电子科技大学

University of Electronic Science and Technology of China

学 士 学 位 论 文

BACHELOR DISSERTATION

论文题目 有限矩发散条件下的统计分析

学生姓名 郭方健

学　　号 2901307007

专　　业 计算机科学与技术

学　　院 英才实验学院

指导教师 周 涛

指导单位 电子科技大学

2013 年 6 月 2 日

摘要

矩, 诸如均值、方差、协方差, 是刻画概率分布形态的重要统计量。每个矩同时对应于一个样本矩和一个总体矩。当总体矩收敛时, 随着样本容量的增大, 样本矩依概率收敛于总体矩; 然而, 对于重尾分布, 总体矩可能发散, 此时样本矩会随着样本容量的增加而不断增大。本文提出等概率分割方法 (EPM) 用于系统地分析发散情形下有限个样本的矩随样本容量增大的渐近行为。本文证明了 EPM 在收敛条件下的准确性, 并对其给出的发散条件下的渐近线进行了数值分析。最后, 针对幂律分布时间序列的自相关，运用 EPM 处理其中的发散矩项，从而得出其一阶自相关的非平凡上下界，该上下界与数值模拟和实证数据相吻合。

关键词： 发散矩，渐近，等概率分割，幂律，记忆性

ABSTRACT

Moments, such as mean, variance and covariance, are important statistics that characterize the shape of distributions. Every moment simultaneously corresponds to a sample moment and a population moment. A specific population moment may diverge for heavy-tailed distributions. When population moments do not diverge, sample moments will converge to their corresponding population moments by probability. However, when population moments diverge, sample moments also keep growing with the sample size. Firstly in this thesis, the *equiprobable partition method* (EPM) is presented for systematically analyzing the behavior of diverging sample moments. Estimators for sample moments are constructed from EPM, which are then used to theoretically derive the asymptotics for sample moments when they diverge. EPM estimators are shown to be unbiased under convergence and their accuracy is also numerically studied. Finally, we study the autocorrelation of power-law series, where EPM is applied to the analysis of diverging moments involved. We find non-trivial bounds for the first-order autocorrelation as a function of the power-law exponent, which agree with numerical experiments and empirical data.

Keywords: diverging moments, asymptotics, equiprobable partition, power-law, memory

Contents

Chapter 1 Introduction

The tasks of estimation, inference and prediction, as an essential constituent of human activities, have been embraced by researchers in computer science, especially in the field of *Artificial Intelligence (AI)*. To solve these tasks, researchers make use of both the mathematical formalism and the computational power realized by modern hardware and programming tools. Unfortunately, due to the continuing influences of the early pioneers and their paradigms, for a while AI scientists equated reasoning with "rational" deduction; and tried to come up with if-then-else rules for various prediction tasks [1].

However, in the recent several decades, a noticeable paradigm shift occurred in the field of AI. More modern influences such as quantum physics and statistics made many appreciate the use of probabilistic machinery, not just for modeling uncertainty but for making efficient estimations and predictions possible at all. From then on, scientists in AI growingly rely on probability theory and statistics as the formalism for seeking both theoretical insights into problems and justifications for methods and algorithms [2].

Even a sub-field called *machine learning* emerged from the general stream of AI. While AI emphasizes the study and design of intelligent agents, including a group of interacting agents, as the central topic, machine learning is more specific in both its objective and methodology. Machine learning is concerned with discovering patterns, making inferences and predictions from *data* and it shares the same formal framework with probability and statistics, which is also the paradigm adopted in this thesis.

In a probabilistic perspective, a central and fundamental task would be the characterization of the distribution of random variables, including the marginal of each variable and the interdependence among them. Ideally, the fullest characterization would be given by writing down the joint distribution of the all the variables involved. Yet, among others, due to the usually unknown or partially observable mechanism underlying real-world data or the "curse of dimensionality" coming with

a large number of variables, one has to assume less accurate but more compact forms of representation of the joint distribution, such as log-linear models, Markov random fields and Bayesian networks [3].

Once we have data and the estimated distribution of the variables in concern, it would be desirable to know the properties of the distribution, for the purpose of developing learning algorithms that utilizes these properties or summarizing the large-volume data into a straightforward, human-perceptible form. Among others, moments are the simplest measures for quantitatively characterizing the shape of distributions. For example, if we have the data tracking a noisy signal and find that the signal follows a Gaussian distribution, we would then like to know the mean and standard deviation of the noise, for the first measure tells us around what value the signal fluctuates and the second informs us the amplitude of the fluctuation.

Moments are also important in their role connecting data and the population. In a statistical point of view, the data collected are considered as a part out of an infinite sea of data that are not fully observed, which are called *statistical population*. On one hand, we use the data to estimate the distribution of population; on the other hand, we assume that future, unobserved data and the data available now are sampled from *the same* population so that our model can be *generalized* to unknown or future cases.

Specifically, a moment simultaneously corresponds to a sample moment and a population moment, where the former characterizes how data in hands are distributed and the latter characterizes how the population are distributed. A sample moment is computed by performing simple algebraic operations on data, e.g. the arithmetic mean is the sum of all observation of a variable divided by the number of observations; a population moment is obtained by getting the mathematical expectation of a polynomial function of the random variable, which is usually done in the form of integral, e.g. the population mean is the mathematical expectation of the random variable itself.

The relation between the sample moments and the corresponding population moments are known when the population moments converge — a sample moment converges to the corresponding population moment by probability when the size of

samples extends to infinity. Therefore, in the convergent case, we can use the sample moments calculated with a large sample to estimate the shape of the population distribution. In the traditional statistical literature, because the distributions encountered (e.g. Gaussian, exponential) are usually exponentially bounded in tails, population moments of all positive orders would naturally converge. As a result, the convergence of moments are often assumed without explicit mention and the relation between the two kinds moments is largely "taken for granted" in practice.

However, such a naive notion towards moments and the general assumption of moments being convergent should be brought to serious examination when dealing with today's ever diversified data. A distinctive exception to the moment-convergent assumption is a family of distributions called *heavy-dutyed distributions*, where the density functions in this family have heavier tails than the exponential distributions [4], rendering their moments with orders higher than a threshold divergent. Heavy-tailed distributions "appear" more common than before with the development of information technology, as data can now be collected and assembled from an unprecedented large scale, e.g. data collected from the whole Internet, from millions of online users and from hundreds of countries around the world. For example, *power-law distributions*, as a member of the heavy-tail family, are found in the distributions for the degrees of the Internet, the population of cities around the world and the citations of papers in academic communities [5].

When dealing with data exhibiting these distributions, it is possible that a certain moment of interest is divergent. Then, the sample moment do not converge when the sample size grows bigger, but instead grows with the size. In this thesis, I will first seek to clarify the underlying mechanism beneath the divergence of moments and the relation between sample and population moments in this case. Then I will propose a method that systematically analyzes the asymptotic behavior of sample moments under convergence as the sample size extends to infinity. Furthermore, I will discuss the application of this method in analyzing heavy-tailed distributions, especially power-law distributions. Before addressing the method for analyzing divergent moments, I would like to outline some key concepts involved as background materials for this work.

1.1 Random variables and probability distributions

1.1.1 Random variables

Probability theory deals with the analysis of the likelihood of random events defined over a sample space Ω . For example, consider rolling a single fair die, where we denote the value rolled as X and the sample space as Ω. The sample space of a single roll can be defined as $\Omega = \{1, 2, 3, 4, 5, 6\}$ and the possible random events are subsets of Ω . For example, we can consider the event that $X = 4$ is rolled as well as the event that an even value is rolled. If an event is a single element of the sample space $X \in \Omega$ then we refer to the event as an atomic event. The event that $X = 4$ is rolled is an example of an atomic event.

When dealing with random events we are often interested in numerical descriptions of the events and not just the probability of the event occurring. In order to handle this we can assign numerical descriptions of events to what are referred to as random variables. For example, X is a random variable in rolling one fair die. And supposing that we roll two fair dice, we also can define a random variable Y to take on the numerical result of the sum of the dice.

Definition 1.1. A random variable X defined over a sample space Ω is a function $X : \Omega \to \mathbb{R}$ that maps an event $X \in \Omega$ to a real value.

Random variables that only take on values from a countable set (such as the integers) are referred to as discrete random variables. The random variable defined as the sum of the roll of two fair dice is an example. Random variables that take on values from an uncountable set (such as the reals) are referred to as continuous random variables. For example, the arrival time for the next car in a highway intersection is a continuous random variable.

1.1.2 Probability distributions

In analyzing random variables we are often interested in the probability that a random variable takes on certain values. To formally describe the different probabilities of a random variable taking on various values we define a probability distribution.

Definition 1.2. A probability distribution P defined on random variable X over a sample space Ω is a mapping from events $\alpha \in \Omega$ to real values on the interval $[0, 1]$ such that $P(\alpha) \geq 0$ and $P(\Omega) = 1$.

1.2 Moment

In statistics, a moment is, loosely speaking, a quantitative measure of the shape of a distribution. For example, mean, as the simplest moment, measures around what value the distribution is centered around; and variance, as the second central moment, measures how much the distribution disperses around the mean; while some moments of higher orders describe other aspects of a distribution such as how the distribution is skewed from its mean, or peaked. While moments can be extended to multivariate distributions, such as *image moments* that are used in 2-D image processing [6], we will focus on univariate distributions due to the scope of this thesis.

1.2.1 Expectation and Riemann-Stieltjes integral

A population moment with regard to a probability distribution is formally defined as the *mathematical expectation* of a polynomial function of the random variable. And a rigorous, unified treatment of expectation relies *Riemann-Stieltjes integral*, which is a generalized form for the *Riemann integral*. Therefore, we will first briefly outline these concepts before moments are introduced.

Whereas the usual Riemann integral of a real-valued function $g(x)$ with regard to the variable x on a range $[a, b]$ is usually denoted by $\int_a^b g(x)dx$, a Riemann-Stieltjes integral for x involves two functions. The Riemann-Stieltjes integral for $g(x)$ with regard to another real function $F(x)$ is denoted by $\int_a^b g(x)dF(x)$. It is well-known that $\int_a^b g(x)dx$ can be interpreted as the area under $g(x)$ with $a \leq x \leq b$. So what does Riemann-Stieltjes integral mean? In fact, $\int_a^b g(t)dF(t)$ (the variable is changed for convenience) can still interpreted as the area under a curve. Let us imagine t as a parameter and we now track the point t moving from a to b, meanwhile drawing the curve $(x, y) = (F(t), g(t))$. Then the integral is simply the

area under the curve, as a sum of rectangles each resulted by a every tiny move of t.

Now let us turn to a more formal definition of $\int_a^b g(x)dF(x)$. First the range $[a, b]$ is partitioned into n intervals with $a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b$. Then we tag each small interval $[x_i, x_{i+1}]$ with a representative number c_i, with $c_i \in [x_i, x_{i+1}]$. Next, we write down the sum of the area of small rectangles as

$$S = \sum_{i=0}^{n-1} g(c_i)(F(x_{i+1}) - F(x_i)). \tag{1–1}$$

The integral can obtained by taking the limit of S while "densifying" the partition. The density can be measured by a *mesh* of the partition, and here we just take the maximum length of the interval $\Delta = \max_{0\leq i\leq n-1}(x_{i+1} - x_i)$ as the mesh. As the mesh $\Delta \to 0$, the partition is continually densified and the sum would converge to a limit, if it exists, which is defined to be the Riemann-Stieltjes integral. Hence, we outline the following definition.

Definition 1.3. If there is some $A \in \mathbb{R}$, such that for any $\epsilon > 0$, there exists $\delta > 0$ so that any $S = \sum_{i=0}^{n-1} g(c_i)(F(x_{i+1}) - F(x_i))$ with the mesh of partition $\Delta < \delta$ satisfies $|S - A| < \epsilon$, then the Riemann-Stieltjes integral $\int_a^b g(x)dF(x)$ exists and we define $\int_a^b g(x)dF(x) \triangleq A$.

Now we turn to the definition for mathematical expectation. For a discrete random variable X, which takes values $x_1, x_2, \cdots, x_n$ with probabilities $p_1, p_2, \cdots, p_m$, then the expectation of X is defend to be the weighted sum, i.e.

$$\mathrm{E}[X] = \sum_{i=1}^{n} x_i p_i. \tag{1–2}$$

For a continuous random variable X with a well-defined PDF $f(x)$, then the expectation is yield with a Riemann integral, i.e.

$$\mathrm{E}[X] = \int_{-\infty}^{+\infty} x f(x)dx. \tag{1–3}$$

It seems that expectation needs to be treated differently for discrete and continuous variables when the Riemann-Stieltjes integral comes to the rescue.

Definition 1.4. Given X is a random variable with its cumulative distribution function (CDF) $F(X)$, then its expectation value is defined to be

$$\mathrm{E}[X] = \int_{-\infty}^{+\infty} x dF(x), \tag{1–4}$$

regardless of X being discrete or continuous.

Similarly, the expectation of the random variable $g(X)$ as a function of X is given by

$$\mathrm{E}[g(X)] = \int_{-\infty}^{+\infty} g(x) dF(x). \tag{1–5}$$

1.2.2 Definition of moment

Every moment is defined with its "double identity" — a population moment and sample moment. Both of them are quantitative measures for the shape of the distributions, but the former is for describing the shape of the "empirical distribution" (the distribution for samples), i.e. the CDF constructed from data, while the latter is for characterizing the distribution of population, where a well-defined, closed-form distribution function is usually assumed. When the sample size is small, the two distributions may be different and the two moments may also differ noticeably in their values; but when the sample size grows big, the empirical distribution will converge to that of the population by the *Strong Law of Large Numbers*, and the two moments will be identical in the limit.

But the sample moment and its corresponding population moment are different in their calculation. Without explicitly getting the CDF for the empirical distribution, one can directly calculate sample moments by performing simple algebraic operations (addition, multiplication and division) on data. But the calculation of population moment is analytic, rather than algebraic. The population distribution is usually given (or assumed to be) as a analytic function and its population moments, if exist, can be acquired by integrating a polynomial function with regard to the PDF. These two moments are also different in their existence: a population moment does not exist if the integral diverges; a sample moment *always* exists because it is a algebraic outcome computed from data.

1.2.2.1 Population moment

A *population moment* can be understood as a particular distance measure defined by the distribution with a point as its reference. When there is no ambiguity, we also refer to *population moment* as *moment* for short.

Definition 1.5. Given a univariate PDF $f(x)$ for the distribution of a random variable X , the n-th moment taken about a point c is defined as $\mu_n(c) = \int_{-\infty}^{+\infty}(x - c)^n f(x)dx$ [1] , if the integral converges.

The reference point c is often chosen to be 0 or the mean. When $c = 0$, the moment is called *raw moment*.

Definition 1.6. The n-th raw moment is defined as $\mu'_n = \mu_n(0)$, where $\mu_n(c)$ is the moment taken about c.

μ'_1 is simply the *mean* of the distribution, and we have $\mu'_0 = 1$ due to the normalization of a distribution. For moments of higher orders, we are often concerned with the moment taken around the mean, which is called *central moment*.

Definition 1.7. The n-th central moment is defined as $\mu_n = \mu_n(\mu'_1)$, where $\mu_n(c)$ is the moment taken about c and μ'_1 is the mean.

It shall be noted the following definitions rely on the condition that the integral for defining the moment is convergent, which is always the case for distributions with exponentially bounded tails, but not necessarily true for other distributions.

1.2.2.2 Sample moment

Sample moments are the same metrics as population moments, except that they are defined by the *empirical distributions* of data, rather than the assumed the distribution function. Sample moments can be computed by performing simple algebraic operations on data and every moment $\mu_n(c)$ has one sample counterpart $m_n(c)$.

[1] When addressing the definition of moments, we assume X is a continuous variable for simplicity. The discrete case can be similarly defined by replacing the integral with sum.

Definition 1.8. Given a sample $\{x_1, x_2, \cdots, x_N\}$ with size N, the n-th sample moment with reference to c is defined as $m_n(c) = \frac{1}{N}\sum_{i=1}^{N}(x_i - c)^n$.

As samples are always finite numbers, a sample moment *always* exists regardless of existence of the corresponding population moment. *Sample raw moments* can be defined in a parallel fashion.

Definition 1.9. Given a sample $\{x_1, x_2, \cdots, x_N\}$ with size N, the n-th sample raw moment is defined as $m'_n = m_n(0)$.

Sample moments can be used as estimators for the corresponding population moments, which we would discuss later.

1.2.3 Moments in common uses

We now describe some population moments that are widely used in statistics for characterizing the shape of distributions.

1.2.3.1 Mean

Mean is the first raw moment μ'_1, which gives the value around which the distribution is centralized around.

1.2.3.2 Variance

Variance is the second central moment μ_2, which describes how far the distribution spreads out around the mean. Its positive squared root σ is referred to as the *standard deviation*. For the special case of Gaussian distributions, the distribution can be fully characterized by the mean and the variance, as shown in Fig. 1–1.

1.2.3.3 Skewness

The third central moment μ_3 is a measure of the lopsidedness of the distribution; any *symmetric* distribution will have a third central moment, if defined, of zero. When using higher order moments, it is a common practice to normalize the k-th moment by the standard deviation to the power of k.

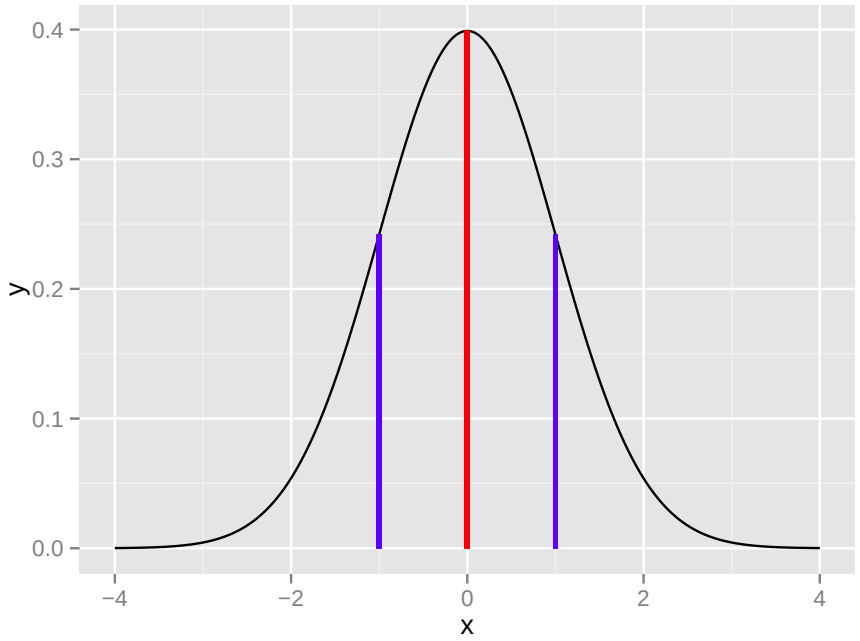

Figure 1–1 The PDF for a standard Gaussian distribution, where the red vertical marks the value of μ_1' and two blue lines mark $\pm\sigma$.

Definition 1.10. The k-th standardized moment for a probability distribution is defined as μ_k/σ^k, where μ_k is the k-th central moment and σ is the standard deviation.

In other words, the k-th standardized moment is simply a *normalzied* version of the k-th moment with respect to the standard deviation. The power of k is used, because moments scale as x^k , meaning that $\mu_k'(\lambda X) = \lambda^k \mu_k'(X)$: they are homogeneous polynomials of degree k , thus the standardized moment is *scale invariant*. This can also be understood as being because moments have dimension; in the above ratio defining standardized moments, the dimensions cancel, so they are dimensionless numbers.

Specifically, the *standardized* third central moment μ_3/σ^3 is called the skewness, often denoted by γ_1, i.e.

Definition 1.11. The skewness for a random variable X is defined as $\gamma_1 = \mathrm{E}[(\frac{x-\mu}{\sigma})^3] = \frac{\mu_3}{\sigma^3}$, where μ

As plotted by Fig. 1–2, a distribution that is skewed to the left (the tail of

the distribution is heavier on the left) will have a negative skewness. A distribution that is skewed to the right (the tail of the distribution is heavier on the right), will have a positive skewness.

