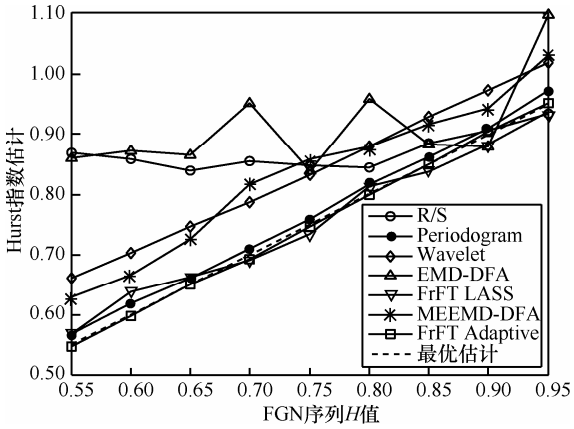
4.1.1　对已知参数的 FGN 序列的估计

对生成的已知 H 值的 FGN 样本序列进行估计，表 3 给出了各算法的估计结果，同时还给出了 FrFT 自适应估计法的最优尺度区间$[j_1, j_2]$，由于样本序列长度 $N=2^{16}=65\,536$，因此自适应方法中最大尺度阶数 J 为 16。从表 3 中可以看出，整体而言，本文提出的 MEEMD-DFA 法与自适应估计法的估计值与 FGN 序列 H 值比较吻合，周期图法与 FrFT LASS 法次之，小波估计法与 EMD-DFA 法会对 Hurst 指数高估，而 R/S 分析法在 Hurst 指数大于 0.8 时会产生较大偏差。

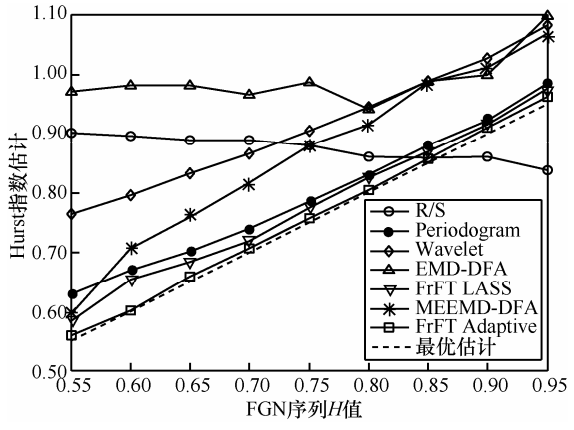
4.1.2　对 FGN 序列与余弦周期信号混合序列的估计

为了考察周期性对算法的影响，采用各估值算法对已知 H 值的 FGN 样本序列与余弦周期信号组成的混合序列进行 Hurst 指数估计，其中周期信号为 $A\cos(0.05x)$。改变周期信号的幅值 A，图 3(a)和图 3(b)分别是 A 为 1 和 5 时的 Hurst 指数估计结果，估计置信度为 95%。最优估计表示对 Hurst 指数的理想估计结果，即其估计值与已知 Hurst 指数完全吻合。

通过比较发现：周期信号对 R/S 法、EMD-DFA 法、小波估计法与本文提出的 MEEMD-DFA 估计法的影响较大，而对周期图法、FrFT LASS 法以及本文的 FrFT 自适应估计法则受到的影响较小，且自适应估计法要具有更高的估计精度。



(a) FGN+ cos(0.05x)



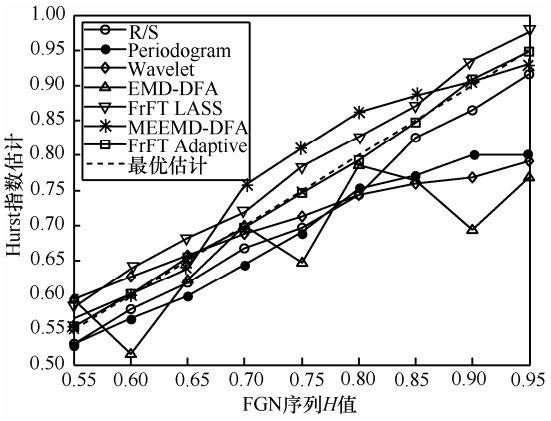
(b) FGN+ 5cos(0.05x)

