The Second International Symposium on Optimization and Systems Biology (OSB'08)
Lijiang, China, October 31– November 3, 2008
Copyright © 2008 ORSC & APORC, pp. 381–388

Simpson's rule based FFT method to compute densities of stable distribution

Li Wang[1] Ji-Hong Zhang[2,*]

[1]School of Economics, RenMin University of China, Beijing 100872, China.
[2]School of International Business, Beijing Foreign Studies University, Beijing 100089, China.

Abstract In recent years, more and more kinds of heavy-tailed distributions are used to model the distribution of microarray gene expression data. Stable distribution is an important type of heavy-tailed distributions. However, lack of closed-form density function blocks its application. In this paper, we derive the Simpson's rule based FFT method for computing the density of stable distribution and compare its accuracy with S. Mittnik's rectangle rule based FFT method. Results show that great improvement can be made using Simpson's rule.

Keywords Stable distribution; Simpson's rule; Fast Fourier Transformation(FFT).

1 Introduction

Large-scale gene expression data sets have been analyzed in order to identify the probability distribution of gene expression levels [1]-[4]. Determining such functions may provide a theoretical basis for accurately counting all expressed genes in a given cell and for understanding gene expression control. Gene expression distribution has been modeled using several densities: cauchy distribution [1], pareto distribution [2], t-student distribution [3] and log-normal distribution [4]. All kinds of these densities are heavy-tailed and have some certain similarities with stable distribution: cauchy density is a particular case of stable distribution, stable distribution show the same paretian tail behavior as pareto density, t-student and log-normal density are all heavy-tailed. However, the critical difference between stable distribution and these distributions is the infinite variance of the stable distribution, which is consistent with the fact that the variance of any given array increases concomitantly with an increase in the number of genes studied.

Stable distribution is first put forward by Lěvy in 1920s, it is an extension of normal distribution, permitting pareto-like heavy tail, possible skewness and stability under addition. These properties make it an useful candidate for modeling a variety of data which exhibit such characteristics[5]-[7]. While the main drawback which prevents its widely use is the lack of closed-form density function. Stable distribution is defined by characteristic function and has many different forms, we use that of Samorodnitsky and Taqqu [8]:

$$\varphi(t) = \exp(i\mu t - |ct|^{\alpha}(1 - i\beta\frac{t}{|t|}\omega(|t|,\alpha))),$$

[*] Corresponding author. Email: zhangjihong@bfsu.edu.cn.

where $\omega(|t|,\alpha) = \begin{cases} \tan\frac{\pi\alpha}{2} & \alpha \neq 1 \\ -\frac{2}{\pi}\log|t| & \alpha = 1 \end{cases}$, $0 < \alpha \leq 2$ is called the characteristic exponent or tail index, $-1 \leq \beta \leq 1$ the skewness parameter, $c > 0$ the scale parameter, and $\mu \in R$ the location parameter. When$\alpha = 2$, the normal distribution results.

Density function $s(x)$ relates to the characteristic function by the inverse Fourier transformation,

$$s(x) = \frac{1}{2\pi}\int\limits_{-\infty}^{+\infty} e^{-itx}\varphi(t)dt.$$

Many applications involve calculation of densities and maximum likelihood estimation of parameters of stable distribution. To compute densities, two kinds of numerical methods can be used, direct integration method of John.p Nolan [9] and FFT method presented by S. Mittnik et al. [10], thus two kinds of methods to construct likelihood function given a data set. Direct numerical integration is nontrivial and burdensome from a computational viewpoint, which makes maximum likelihood estimation based on such a method difficult to implement and time-consuming. The computational efficiency of FFT method may be appropriate for this task. But the accuracy of S. Mittnik's FFT method is not satisfying enough. In our paper, we make an improvement of the FFT methods of S. Mittnik by using a different numerical rule-Simpson's rule, and compare it with S. Mittnik's method.

This paper is organized as follows: in section 2, we have a brief revision of the FFT method and derive the Simpson's rule based FFT method. In section 3, relative error of Simpson's rule based FFT method will be analyzed, section 4 makes a comparison with S. Mittnik's FFT method . Section 5 concludes the paper with final remarks and suggestions for further research.