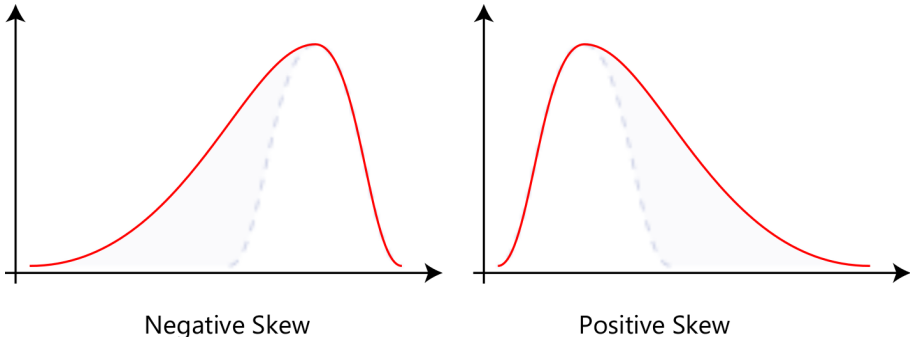

Figure 1–2 Probabilty distributions with a negative skewness has a fatter or longer tail on the left side, and distributions with a positive skewness has a fatter or longer tail on the right side. (figure excerpted from [7])

1.2.3.4 Kurtosis

The fourth central moment is a measure of whether the distribution is tall and skinny or short and squat, compared to the normal distribution of the same variance. Since it is the expectation of a fourth power, the fourth central moment, where defined, is always non-negative.

One common measure of kurtosis, originating with Karl Pearson, is based on a scaled version of the fourth moment of the data or population, but it has been argued that this measure really measures heavy tails, and not peakedness. It is common practice to use an adjusted version of Pearson's kurtosis, the excess kurtosis, to provide a comparison of the shape of a given distribution to that of the normal distribution (μ_4 for a normal distribution is $3\sigma^4$). Fig. 1–3 shows three distributions of the same family with different values of kurtosis.

Definition 1.12. The excess kurtosis for a random variable, if it exists, is defined as $\gamma_2 = \mu_4/\sigma^4 - 3$, where μ_4 is the fourth central moment and σ is the standard deviation.

1.2.3.5 Covariance and correlation coefficient

Covariance and *correlation* are *mixed moments* of most frequent uses. While the aforementioned moments only involve a single random variable, *mixed moments*

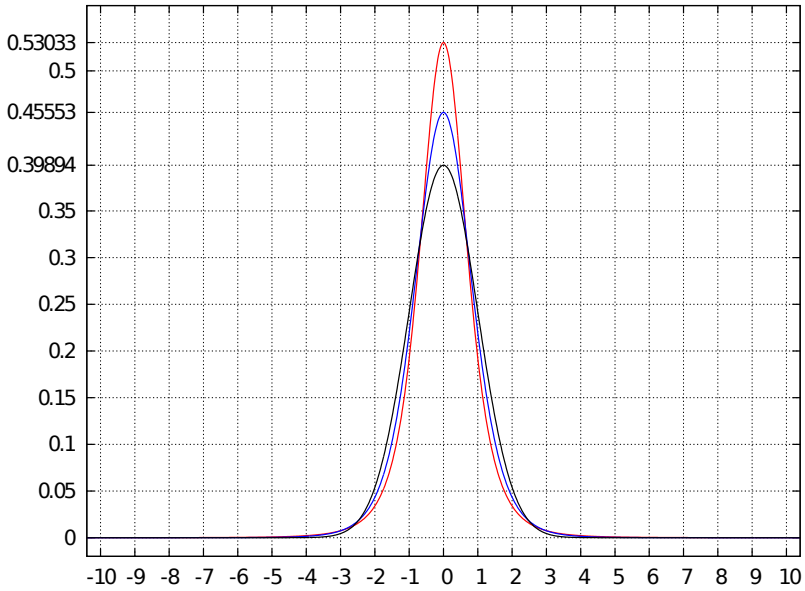

Figure 1–3 PDF for the Pearson type VII distributions with kurtosis of infinity (red), 2 (blue) and 0 (black).

are moments of multiple random variables.

In probability theory and statistics, covariance is a measure of how much two random variables change together. If the greater values of one variable mainly correspond with the greater values of the other variable, and the same holds for the smaller values, i.e., the variables tend to show similar behavior, the covariance is positive. In the opposite case, when the greater values of one variable mainly correspond to the smaller values of the other, i.e., the variables tend to show opposite behavior, the covariance is negative. The sign of the covariance therefore shows the tendency in the linear relationship between the variables.

Definition 1.13. The covariance between two jointly distributed real-valued random variables X and Y with finite second moments is defined as

$$\sigma(x, y) = \mathrm{E}\left[(X - \mathrm{E}(X))(Y - E(Y))\right]. \tag{1–6}$$

After some algebra, covariance can be rewritten in the form as

$$\begin{aligned}\sigma(X,Y) &= \mathrm{E}\left[(X-\mathrm{E}\left[X\right])\left(Y-\mathrm{E}\left[Y\right]\right)\right] \\ &= \mathrm{E}\left[XY - X\,\mathrm{E}\left[Y\right] - \mathrm{E}\left[X\right]Y + \mathrm{E}\left[X\right]\mathrm{E}\left[Y\right]\right] \\ &= \mathrm{E}\left[XY\right] - \mathrm{E}\left[X\right]\mathrm{E}\left[Y\right] - \mathrm{E}\left[X\right]\mathrm{E}\left[Y\right] + \mathrm{E}\left[X\right]\mathrm{E}\left[Y\right] \\ &= \mathrm{E}\left[XY\right] - \mathrm{E}\left[X\right]\mathrm{E}\left[Y\right]. \end{aligned} \tag{1–7}$$

The definition above can also be extended to the case of more than two random variables, where covariance is formed as a matrix.

Definition 1.14. Supposing $\boldsymbol{X}$ and $\boldsymbol{Y}$ are two random vectors of dimension m and n respectively, each with finite second moment, then $m \times n$ *covariance matrix* is defined as

$$\boldsymbol{\sigma}(\boldsymbol{X},\boldsymbol{Y}) = \mathrm{E}\left[(\boldsymbol{X}-\mathrm{E}(\boldsymbol{X}))(\boldsymbol{Y}-\mathrm{E}(\boldsymbol{Y}))^T\right] \tag{1–8}$$

$$= \mathrm{E}\left[\boldsymbol{X}\boldsymbol{Y}^T\right] - E[\boldsymbol{X}]E[\boldsymbol{Y}]^T. \tag{1–9}$$

Whereas one can learn the consistency between two random variables (or two random vectors) from the sign of covariance, the value itself is often hard to interpret. Therefore, its standardized version, the *correlation coefficient* is introduced as a measure lying in $[-1,+1]$ to show the strength of linear relation.

In statistics, the *Pearson product-moment correlation coefficient* (sometimes referred to as PCC, Pearson's r, or correlation coefficient) is a measure of the *linear correlation* (dependence) between two variables X and Y , giving a value between -1 and $+1$ inclusive. It is widely used in the sciences as a measure of the strength of linear dependence between two variables.

Pearson's correlation coefficient between two variables is defined as the covariance of the two variables normalized by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables.

Definition 1.15. The correlation coefficient $\rho_{X,Y}$ for two random variables X and Y, each with finite second moment, is defined as

$$\rho_{X,Y} = \frac{\sigma(X,Y)}{\sigma_X\sigma_Y} = \frac{\mathrm{E}\left[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])\right]}{\sigma_X\sigma_Y}. \tag{1–10}$$

The correlation coefficient can also be viewed as the product of the two *standardized* random variables, i.e.

$$\rho_{X,Y} = \mathrm{E}\left[\left(\frac{X-\mathrm{E}(X)}{\sigma_X}\right)\left(\frac{Y-\mathrm{E}(Y)}{\sigma_Y}\right)\right]. \tag{1–11}$$

By regarding the expectation of the product of two variables as an *inner product* between the two, i.e. $\langle X, Y\rangle \triangleq \mathrm{E}(XY)$, then due to the famous *Cauchy-Schwartz inequality*, we have

$$\begin{aligned}\rho_{X,Y}^2 &= \langle\frac{X-\mathrm{E}(X)}{\sigma_X}, \frac{Y-\mathrm{E}(Y)}{\sigma_Y}\rangle^2 \\ &\leq \frac{\langle X-\mathrm{E}(X), X-\mathrm{E}(X)\rangle}{\sigma_X^2}\frac{\langle Y-\mathrm{E}(Y), Y-\mathrm{E}(Y)\rangle}{\sigma_Y^2} = 1,\end{aligned} \tag{1–12}$$

from which we see that ρ is naturally bounded inside $[-1, +1]$. And there is another advantage to standardize the covariant — to make correlation coefficient a *scale invariant* measure such that any *linear transform* applied to one variable does not affect the outcome.

By replacing the population moments with the corresponding sample moments, we can write down the correlation coefficient for two samples $(x_1, x_2, \cdots, x_n)$ and $(y_1, y_2, \cdots, y_n)$ as

$$\rho_{x,y} = \frac{\sum_{i=1}^n (x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^n (x_i-\overline{x})^2}\sqrt{\sum_{i=1}^n (y_i-\overline{y})^2}}. \tag{1–13}$$

$\rho_{x,y}$ is usually computed from data to reflect the linear relation between two variables of observation, which could be interpreted as the the cosine of the angle between both possible regression lines $y = g_x(x)$ and $x = g_y(y)$ geometrically. Fig. 1–4 plots several sets of points and their corresponding values of correlation coefficient. It shall be noted that the correlation coefficient only reflects the *strength* of the linear relation, not to be confused with the *slope* of linearity.

1.2.4 Properties

1.2.4.1 The relation between raw moment and central moment

The central moments μ_n can be expressed as terms of the raw moments μ'_n (i.e. those taken about zero) using the binomial transform [9].

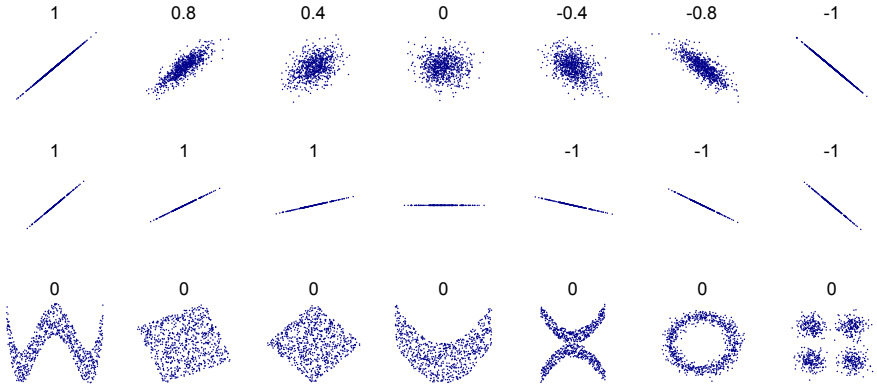

Figure 1–4 Several sets of (x, y) points, with the correlation coefficient of x and y for each set. Note that the correlation reflects the non-linearity and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). (figure excerpted from [8])

Theorem 1.1. *The n-th population central moment μ_n can be expressed as*

$$\mu_n = \sum_{k=0}^{n} \binom{n}{k} (-1)^{n-k} \mu'_k \mu_1'^{n-k},$$

where μ'_k is the k-th raw moment.

1.2.4.2 Moment-generating function

Population moments, if exist, can also be obtained with *moment-generating functions*.

Definition 1.16. The moment-generating function for a random variable X is defined as

$$M_X(t) = \mathrm{E}[e^{tX}], \quad t \in \mathbb{R},$$

whenever the expectation exists.

The moments inside the function can be seen via the Taylor expansion as

$$\begin{aligned} M_X(t) = \mathrm{E}[e^{tX}] &= \mathrm{E}\left[1 + tX + \frac{(tX)^2}{2!} + \frac{(tX)^3}{3!} + \cdots\right] \\ &= 1 + t\mu'_1 + \frac{t^2}{2!}\mu'_2 + \frac{t^3}{3!}\mu'_3 + \cdots . \end{aligned}$$

By taking the n-th order derivative and setting $t = 0$, the term with μ'_n with be left out while other terms vanish. Therefore, we have

$$\mu'_n = \mathrm{E}[X^n] = \frac{d^n M_X(0)}{dt^n}. \tag{1–14}$$

Chapter 2 Diverging Moments and Their Asymptotics

In the previous chapter, we have revisited some of the basic concepts in probability and outlined the definition for sample and population moments, along with some of their mathematical properties. In this chapter, we will first clarify the relation between the sample and population moments under convergence and use it as an intuition towards the central topic in this thesis, i.e. the analysis of diverging moments. Then, in the scenario of diverging population moments, we will analyze the behavior of sample moments with the growth of population size. Next, the equiprobable-slice method will be proposed, which is to serve as the key to the asymptotic analysis of diverging moments. When moments converge, the method will be shown to be consistent with the true value of moments; when moments diverge, we will see how to the method to reproduce the asymptotics of divergence.

2.1 The relation between sample and population moment under convergence

For the sake of simplicity, we restrict the moments discussed here to *raw moments*, i.e. moments taken about $c = 0$. The conclusions can also be applied to moments taken about any other value. Let us consider a random variable X and assume that its k-th population raw moment μ'_k is convergent. Supposing the population distribution is characterized by a PDF $f(x)$, then $\mu'_k = \int_{-\infty}^{+\infty} x^k f(x)dx$.

Now supposing we have N independent and identical distributed (i.i.d.) samples $X_1, X_2, \cdots, X_N$, we can compute the corresponding sample raw moment as $m'_k = \frac{1}{N}\sum_{i=1}^{N} X_i^k$. The distribution of each sample is obviously identical to the population distribution, therefore we have

$$\mathrm{E}[m'_k] = \frac{1}{N}\sum_{i=1}^{N}\mathrm{E}[X_i^k] = \frac{1}{N}\sum_{i=1}^{N}\int_{-\infty}^{+\infty} x_i^k f(x)dx = \mu'_k,$$

which can be equally stated as the theorem below.

Theorem 2.1. *The sample moment m'_k is an unbiased estimator for the corresponding population moment μ'_k, given that μ'_k exists (is convergent).*

As sample moments are computed from samples, they are random variables by themselves. Then how can one make the estimation for population moment as precise as possible? It is straightforward to imagine that a sample moment coming from samples of a large size would be a better estimation than one from a small size. Hence, when the sample size $N \to \infty$, can we guarantee that the sample moment equates with population moment? The answer is *yes* due to the *Law of Large Numbers*.

Theorem 2.2. *(The Weak Law of Large Numbers) Let $X_1, X_2, \cdots, X_N$ be a sequence of independent and identically distributed random variables, each having the same expectation $\mathrm{E}[X_i] = \mu$* [1] *, then their arithmetic mean $\overline{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$ converges in probability towards μ, i.e. $\lim_{N\to\infty} P(|\overline{X}_N - \mu| > \epsilon) = 0$ for any positive ϵ.*

By the two theorems above, we arrive at the following lemma, which formally states the simple relation between the sample and population moment under convergence.

Lemma 2.1. *A sample moment m'_k converges in probability towards the corresponding population moment μ'_k when the sample size $N \to \infty$, given that μ'_k exists (is convergent).*

Fig. 2–1 is an example that illustrates how this convergence is realized when the sample size grows bigger. We draw i.i.d. samples from a standard Gaussian distribution and compare the sample μ_4 with its population counterpart $\mu_4 = 3$. The convergence is guaranteed when the sample size extends to infinity, as indicated by the trend shown here.

The convergence of sample moments towards their population counterparts when they exist can also be interpreted as a consequence of the convergence of empirical distribution towards the population distribution. The empirical distribution

[1] An assumption of finite variance $\sigma(X_1)^2 = \sigma(X_2)^2 = \cdots = \sigma^2 < \infty$ is not necessary. Large or infinite variance will make the convergence slower, but the law holds anyway. This assumption is often used because it makes the proofs easier.

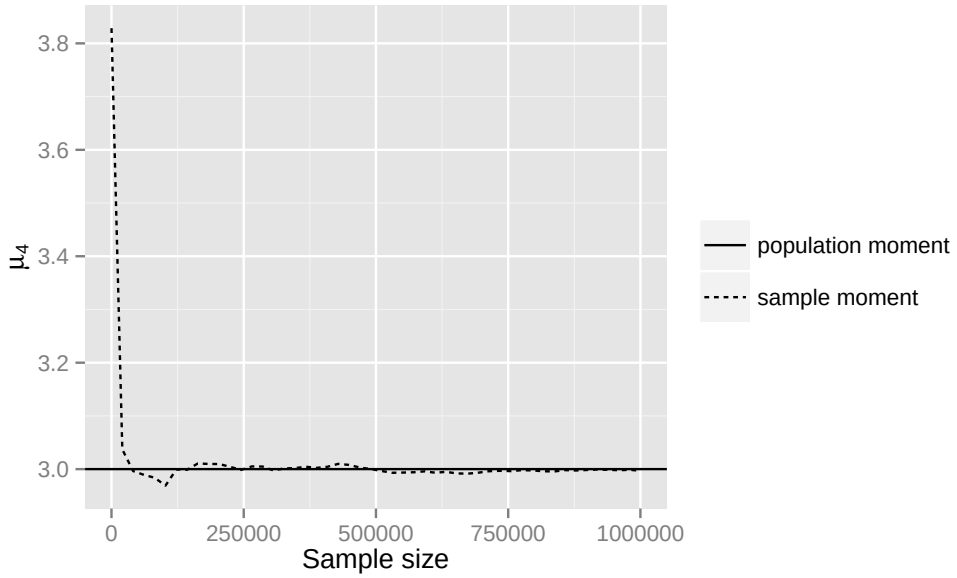

Figure 2–1 Moment μ_4 computed from i.i.d. samples of growing sample sizes drawn from a standard Gaussian distribution. It can be seen that the sample moment converges to the population moment μ_4 as the population size grows larger.

of a sample with size n can be characterized by constructing an empirical cumulative distribution function (empirical CDF) from data.

Definition 2.1. Let $x_1, x_2, \cdots, x_n$ be i.i.d. samples *already* independently drawn from the same population CDF $F(x)$, then the empirical CDF $\hat{F}_n(x)$ is defined as

$$\hat{F}_n(x) = \frac{\text{number of samples} \leq x}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(x_i \leq x), \tag{2–1}$$

where $\mathbf{1}(\cdot)$ is the indicator function.

By introducing the empirical CDF, sample moments and population moments can be unified — both of them are results of an Riemann-Stieltjes integral with the same integrand (a polynomial) but with regard to different CDFs. For example, the 4-th population raw moment is $\int_{-\infty}^{+\infty} x^4 dF(x)$, whereas the corresponding sample moment is integrated with the empirical CDF as $\int_{-\infty}^{+\infty} x^4 d\hat{F}_n(x)$. Meanwhile, the empirical CDf $\hat{F}_n(x)$ is guaranteed to converge to the population CDF $F(x)$ almost surely by the *Strong Law of Large Numbers*.

Theorem 2.3. *(The Strong Law of Large Numbers) Let $X_1, X_2, \cdots, X_n$ be a sequence of i.i.d. samples with the sample expectation μ, then their arithmetic*

mean $\overline{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$ *converges almost surely to the expected value, that is* $P(\lim_{N\to\infty} \overline{X}_N = \mu) = 1$.

By the Strong Law of Large Numbers, we can derive the convergence of $\hat{F}_n(x)$ towards $F(x)$ as below (readers interested in the proof may refer to [10]).

Theorem 2.4. *The empirical CDF* $\hat{F}_n(x)$ *converges to the population CDF* $F(x)$ *as* $n \to \infty$ *almost surely, i.e.*

$$P(\lim_{n\to\infty} \hat{F}_n(x) = F(x)), \text{ for any } x \in \mathbb{R}. \tag{2–2}$$

As a consequence of $\hat{F}_n(x) \xrightarrow{a.s.} F(x)$, the convergence of sample moment towards the population counterpart is guaranteed, also in the *almost surely* sense. Fig. 2–2 is an illustrative example that plots three empirical CDFs constructed with different sample sizes along with the population CDF. The empirical CDF with a bigger sample size conforms better to the population CDF.