图 3　各算法对 FGN 序列与余弦信号混合序列的 Hurst 指数估计

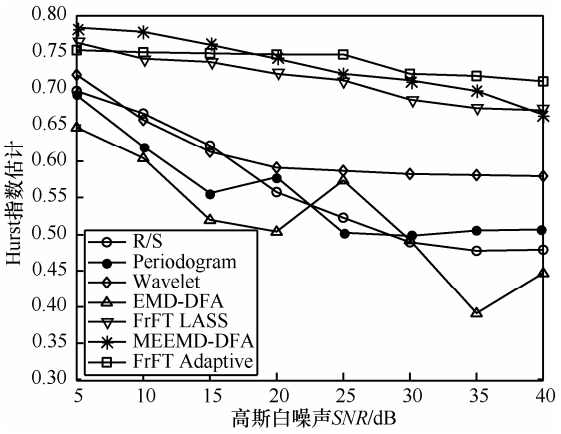
4.1.3　对 FGN 序列与高斯白噪声的混合序列的估计

为了考察采样噪声对算法 Hurst 指数估计准确性的影响，分别用上述 7 种估计算法对已知 H 值的

FGN 样本序列与高斯白噪声组成的混合序列进行 Hurst 指数估计。图 4(a)是对信噪比（*SNR*）为 5 dB 的高斯白噪声与 *H* 值在 0.55~0.95 范围内变化的 FGN 序列组成的混合序列的 Hurst 指数估计情况，图 4(b)所示为对 *H* 值是 0.75 的 FGN 序列与信噪比（*SNR*）在 5 dB~40 dB 范围内变化的高斯白噪声组成的混合序列的 Hurst 指数估计结果，其中，估计置信度为 95%。



(a) FGN 序列 *H* 值变化时



(b) 高斯白噪声信噪比变化时

图 4　各算法对 FGN 序列与高斯白噪声的混合序列的 Hurst 指数估计

从图 4 中可以看出，在 *SNR* 保持不变时，除自适应估计法外，其他 6 种方法对 Hurst 的估计都会产生较大的误差，随着原 FGN 序列 *H* 值的不断增大，周期图法、EMD-DFA 法与小波估计法的估计误差还会增大，并且出现低估的现象。在 FGN 序列的 *H* 值固定时，随着 *SNR* 的不断增大，各算法的估计值也会出现不同程度的减小，相较于其他方法而言，自适应估计法与实际 *H* 值的偏差最小。

4.1.4　对 FGN 序列与趋势成分混合序列的估计

为了研究非平稳性对估计算法的影响，为此笔者采用 FGN 样本序列与各种不同的衰减或渐增趋势项进行叠加，并用 7 种估计算法对混合序列进行 Hurst 指数估计。为了保证趋势项影响不失一般性，分别使用平滑多项式趋势项和高频振荡趋势项[29]与 FGN 序列合成组合序列，并进行 Hurst 指数估计，对得到的 2 组估计 *H* 值进行平均，获得的估计结果如图 5 所示。