2 FFT method to compute stable densities

The main idea of S. Mittnik's FFT method contains three steps. First, restrict the infinite integral $\int\limits_{-\infty}^{+\infty} e^{-itx}\varphi(t)dt$ onto a finite interval $[-a,a]$:

$$s(x) = \frac{1}{2\pi}\int\limits_{-\infty}^{+\infty} e^{-itx}\varphi(t)dt \approx \frac{1}{2\pi}\int\limits_{-a}^{+a} \exp(-itx)\varphi(t)dt \overset{\Delta}{=} s_T(x)$$

Then, $[-a,a]$ is divided into N equal-length interval with endpoint $t_j = -a + jh, j = 0,1,...,N$, $h = \frac{2a}{N}$. So we have $\int\limits_{-a}^{+a} \exp(-itx)\varphi(t)dt = \sum\limits_{j=0}^{N-1}\int\limits_{t_j}^{t_{j+1}} \exp(-itx)\varphi(t)dt.$

In the third step, S. Mittnik uses rectangle rule $\int\limits_{a}^{b} f(x)dx \approx (b-a)f(a)$ to approximate the integration on each subinterval $[t_j,t_{j+1}]$, while as we can see, rectangle rule is too rough to give accurate approximation. To improve accuracy, we use Simpson's rule $\int\limits_{a}^{b} f(x)dx \approx \frac{b-a}{6}(f(a) + 4f(\frac{a+b}{2}) + f(b))$ instead to make a better approximation.

Set $x_k = -\frac{N\pi}{2a} + \frac{\pi}{a}k, k = 0,1,...,N-1$, we first compute densities on x_k as follows:

$s_T(x_k) \approx \frac{1}{2\pi}\sum\limits_{j=0}^{N-1}\frac{h}{6}(\exp(-it_jx_k)\varphi(t_j)+4\exp(-it_j^*x_k)\varphi(t_j^*)+\exp(-it_{j+1}x_k)\varphi(t_{j+1}))\overset{\Delta}{=}\widetilde{s}(x_k),$
where $t_j^* = \frac{t_j+t_{j+1}}{2}$. Using the following facts (1)-(3),

$$t_jx_k = (-a+j\frac{2a}{N})(-\frac{N\pi}{2a}+\frac{\pi}{a}k) = \frac{N\pi}{2}-\pi k - j\pi + \frac{2\pi kj}{N}, \tag{1}$$

$$t_{j+1}x_k = (t_j+\frac{2a}{N})x_k = t_jx_k + \frac{2a}{N}x_k, \tag{2}$$

$$t_j^*x_k = (t_j+\frac{h}{2})x_k = t_jx_k+\frac{a}{N}x_k, \tag{3}$$

with simple computations, we find

$$\widetilde{s}(x_k) = \frac{(-1)^k}{2\pi}\frac{a}{3N}DFT((-1)^j\varphi(t_j)) + \frac{(-1)^{k+1}}{2\pi}\frac{a}{3N}\exp(-i\frac{2\pi k}{N})DFT((-1)^j\varphi(t_j+\frac{2a}{N}))$$

$$+\frac{(-1)^k}{2\pi}\frac{4a}{3N}\exp(-i\frac{\pi k}{N})iDFT((-1)^j\varphi(t_j^*)). \tag{4}$$

Denote $y_j^1 = (-1)^j\varphi(t_j), y_j^2 = (-1)^j\varphi(t_j+\frac{2a}{N}), y_j^3 = (-1)^j\varphi(t_j^*)$, $C_1 = \frac{(-1)^k}{2\pi}\frac{a}{3N}, C_2 = \frac{(-1)^{k+1}}{2\pi}\frac{a}{3N}\exp(-i\frac{2\pi k}{N})$, $C_3 = \frac{(-1)^k}{2\pi}\frac{4a}{3N}\exp(-i\frac{\pi k}{N})i$, then (4) becomes

$$\widetilde{s}(x_k) = C_1DFT(y_j^1)+C_2DFT(y_j^2)+C_3DFT(y_j^3), k=0,1,\cdots,N-1. \tag{5}$$

DFT stands for discrete fourier transformation , we use the Cooley-Tukey FFT algorithm to compute equation (5). For an arbitrary point x, density can be calculated through interpolating, linear or nonlinear, according to required accuracy.