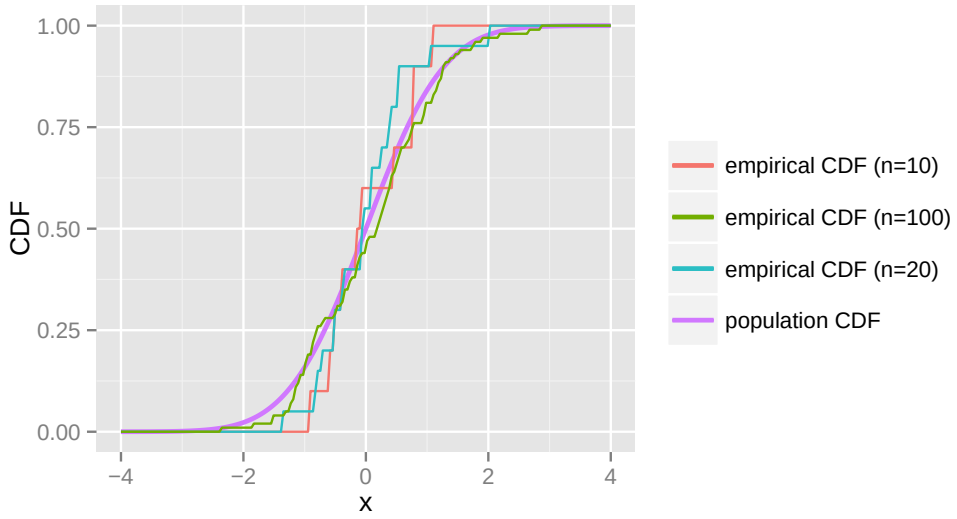

Figure 2–2 The CDF for the standard Gaussian distribution, accompanied with empirical CDF constructed from sample drawn from the distribution of sizes $n = 10$, $n = 20$ and $n = 100$. As one can observe from the trend, the gap between the population CDF and the empirical CDF diminishes.

Hence, under convergence, sample moments obtained from a sample with large size are precise estimators for the corresponding population moments, which can be further used to estimate the parameter for the population distribution, which is called *the method of moments*.

2.1.0.3 The method of moments

Supposing the population distribution for X is believe to be $f(x;\theta_1,\theta_2,\cdots,\theta_p)$, where $\theta_1,\theta_2,\cdots,\theta_p$ are the parameters to be estimated. From the distribution function, we can calculate the population moments $\mu_1',\mu_2',\cdots,\mu_p'$, each expressed as a function of $\theta_1,\theta_2,\cdots,\theta_p$, i.e.

$$\begin{aligned}
\mu_1'(\theta_1,\theta_2,\cdots,\theta_p) &= \int_{-\infty}^{+\infty} x f(x;\theta_1,\theta_2,\cdots,\theta_p)dx,\\
\mu_2'(\theta_1,\theta_2,\cdots,\theta_p) &= \int_{-\infty}^{+\infty} x^2 f(x;\theta_1,\theta_2,\cdots,\theta_p)dx,\\
&\vdots\\
\mu_p'(\theta_1,\theta_2,\cdots,\theta_p) &= \int_{-\infty}^{+\infty} x^p f(x;\theta_1,\theta_2,\cdots,\theta_p)dx.
\end{aligned} \tag{2–3}$$

Then we collect the sample moments $m_1',m_2',\cdots,m_p'$ from data with a large sample size. By equating every population with the corresponding sample moment, we have the equations

$$\begin{aligned}
\mu_1'(\theta_1,\theta_2,\cdots,\theta_p) &= m_1',\\
\mu_2'(\theta_1,\theta_2,\cdots,\theta_p) &= m_2',\\
&\vdots\\
\mu_p'(\theta_1,\theta_2,\cdots,\theta_p) &= m_p',
\end{aligned} \tag{2–4}$$

the solution to which provide an estimation for the parameters. And this method of parameter estimation is called the method of moments.

However, the method of moments have been largely superseded by the method of *maximum likelihood estimation* (MLE), which generally gives estimators with a bigger probability of being close to the parameters to be estimated. Yet, the method of moments can yield the results more quickly and easily by solving the equation, while the MLE may take more time solving an optimization.

2.2 Diverging moments caused by heavy tails

Previously, we have studied the asymptotic convergence (*asymptotic* in the sense that the sample size $n \to \infty$) relation between sample and population mo-

ments when they exist. Now it is time to move on to exploring the case when population moments diverge. We will see that the divergence of moments is caused by the heavy tails of a distribution, where the probability mass declines to zero with a pace slower than an exponential tail (e.g. the tails in Gaussian). An important member in the heavy-tailed family is the power-law distribution and we will discuss its definition, properties and empirical studies in detail. Meanwhile, it is possible for a heavy-tailed distribution has convergent lower-order moments but divergent higher-order moments. So let us first explore how the *order* of a moment matters.

2.2.1 Moment order

The theorem below states that for a distribution, if a moment of higher order exists, then every moment of lower order also exists.

Theorem 2.5. *For a random variable, if the n-th population moment μ'_n exists, then for every $0 < k \leq n$, μ'_k also exists.*

Proof. Supposing a random variable X with PDF $f(x)$, then

$$\mu'_k = \int_{-\infty}^{+\infty} x^k f(x)dx \leq \int_{-\infty}^{+\infty} |x|^k f(x)dx = \int_{|x|>1} |x|^k f(x)dx + \int_{|x|\leq 1} |x|^k f(x)dx,$$

where the first term satisfies

$$\int_{|x|>1} |x|^k f(x)dx \leq \int_{|x|>1} |x|^n f(x)dx = \int_1^{+\infty} x^n f(x)dx \pm \int_{-\infty}^{-1} x^n f(x)dx,$$

where both terms on RHS are convergent due to the existence of μ'_n. Meanwhile, it follows that

$$\int_{|x|\leq 1} |x|^k f(x)dx \leq \int_{-\infty}^{+\infty} f(x)dx = 1.$$

Therefore, μ'_k is convergent. □

More generally, the existence of μ'_n guarantees the convergence of $\mu_k(c),\ 0 < k \leq n$ for any c. Hence, for a distribution, the convergence of moments falls into either category:

1. There exists a maximum order n_c for moment to be convergent. Moments with orders less than n_c are all convergent, while moments with orders higher than n_c are all divergent.

2. Moments of all orders are convergent ($n_c = \infty$).

The convergence of moment is only a problem when a distribution has a tail (the range of distribution extends to infinity on one side) or has two tails (the range extends to infinity on both sides); distributions with x bounded to a finite range always have convergent moments. By recalling how an improper integral is defined, we can see that the convergence of a population moment solely depends on the *shape of its tail (or tails)*. For distributions with exponentially bounded tails, $n_c = \infty$; for heavy-tailed distributions, we can expect $n_c < \infty$.

2.2.2 Heavy-tailed distribution

Supposing the CDF for a function is $F(x)$, then a distribution (not limited to a finite range) is said to have an exponentially bounded *right tail* if there exists some $\lambda > 0$ such that

$$\lim_{x\to+\infty} e^{\lambda x}(1 - F(x)) < \infty, \tag{2–5}$$

or an exponentially bounded *left tail* if there exists some $\lambda > 0$ such that

$$\lim_{x\to-\infty} e^{-\lambda x}F(x) < \infty. \tag{2–6}$$

As the definitions above imply, an exponentially bounded tail vanishes to zero as fast as an exponential function with some $\lambda > 0$ or faster than that. Either left or right, exponentially bounded tails guarantee the existence of moments of all positive orders, i.e. $n_c = \infty$, which is an immediate result from that $\lim_{x\to\infty} e^{-x}x^n = 0$ holds for any n.

However, if a tail vanishes to zero slower than *any* exponential function with $\lambda > 0$, we call it a heavy-tail and it causes moments with order higher than some n_c diverge. For the sake of simplicity, we restrict tails discussed here to be *right tails*, which can be extended to left tails in a similar fashion.

Definition 2.2. The distribution of a random variable X with its CDF $F(x)$ is said to have a heavy tail if

$$\lim_{x\to+\infty} e^{\lambda x}(1 - F(x)) = \infty \quad \text{for all } \lambda > 0. \tag{2–7}$$

This definition is equivalent to the statement that the moment generating function $M(t)$ is infinite for any $t > 0$ [11].

Some common distributions with heavy-tails are listed as below.
Those that are one-tailed include

- the power-law distribution (also referred to as *Pareto distribution*),
- the log-normal distribution,
- the Levy distribution.

Those that are double-tailed include

- the Cauchy distribution,
- the t-distribution.

As a typical example of heavy-tail with both abundant empirical evidence and theoretical studies, we now briefly outline the definition and properties of *power-law distributions*, which are also referred to as *Pareto distributions*.

2.2.2.1 Power-law distribution

For the sake of simplicity, we assume the random variable here to be continuous. For a random variable X obeying power-law distribution, its PDF is in the form of

$$f(x) = \frac{\alpha - 1}{x_{\min}}(\frac{x}{x_{\min}})^{-\alpha}, \tag{2–8}$$

where x is defined on the range $[x_{\min}, +\infty)$ and $x_{\min} > 0$ is a cut-off on the minimum value for X. $\alpha > 1$ is called the *scaling exponent*, its value being bigger than 1 is required for power-law to be normalizable.

The name *scale exponent* is due to a property of power-law called *scale invariance*. Scaling the variable X by a positive constant c only causes a proportionate scaling of the PDF itself, that is

$$f(cx) = c^{-\alpha} f(x), \tag{2–9}$$

where α characterizes the power of the homogeneous function. Therefore, two power-laws with the same scaling exponent are equivalent up to a constant factor, with one being a *scaled* version of the other. To simplify the expression for the

PDF, one can scale X by dividing it by x_{min}. Then the resulting random variable $X' = X/x_{min}$ has the PDF with the same α, i.e.

$$f_{X'}(x) = (\alpha - 1)x^{-\alpha}. \tag{2–10}$$

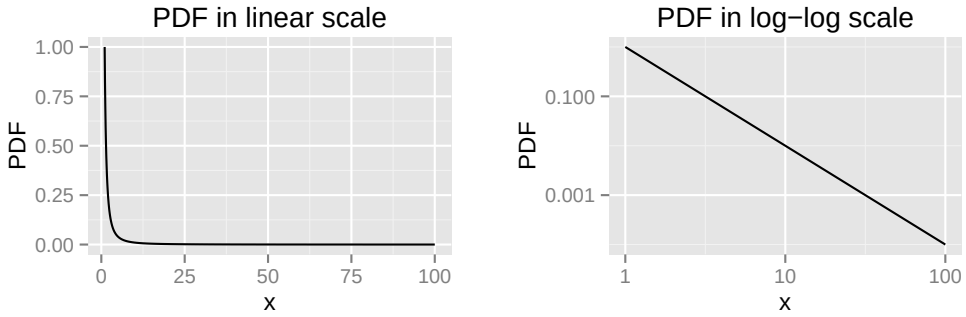

Figure 2–3 The PDF for a power-law with $x_{\min} = 1$ and $\alpha = 2$. The figure on the left in plotted in linear-linear scale and the figure on the right is plotted in log-log scale.

Fig. 2–3 shows the PDF for a power-law distribution in two drawing scales. As the tail for power-law is heavier than a exponential one, the length of the power-law tail can be very long (the maximum of x in data can be very big) compared to the length of the head, where most of the "probability mass" is concentrated. Hence, a power-law is often preferred to be drawn in a log-log scale. By taking logarithm on both sides of the PDF, we can see that the logarithm of the PDF is related to the logarithm of the variable in a linear relation, with the slope being just $-\alpha$, i.e.

$$\log f(x) = -\alpha \log(x) + \log[(\alpha - 1)x_{\min}^{\alpha-1}]. \tag{2–11}$$

Meanwhile, to see if a variable follows a power-law from data, it is often handy to plot the probability versus the value in log-log scale and see if is basically a straight line. However, without a following rigorous treatment for parameter estimation and statistical test, a naive graphical method for identifying power-law and estimating the exponent can be misleading [5].

To observe the heavy tail for power-law more straightforwardly, we can integrate the PDF to arrive at the CDF as

$$F(x) = 1 - (\frac{x}{x_{\min}})^{1-\alpha}. \tag{2–12}$$

It follows that for any $\lambda > 0$ we have

$$\lim_{x\to+\infty} e^{\lambda x}(1 - F(x)) = \lim_{x\to+\infty} \frac{e^{\lambda x}}{(x/x_{\min})^{\alpha-1}} = 0,$$

thus proving that the tail is indeed heavier than an exponential one. The *tail distribution* for power-law is given by

$$P(X > x) = 1 - F(x) = (x/x_{\min})^{1-\alpha}, \tag{2–13}$$

which is also a relation expressed by multiplicative power. This form of characterization is often seen when the power-law is called the *Pareto distribution* and the constant $\beta = \alpha - 1 > 0$ is called the *Pareto index*.

The scaling exponent α (or the Pareto index $\beta = \alpha - 1$) characterizes the *heterogeneity* of a power-law distribution. As drawn in Fig. 2–4, when $\alpha \to \infty$, the distribution approaches to a Dirac delta function, with all probability mass places on a single point, which corresponds to maximal heterogeneity; when $\alpha \to 1$, loosely speaking, the distribution approaches a "uniform", where the probability mass is equally places from $x_{\min}$ to ∞, which corresponds to maximal homogeneity.

Power-law distributions (i.e. Pareto distributions) has been empirically found in an extraordinarily diverse phenomena. Power-law is often connected with the intuition that the data is not centered around a "typical value", which is often referred to as *heterogeneity* For example, the distribution of population of cities and towns is found to follow a power-law: the largest city in US is New York City with more than 8 million residents but the smallest one in US is a small town with about 50 residents only [12]. The distribution for the population of cities from the US Census in 2000 is drawn by Fig. 2–5, where the straight line in log-log scale matches the form of power-law. Aside from city population, the distributions for the sizes of earthquakes [13], solar flares [14], the frequency of words in any human language [15], the number of citations received by papers [16], the degrees of web pages on the Internet [17, 18], the sales of books and recordings [19], people's annual incomes [20] have also been found to follow the power-law.

We now take a look at the maximum order n_c allowed in a power-law for a

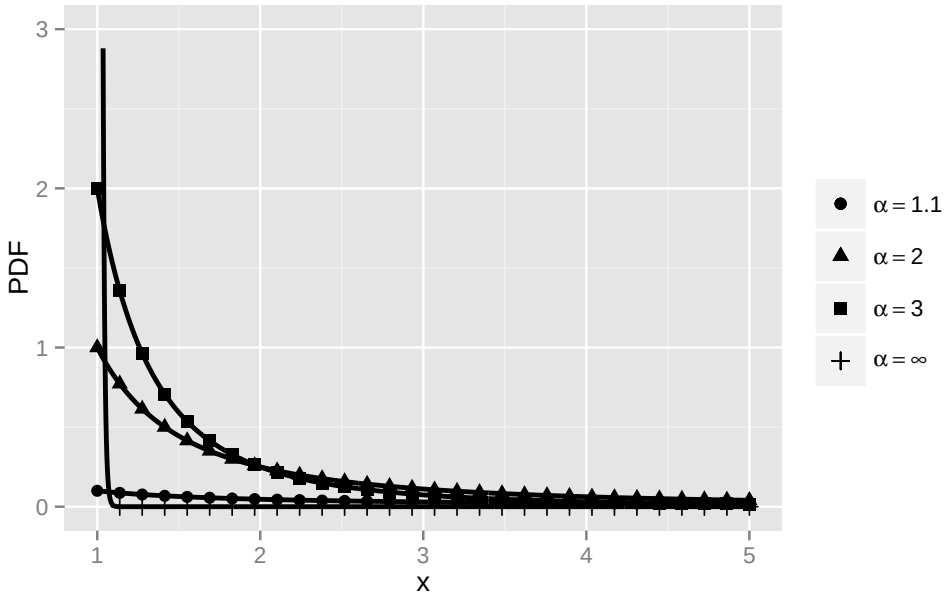

Figure 2–4 PDFs of power-law distributions with $x_{\min} = 1$ and different values of α. It can be seen that the distribution approaches an uniform distribution when $\alpha \to 1$ (more homogenous), while it approaches a Dirac delta function centered at a single point when $\alpha \to \infty$ (more heterogeneous).

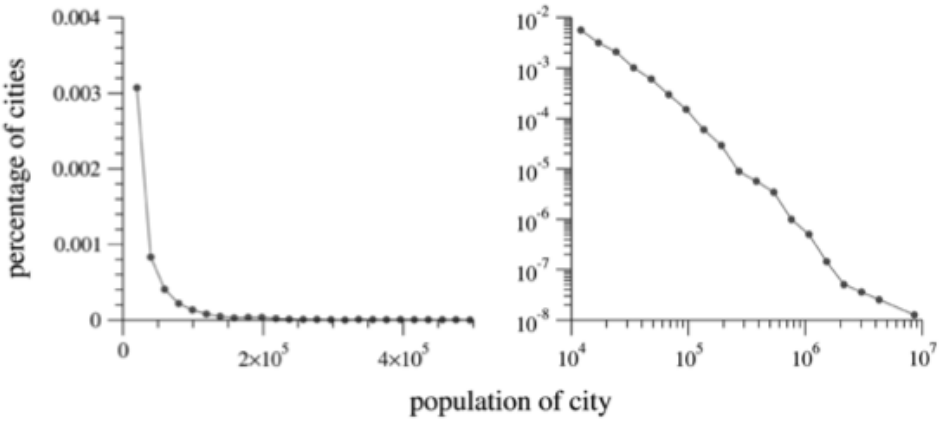

Figure 2–5 Left: histogram of the populations of all US cities with population of 10000 or more. Right: another histogram of the same data, but plotted on logarithmic scales. The approximate straight-line form of the histogram in the right panel implies that the distribution follows a power law. Data is collected from the 2000 US Census. (figure excerpted from [12]))

moment to converge. The k-th raw moment is given by

$$\begin{aligned}\mu'_k &= \int_{x_{\min}}^{+\infty} \frac{\alpha-1}{x_{\min}} \left(\frac{x}{x_{\min}}\right)^{-\alpha} x^k dx \\ &= (\alpha-1) x_{\min}^{\alpha-1} \int_{x_{\min}}^{+\infty} x^{k-\alpha} dx = (\alpha-1) x_{\min}^{\alpha-1} \frac{x^{k-(\alpha-1)}}{k-(\alpha-1)} \Bigg|_{x_{\min}}^{+\infty},\end{aligned}$$

which requires $\lim_{x\to+\infty} x^{k-(\alpha-1)}$ to converge. Therefore, we have

$$k < n_c = \alpha - 1 \tag{2–14}$$

for a power-law with scaling exponent $\alpha > 1$. Notably, when $2 < \alpha \leq 3$, the power-law has finite mean but diverging variance; when $1 < \alpha \leq 2$, even the mean diverges.

2.3 Equiprobable partition method

In this section, we will develop the *equiprobable partition method* (EPM) as a theoretical tool for analyzing the asymptotics of diverging moments. The equiprobable partition method partitions the area under the PDF into slices with equal sizes (i.e. equal probabilities) and use these slices to represent or approximate the samples. We will see that when the moment is convergent, the sample moments given by EPM is consistent with its expectation, i.e. the corresponding population moment (that is to say, EPM is precise in the convergent case); and when the moment is divergent, we can use EPM to reproduce the asymptotic behaviors of the sample moments as the sample size grows.

Before we go in depth to the method, we shall first explain what happens to sample moments when the corresponding population moments diverge and what we mean by using the word *asymptotics* for sample moments.