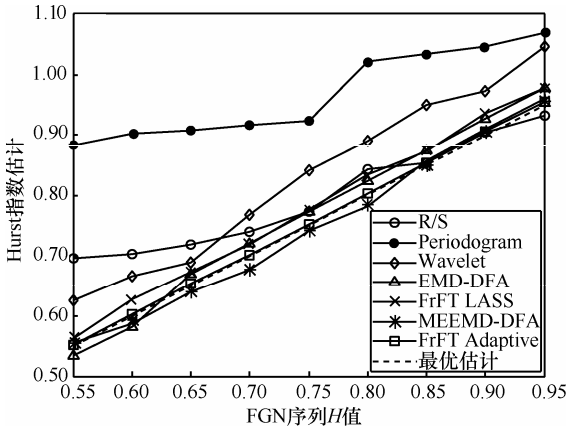


图 5　各算法对 FGN 序列叠加趋势项的混合序列的 Hurst 指数估计

从图 5 中的变化曲线可知，当原 FGN 序列的 *H* 值大于 0.8 时，周期图法的 Hurst 指数估计值均要大于 1，而自相似性的定义则是 *H*<1；随着 Hurst 指数的不断增大，小波估计法与 FrFT LASS 法的估计误差在不断增大，R/S 法的估计值虽然逐渐接近于实际 *H* 值，但其整体误差仍较大；而 EMD-DFA 法、MEEMD-DFA 估计法与自适应估计法的估计结果均稳定地保持在实际 *H* 值附近。

4.1.5　顽健性定量比较

为了更准确地量化各估值算法的顽健性，笔者采用 *S* 来计算不同估计结果的标准误差，*S* 定义为

$$S=\sqrt{\frac{\sum_{i=1}^{n}\left(H_{\mathrm{E}}-H_{\mathrm{T}}\right)^{2}}{n-1}} \tag{20}$$

其中，H_{T} 为 Hurst 指数的真实值，H_{E} 为 Hurst 指数的估计值。

表 4 列出了各种算法对上述 4 类数据（FGN 序列 *H* 值的范围为 0.55~0.95）估计值的标准误差。从表中可以更直观地看到，对于已知参数的 FGN 序列和 FGN 序列与趋势项的混合序列，本文提出的 2 种估计方法会得到最准确的估计值；而对于 FGN 序列与余弦周期信号的混合序列和 FGN 序列与高斯白噪声的混合序列，FrFT LASS 法与 FrFT 自适应估计法会得到最准确的估计值。综合来讲，

FrFT 域的 LASS 法、MEEMD-DFA 法以及自适应估计法具有比传统时域或频域估计法更强的顽健性；LASS 法与自适应估计法的估计精度要高于 MEEMD-DFA 法，其中又以 FrFT 自适应估计法的估计性能最优。这说明：在分数阶 Fourier 变换域中，“频域”法的估计精度仍然要好于“时域”估计法。相比于现有的估值算法，FrFT 域中基于 WLSR 的 Hurst 指数自适应估计法要具有更令人满意的整体估计性能，能够有效克服网络流量序列中存在的周期性、采样噪声以及趋势成分等干扰因素的影响。

表 4 各估值算法对 4 类数据的 Hurst 指数估计的顽健性比较

Hurst 指数估计方法	FGN	FGN+余弦周期信号	FGN+高斯白噪声	FGN+趋势项
R/S	0.026 4	0.205 2	0.038 6	0.071 4
Periodogram	0.015 5	0.049 9	0.080 4	0.241 2
Wavelet	0.023 2	0.174 2	0.085 2	0.084 6
EMD-DFA	0.023 9	0.280 6	0.114 1	0.023 9
FrFT LASS	0.022 0	0.031 3	0.032 0	0.028 1
MEEMD-DFA	0.013 7	0.119 8	0.039 7	0.013 7
FrFT Adaptive	0.003 5	0.008 3	0.005 1	0.003 4

4.2 算法的计算复杂度分析

由于实现原理和实现方法不同，对于相同长度的序列，各种算法计算速度不同。从原理上分析各种算法的计算复杂度，表 5 给出了上述 7 种算法的计算复杂度。其中，n 表示长度为 N 的序列的二进制尺度大小，即 $n=\mathrm{lb}(N)$，m 为改进的 EEMD 算法的整体实验次数。