3 Relative error analysis

We assume $\alpha > 1$, $\beta \geq 0$, $\sigma = 1$ and $\mu = 0$, this is because the following facts, $s(x;\alpha,\beta,\sigma,\mu) = \frac{1}{\sigma}s(z;\alpha,\beta,1,0)$, $s(z,\alpha,\beta,1,0) = s(-z,\alpha,-\beta,1,0)$, where $z = \frac{x-\mu}{\sigma}$. $\alpha > 1$ is suitable for most real-world applications.

Now, we analyze the relative errors of our Simpson's rule FFT based method. We only consider densities on x_k. The relative error $\varepsilon(x_k) = s(x_k) - \widetilde{s}(x_k)$ can be decomposed into two parts, $\varepsilon(x_k) = s(x_k)-\widetilde{s}(x_k) = s(x_k)-s_T(x_k)+s_T(x_k)-\widetilde{s}(x_k)\overset{\Delta}{=}\varepsilon_1(x_k)+\varepsilon_2(x_k)$. The first part $\varepsilon_1(x_k)$ is brought by truncating the infinite integral region to a finite interval; the second part $\varepsilon_2(x_k)$ comes from applying numerical integration rule-Simpson's rule. For $\varepsilon_1(x_k)$, we have:

$$|2\pi\varepsilon_1(x_k)| = |\int\limits_{-\infty}^{+\infty}\exp(-itx_k)\varphi(t)dt - \int\limits_{-a}^{a}\exp(-itx_k)\varphi(t)dt|$$

$$= |\int_{-\infty}^{-a} \exp(-itx_k)\varphi(t)dt + \int_{a}^{+\infty} \exp(-itx_k)\varphi(t)dt|$$

$$= |\int_{-\infty}^{-a} \exp(-|t|^{\alpha})\exp(-itx_k + i\beta t|t|^{\alpha-1}\tan(\frac{\pi\alpha}{2}))dt +$$

$$\int_{a}^{+\infty} \exp(-|t|^{\alpha})\exp(-itx_k + i\beta t|t|^{\alpha-1}\tan(\frac{\pi\alpha}{2}))dt|$$

$$= |\int_{-\infty}^{-a} \exp(-|t|^{\alpha})\cos(tx_k - \beta t|t|^{\alpha-1}\tan(\frac{\pi\alpha}{2}))dt +$$

$$\int_{a}^{+\infty} \exp(-|t|^{\alpha})\cos(tx_k - \beta t|t|^{\alpha-1}\tan(\frac{\pi\alpha}{2}))dt|$$

$$= 2|\int_{a}^{+\infty} \exp(-t^{\alpha})\cos(tx_k - \beta t^{\alpha}\tan(\frac{\pi\alpha}{2}))dt| < 2\int_{a}^{+\infty} \exp(-t)dt = 2e^{-a}$$

Note that the last inequation comes from the assumption $\alpha > 1$, if we choose a enough large a, the error $\varepsilon_1(x_k)$ can be negligible.

The second error $\varepsilon_2(x_k) = s_T(x_k) - \widetilde{s}(x_k)$ involves numerical integration of $s_T(x_k)$. $\widetilde{s}(x_k)$ is obtained by equally dividing the interval $[-a,a]$ into N subintervals and then applying Simpson's rule to every subinterval integral without considering the characteristics of the integrant. Let's take a look at the integrant first. We rewrite the finite interval integral as follows:

$$\int_{-a}^{a} \exp(-itx)\varphi(t)dt = \int_{-a}^{a} \exp(-|t|^{\alpha})\exp(-i(tx + \beta t|t|^{\alpha-1}\tan(\frac{\pi\alpha}{2})))dt$$

$$= 2\int_{0}^{a} \exp(-t^{\alpha})\cos(tx + \beta t^{\alpha}\tan(\frac{\pi\alpha}{2}))dt \qquad (6)$$