As an illustrative example, we draw "bags" of samples from a power-law distribution with $x_{\min} = 1$ and $\alpha = 2.5$, where the sample size n for the "bag" is growing. Fig. 2–6 draws the sample raw moments m'_1, m'_2, m'_3 and m'_4 versus size of samples for computing them. Due to $\alpha = 2.5$, the maximum order for moment to converge is $n_c = \alpha - 1 = 1.5$. Therefore, as shown in the figure, only m'_1 converges to its expectation $\frac{\alpha-1}{\alpha-2} = 3$. Other moments m'_1, m'_2 and m'_3 clearly *grows with the sample size* (although the growth may look unsteady with fluctuations, the general increasing trend is obvious by referring to the exponentially spaced numbers on the y-axis). In other words, instead of converging to a fixed value, these sample moments continue to grow as more data is accumulated. The growth of a divergent sample moment relying on the sample size n when $n \to \infty$ is what we call the *asymptotic behavior*

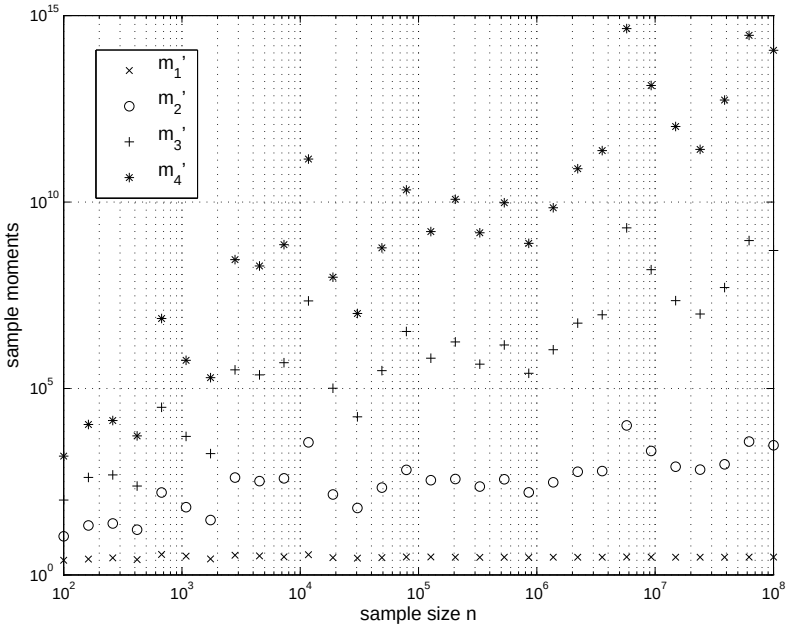

Figure 2–6 Sample raw moments m'_1, m'_2, m'_3 and m'_4 computed from samples drawn from a power-law distributions with $x_{\min} = 1$ and $\alpha = 2.5$, plotted against the sample size n in the log-log scale .

of the sample moment, or its *asymptotics*. With the aid of *equiprobable partition method*, we can approximate a diverging moment in the form of $n^\gamma g(n)$, where n^γ characterizes the *leading order* of divergence (thus answering: *how fast is the divergence?*) and $g(n) < \infty$ $(n \to \infty)$ is a function of n that characterizes the sample moment's "convergence" by cutting off the divergence (thus answering: *what is left in the moment other than the leading order?*).

2.3.1 Equiprobable partitions

Similar to the *partition* defined in Riemann integrals (or Riemann-Stieltjes integrals), a *equiprobable partition* is a list of points that cut the range for x into consecutive intervals. But, as its name suggests, the equiprobable partition is special in that it cuts the x-axis in the way that the area of the PDF above every interval is *the same*, so that a random sample falls into every interval with *equal probability*. We will discuss why we choose such a partition scheme later.

Let us denote a partition that produces n intervals with $\mathcal{P}_n$, which is called a *n-separated partition*. The partition $\mathcal{P}_n$ for a random variable defined on the interval

$[a, b]$ is defined as below.

Definition 2.3. Let X be a continuous random variable defined on the interval $[a, b] \in \mathbb{R}$ with the PDF $f(x)$ $(a \leq x \leq b)$, then its n-separated partition $\mathcal{P}_n(a, b)$ is defined as

$$\mathcal{P}_n(a, b) = (\hat{t}_0, \hat{t}_1, \cdots, \hat{t}_n), \tag{2–15}$$

where $\hat{t}_0 = a$, $\hat{t}_n = b$ and it satisfies that

$$\int_{\hat{t}_i}^{\hat{t}_{i+1}} f(x)dx = \frac{1}{n} \quad (i = 1, 2, \cdots, n-1). \tag{2–16}$$

It directly follows from the definition that the interval points can be obtained from the CDF $F(x)$ as

$$F(\hat{t}_i) = \frac{i}{n} \quad (i = 0, 1, \cdots, n) \tag{2–17}$$

If we assume that the PDF $f(x)$ is a well-defined function, then its integral $F(x)$ should be an increasing function, of which the inverse function $F^{-1}(x)$ must exist. The interval points can thus be uniquely determined by the inverse function of CDF, i.e.

$$\hat{t}_i = F^{-1}(\frac{i}{n}) \quad (i = 0, 1, \cdots, n). \tag{2–18}$$

Fig. 2–7 is a schematic illustration of a partition $\mathcal{P}_n(a, b)$ with $n = 9$. Clearly, the interval points are not evenly spaces. Instead, the density for interval points is high when $f(x)$ is large and is low when $f(x)$ is small, which is consistent with how samples are distributed in regions with different probability densities.

Hence, we can use the interval points in equiprobable partitions to *approximate sample points*. In the Riemann integral, one point picked in each interval in the partition as the *representative point* for that interval, and the partition is said to be *tagged* by these representative points. How one picks the representative points does not matter as long as the Riemann sum converges when the maximum length of intervals diminishes to zero. Similarly, in our case, to every interval we assign a representative point placed in the interval. As the positioning of this point inside the interval does not matter when $n \to \infty$, for convenience, we just pick the left end $\hat{t}_i$ for the interval $[\hat{t}_i, \hat{t}_{i+1}]$ as the *representative point* .

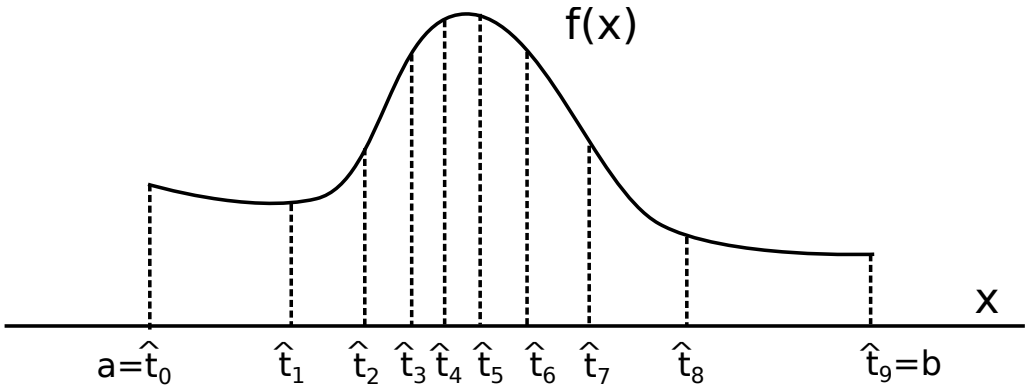

Figure 2–7 A schematic of illustration of $P_9(a, b)$. Instead, the density for interval points is high when f(x) is large and is low when f(x) is small, which is consistent with how samples are distributed in regions with different probability densities.

Now, we use n representative points $\{\hat{t}_0, \hat{t}_1, \cdots, \hat{t}_{n-1}\}$ derived from an n-separated equiprobable partition to approximate a sample $\{x_1, x_2, \cdots, x_n\}$ of size n. That is to say, the set of points $\{\hat{t}_0, \hat{t}_1, \cdots, \hat{t}_{n-1}\}$ as a whole is an approximation for the set of samples $\{x_1, x_2, \cdots, x_n\}$ as a whole, while the ordering among the elements in each does not matter.

Meanwhile, because we know that $\hat{t}_0 \leq \hat{t}_1 \leq \cdots \leq \hat{t}_{n-1}$, we first sort the samples and then use $\hat{t}_i$ as an approximation for the $(i+1)$-th smallest sample. In other words, the determinant points $(\hat{t}_0, \hat{t}_1, \cdots, \hat{t}_{n-1})$ can be used to estimate the *order statistics* for a distribution, where both the ordering among $\{\hat{t}_i\}$ and the ordering among samples after sorting shall be taken into account.

For now, we focus on the first case where ordering does not matter. By using $\{\hat{t}_0, \hat{t}_1, \cdots, \hat{t}_{n-1}\}$ to approximate the samples $\{x_1, x_2, \cdots, x_n\}$, we seek to construct estimators for sample moments.

2.3.2 EPM estimators for sample moments

Every sample moment taken at point c is a polynomial function of sample points, i.e.

$$m_k(c) = \frac{1}{n}\sum_{i=1}^{n}(x_i - c)^k. \tag{2–19}$$

The EPM (equiprobable partition method) estimator for the sample moment is constructed by simply *replacing the samples* $\{x_1, x_2, \cdots, x_n\}$ *with the representative*

points $\{\hat{t}_0, \hat{t}_1, \cdots, \hat{t}_{n-1}\}$.

Definition 2.4. Given a random variable X and a sample of size n independently drawn from its distribution, the EPM (equiprobable partition method) estimator for the k-th sample moment $m_k(c)$ is defined to be

$$\hat{m}_k(n; c) = \frac{1}{n}\sum_{i=1}^{n}(\hat{t}_{i-1} - c)^k, \tag{2–20}$$

which is a function of n. The interval points $\{\hat{t}_0, \hat{t}_1, \cdots, \hat{t}_n\}$ are given by the n-separated equiprobable partition $\mathcal{P}_n$ for X.

We now show that when X is a random variable defined on a finite interval and its PDF is bounded. Then the EPM moment estimator is *unbiased* when $n \to \infty$, that is it coincides with population moment. We will show that, in fact, an EPM moment estimator is in fact the same as the Riemann-Stieltjes integral for the corresponding population moment when n extends to infinity.

Theorem 2.6. *Let X be a random variable defined on the interval $[a, b]$ with its PDF $f(x)$ $(a \le x \le b)$. Supposing $f(x)$ is bounded, i.e. there exists some M that it holds that $\forall x \in [a, b]$, $f(x) < M$, we have*

$$\lim_{n\to\infty} \hat{m}'_k(n) = \mu'_k, \tag{2–21}$$

where $\hat{m}'_k(n) = \frac{1}{n}\sum_{i=1}^{n}(\hat{t}_{i-1})^k$ is the EPM estimator for k-th raw moment, μ'_k is the k-th population moment.

Proof. From the definition of $\mathcal{P}_n(a, b)$, we have

$$\begin{aligned}\frac{1}{n} &= F(\hat{t}_i) - F(\hat{t}_{i-1}) \quad (i = 1, 2, \cdots, n)\\ &= f(\hat{t}_{i-1})(\hat{t}_i - \hat{t}_{i-1}) + o(\hat{t}_i - \hat{t}_{i-1}).\end{aligned}$$

For an $\epsilon > 0$ that is small enough, if $\hat{t}_i - \hat{t}_{i-1} < \epsilon$, then we have $o(\hat{t}_i - \hat{t}_{i-1}) < \delta(\hat{t}_i - \hat{t}_{i-1})$ for some $\delta > 0$. Meanwhile, due to the bound $f(x) < M$, we have

$$\frac{1}{n} < (M + \delta)(\hat{t}_i - \hat{t}_{i-1}).$$

Therefore, for any small $\epsilon > 0$, when $n > \frac{1}{(M+\delta)\epsilon}$, we shall have $(\hat{t}_i - \hat{t}_{i-1}) < \epsilon$. That is,

$$\lim_{n\to\infty} \max_{1\le i\le n} (\hat{t}_i - \hat{t}_{i-1}) = 0,$$

showing that the *mesh* for the partition vanishes to zero when $n \to \infty$. Meanwhile, the EPM estimator can be rewritten as

$$\begin{aligned}\lim_{n\to\infty} \hat{m}'_k(n) &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (\hat{t}_{i-1})^k \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} (\hat{t}_{i-1})^k [F(\hat{t}_i) - F(\hat{t}_{i-1})].\end{aligned}$$

Hence, by the definition of Riemann-Stieltjes integral, we have

$$\lim_{n\to\infty} \hat{m}'_k(n) = \int_a^b x^k dF(x) = \mu'_k.$$

□

By expanding a moment taken about c, we arrive at the following conclusion immediately.

Lemma 2.2. *For a random variable X with bounded PDF $f(x) < M$ $(a \le x \le b)$, we have*

$$\lim_{n\to\infty} \hat{m}_k(n) = \mu_k, \tag{2–22}$$

$$\lim_{n\to\infty} \hat{m}_k(n; c) = \mu_k(c). \tag{2–23}$$

That is to say, the EPM estimators for sample central moments or moments taken about any point coincide with population moment, if X is a random variable with finite range and bounded PDF. To arrive at these conclusions, we have assumed that the PDF is bounded, which applies to most well-defined distributions and only rules out those with *generalized functions* (e.g. PDFs containing Dirac's delta).

2.3.3 Asymptotics for diverging moments

Having proven that EPM estimators are *precise* when moments converge, we now extend the usage of EPM estimators to deriving the asymptotics of diverging sample moments. For convenience, in this section we assume that X is a random

variable *with a heavy tail on the right side*, while right-tailed or double-tailed distributions can be analyzed in a similar fashion. Let $f(x)$ $(a \leq x < +\infty)$ be the PDF for X and $f(x)$ is bounded by some M. The equiprobable partition for X is defined by changing $\mathcal{P}_n(a, b)$ to $\mathcal{P}_n(a, +\infty)$, that is

$$\mathcal{P}_n(a, +\infty) = (\hat{t}_0, \hat{t}_1, \cdots, \hat{t}_{n-1}, \hat{t}_n), \tag{2–24}$$

where $\hat{t}_0 = a$ and $\hat{t}_n = +\infty$. The representative points $\{\hat{t}_0, \hat{t}_1, \cdots, \hat{t}_{n-1}\}$ are given by

$$F(\hat{t}_i) = \frac{i}{n} \quad (i = 0, 1, \cdots, n-1), \tag{2–25}$$

where it shall be noted that the probability mass from the maximum representative to infinity is also $1/n$. That is to say, the EPM imposes an upper cut-off on approximated samples and the probability of drawing a random sample beyond the cut-off is $1/n$, which vanishes to zero when $n \to \infty$.

The k-th sample moment $m_k(c)$ computed from a sample of size n is approximated by the EPM estimator $\hat{m}_k(n; c)$, which is constructed by substituting the sample points with representative points from the n-separated equiprobable partition, i.e.

$$\hat{m}_k(n; c) = \frac{1}{n} \sum_{i=1}^{n} (\hat{t}_{i-1} - c)^k. \tag{2–26}$$

Then, we rewrite $\hat{m}_k(n)$ (c is left out for simplicity) into the form of

$$\hat{m}_k(n) = n^\gamma g(n), \tag{2–27}$$

where n^γ is the *leading order* that characterizes the speed of convergence. And $g(n)$ is a function that satisfies

$$0 < \lim_{n \to \infty} |g(n)| < \infty, \tag{2–28}$$

which is the "convergent term" for sample moments formed by deducting the effect of divergence from the sample moment.

To showcase the analytical procedure above, we use this machinery to find the asymptotics for the first several diverging moments of the power-law distributions with $\alpha < 3$ and compare them to numerical results.

Here we assume the lower cut-off $x_{\min} = 1$, and the PDF is given by

$$f(x) = (\alpha - 1)x^{-\alpha}, \tag{2–29}$$

and the CDF is given by

$$F(x) = 1 - x^{1-\alpha}. \tag{2–30}$$

We use the CDF to construct the n-separated equiprobable partition $\mathcal{P}_n(1, +\infty)$. From the relation

$$F(\hat{t}_i) = 1 - (\hat{t}_i)^{1-\alpha} = \frac{i}{n} \quad (i = 0, 1, \cdots, n-1), \tag{2–31}$$

we have the representative point as

$$\hat{t}_i = (1 - \frac{i}{n})^{-c}, \tag{2–32}$$

where we adopt a shorthand $c = \frac{1}{\alpha-1} > \frac{1}{2}$. When $\alpha \leq 3$, the maximum order for convergent moment is $n_c = \alpha - 1 < 2$, here we present the asymptotics for diverging sample raw moments m'_k with $k \geq 2$.

The EPM estimator for m'_k is

$$\hat{m}'_k(n) = \frac{1}{n}\sum_{i=0}^{n-1}(\hat{t}_i)^k = \frac{1}{n}\sum_{i=0}^{n-1}(1 - \frac{i}{n})^{-ck},$$

which can be rewritten as the product of a leading order term and a "convergent" term, i.e.

$$\hat{m}'_k(n) = \frac{1}{n}\sum_{i=0}^{n-1} n^{ck}(n-i)^{-ck} = n^{ck-1}\sum_{i=1}^{n}\frac{1}{i^{ck}}. \tag{2–33}$$

Therefore, the leading order for the k-th diverging raw moment is n^{ck-1} $(ck > 1)$, and a higher order k means a faster divergence with n. The "convergent" term is

$$g(n) = \sum_{i=1}^{n}\frac{1}{i^{ck}} \quad (ck > 1), \tag{2–34}$$

and its limit is the famous *Riemann zeta function*

$$\lim_{n\to\infty} g(n) = \zeta(ck). \tag{2–35}$$

Therefore in limit of $n \to \infty$, the EPM asymptotics for sample moments are given by

$$\hat{m}'_k(n) \approx n^{ck-1}\zeta(ck). \tag{2–36}$$

In Fig. 2–8, we compare the sample moments computed from numerical experiments with their asymptotics given by EPM for $k = 2, 3, 4$. That the diverging

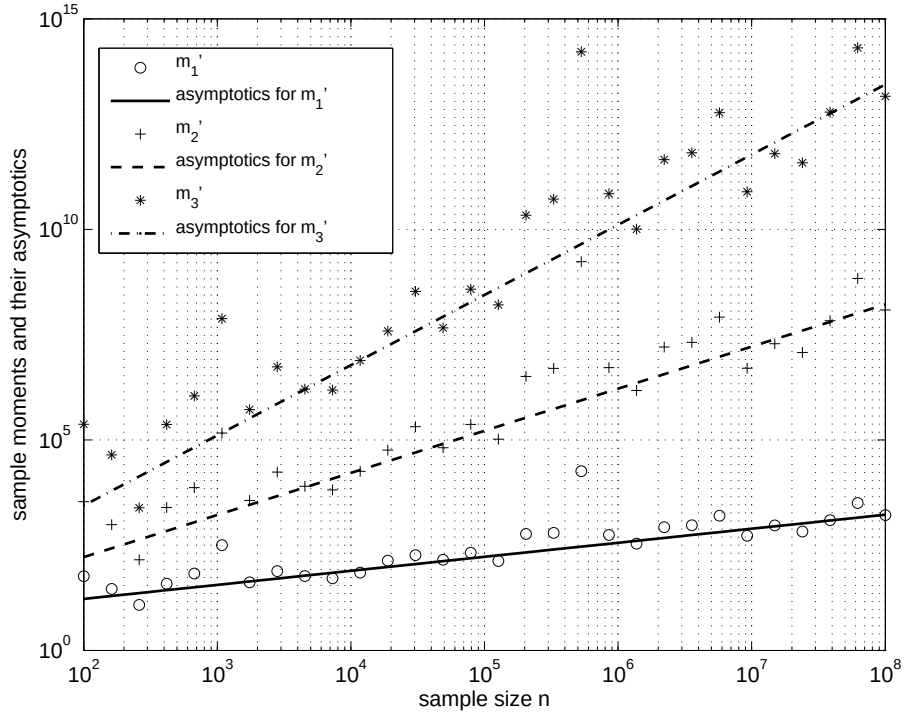

Figure 2–8 The comparison between numerical sample moments $\hat{m}_3$, $\hat{m}_4$ $\hat{m}_5$ (cross) and their EPM asymptotics (line). The divergence can be effectively characterized by the asymptotics. Sample moments fluctuate around the theoretical line due to randomness.

sample moments fluctuate around their asymptotics shows that the equiprobable partition method is valid in two aspects: the leading order is correct, so that the slope matches the speed of divergence with different orders; the remaining "convergent" term is correct, so that the intercept match the positioning of points.