通过比较可以看出，R/S 法是最普遍使用的方法，但速度较慢；小波估计法虽然速度较快，但实现较为复杂；周期图法也表现出比较快的计算速度，但其稳定性较差；FrFT LASS 法虽然具有较好的估计顽健性，但计算速度还较慢；本文在 EMD-DFA 法基础上提出的 MEEMD-DFA 法的估计精度虽然得到了提高，但计算复杂度较高；相比于上述算法，本文提出的 FrFT 自适应方法计算复杂度较低，且能够获得最优的估计精度，因此，综合考虑算法在执行速度与估计顽健性方面的优势，FrFT 自适应法可作为一种在线估计 Hurst 指数的方法。

表 5 各估值方法的计算复杂度

Hurst 指数估计方法	计算复杂度
R/S	$O(N^2)$
Periodogram	$O(N\log N)$
Wavelet	$O(N\log N)$
EMD-DFA	$O(N^2)$
FrFT LASS	$O(nN\log N)$
MEEMD-DFA	$O(mN^2)$
FrFT Adaptive	$O(N\log N)$

4.3 实际网络流量分析

为了检验本文方法对真实网络流量数据 Hurst 指数估计的准确性，选取 Bellcore 测得的 4 组数据（BC_pAug89.TL、BC_pOct89.TL、BC_Oct89Ext.TL 以及 BC_Oct89Ext4.TL），将它们分别处理到 1 s、5 s 和 10 s 的时间尺度上，利用本文提出的 FrFT 域的“时域”和“频域”估计算法以及小波估计法分别估计其 H 值，估计结果如表 6 所示。可以看出，FrFT 自适应估计法在相同流量不同时间尺度下的估计值相差很小，这说明 FrFT 自适应估计法不易受时间尺度变化的影响；而小波估计法与 MEEMD- DFA 法的估计值则相差较大，尤其是对于数据 BC_pAug89.TL 和 BC_pOct89.TL 在 10 s 级时间尺度上的累积序列，MEEMD-DFA 法的 Hurst 指数估计值均大于 1，分析发现这 2 个累积序列的数据长度分别为 315 和 176，可见 MEEMD-DFA 算法要求序列较长，同时这也说明 FrFT 自适应估计法在序列长度较小时就可以得到较为准确的估计结果。

表 6 4 组实际网络流量数据的估计值

流量数据	尺度/s	Wavelet	MEEMD-DFA	FrFT Adaptive
BC_pAug89.TL	1	0.834	0.847	0.848
	5	0.855	0.858	0.854
	10	0.887	1.026	0.858
BC_pOct89.TL	1	0.804	0.926	0.947
	5	0.877	0.931	0.952
	10	0.963	1.061	0.957
BC_Oct89Ext.TL	1	0.905	0.886	0.917
	5	0.921	0.907	0.921
	10	0.954	0.919	0.926
BC_Oct89Ext4.TL	1	0.861	0.913	0.942
	5	0.909	0.927	0.946
	10	0.965	0.976	0.948

5　结束语

本文分析了网络流量数据在 FrFT 域的自相似特性，并对表征网络流量自相似程度的 Hurst 指数估计方法进行了研究。针对网络流量在 FrFT 域的“时域”展开，给出了基于 MEEMD-DFA 的估计方法；针对网络流量在 FrFT 域的“频域”展开，给出了基于 WLSR 的自适应估计方法。仿真结果表明，相比于现有估值算法，MEEMD-DFA 法具有较高的估计精度，但其计算复杂度高；而自适应估计法则具有更强的估计顽健性，能够有效克服网络流量序列中存在的周期性、采样噪声以及趋势成分等干扰因素的影响，不易受时间尺度变化影响且计算复杂度较低，可作为一种实时在线估计真实网络数据 Hurst 指数的方法。进一步的研究工作是如何将分数阶 Fourier 变换用于网络流量分形特性研究和异常检测等方面中。

参考文献：

[1] LELAND W E, TAQQU M S, WILLINGER W, *et al*. On the self-similar nature of Ethernet traffic (extended version)[J]. IEEE/ACM Trans on Networking, 1994, 2(1):1-15.