The integrand is $\exp(-t^{\alpha})\cos(tx + \beta t^{\alpha}\tan(\frac{\pi\alpha}{2}))$, $\cos(tx + \beta t^{\alpha}\tan(\frac{\pi\alpha}{2}))$ is a periodic and oscillating function with decreasing period when $|x|$ increase, that is to say when $|x|$ becomes large, the integrand becomes oscillate much. In this case, neither the Simpson's rule nor the rectangle rule gives accurate approximation, ε_2 will increase when $|x|$ becomes large.

We will not estimate $\varepsilon_2(x_k)$ using numerical methods because first we are aiming at comparing our method with S. Mittnik's method, second numerical integration methods, like adaptive quadrature, are not always accurate for the integration of an oscillating function and are very time-consuming for estimating errors on $x_k,\ k = 0, 1, \cdots, N$ when N is large.

4 Comparison with S. Mittnik's method

S. Mittnik et al [10]. in their paper use densities calculated by Nolan's method [9] as approximation to true densities. They uses two kinds of measurements to assess the proximity of the density values generated by these two algorithms:

$$d_1 = \frac{1}{N}\sum_{k=0}^{N-1} |s_N(x_k) - \widetilde{s}(x_k)|,$$

$$d_2 = \max_{k=0,1,\ldots,N-1} |s_N(x_k) - \widetilde{s}(x_k)|.$$

Table 1 in their paper reported the magnitudes of these two measurements: 10^{-5} when $N = 2^{13},\ h = 0.01$; 10^{-6} when $N = 2^{16},\ h = 0.01$ approximately. We use the same measurements to make a comparison. Set $N = 2^{15}, h = \frac{\pi}{2^8}$, then $a = 2^6\pi$. For each combination of $\alpha = 1.25, 1.5, 1.75,\ \beta = 0, 0.5, 1,\ d_1$ and d_2 are computed respectively. Results are given in the following table:

Table 1: Two measurements computed with Simpson's rule based FFT method.

α	β	$d_1 \times 10^{-7}$	$d_2 \times 10^{-7}$
1.25	0	1.0732192	2.9426751
	0.5	1.0711204	4.4866181
	1	1.1211393	6.0178983
1.5	0	0.29064173	0.75337895
	0.5	0.29059912	1.1310383
	1	0.29054025	1.5083377
1.75	0	0.050937323	0.13307321
	0.5	0.050938599	0.19849229
	1	0.050938433	0.263906

From table 1, the magnitude is $10^{-7}$when $N = 2^{15}, h = \frac{\pi}{2^8} \approx 0.012$, which is better than S. Mittnik's result. Figures of the relative errors $e(x_k) = s_N(x_k) - \widetilde{s}(x_k)$ are also given as follows(Figure 1-3),

From these figures, we can come to three conclusions. First, almost all of the relative errors are larger than zero, which means that the Simpson's rule based FFT method tends to underestimate true densities of the stable distribution. Second, when $|x|$ becomes large, the relative error becomes large too, which is consistent with the increasingly oscillating property of the integrand. Third, the larger α is, the smaller the relative errors are. This is not beyond our expectation. On the one hand, $\exp(-t^\alpha)$ in equation (6) is a descending function, its velocity of descending is positively correlated to α; on the other hand,