2.3.4 EPM estimators for order statistics

As mentioned earlier, the ordering among representative points given by EPM can also be utilized to construct estimators for order statistics, i.e. the sorted samples.

Definition 2.5. Given X as a random variable and supposing $X_1, X_2, \cdots, X_n$ are n i.i.d. samples drawn from the distribution, then the *order statistics* $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ for the samples are defined by sorting the samples in the increasing order. That is

$X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ is a permutation of the samples that satisfies

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}. \tag{2–37}$$

The representative points given by an n-seperated equiprobable partition naturally satisfies

$$t_0 < t_1 < \cdots < t_{n-1}. \tag{2–38}$$

Therefore, we can use the i-th smallest EPM representative point to estimate the i-th smallest sample, i.e.

$$\hat{X}_{(i)} = t_{i-1} \quad (i = 1, 2, \cdots, n). \tag{2–39}$$

For most cases in asymptotic analysis for sample moments, it is *unnecessary* to use the estimators for order statistics, as sample moments are defined to be *permutational invariant*. For example, the variance is the same for samples X_1, X_2, X_3 and X_3, X_2, X_1, where the ordering among samples does not matter.

Yet, the EPM estimators for order statistics is useful when addressing the statistics where the ordering among elements is concerned, e.g. those in *time series*. In the following Chapter, we will see a scenario that applies EPM estimators to obtaining the bounds of a statistic with regard to all permutations of the time series.

2.3.5 Why equiprobable

We have introduced the equiprobable partition method and used it as an estimation for samples and order statistics. The reader may be curious about the reason behind the principle of cutting the probability mass into slices of *equal probabilities*, which plays a central role in the magic of EPM. Now, instead of a rigorous argument towards its mathematical necessity, we just briefly outline the intuitive *rationale* behind this principle.

1. A equiprobable partition ensures that the *mesh* of the partition vanishes to zero as the number of partitions increases.

   This has been proven by Theorem 2.6 and is important to guarantee that the estimator is unbiased when the moment converges. Of course, it is possible to propose other schemes of partitioning with the same mesh-vanishing property.

2. The equiprobable partition reaches maximum entropy.

   Let us view a partition as a way to *discretize* a continuous distribution. Then, for a n-separated partition, there are n outcomes with one outcome corresponding to one interval. Let p_i $(i = 1, 2, \cdots, n)$ be the probability of a random sample falls into the i-th interval, then the entropy for this discrete distribution would be

   $$H = -\sum_{i=1}^{n} p_i \log(p_i). \tag{2–40}$$

   And the configuration $p_1 = p_2 = \cdots = p_n = \frac{1}{n}$, as used in EPM, is the only one that maximizes the entropy. The principle of maximum entropy has been applied in statistics, information theory and statistical mechanics and reader may refer to the discussions in [21–23] by Jaynes.

3. It is easy to determine interval ends in the equiprobable partition.

   For most well-defined distributions, the interval ends and therefore the representative points can be easily and uniquely determined with $F^{-1}(x)$. This is important in practical applications.

Chapter 3 Memory Constraints for Power-law Series

In this Chapter, we will present an application of EPM (*equiprobable partition method*) to time series analysis. With a permutational technique, we will explore the bounds for the 1st-order autocorrelation for series with different (marginal) distributions, including Gaussian, uniform and power-law.

The 1st-order autocorrelation is a moment defined for multiple variables for characterizing the interdependence among them. Specifically, it can be used to measure the temporal memory effect for time series and it is therefore also referred to as *memory* by some physicists. We found that, very interestingly, for series that follow Gaussian and uniform distributions, the bounds for memory are just the bounds allowed by its definition (natural bounds). However, interestingly, the memory for power-law series is constrained to a much narrower region that relies on the scaling exponent. With EPM, we obtained these bounds when the sample moments involved are divergent, a case that cannot be readily tackled with traditional probabilistic methods.

3.1 Memory for time series

From a statistical point of view, a time series $\{t_1, t_2, \cdots, t_n\}$ is a ordered sequence of random variables. Ideally, the distribution for a time series of length n can be fully characterizes by the joint distribution $f(t_1, t_2, \cdots, t_n)$. However, for a long series, the space for this distribution is very huge that it can neither be effectively determined or compactly represented. Therefore, we split the characterization for the distribution into two aspects — the marginal and the interdependence.

The marginal distribution refers to how one element is distributed regardless of other elements in the series, denoted by $f(t_1)$, $f(t_2)$, $\cdots$, $f(t_n)$. Practically, it is often assumed that the marginal is *stable*, i.e. the marginal distribution for every element is the same and does not change with time. Hence, for example, we can plot all the elements in the series in a histogram and by its distinguished bell curve

we learn the marginal is Gaussian. When we speak of Gaussian series or power-law series, we are referring to the marginal distribution.

Aside from the marginal distribution, the elements in the series are intercorrelated by themselves. That is to say, a time series is more than a collection of samples independently drawn from the same marginal. Instead, their temporal ordering affects what values the elements may take. For example, a rising or falling trend is often observed for time series in financial markets. And memory (1st-order autocorrelation) is introduced as a simplest measure to characterize the interdependence among these elements [24].

For a time series $\{t_1, t_2, \cdots, t_n\}$, memory is defined as the Pearson's correlation coefficient between $\{t_2, t_3, \cdots, t_n\}$ and its lag-1 counterpart $\{t_1, t_2, \cdots, t_{n-1}\}$, i.e.

$$M = \frac{1}{n-1} \sum_{i=1}^{n-1} \frac{(t_i - m_1)(t_{i+1} - m_2)}{\sigma_1 \sigma_2}, \tag{3–1}$$

where m_1, m_2 and σ_1, σ_2 refer to the mean and standard deviation of the two series. Intuitively from the definition, memory describes the temporal *consistency* of a series: if a big (compared to average) element tends to follow a big element and a small one tends to follow a small one, then the series is *consistent* and its memory is positive; on the contrary, if a big one tends to follow a small one and a small one tends to follow a big one, then the series is *inconsistent* and its memory is negative.

Due to Cauchy-Schwartz inequality

$$|M| \leq \frac{1}{\sigma_1 \sigma_2} \sqrt{\left[\frac{1}{n-1} \sum_{i=1}^{n-1} (t_i - m_1)^2\right] \left[\frac{1}{n-1} \sum_{i=1}^{n-1} (t_{i+1} - m_2)^2\right]} = 1, \tag{3–2}$$

the natural bounds $-1 \leq M \leq 1$ hold for all series. The upper extreme $+1$ marks the maximum consistency and the minimum -1 marks the power extreme -1 marks the minimum consistency. And $M = 0$ suggests that the series is neither consistent or inconsistent — the series is similar to independently samples in this regard and the short-range interdependence among elements is weak.

3.2 Modeling correlated series with a specified marginal

Clearly, when every t_i is independently sampled from the same marginal distribution $f(x)$, there is no interdependence among elements and the expectation value of M is zero. To modify the interdependence structure while preserving the marginal, we first sort the independently sampled series and then reorder the samples by a permutation θ with regard to every element's ranking in the sorted series. The sorted series directly give us the order statistics $\{t_{(i)}\}$, where $0 < t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)}$. Then, by applying a permutation θ (a one-to-one mapping from the set $\{1, 2, \cdots, n\}$ to itself) to $\{t_{(i)}\}$, we now have a sequence $\{t_{(\theta_i)}\}$ with a different interdependence structure but the same marginal marginal distribution.

For a series samples from a specified marginal, by changing θ, i.e. the way we permute them, we can expect series produced with different values of memory M. For example, if series are permuted such that big elements tend to be followed by big ones and small followed by small ones, M would be positive; on the contrary, if big elements followed by small ones and small elements followed by big ones, M would be negative.

3.2.1 Permutational extremes for memory

Interestingly, there exist fixed $\theta_{\max}$ and $\theta_{\min}$ that respectively maximizes and minimizes M among all possible permutations. And we can use these two extremes to derive the bounds for memory in the sense of all permuted independent samples.

To see this, we need to find out how a permutation affects M. As the values of the two aforementioned series are different in only one element (t_1 in head and t_n in tail), we assume $m_1 = m_2 = m$ and $\sigma_1 = \sigma_2 = \sigma$ when n is large, where m and σ are the mean and standard deviation of the whole series. Then the memory of $\{t_{(\theta_i)}\}$ can be rewritten as

$$M = \frac{1}{\sigma^2}\Big(\frac{1}{n-1}\sum_{i=1}^{n-1} t_{(\theta_i)} t_{(\theta_{i+1})} - m^2\Big), \tag{3–3}$$

where the reordering of the series only affects the summed products of adjacent terms $S_\theta = \sum_{i=1}^{n-1} t_{(\theta_i)} t_{(\theta_{i+1})}$, while leaving m and σ unchanged. The desired extreme

permutations for M are just those maximize/minimizes S_θ, denoted by $\theta_{\max}$ and $\theta_{\min}$. It has been shown that, for any n, there are fixed $\theta_{\max}$ and $\theta_{\min}$ for any real numbers $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)}$ [25] [1].

$\theta_{\max}$ obtains the maximum memory by first arranging the odd elements of order statistics in the increasing order, followed by even elements in the decreasing order, which is

$$\begin{aligned} &t_{(1)}, t_{(3)}, \cdots, t_{(2l-1)}, t_{(2l)}, t_{(2l-2)}, \cdots, t_{(4)}, t_{(2)} \quad (n = 2l), \\ &t_{(1)}, t_{(3)}, \cdots, t_{(2l-1)}, t_{(2l+1)}, t_{(2l)}, \cdots, t_{(4)}, t_{(2)} \quad (n = 2l+1). \end{aligned} \tag{3–4}$$

For simplicity, we only address the case when $n = 2l$, the sum with order statistics is expressed as

$$S_{\theta_{\max}} = \sum_{i=1}^{2l-2} t_{(i)} t_{(i+2)} + t_{(2l)} t_{(2l-1)} \quad (n = 2l), \tag{3–5}$$

while the case of odd n can be handled similarly.

On the contrary, $\theta_{\min}$ arranges the order statistics by alternating small and big terms, i.e.

$$\begin{aligned} &t_{(2l)}, t_{(1)}, t_{(2l-2)}, \cdots, t_{(2l-3)}, t_{(2)}, t_{(2l-1)} \quad (n = 2l), \\ &t_{(2l)}, t_{(2)}, \cdots, t_{(l)}, \cdots, t_{(1)}, t_{(2l+1)} \quad (n = 2l+1). \end{aligned} \tag{3–6}$$

Again, for even n, we have

$$S_{\theta_{\min}} = \sum_{i=1}^{l-1} \left(t_{(i)} t_{(2l+1-i)} + t_{(i)} t_{(2l-1-i)} \right) + t_{(l)} t_{(l+1)} \quad (n = 2l). \tag{3–7}$$

We then define the upper bound and lower bound for memory as the mathematical expectation of M under $\theta_{\max}$ and $\theta_{\min}$ respectively. And in the limit of large length of series, the bounds

$$M_{\max} = \lim_{n \to \infty} \mathrm{E}\left[M(t_{\theta_{\max}}) \right] \tag{3–8}$$

and

$$M_{\min} = \lim_{n \to \infty} \mathrm{E}\left[M(t_{\theta_{\min}}) \right] \tag{3–9}$$

[1] Whereas [25] uses an objective function that sums the products in a circle, i.e. $S'_\theta = \sum_{i=1}^{n-1} t_{\theta_i} t_{\theta_{i+1}} + t_{\theta_1} t_{\theta_n}$, the results can be reduced to our case by introducing an additional $S_0 = 0$ to the series, which makes zero contribution to the sum.

can be derived in a closed form or effectively approximated. It can be shown that $M_{\max} = 1$ and $M_{\min} = -1$ hold for Gaussian and uniform distributions, corresponding the natural range of M. However, a much narrower range is found for power-law distributions, where the bounds rely on the exponent α. It shall be noted that the bounds discussed here are tight as they are the values of M under permutations $\theta_{\max}$ and $\theta_{\min}$.

3.3 The bounds for memory

3.3.1 Adjusting memory by iterative rearrangement

Before an in-depth theoretical treatment, we first explore the bounds for uniform, Gaussian and power-law samples by iteratively rearranging them towards $\theta_{\max}$ or $\theta_{\min}$.

To do this, we construct an approximation of the order statistic iteratively. The process contains n steps, the same as the length of the series. In each step, the series is rearranged with one pass of bubble sort on the series produced after the last step, which means stepping through the series from the first element to the last, comparing each pair of adjacent elements and swap them if they are not in increasing order. The series obtained after i such steps is denoted as $\{\tilde{t}^{(i)}\}$, as an approximation to the order statistics, with $\{\tilde{t}^{(n)}\}$ guaranteed to be the same as the order statistics because a bubble sort always finishes in n passes.

In each step, treating $\{\tilde{t}^{(i)}\}$ as approximate order statistics, series with positive and negative memory are obtained by rearranging $\{\tilde{t}^{(i)}\}$ according to (3–4) and (3–6) respectively.

In our experiment, series are of length $n = 10,000$, independently drawn from uniform distribution on $[0, 1]$, standard Gaussian distribution (all samples are added by the same positive constant afterwards to ensure being positive), and power-law distribution whose probability density function is

$$f(t) = \frac{\alpha - 1}{t_{\min}} (\frac{t}{t_{\min}})^{-\alpha}, \tag{3–10}$$

with $t_{\min} = 1$ and $\alpha = 3.5$. Results are obtained by averaging 5 independent

runs and reported by Figure 3–1 for positive memory and Figure 3–2 for negative memory.

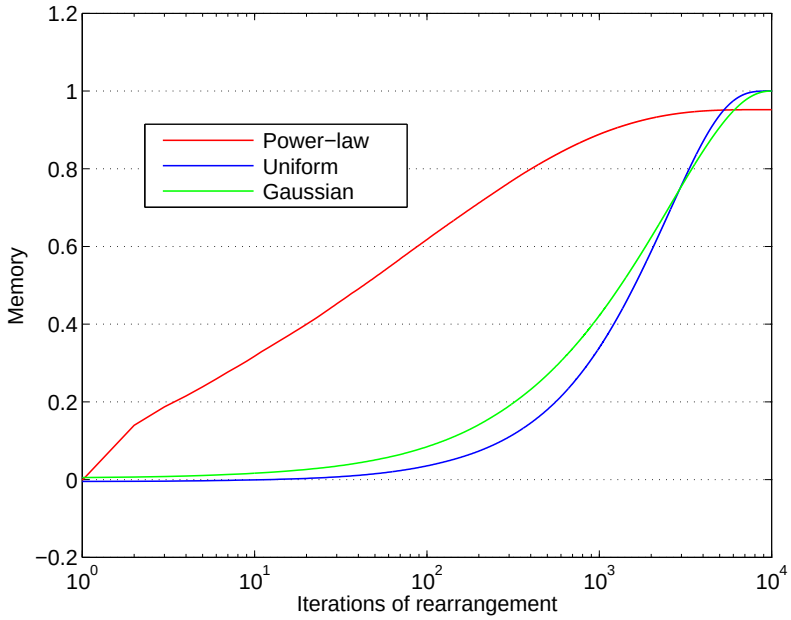

Figure 3–1 Memory tuned up by iteratively rearranging independently sampled series.

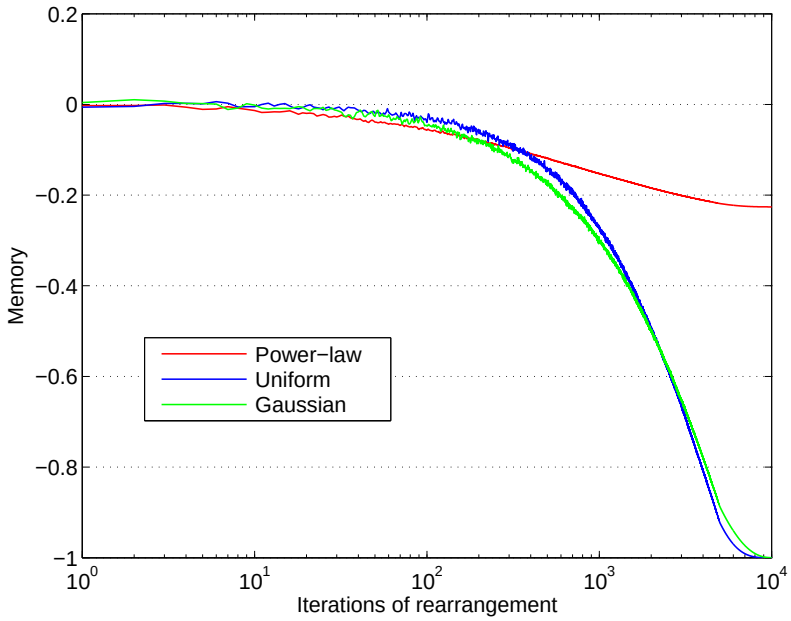

Figure 3–2 Memory tuned down by iteratively rearranging independently sampled series.

As can be seen, by rearranging the series towards $\theta_{\max}$, the memory goes to $+1$ for all three distributions. As is noted, the maximum memory of power-law series after the last iteration is slightly below $+1$, due to the effect of finite n, which we will see later. However, while the memory can be tuned down to -1 by rearrangement

towards $\theta_{\min}$ for uniform and Gaussian distributed series, it is bounded above about -0.2 for the power-law series. In the following sections, we will further analyze how this non-trivial lower bound depends on the exponent α.

3.3.2 Probablistic method for the $\alpha > 3$ case

To derive how the bounds rely on α, we first consider the case where $\alpha > 3$, which is necessary for the population variance $\sigma(\alpha)^2$ to converge, and the corresponding sample variance σ^2 appear in the denominator of M. When $\alpha > 3$, by simply equating the sample moments with the corresponding population moments, we have

$$\mathrm{E}[M(t_\theta)] = \frac{1}{\sigma(\alpha)^2}\left(\frac{1}{n-1}\sum_{i=1}^{n-1}\mathrm{E}[t_{\theta_i}t_{\theta_{i+1}}] - m(\alpha)^2\right), \tag{3–11}$$

where

$$m(\alpha) = \int_1^{+\infty} x f(x)dx = \frac{\alpha-1}{\alpha-2}, \tag{3–12}$$

$$\sigma(\alpha)^2 = \int_1^{+\infty} x^2 f(x)dx - m(\alpha)^2 = \frac{\alpha-1}{\alpha-3} - m(\alpha)^2. \tag{3–13}$$

Here, $f(x) = (\alpha-1)x^{-\alpha}$ is the PDF for power-law, and without loss of generality, we assume $x_{\min} = 1$ because M would remain the same if every t_i is divided by the same constant.

The upper bound directly relates to $\lim_{n\to\infty}\frac{1}{n-1}E[S_{\theta_{\max}}]$, where the expected value of $S_{\theta_{\max}}$ is composed of the expected value of products of adjacent items according to equation (3–5), i.e.

$$\frac{1}{n-1}\,\mathrm{E}[S_{\theta_{\max}}] = \frac{1}{2l-1}\sum_{i=1}^{2l-2}\mathrm{E}[t_{(i)}t_{(i+2)}] + \frac{1}{2l-1}\,\mathrm{E}[t_{(2l)}t_{(2l-1)}], \tag{3–14}$$

assuming $n = 2l$. The case when $n = 2l+1$ can be worked out in a similar fashion to arrive at the same results.