[2] BORGNAT P, DAWAELE G, FUKUDA K, *et al.* Cho seven years and one day: sketching the evolution of Internet traffic[A]. Proc of the 28th Conf on Computer Communications (INFOCOM 2009)[C]. Rio de Janeiro: IEEE, 2009. 711-719.

[3] GUPTA H, RIBEIRO V J, MAHANTI A. A longitudinal study of small-time scaling behavior of Internet traffic[A]. Networking 2010 Springer Berlin Heidelberg[C]. 2010. 83-95.

[4] CIFLIKLI C, GEZER A, ÖZSAHIN A T, *et al*. Bittorrent packet traffic features over IPv6 and IPv4[J]. Simulation Modelling Practice and Theory, 2010, 18(9):1124-1214.

[5] CIFLIKLI C, GEZER A, ÖZSAHIN A T, *et al*. Packet traffic features of IPv6 and IPv4 protocol traffic[J]. Turkish Journal of Electrical Engineering and Computer Sciences, 2012, 20(5):727-749.

[6] BĂRULESCU A, SERBAN C, MAFTEL C. Evaluation of Hurst exponent for precipitation time series[A]. Proceedings of the 14th WSEAS CSCC Multiconference[C]. 2010. 590-595.

[7] 张宾, 杨家海, 吴建平. Internet 流量模型分析与评述[J]. 软件学报, 2011, 22(1):115-131.
ZHANG B, YANG J H, WU J P. Survey and analysis on the Internet traffic model[J]. Journal of Software, 2011, 22(1):115-131.

[8] REA W, OXLEY L, REALE M, *et al*. Estimators for long range dependence: an empirical study[J]. Electronic Journal of Statistics, 2009, 3:1-16.

[9] 陶然, 邓兵, 王越. 分数阶Fourier变换在信号处理领域的研究进展[J]. 中国科学, 2006, 36(2):113-136.
TAO R, DENG B, WANG Y. The research evolution of fractional order Fourier transform in signal processing domain[J]. Science in China Ser E Information Sciences, 2006, 36(2):113-136.

[10] CIFLIKLI C, GEZER A. Self similarity analysis via fractional Fourier transform[J]. Simulation Modeling Practice and Theory, 2011, 19(3): 986-995.

[11] SUN R, CHEN Y, ZAVERI N, *et al*. Local analysis of long-range dependence based on fractional Fourier transform[A]. Proceedings of the 2006 IEEE Mountain Workshop on Adaptive and Learning Systems[C]. 2006. 13-18.

[12] CHEN Y, SUN R, ZHOU A. An improved Hurst parameter estimator based on fractional Fourier transform[J]. Telecommun System, 2010, 43(3-4):197-206.

[13] STOEV S, TAQQU M S, PARK C, *et al*. LASS: a tool for the local analysis of self-similarity[J]. Computational Statistics & Data Analysis, 2006, 50(1):2447-2471.

[14] BREGNI S, PRIMERANO L. The modified allan variance as time-domain analysis tool for estimating the hurst parameter of long-range dependent traffic[A]. IEEE GLOBECOM '04 Global Telecommunications Conference[C]. 2004. 1406-1410.

[15] WANG Y G, YU H L, LIANG X Y. Time delay model of fractional Fourier transform and the application in signal filtering[J]. Applied Mechanics and Materials, 2012, 121:3637-3641.

[16] LAB B. Traces available in the internet traffic archive[EB/OL]. http://ita.ee.lbl.gov/html/traces.html, 1989.

[17] MAWI working group traffic archive[EB/OL]. http://tracer.csl.sony.co.jp/mawi, 2011.

[18] BARDET J M, KAMMOUN I. Asymptotic properties of the detrended fluctuation analysis of long range dependence processes[J]. IEEE Transactions on Information Theory, 2008, 54(5):2041-2052.