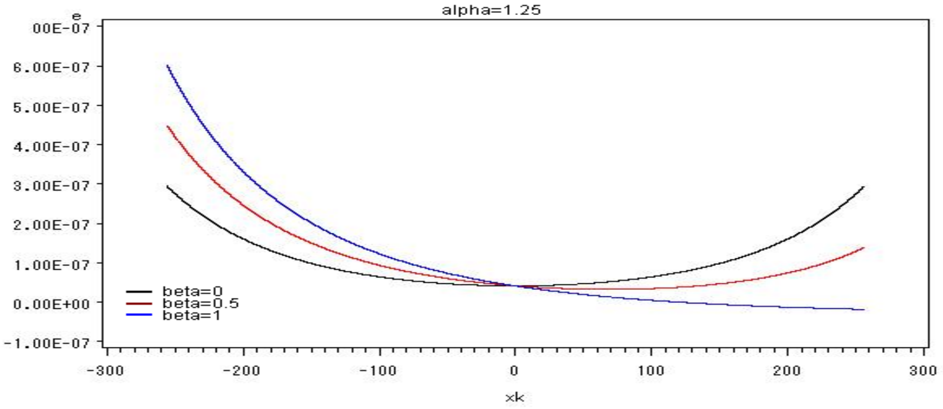

Figure 1: Relative errors when $\alpha = 1.25,\ \beta = 0, 0.5, 1$.

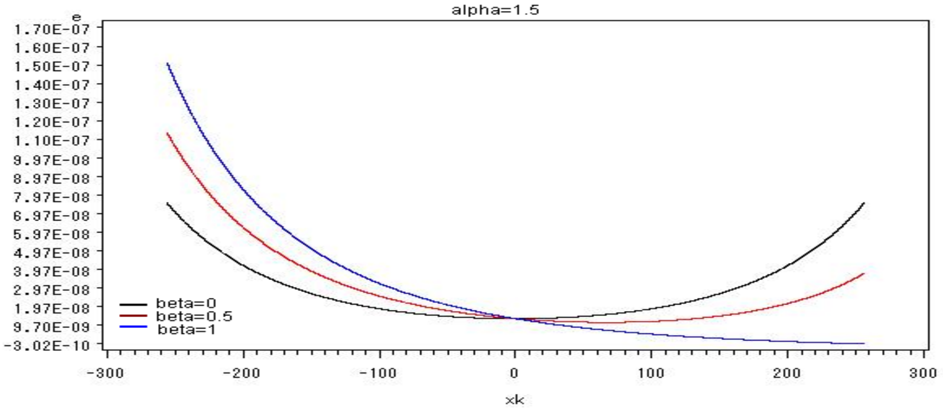

Figure 2: Relative errors when $\alpha = 1.5,\ \beta = 0, 0.5, 1$.

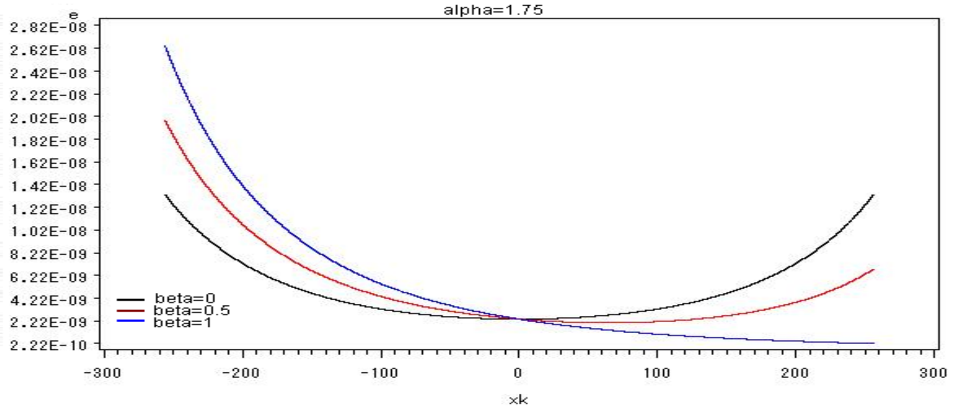

Figure 3: Relative errors when $\alpha = 1.75,\ \beta = 0, 0.5, 1$.