The expected value of each term can be obtained by using the joint distribution of order statistics. The probability density function for the joint distribution of two order statistics $t_{(j)}, t_{(k)}$ $(j < k)$ is given by [26]

$$f_{t_{(j)},t_{(k)}}(x,y) = n!\frac{[F(x)]^{j-1}}{(j-1)!}\frac{[F(y)-F(x)]^{k-1-j}}{(k-1-j)!}\frac{[1-F(y)]^{n-k}}{(n-k)!}f(x)f(y) \quad (x \le y), \tag{3–15}$$

where $F(x)$ is the cumulative distribution function for power-law, i.e.

$$F(x) = \int_x^{+\infty} f(u)du = 1 - x^{1-\alpha}. \tag{3–16}$$

Therefore, we have

$$\begin{aligned}\mathrm{E}[t_{(i)}t_{(i+2)}] &= \iint_{1\le x\le y<\infty} xy f_{t_{(i)},t_{(i+2)}}(x,y)dxdy \\ &= \frac{\Gamma(2l+1)}{\Gamma(2l+1-2c)}\frac{\Gamma(2l-i+1-2c)}{\Gamma(2l-i-1)}\frac{1}{[(2l-i)-c][(2l-i)-\alpha c]} \quad (1\le i\le 2l-2),\end{aligned} \tag{3–17}$$

and

$$\begin{aligned}\mathrm{E}[t_{(2l-1)}t_{(2l)}] &= \iint_{1\le x\le y<\infty} xy f_{t_{(2l-1)},t_{(2l)}}(x,y)dxdy \\ &= \frac{(\alpha-1)^2}{2(\alpha-2)^2}\Gamma(3-2c)\frac{\Gamma(2l+1)}{\Gamma(2l+1-2c)},\end{aligned} \tag{3–18}$$

where we use a shorthand $c = \frac{1}{\alpha-1}$ $(0 < c < \frac{1}{2})$.

In the limit of $n \to \infty$, the first term in (3–14) would be

$$\begin{aligned}&\lim_{l\to\infty}\frac{1}{2l-1}\sum_{i=1}^{2l-2}E[t_{(i)}t_{(i+2)}] \\ =&\lim_{l\to\infty}\frac{1}{2l-1}\frac{\Gamma(2l+1)}{\Gamma(2l+1-2c)}\sum_{i=1}^{2l-2}\frac{1}{[2l-i-2c][2l-i-\alpha c]}\frac{\Gamma(2l-i+1-2c)}{\Gamma(2l-i-1)}.\end{aligned} \tag{3–19}$$

Reducing the prefactor by using a property of Gamma function, namely

$$\lim_{x\to\infty}\frac{\Gamma(x)}{\Gamma(x+\gamma)x^\gamma} = 1 \text{ for real } \gamma, \tag{3–20}$$

and rewriting the summation with index $k = 2l - i$, we have

$$\lim_{l\to\infty}\frac{1}{2l-1}\sum_{i=1}^{2l-2}E[t_{(i)}t_{(i+2)}] = \lim_{l\to\infty}(2l+1)^{-(1-2c)}\sum_{k=2}^{2l-1}\frac{1}{(k-c)(k-c-1)}\frac{\Gamma(k+1-2c)}{\Gamma(k-1)}. \tag{3–21}$$

As $c < \frac{1}{2}$, the prefactor approaches zero when $n \to \infty$, making the result depends on the limiting behavior of the summing terms only. Now we have

$$\lim_{l\to\infty}\frac{1}{2l-1}\sum_{i=1}^{2l-2}E[t_{(i)}t_{(i+2)}] = \lim_{l\to\infty}\sum_{k=2}^{2l-1}(\frac{k}{2l+1})^{-2c}\frac{1}{2l+1} = \int_0^1 t^{-2c}dt = \frac{\alpha-1}{\alpha-3}. \tag{3–22}$$

Meanwhile, the second term in the RHS of (3–14) vanishes in the limit of large n, i.e.

$$\lim_{l\to\infty} \frac{1}{2l-1} \mathrm{E}[t_{(2l)} t_{(2l-1)}] = 0. \tag{3–23}$$

Substituting (3–22) and (3–23) into (3–11), we arrive at

$$M_{\max} = \lim_{n\to\infty} \mathrm{E}[M_{\theta_{\max}}] = 1 \quad (\alpha > 3). \tag{3–24}$$

Similarly, to obtain the lower bounds, we rewrite the summed products as

$$\frac{1}{n-1} \mathrm{E}[S_{\theta_{\min}}] = \frac{1}{2l-1} \sum_{i=1}^{l-1} \left(\mathrm{E}[t_{(i)} t_{(2l+1-i)}] + \mathrm{E}[t_{(i)} t_{(2l-1-i)}] \right) + \frac{1}{2l-1} \mathrm{E}[t_{(l)} t_{(l+1)}], \tag{3–25}$$

again assuming $n = 2l$ for convenience. Taking the large n limit, we have

$$\begin{aligned}
&\lim_{l\to\infty} \frac{1}{2l-1} \sum_{i=1}^{l-1} \mathrm{E}[t_{(i)} t_{(2l+1-i)}] \\
&= \lim_{l\to\infty} \frac{1}{2l-1} \frac{\Gamma(2l+1)}{\Gamma(2l+1-2c)} \sum_{i=1}^{l-1} \frac{\Gamma(i+2-c)}{\Gamma(i+2)} \frac{\Gamma(2l-i+1-2c)}{\Gamma(2l-i+1-c)} \\
&= \lim_{l\to\infty} (2l+1)^{-(1-2c)} \sum_{i=1}^{l-1} (i+2)^{-c} (2l-i+1)^{-c} \\
&= \lim_{l\to\infty} \sum_{i=1}^{l-1} (\frac{i+2}{2l+1})^{-c} (1 - \frac{i}{2l+1})^{-c} (\frac{1}{2l+1}) \\
&= \int_0^{\frac{1}{2}} u^{-c} (1-u)^{-c} du \\
&= B(\frac{1}{2}; \frac{\alpha-2}{\alpha-1}, \frac{\alpha-2}{\alpha-1}),
\end{aligned} \tag{3–26}$$

where B is the *incomplete beta function*, defined as

$$B(x; a, b) = \int_0^x t^{a-1} (1-t)^{b-1} dt. \tag{3–27}$$

We also have the other term with the same limit, namely

$$\lim_{l\to\infty} \frac{1}{2l-1} \sum_{i=1}^{l-1} \mathrm{E}[t_{(i)} t_{(2l-1-i)}] = B(\frac{1}{2}; \frac{\alpha-2}{\alpha-1}, \frac{\alpha-2}{\alpha-1}), \tag{3–28}$$

and again the remaining term vanishes in the limit, i.e.

$$\lim_{l\to\infty} \frac{1}{2l-1} \mathrm{E}[t_{(l)} t_{(l+1)}] = 0. \tag{3–29}$$

Therefore, the minimum memory is derived as

$$M_{\min} = \lim_{n\to\infty} \mathrm{E}[M_{\theta_{\min}}] = \frac{1}{\sigma(\alpha)^2}\left[2B\left(\frac{1}{2}; \frac{1}{m(\alpha)}, \frac{1}{m(\alpha)}\right) - m(\alpha)^2\right] \quad (\alpha > 3). \tag{3–30}$$

3.3.3 EPM approximations for the $\alpha < 3$ case

Having obtained the upper bound and the lower bound for $\alpha > 3$, we now turn to the case for $1 < \alpha \leq 3$ ($\alpha > 1$ is necessary for power-law to be normalized). In this case, the corresponding population moment for σ^2 would diverge ($m(\alpha)$ also diverges when $\alpha < 2$), rendering the moment-substitution technique infeasible. Meanwhile, the probability distributions of $M(t_{\theta_{\max}})$ and $M(t_{\theta_{\min}})$ themselves are intractable. Therefore, we present an approximation method that recovers the asymptotic behavior of statistics when $n \to \infty$ by substituting random variables with determinant values.

To do so, we pick the points $\{\hat{t}_1, \hat{t}_2, \cdots, \hat{t}_n\}$ that cut the area under the probability density function $f(t)$ into slices of equal area $\frac{1}{n}$, with $\hat{t}_1 = x_{\min} = 1$ and the area from t_n extending to infinity also being $\frac{1}{n}$. Then we approximate the random samples $\{t_{(1)}, t_{(2)}, \cdots, t_{(n)}\}$ with these determinant points $\{\hat{t}_1, \hat{t}_2, \cdots, \hat{t}_n\}$. It should be noted that such approximation imposes a cut-off on the maximum value of $\{t_i\}$ and the probability of drawing a sample exceeding the cut-off is $\frac{1}{n}$, which diminishes to zero as $n \to \infty$. By setting $\int_1^{\hat{t}_i} = \frac{i-1}{n}$, we have

$$\hat{t}_i = (1 - \frac{i-1}{n})^{-c} \quad (i = 1, 2, \cdots, n), \tag{3–31}$$

where $c = \frac{1}{\alpha - 1} > \frac{1}{2}$ in this case.

Rewriting M in terms of samples as

$$M = \frac{s - m^2}{\overline{t^2} - m^2}, \tag{3–32}$$

where $s = \frac{1}{n-1}S = \frac{1}{n-1}\sum_{i=1}^{n-1} t_i t_{i+1}$, $\overline{t^2} = \frac{1}{n}\sum_{i=1}^{n} t_i^2$ and $m^2 = (\frac{1}{n}\sum_{i=1}^{n} t_i)^2$ are the statistics in concern. By substituting $t_{(i)}$ with $\hat{t}_i$, we seek approximations for these statistics in the form of $n^{\mu(\alpha)}g(n, \alpha)$, where when $n \to \infty$, $g(n, \alpha)$ converges to a function of α while the divergence is characterized by the term $n^{\mu(\alpha)}$.

Then, for the upper bound, we have

$$\hat{s}_{\theta_{\max}} = \frac{n^{2c}}{n-1}\Big[\sum_{k=2}^{n-1}(k^2-1)^{-c} + 2^{-c}\Big] \tag{3–33}$$

and

$$\overline{\hat{t}^2}_{\theta_{\max}} = \frac{1}{n}\sum_{i=1}^{n}\hat{t}_i^2 = n^{2c-1}\sum_{k=0}^{n-1}(k+1)^{-2c}, \tag{3–34}$$

both of which diverge with the order of n^{2c-1}, while it is found that $\hat{m}^2$ either converges when $2 < \alpha \leq 3$ or diverges with a lower order n^{2c-2} when $1 < \alpha \leq 2$. Hence, in the limit of large n, $M_{\theta_{\max}}$ can be approximated by neglecting $\hat{m}^2$. Therefore, we have

$$M_{\max} \approx \lim_{n\to\infty}\hat{M}_{\theta_{\max}} = \lim_{n\to\infty}\frac{\sum_{k=2}^{n-1}(k^2-1)^{-c}+2^{-c}}{\sum_{k=0}^{n-1}(k+1)^{-2c}} \quad (c = \frac{1}{\alpha-1},\ 1<\alpha\leq 3), \tag{3–35}$$

where both the numerator and the denominator are convergent and can thus be approximately computed by taking a large n.

Meanwhile, when analyzing the lower bound with the same method, it is found that $\overline{\hat{t}^2}_{\theta_{\min}}$ is the term diverging with the biggest order of n, while the orders for $\hat{m}^2$ and $\hat{s}_{\theta_{\min}}$ are smaller if they diverge. Because the term with the biggest order only appears in the denominator, we have

$$M_{\min} \approx \lim_{n\to\infty}\hat{M}_{\theta_{\min}} = 0. \tag{3–36}$$

To summarize, as shown by the solid curves in Fig. 3–3, when α grows from 1 to 3, the lower bound remains 0 while the upper bounds increases from 0 to +1, constraining M to the positive region; when α grows above 3, the upper bound remains +1 while the lower bound slides down to the negative region as a decreasing function of α, but with the limit $M_{\min} \to -0.64$ ($\alpha \to \infty$). Fig. 3–3 also reports the numerical values of $M_{\min}$ and $M_{\max}$ of series with finite lengths, produced by drawing independent samples and permuting them by $\theta_{\min}$ and $\theta_{\max}$. The gap resulted from finite system size diminishes with larger n.

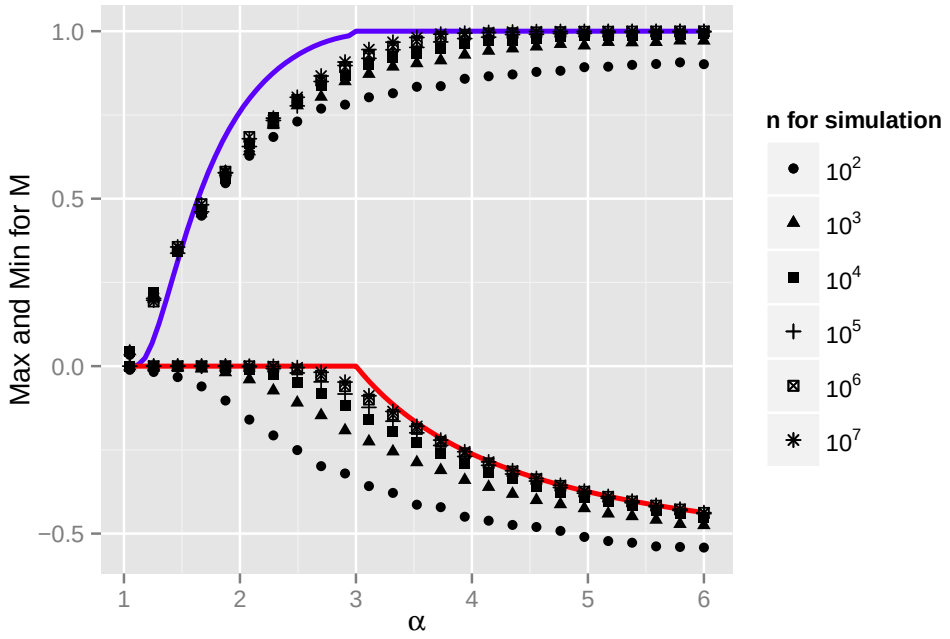

Figure 3–3 Theoretical bounds for $M_{\min}$ and $M_{\max}$ and numerical simulations with different series lengths. Each point in simulation is produced by averaging 1000 independent runs.

3.4 Empirical studies

We study the distribution of memory against the scaling exponent from empirical time series data that follow power-law distributions. To this end, we use *inter-event time series* data collected from online human activities. Inter-event time series refer to the series made up of time intervals between consecutive events and have been widely found to follow power-law distributions.

The *Movielens* dataset collects the time stamps for the users on its website[2] when they make a rating for movies. Then, one user corresponds to one inter-event time series, which is composed of time intervals between every two consecutive ratings. And the *Twitter* dataset collects the time stamps when users send a tweet. Then the inter-event time series for a user is made up of time intervals between every two tweeting actions of him or her.

Because the datasets do not explicitly contain the parameters for power-law,

[2] www.movielens.org

we estimate the scaling exponent α with the MLE method proposed by [5]. As we are examining the property of long power-law series, we rule out those series that are either too short ($n < 190$) or are unlikely to follow a power-law ($p - value < 0.1$). $p - value$ is used as a goodness-of-fit test based on the Kolmogorov-Smirnov (KS) statistic and likelihood ratios, which is also suggested by [5].

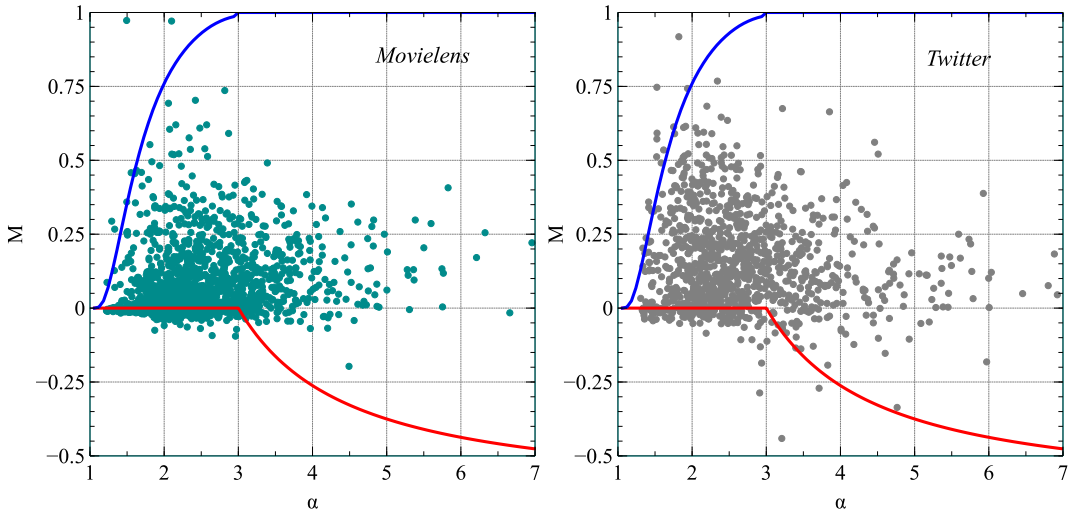

Figure 3–4 Memory for power-law distributed inter-event time series from empirical inter-event time series, where each series is represented by a point and theoretical bounds are drawn as solid curves. The regions of memory from empirical data agree with theoretical bounds. Left: *MovieLens* dataset for online movie rating. Right: *Twitter* dataset for sending tweets.

In agreement with theoretical values, we find the memory constraints existing in empirical power-law data. Fig. 3–4 plots (α, M) for power-law distributed inter-event time series selected from the *MovieLens* dataset and the *Twitter* dataset. Data points in both datasets fall into the regions predicted by theoretical bounds. A few outliers are either due to the last interval being exceptionally large (beyond the upper bounds) or the fact that they are programmed robots that perform actions at fixed times (beyond the lower bound).

Chapter 4 Conclusion and Future Work

4.1 Summary of contributions

The major methodological contribution of this work is the *Equiprobable Partition Method* (EPM) presented in this thesis. We have showed that, in the convergent moment case, EPM is accurate when the number of partitions extends to infinity. And we have demonstrated that, in the divergent moment case, how EPM can be used as a method to effectively obtain the asymptotics of moments as a function of the sample size. Specifically, an EPM estimator for diverging sample moments consists two terms — the leading order term characterizes the speed of polynomial divergence and the remaining term describes the details when the divergence is taken away. This is useful in treating statistics made up of several moments: by comparing the leading orders, we can preserve those with highest diverging orders only and neglecting others in large sample size; after canceling moments with the same leading order, we can learn the behavior of the $\frac{\infty}{\infty}$-form statistics by referring to the remaining terms.

As another major contribution of this thesis, we pointed out a special statistical property of power-law distributions that has not been presented before. With a permutational approach, by using both EPM for the divergent moment case and the traditional probabilistic method for the convergent moment case, we have derived the non-trivial bounds for the 1st-order autocorrelation (memory) of power-law series. The bounds are narrower than the natural bounds and are dependent on the power-law exponent, which suggests that the space for interdependence among series allowed by power-law is smaller than that allowed by Gaussian or uniform distributions and this effect is even related to the scaling exponent. This discovery also points out the risk of comparing the memory effect of different systems with the same quantity. Although this is a common practice, it might be unfair since the interdependence space allowed by different systems may be different.

4.2 Future work

For the *Equiprobable Partition Method*, we still need to prove the accuracy of EPM estimators for diverging sample moments, or to construct the bounds for the errors. Ideally, I feel this can be done by proving an asymptotic convergence property in parallel to what we have done for the convergent moment case.

To be specific, supposing the EPM estimator for the diverging sample moment μ'_k is $n^\gamma g(n)$, we need to prove that, for any $\epsilon > 0$, it holds that

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n^{\gamma+1}}\sum_{i=1}^{n} x_i^k - g(n)\right| \geq \epsilon\right) = 0. \tag{4–1}$$

Another direction of future work would be a further analysis of the interdependence structure of the time series with heavy-tailed distributions. We need a deeper understanding for the interdependence structure beyond the simplest characterization given by the 1st-order autocorrelation, and for a wider family distributions more than power-law. Meanwhile, it would be useful to develop quantities for describing interdependence that does not cause marginal-specific bias.

References

[1] P. Ravikumar, et al. Approximate inference, structure learning and feature estimation in markov random fields[D]. Pittsburgh: Carnegie Mellon University, School of Computer Science, Machine Learning Department, 2007.

[2] S. J. Russell, P. Norvig, E. Davis. Artificial intelligence: a modern approach[M]. Vol. 2: Prentice hall Englewood Cliffs, 2010.

[3] C. M. Bishop, et al. Pattern recognition and machine learning[M]. Vol. 1. New York: Springer, 2006.

[4] S. Asmussen, S. Asmussen, S. Asmussen. Applied probability and queues[M]. Vol. 2: Springer New York, 2003.