[19] WERON R. Estimation long-range dependence: finite sample properties and confidence intervals[J]. Physica A:Statistical Mechanics and Its Applications, 2002, 312(1):285-299.

[20] QIAN X Y, ZHOU W X, GU G F. Modified detrended fluctuation analysis based on empirical mode decomposition[J]. Physica A: Statistical Mechanics and Its Applications, 2011, 390(23):4388-4395.

[21] WU Z, HUANG N E. Ensemble empirical mode decomposition: a noise assisted data analysis method[J]. Advances in Adaptive Data Analysis, 2009, 1(1):1-41.

[22] LIN J. Improved Ensemble Empirical Mode Decomposition Method and Its Simulation[M]. Advances in Intelligent Systems, 2012.

[23] TARRIO P, BERNARDOS A M, CASAR J R. Weighted least squares techniques for improved received signal strength based localization[J]. Sensors, 2011, 11(9):8569-8592.

[24] YARDIBI T, LI J, STOICA P, *el al*. Source localization and sensing: a nonparametric iterative adaptive approach based on weighted least squares[J]. IEEE Transactions on Aerospace and Electaonic Systems, 2010, 46(1):425-433.

[25] BELSLEY D A, KUH E, WELSCH R E. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity[M]. 2005.

[26] KUO C C, CHUI L C, CHUNG B L. The comparison of algorithms in change-points problem[J]. Journal of Applied Science and Engineering, 2012, 15(1):11-19.

[27] BERTRAND P R, FHIMA M, GUILIN A. Fast change point analysis on the Hurst index of piecewise fractional brownian motion[J]. Journée de Statistiques 2011 (JDS 2011), 2011,(3):1-6.

[28] HUAT N K, MIDI H. Change point detection with robust control chart[J]. Mathematical Problems in Engineering, 2011. 2011:1-20.

[29] SHENG H, CHEN Y Q, QIU T. On the robustness of Hurst estimators[J]. IET Signal Process, 2011, 5(2):209-225.

[30] Stilian stoev's new web page[EB/OL]. http://www.stat.lsa.umich.edu/~sstoev, 2011.

[31] PARK J, PARK C. Robust estimation of the Hurst parameter and selection of an onset scaling[J]. Statistica Sinica, 2009, 19(4):1531- 1555.

作者简介：

郭通（1984-），男，江西南昌人，国家数字交换系统工程技术研究中心博士生，主要研究方向为宽带信息网络、高速网络业务管控等。

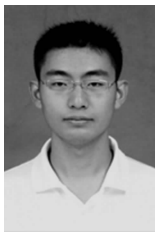
兰巨龙（1962-），男，河北张北人，博士，国家数字交换系统工程技术研究中心教授、博士生导师，主要研究方向为宽带信息网络、高速路由器核心技术等。

黄万伟（1979-），男，江苏盐城人，博士，国家数字交换系统工程技术研究中心讲师，主要研究方向为实时任务调度、可重构计算。

张震（1985-），男，山东济宁人，国家数字交换系统工程技术研究中心博士生，主要研究方向为宽带信息网络、高速网络业务识别等。

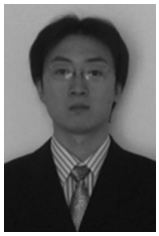
（上接第 37 页）

作者简介：



崔宇（1985-），男，黑龙江哈尔滨人，哈尔滨工业大学博士生，主要研究方向为IPv6、网络安全。

张宏莉（1973-），女，吉林榆树人，哈尔滨工业大学教授、博士生导师，主要研究方向为计算机网络信息安全、并行处理。



田志宏（1978-），男，黑龙江哈尔滨人，博士，哈尔滨工业大学副研究员，主要研究方向为网络安全主动防御、入侵取证。

方滨兴（1960-），男，江西万年人，哈尔滨工业大学教授、博士生导师，主要研究方向为计算机体系结构、信息安全和计算机网络。