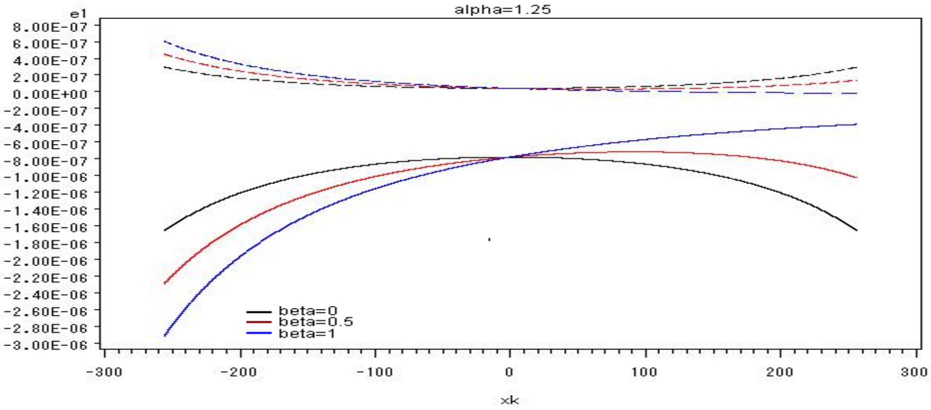

Figure 4: Comparison between relative errors of Simpson's rule based FFT method and that of rectangle rule based FFT method when $\alpha = 1.25, \beta = 0, 0.5, 1$.

when α increase, the tail of the distribution becomes light, the probability of extreme value becomes small, more and more data will concentrate in the central region of the distribution.

Figure 4 gives a comparison of relative errors of the Simpson's rule based FFT method and the rectangle rule based FFT method. The real lines are the relative errors of Simpson's rule, the dashed line are those of rectangle rule. The latter rule tends to overestimate true densities. In a relatively larger scope, densities computed by Simpson's rule FFT based method can be trusted.

5 Conclusions and suggestions

For the FFT based method, the fast fourier transformation is the only tool used to calculate densities. No matter how large a data set is, we just need to compute densities on N equally-spaced points x_k. Densities of the remaining data points can be found by interpolation, linear or nonlinear depending on your requiring accuracy. This method is especially efficient for modeling microarray gene expression data with stable distribution, for this kind of data sets often contain hundreds of thousands of data. When maximum likelihood estimation is used to estimate parameters, the FFT method can be used to construct the approximate likelihood function. However, as we can see, despite the Simpson's rule can improve accuracy, the relative error still becomes large when x increases. To remedy this limitation, we can further consider using the Bergström expansion to estimate tail densities, which is the suggestion of DuMouchel [11].

References

[1] Khondoker M.R., Glasbey C.A., Worton B.J., 2006. Statistical estimation of gene expression using multiple laser scans of microarrays. Bioinformatics 22 (1), pp215–219.

[2] Kuznetsov V.A., 2001. Distribution associated with stochastic processes of gene expression in a single eukaryotic cell. EURASIP Journal on Applied Signal Processing 4, pp285–296.

[3] Lönnstedt I., Speed T., 2002. Replicated microarray data. Statistica Sinica 12, pp31–46.

[4] Hoyle D.C., Rattray M., Jupp R., Brass A., 2002. Making sense of microarray data distributions. Bioinformatics 18 (3), pp576–584.

[5] Gencağa D., Ertüzün A., Kuruoğlu E.E., 2008. Modeling of non-stationary autoregressive alpha-stable processes by particle filters. Digital signal processing 18, pp465-478.

[6] Mikosch T., Resnick S., Rootzén H., Stegeman A., 2002. Is network traffic approximated by stable lévy motion or fractional brownian motion? Annals of Applied Probability 12(1), pp23-68.

[7] Mittnik S., Paolella M.S., 2003. Prediction of financial downside-risk with heavy-tailed conditional distributions. Working paper.

[8] Samorodnitsky G., Taqqu M., 1994. Stable non-gaussian random process: stochastic models with infinite variance. Chapman-Hall, New York.

[9] Nolan J.P., 1997. Numerical calculation of stable densities and distribution functions. Communications in Statistics-Stochastic Models. 13 (3), pp759–774.

[10] Mittnik S., Doganoglu T., Chenyao D., 1999. Computing the probability density function of the stable paretian distribution, Mathematical and computer modeling. 29, pp235-240.

[11] DuMouchel W., 1971. Stable distributions in statistical inference, Ph.D dissertation, Yale University.