[5] A. Clauset, C. Shalizi, M. Newman. Power-Law Distributions in Empirical Data[J]. SIAM Review, 2009, 51(4):661–703.

[6] M.-K. Hu. Visual pattern recognition by moment invariants[J]. Information Theory, IRE Transactions on, 1962, 8(2):179–187.

[7] `http://en.wikipedia.org/wiki/Skewness`.

[8] `http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient`.

[9] A. Papoulis, S. U. Pillai. Probability, random variables and stochastic processes[M]: Tata McGraw-Hill Education, 2002.

[10] A. W. Van der Vaart. Asymptotic statistics[M]. Vol. 3: Cambridge university press, 2000.

[11] T. Rolski, H. Schmidli, V. Schmidt, et al. Stochastic processes for insurance and finance[M]. Vol. 505: Wiley, 2009.

[12] M. E. Newman. Power laws, Pareto distributions and Zipf's law[J]. Contemporary physics, 2005, 46(5):323–351.

[13] B. Gutenberg, C. Richter. Earthquake magnitude, intensity, energy, and acceleration[J]. Bulletin of the Seismological Society of America, 1942, 32(3):163–191.

[14] E. T. Lu, R. J. Hamilton. Avalanches and the distribution of solar flares[J]. The Astrophysical Journal, 1991, 380:L89–L92.

[15] G. K. Zipf. Human Behaviour and the Principle of Least-Effort[M]: Addison-Wesley, 1949.

[16] D. J. de Solla Price. Networks of scientific papers[J]. Science, 1965, 149(3683):510–515.

[17] A.-L. Barabási, R. Albert. Emergence of scaling in random networks[J]. Science, 1999, 286(5439):509–512.

[18] M. Faloutsos, P. Faloutsos, C. Faloutsos. On power-law relationships of the internet topology[C] ACM SIGCOMM Computer Communication Review. Cambridge, Massachusetts: ACM, 1999，29:251–262.

[19] R. A. Cox, J. M. Felton, K. H. Chung. The concentration of commercial success in popular music: an analysis of the distribution of gold records[J]. Journal of cultural economics, 1995, 19(4):333–340.

[20] V. Pareto. Cours d'economie politique[M]: Librairie Droz, 1964.

[21] E. T. Jaynes. Information theory and statistical mechanics[J]. Physical review, 1957, 106(4):620.

[22] E. T. Jaynes. Information theory and statistical mechanics. II[J]. Physical review, 1957, 108(2):171.

[23] E. Jaynes. The relation of Bayesian and maximum entropy methods[M]//Maximum-entropy and Bayesian methods in science and engineering: Springer, 1988:25–29.

[24] K. I. Goh, A. L. Barabási. Burstiness and memory in complex systems[J]. EPL (Europhysics Letters), 2008.

[25] M. Hallin, G. Melard, X. Milhaud. Permutational extreme values of autocorrelation coefficients and a Pitman test against serial dependence[J]. Annals of Statistics, 1992, 20(1):523–534.

[26] H. A. David, H. N. Nagaraja. Order Statistics[M]. 3rd Edition. New Jersey: Wiley, 2003.

[27] S. Zhou, R. J. Mondragón. Structural constraints in complex networks[J]. New Journal of Physics, 2007, 9(6):173.

Acknowledgements

This thesis and completing my bachelor's degree would not be possible without the support, assistance and advice of many people. And I would like to take this opportunity to acknowledge my family, friends, classmates and teachers who have helped me so much along the way.

First, I would like to thank my advisor, *Dr. Tao Zhou*, for guiding me through my undergraduate research. I remember how I walked to his lab during a group meeting and I was at that time a kid knowing nothing about research. Tao taught me how to focus on a problem, and how to break it down to smaller parts with better formulation and meanwhile seeking many ways to solve it. I greatly appreciate his encouragement, helpfulness and his easy-going style with students and lab members. I am also indebted to Tao for his generous recommendation for my academic visiting and applying to graduate schools. I would cherish my memories spent with Tao and other members in the Web Sciences Center.

I would also like to thank my teachers as I learned to reflect and grow up due to their influence. *Mr. Heping Pu* led me into mathematical analysis, the first math course in college. I learned how to transit from high school trick-playing to a new way of thinking in college from his inspiration and I respect his true dedication to teaching and education. I am also indebted to *Mr. Baohua Teng* for his encouragement and long discussions with him outside class. Among others, I enjoyed a lot in the quantum mechanics course taught by *Mr. Haitang Yang*, whose dedication, frankness and insight made him a charming physicist that I truly admire.

For this thesis, I would like to thank *Chenmin Sun*, my high school classmate, for talking with me over a math conclusion, and two users *Alfred Chern* and *Did* on math.StackExchange for helping me out of stuck.

Then, I want to thank my friends and fellow classmates, who let me know I am never walking alone.

Finally, I want to express my deepest thanks to my parents for their unconditional love and support. This thesis is dedicated to them.

Burstiness and Memory in Complex Systems

Kwang-Il Goh[1,2] and Albert-László Barabási[1]
[1] *Center for Complex Network Research and Department of Physics, 225 Nieuwland Science Hall, University of Notre Dame, Notre Dame, IN 46556, USA*
[2] *Department of Physics, Korea University, Seoul 136-713, Korea*
(Dated: May 29, 2013)

The dynamics of a wide range of real systems, from email patterns to earthquakes, display a bursty, intermittent nature, characterized by short timeframes of intensive activity followed by long times of no or reduced activity. The understanding of the origin of such bursty patterns is hindered by the lack of tools to compare different systems using a common framework. We introduce two measures to distinguish the mechanisms responsible for the bursty nature of real signals, changes in the interevent times and memory. We find that while the burstiness of natural phenomena is rooted in both the interevent time distribution and memory, for human dynamics memory is weak, and the bursty character is due to changes in the interevent time distribution. Finally, we show that current models lack in their ability to reproduce the activity pattern observed in real systems, opening up new avenues for future work.

PACS numbers: 89.75.-k, 05.45.Tp

The dynamics of most complex systems is driven by the loosely coordinated activity of a large number of components, such as individuals in the society or molecules in the cell. While we witnessed much progress in the study of the networks behind these systems [1], advances towards understanding the system's dynamics has been slower. With increasing potential to monitor the time-resolved activity of most components of selected complex systems, such as time-resolved email [2–4], web browsing [5], and gene expression [6] patterns, we have the opportunity to ask an important question: is the dynamics of complex systems governed by generic organizing principles, or each system has its own distinct dynamical features? While it is difficult to offer a definite answer to this question, a common feature across many systems is increasingly documented: the burstiness of the system's activity patterns.

Bursts, vaguely corresponding to significantly enhanced activity levels over short periods of time followed by long periods of inactivity, have been observed in a wide range of systems, from email patterns [3] to earthquakes [7, 8] and gene expression [6]. Yet, often burstiness is more of a metaphor than a quantitative feature, and opinions about its origin diverge. In human dynamics, burstiness has been reduced to the fat-tailed nature of the response time distribution [3, 4], in contrast with earthquakes and weather patterns, where memory effects appear to play a key role [9, 10]. Once present, burstiness can affect the spreading of viruses [3] or resource allocation [11]. Also, deviations towards regular, "anti-bursty" behavior in heartbeat may indicate disease progression [12]. Given the diversity of systems in which it emerges, there is a need to place burstiness on a strong quantitative basis. Our goal in this paper is to make a first step in this direction, by developing measures that can help quantify the magnitude and potential origin of the bursty patterns seen in different real systems.

Let us consider a system whose components have a measurable activity pattern that can be mapped into a discrete signal, recording the moments when some events take place, like an email being sent, or a protein being translated. The activity pattern is random (Poisson process) if the probability of an event is time-independent. In this case the interevent time between two consecutive events (τ) follows an exponential distribution, $P_{\rm P}(\tau) \sim \exp(-\tau/\tau_0)$ (Fig. 1a). An apparently bursty (or anti-bursty) signal could emerge if $P(\tau)$ is different from the exponential, such as the bursty pattern of Fig. 1b, or the more regular pattern of Fig. 1c. Yet, changes in the interevent time distribution is not the only way to generate a bursty signal. For example, the signals shown in Fig. 1d,e have exactly the same $P(\tau)$ as in Fig. 1a, yet they have a more bursty or a more regular character. This is achieved by introducing memory: in Fig. 1d the short interevent times tend to follow short ones, resulting in a bursty look. In Fig. 1e the relative regularity is due to a memory effect acting in the opposite direction: short (long) interevent times tend to be followed by long (short) ones. Therefore, the apparent burstiness of a signal can be rooted in two, mechanistically quite different deviations from a Poisson process: changes in the interevent time distribution or memory. To distinguish these effects, we introduce the burstiness parameter

Δ and the memory parameter μ, that quantify the relative contribution of each in real systems.

The *burstiness parameter* Δ is defined as

$$\Delta \equiv \frac{\mathrm{sgn}(\sigma_\tau - m_\tau)}{2} \int_0^\infty |P(\tau) - P_{\mathrm{P}}(\tau)| d\tau \ , \qquad (1)$$

where m_τ and σ_τ are the mean and the standard deviation of $P(\tau)$ [13]. The meaning of Δ is illustrated in Fig. 1f–h, where we compare $P(\tau)$ for a bursty- (Fig. 1f) and an anti-bursty (Fig. 1g) signal with a Poisson interevent time distribution. A signal will appear bursty if the frequency of the short and long interevent times is higher than in a random signal (Fig. 1f), resulting in many short intervals separated by longer periods of inactivity. Therefore, there are fewer interevent times of average length than in a Poisson process. A signal displays anti-bursty character, however, if the frequency of the interevent times is enhanced near the average and depleted in the short and long interevent time region (Fig. 1g). Δ is bounded in the range $(-1, 1)$, and its magnitude correlates with the signal's burstiness: $\Delta = 1$ is the most bursty signal, $\Delta = 0$ is completely neutral (Poisson), and $\Delta = -1$ corresponds to a completely regular (periodic) signal. For example, in Fig. 1h we show Δ for the stretched exponential distribution, $P_{\mathrm{SE}}(\tau) = u(\tau/\tau_0)^{u-1} \exp[-(\tau/\tau_0)^u]/\tau_0$, often used to approximate the interevent time distributions of complex systems [14]. The smaller the u is, the burstier is the signal, and for $u \to 0$, $P_{\mathrm{SE}}(\tau)$ follows a power law with the exponent -1, for which $\Delta = 1$. For $u = 1$, $P_{\mathrm{SE}}(\tau)$ is simply the exponential distribution with $\Delta = 0$. Finally, for $u > 1$, the larger u is, the more regular is the signal, and for $u \to \infty$, $P(\tau)$ converges to a Dirac delta function with $\Delta = -1$.

Most complex systems display a remarkable heterogeneity: some components may be very active, and others much less so. For example, some users may send dozens of emails during a day, while others only one or two. To combine the activity levels of so different components, we can group the signals based on their average activity level, and measure $P(\tau)$ only for components with similar activity level. As the insets in Fig. 2 show, the obtained curves are systematically shifted. If we plot, however, $\tau_0 P(\tau)$ as a function of τ/τ_0, where τ_0 being the average interevent time, the data collapse into a single curve $\mathcal{F}(x)$ (Fig. 2), indicating that the interevent time distribution follows $P(\tau) = (1/\tau_0)\mathcal{F}(\tau/\tau_0)$, where $\mathcal{F}(x)$ is independent of the average activity level of the component, and represents an universal characteristic of the particular system [8, 15]. This raises an impor-

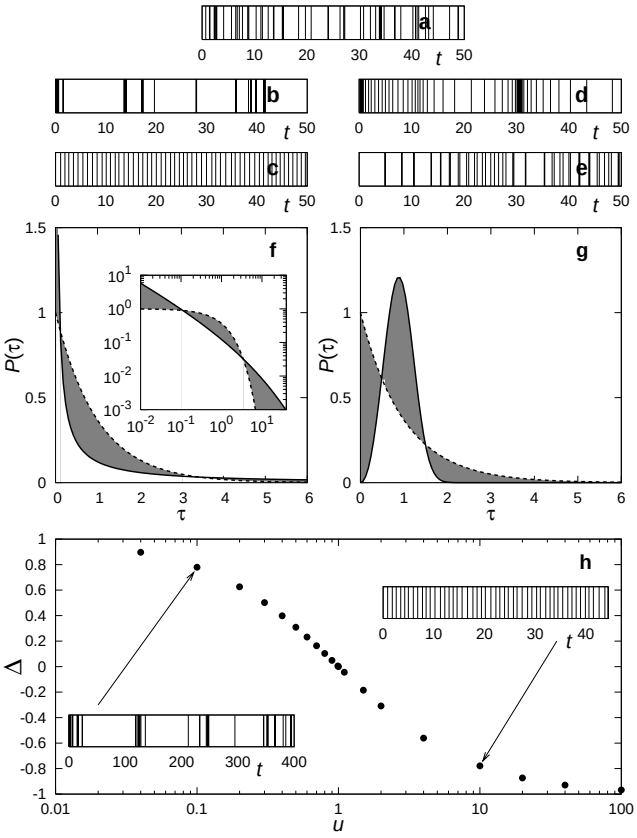

FIG. 1: (a) A signal generated by a Poisson process with a unit rate. (b,c) Bursty character through the interevent time distribution: A bursty signal (b) generated by the power-law interevent time distribution $P(\tau) \sim \tau^{-1}$, and an anti-bursty signal (c) generated by the Gaussian interevent time distribution with $m = 1$ and $\sigma = 0.1$. A bursty signal can emerge through memory as well. For example, the bursty signal shown in (d) is obtained by shuffling the Poisson signal of (a) to increase the memory effect. A more regular looking signal, with negative memory, is obtained by the same shuffling procedure (e). Note that signals in (a), (d) and (e) have identical interevent time distribution. (f) An interevent time distribution (solid line) will appear bursty ($\Delta > 0$) if it has a higher frequency of long or short interevent times than predicted for a Poisson process (dotted line). Inset shows the same curves in log-log scale. (g) The signal will appear to be regular (anti-bursty, $\Delta < 0$) if $P(\tau)$ is higher in the average interevent time region than a Poisson process. The burstiness parameter Δ is half of the shaded area between the corresponding interevent time distribution (solid) and the reference distribution (dotted). (h) The stretched exponential interevent distribution interpolates between a highly bursty ($\Delta = 1$), a Poisson ($\Delta = 0$), and a regular ($\Delta = -1$) signal. The figure shows Δ as a function of the parameter u.

tant question: will Δ depend on τ_0? The burstiness parameter Δ is invariant under the time rescaling as $\tilde{\tau} \equiv \tau/\tau_0$ and $\tilde{P}(\tilde{\tau}) \equiv \tau_0 P(\tau)$ with a constant τ_0, since $\tilde{\Delta} \equiv \int_0^\infty |\tilde{P}(\tilde{\tau}) - \tilde{P}_0(\tilde{\tau})| d\tilde{\tau} = \int_0^\infty |\tau_0 P(\tau) - \tau_0 P_0(\tau)| d(\tau/\tau_0) = \int_0^\infty |P(\tau) - P_0(\tau)| d\tau \equiv \Delta$, *i.e.*, it characterizes the universal function $\mathcal{F}(x)$. Such invariance of Δ enables us to assign to each system a single burstiness parameter, despite the different activity level of its components.

The *memory coefficient* μ of a signal is defined as the correlation coefficient of all consecutive interevent time values in the signal over a population. That is, given all pairs of consecutive interevent times $(\tau_{k,i}, \tau_{k,i+1})$ for all components $\{k = 1, \cdots, N\}$,

$$\mu \equiv \frac{1}{N} \sum_{k=1}^{N} \sum_{i=1}^{n_k - 1} \frac{(\tau_i - m_{k1})(\tau_{i+1} - m_{k2})}{\sigma_{k1}\sigma_{k2}} , \quad (2)$$

where N is the number of components in the system, n_k is the number of events recorded for component k, and $m_{k1}(m_{k2})$ and $\sigma_{k1}(\sigma_{k2})$ are the mean and standard deviation of $\tau_{k,i}$'s ($\tau_{k,i+1}$'s), respectively. The memory coefficient is positive when a short (long) interevent time tends to be followed by a short (long) one, and it is negative when a short (long) interevent time is likely to be followed by a long (short) one. The measurements indicate that μ is independent of τ_0.

Mapping complex systems on the (μ, Δ) *space*— Given that the burstiness of a signal can have two qualitatively different origins, the best way to characterize a real system is to identify its μ and Δ parameters, placing them in a (μ, Δ) phase diagram (Fig. 3). As a first example, we measured the spacing between the consecutive occurrence of the same letter in texts of different kind, era, and language [16]. For these signals, we find $\Delta \approx 0$, *i.e.*, the interevent time distribution follows closely an exponential (Fig. 2b) and $\mu \approx 0.01$, indicating the lack of memory. Thus this signal is best described by a Poisson process, at the origin of the phase diagram (Fig. 3). In contrast, natural phenomena, like earthquakes [17] and weather patterns [18] are in the vicinity of the diagonal, indicating that $P(\tau)$ and memory equally contribute to their bursty character. The situation is quite different, however, for human activity, ranging from email and phone communication to web browsing and library visitation patterns [2, 4, 5, 20]. For these we find a high Δ and small or negligible μ, indicating that while these systems display significant burstiness rooted in $P(\tau)$, memory plays a small role in their temporal inhomogeneity. This lack of memory is quite unexpected, as it suggests the lack of predictability in these systems in contrast with natural phenomena, where strong memory effects could lead to predictive tools. Finally for cardiac rhythms describing the time between two consecutive heartbeats (Fig. 2c) [21], we find $\Delta_{\text{cardiac, healthy}} = -0.73(4)$ for healthy individuals and $\Delta_{\text{cardiac, CHF}} = -0.82(6)$ for patients with congestive heart failure (CHF), both signals being highly regular. Thus the Δ parameter captures the fact that cardiac rhythm is more regular with CHF than in the healthy condition [12]. Furthermore, we find $\mu \approx 0.97$, indicating that memory also plays a very important role in the signal's regularity.

The discriminative nature of the (μ, Δ) phase diagram is illustrated by the non-random placement of the different systems in the plane: human activity patterns cluster together in the high Δ, low μ region, natural phenomena near the diagonal, heartbeats in the high μ, negative Δ region and texts near $\Delta = \mu = 0$, underlying the existence of distinct classes of dynamical mechanisms driving the temporal activity in these systems.

Following the clustering of the empirical measurements in the phase diagram, a natural question emerges: to what degree can current models reproduce the observed quantitative features of bursty processes? Queueing models, proposed to capture human activity patterns, are designed to capture the waiting times of the tasks, rather than interevent times [3, 4, 22]. Therefore, placing them on the phase diagram is not meaningful. A bursty signal can be generated by 2-state model [23], switching with probability p its internal state between Poisson processes with two different rates $\lambda_0 < \lambda_1$. Δ for the model is independent of p in the long time limit as long as $p > 0$, and takes its value in the range $0 < \Delta < 0.5$, approaching 0 when $\lambda_0 \approx \lambda_1$ and 0.5 as $\lambda_1 \to \infty$ and $\lambda_0 \to 0$. The memory coefficient of the model follows $\mu = A(0.5-p)$, where A is a positive constant dependent on λ_0 and λ_1 so that $-1/3 < \mu < 1/3$. The region in the (μ, Δ) space occupied by the model is shown in the light grey area in Fig. 3a, suggesting that by changing its parameters the model could account for all observed behaviors. Yet, the agreement is misleading: for example, for human activities $P(\tau)$ is fat-tailed, which is not the case for the model. This indicates that Δ and μ offer only a first order approximation for the origin of the burstiness, and for a detailed comparison between models and real systems we need to inspect other measures as well, such as the functional form of $P(\tau)$. It also indicates the lack of proper modeling tools to capture the detailed mechanisms responsible for the bursty interevent time distributions seen in real

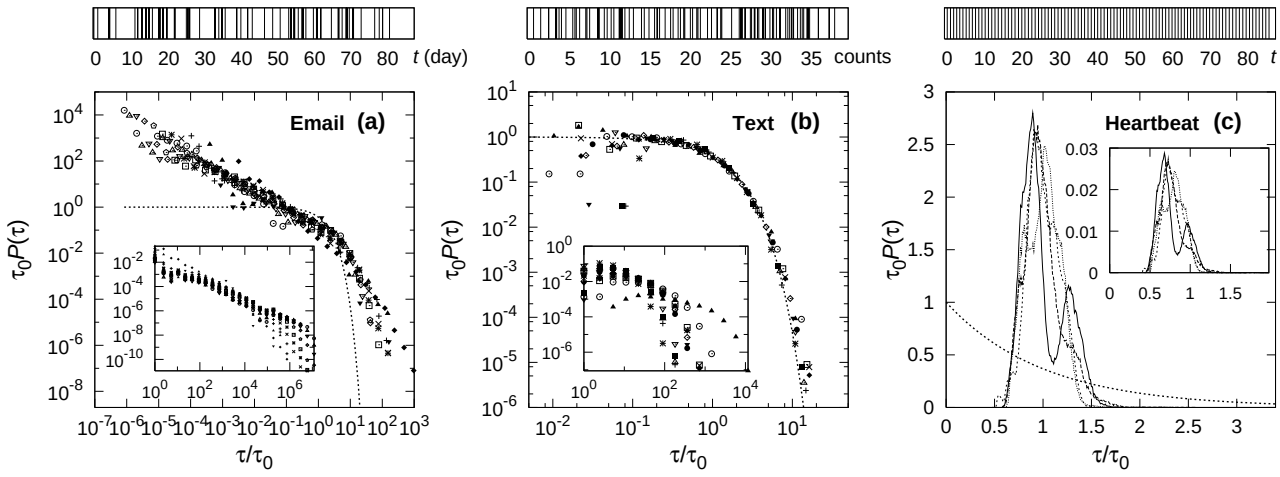

FIG. 2: Interevent time distributions $P(\tau)$ for some real signals. (a) $P(\tau)$ for e-mail activity of individuals from a University [2]. τ corresponds to the time interval between two emails sent by the same user. (b) Interevent time distribution for the occurrence of letter in the text of C. Dickens' *David Copperfield* [16]. (c) Interevent time distribution of cardiac rhythm of individuals [21]. Each event corresponds to the beat in the heartbeat signal. In each panel, we also show for reference the exponential interevent time distribution (dotted). Unscaled interevent time distribution is shown in the inset for each dataset.

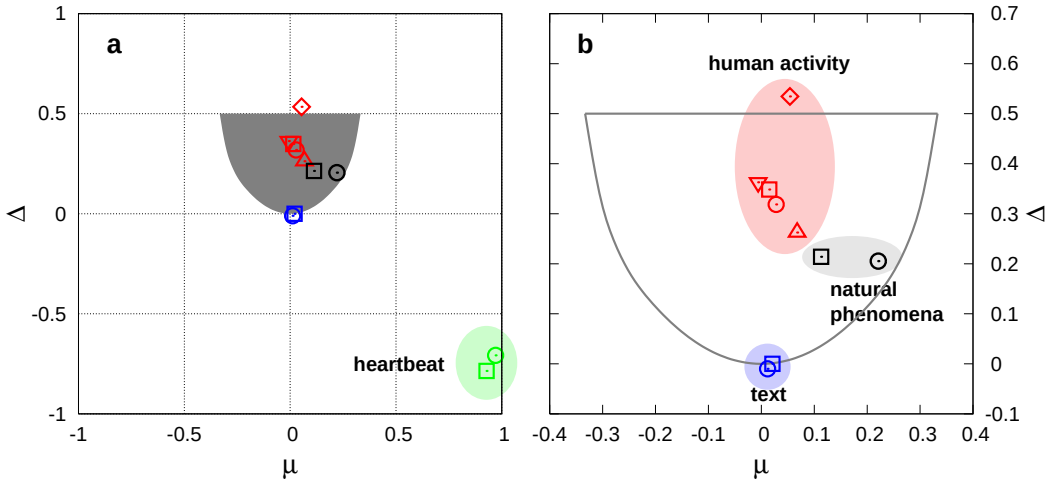

FIG. 3: (Color) (a) The (μ, Δ) phase diagram. Human behaviors (red) are captured by activity patterns pertaining to email (□) [2], library loans (○) [4], and printing (◇) [19] of individuals in Universities, call center record at an anonymous bank (△) [20], and phone initiation record from a mobile phone company (▽). Data for natural phenomena (black) are earthquake records in Japan (○) [17] and daily precipitation record in New Mexico, USA (□) [18]. Data for human texts (blue) [16] are the English text of *David Copperfield* (○) and the Hungarian text of *Isten Rabjai* by Gárdonyi Géza (□). Data for physiological behaviors (green) are the normal sinus rhythm (○) and the cardiac rhythm with CHF (□) of human subjects [21]. Grey area is the region occupied by the 2-state model [23]. (b) Close-up of the most populated region.

systems, opening up possibilities for future work.

[1] S. N. Dorogovtsev and J. F. F. Mendes, *Evolution of Networks* (Oxford University Press, Oxford, 2002); M. E. J. Newman, SIAM Rev. **45**, 167 (2003); R. Pastor-

Satorras and A. Vespignani, *Structure and evolution of the Internet* (Cambridge University Press, Cambridge, 2003); S. Boccaletti, *et al.*, Phys. Rep. **424**, 175 (2006); M. E. J. Newman, D. J. Watts, and A.-L. Barabási (eds.), *Structure and Dynamics of Complex Networks* (Princeton University Press, Princeton, 2006).
[2] J. P. Eckmann, E. Moses, and D. Sergi, Proc. Natl. Acad. Sci. U.S.A. **101**, 14333 (2004).
[3] A.-L. Barabási, Nature (London) **207**, 435 (2005); A. Vázquez, Phys. Rev. Lett. **95**, 248701 (2005).
[4] A. Vázquez, *et al.*, Phys. Rev. E **73**, 036127 (2006).
[5] Z. Dezső, *et al.*, Phys. Rev. E **73**, 066132 (2006).
[6] I. Golding, *et al.*, Cell **123**, 1025 (2005); J. R. Chubb, *et al.*, Curr. Biol. **16**, 1018 (2006).
[7] P. Bak, *et al.*, Phys. Rev. Lett. **88**, 178501 (2002).
[8] A. Corral, Phys. Rev. E **68**, 035102(R) (2003).
[9] A. Bunde, *et al.*, Phys. Rev. Lett. **94**, 048701 (2005).
[10] V. N. Livina, S. Havlin, and A. Bunde, Phys. Rev. Lett. **95**, 208501 (2005).
[11] W. E. Leland, *et al.*, IEEE/ACM Trans. Networking **2**, 1 (1994).
[12] S. Thurner, M. C. Feurstein, and M. C. Teich, Phys. Rev. Lett. **80**, 1544 (1998).
[13] As an alternative, σ_τ/m_τ can also be used intead of Δ to meausre burstiness [A. Vázquez, private communications].
[14] J. Laherrère and D. Sornette, Eur. Phys. J. B **2**, 525 (1998).
[15] A. Saichev and D. Sornette, Phys. Rev. Lett. **97**, 078501 (2006).
[16] Project Gutenberg, http://gutenberg.org.
[17] Japan University Network Earthquake Catalog, http://wwweic.eri.u-tokyo.ac.jp/CATALOG/junec/.
[18] National Resources Conservation Service, http://www.nm.nrcs.usda.gov/snow/data/historic.htm.
[19] U. Harder and M. Paczuski, Physica A **361**, 329 (2006).
[20] I. Guedj and A. Mandelbaum, http://iew3.technion.ac.il/serveng/callcenterdata/.
[21] PhysioBank, http://www.physionet.org/physiobank/.
[22] J. G. Oliveira and A.-L. Barabási, Nature (London) **437**, 1251 (2005).
[23] J. Kleinberg, Proc. ACM SIGKDD '02, pp. 91 (2002).

复杂系统中的阵发性和记忆性

Kwang-Il Goh、Albert-László Barabási

摘要

在包括邮件到地震的各种各样的真实系统的动力学中,经常有阵发性的规律存在,表现为短时间的剧烈活动然后是长时间的停止或减弱。对这类现象的理解不足,是因为缺乏可以把不同系统在同一个框架下比较的统一工具。我们引入两种度量来区别阵发性、记忆性两个机制。我们发现尽管阵发性源于时间间隔的分布和记忆性,但由于人类动力学中的记忆性很弱,所以阵发性是由时间间隔分布引起的。最好,我们表明现在的模型还不足以再现实际观察到的规律,所为未来的工作打开了大门。

复杂系统的很多动力学过程是由彼此之间松散连接的小部分引发的,比如社会和细胞。尽管我们这背后的网络的研究中有很多进展 [1], 但我们对动力学的理解进展很慢。随着现在我们有越来越强的能力监控具有时间分辨度的系统,比如电子邮件 [2, 3, 4]、上网 [5] 以及基因表达 [6],我们现在有机会问一个重要的问题: 复杂系统的动力学是由一个通用的原理支配着,还是每个系统有各自的特点? 尽管要确定地回答这个问题很困难,但有越来越多的证据表明不同系统都有一个共同的特点:系统活动的阵发性。

阵发性,粗略地说,就是短时间的活跃程度增强然后是长时间的沉寂。阵发性在很多系统里都有发现,包括邮件 [3]、地震 [7, 8] 和基因表达 [6]。但是很多时候阵发性没有确切的度量,而且起源有争议。在人类动力学中,阵发性被认为是源于响应时间的胖尾分布 [3, 4]。然而,在地震和气候现象中,记忆性被认为扮演者重要角色 [9, 10]。阵发性的存在可以影响病毒传播 [3] 和资源分配 [11]。并且,规律的、非阵发的心跳是疾病的征兆 [12]. 出现阵发性的系统如此之多,有必要将其量化。我们的这篇文章是这个方向的第一步,我们制定了度量来帮助量化阵发性的强度和原因。

我们假设系统的活动已经被映射成离散的信号。如果事件的几率不随时间变化,那么活动是个泊松过程。这样相邻事件的时间间隔 (τ) 服从指数分布,$P_{\rm P}(\tau) \sim \exp(-\tau/\tau_0)$ (图 1a). 一个明显阵发(或者非阵发)的信号可以当 $P(\tau)$ 不同于指数的时候产生, 比如图 1b 的阵发信号和图 1c 的非阵发信号。但是改动间隔时间分布不是造成阵发的唯一办法。比如图 1d 有和图 1a 一样的 $P(\tau)$ 但是阵发性不一样。这是通过改变记忆性做到的:图 1d 中短的时间间隔倾向于跟随短间隔,从而看起来是阵发的。但在图 1e 中,非阵发性是由于另一个方向的记忆性:短的倾向于跟随长的。因此,阵发性是由两个机制引起的:间隔时间分布的改变和记忆性。为区别,我们引入阵发性参数 Δ 和记忆性参数 μ, 来分别量化两者的作用。

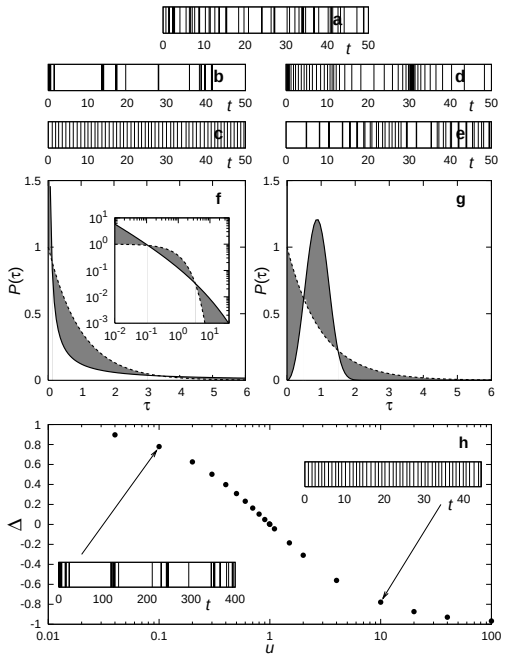

图 1: (a) 泊松分布的信号. (b,c) 改变间隔时间: 阵发信号 (b) 由幂律 $P(\tau) \sim \tau^{-1}$ 产生, 非阵发信号 (c) 由 $m = 1$、$\sigma = 0.1$ 的高斯分布产生. 阵发信号 (d) 通过打乱 (a) 以增加记忆性产生. 非阵发、负记忆性的信号 (e) 也是打乱产生的.(a), (d) 和 (e) 有相同的间隔时间分布。

阵发性参数 Δ 定义为

$$\Delta \equiv \frac{\mathrm{sgn}(\sigma_\tau - m_\tau)}{2} \int_0^\infty |P(\tau) - P_{\rm P}(\tau)| d\tau \ , \qquad (1)$$

其中 m_τ 和 σ_τ 是均值和标准差 [13]. Fig. 1f–h 解释了 Δ, 我们比较了具有泊松间隔时间分布的阵发、非阵发的 $P(\tau)$。当长、短的时间间隔比随机序列出现得更多时,序列看上去是阵发的 (图 1f)。如果平均长度的间隔更多、长短间隔更少,看起来就不是阵发的 (图 1g)。$\Delta \in (-1, 1)$, 其大小对应于阵发性: $\Delta = 1$ 意味着阵发性最强,$\Delta = 0$ 意味着中性(泊松), $\Delta = -1$ 对应于规律(周期)信号。

很多复杂系统具有异质性: 一些部分很活跃,一些不是。比如有的人每天很多邮件,有的人只有一两封。我们可以把活动程度相似的放在一起比较 $P(\tau)$ 。如图 2 所示, 曲线被平移了. 如果我们把 $\tau_0 P(\tau)$ 作为 τ/τ_0 的函数画图, 其中 τ_0 为平均间隔时间, 数据会落在曲线 $\mathcal{F}(x)$ 上 (图 2), 表明间隔时间服从 $P(\tau) = (1/\tau_0)\mathcal{F}(\tau/\tau_0)$, 其中 $\mathcal{F}(x)$ 与平均活动强度无关, 表示系统的一个特征 [8, 15]。这样提出一个重要的问题:Δ 依赖于 τ_0 吗? Δ 随着时间放大缩小不变,随着 $\tilde{\tau} \equiv \tau/\tau_0$ and $\tilde{P}(\tilde{\tau}) \equiv \tau_0 P(\tau)$, 由于 $\tilde{\Delta} \equiv \int_0^\infty |\tilde{P}(\tilde{\tau}) - \tilde{P}_0(\tilde{\tau})| d\tilde{\tau} = \int_0^\infty |\tau_0 P(\tau) - \tau_0 P_0(\tau)| d(\tau/\tau_0) = \int_0^\infty |P(\tau) - P_0(\tau)| d\tau \equiv \Delta$, 也就是说其刻画了 $\mathcal{F}(x)$.

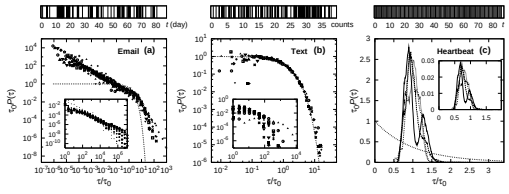

图 2: 真实信号的间隔时间分布 $P(\tau)$。(a) 一个大学电子邮件的 $P(\tau)$ [2]. τ 对应于一个用户两次发邮件的间隔。(b) 相邻两个字母在 *David Copperfield* 中出现的间隔 [16]. (c) 心律的间隔时间 [21]. 我们还画出了指数间隔分布作为参考 (点)。

信号的记忆性系数 μ 被定义为相继两个序列的关联系数。即,给定所有相继的时间间隔对 $(\tau_{k,i}, \tau_{k,i+1})$, $\{k = 1, \cdots, N\}$,

$$\mu \equiv \frac{1}{N} \sum_{k=1}^{N} \sum_{i=1}^{n_k - 1} \frac{(\tau_i - m_{k1})(\tau_{i+1} - m_{k2})}{\sigma_{k1}\sigma_{k2}} , \qquad (2)$$

其中 N 是系统组成部分的数量,n_k 是该部分的事件数目,$m_{k1}(m_{k2})$ 与 $\sigma_{k1}(\sigma_{k2})$ 分别是 $\tau_{k,i}$ $(\tau_{k,i+1})$, 的均值和标准差。当长间隔倾向与跟随长间隔时,记忆性系数是正的;反之是负的。μ 与 τ_0 无关。

将复杂系统映射到 (μ, Δ) 空间 — 我们已经指导阵发性有两个可能的原因, 对他们可以用 μ 和 Δ 两个参数来刻画, 画在 (μ, Δ) 相图中 (图 3)。作为第一个例子,我们测量了不同文本中一个字母相继出现的间隔 [16]. 对于这些信号, 我们发现 $\Delta \approx 0$, 即间隔时间分别接近于指数分布 (图 2b),同时 $\mu \approx 0.01$, 缺乏记忆性。在相图原点附近的信号 (图 3) 可以用泊松过程描述。相反,自然现象, 比如地震 [17] 和天气 [18] 在对角线的边上,表明 $P(\tau)$ 和记忆性对于阵发性的贡献差不多。然而,对于人类活动,例如电子邮件、电话、上网 [2, 5, 4, 20],情况就不同了。对于这些现象我们发现大的 Δ 和可以忽略的小 μ, 表明这些系统的记忆性来自 $P(\tau)$, 记忆性对于异质性的贡献很小。记忆性小是理所应当的,因为它表明它们相对于自然现象的可预测性小,而自然现象的强记忆性使得其可以被预测。最后,对于心率 (图 2c) [21], 我们发现对于健康人 $\Delta_{\text{cardiac, healthy}} = -0.73(4)$, 对于心脏衰竭的人 $\Delta_{\text{cardiac, CHF}} = -0.82(6)$ 两者都很规律。因此 Δ 表明心脏衰竭的人比健康的人心跳更规律 [12]。进一步,我们发现 $\mu \approx 0.97$,,表明记忆性在心律的规律性中有很大作用。

各种系统在 (μ, Δ) 相图上有规律的分布着:人类行为集中在 Δ 高 μ 低的区域, 自然现象在对角线附近, 心律在 μ 高 Δ 为负的区域,语言文字靠近 $\Delta = \mu = 0$, 表明这些系统的时间行为由不同种类的动力学机制支配着。

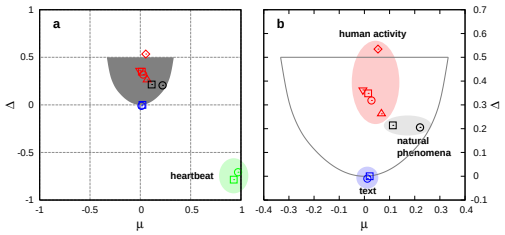

图 3: (a) (μ, Δ) 相图。人类行为 (红色) 包括电子邮件 (□) [2]、图书馆借阅 (○) [4]、大学里的打印 (◇) [19] 、一个银行的通话 (△) [20] 以及一个公司的拨打电话行为 (▽)。自然现象的数据包括日本的地震 (○) [17] 和新墨西哥州的每日降水量 (□) [18]。语言文字数据 [16] 包括英文版的 *David Copperfield* (○) 和匈牙利文版的 *Isten Rabjai* (□). 生理现象的数据 (绿色) 包括正常人 (○) 和心脏衰竭者 (□) 的心律 [21]. 灰色是 2-状态模型的区域 [23]。

实证数据在相图中有不同区块,我们不禁要问: 我们现在的模型对于复现这些现象表现得如何? 排队模型被用来描述等待时间 [3, 4, 22],画在这里没有意义。阵发性信号可以由 2-状态模型产生 [23], 以概率 p 切换对应于两种速率 $\lambda_0 < \lambda_1$ 的泊松过程。长时间下,当 $p > 0$, Δ 独立于 p , 取值范围为 $0 < \Delta < 0.5$, approaching 0 当 $\lambda_0 \approx \lambda_1$ 时接近 0,当 $\lambda_1 \to \infty$ 且 $\lambda_0 \to 0$ 时接近 0.5。模型的记忆性为 $\mu = A(0.5 - p)$, 其中 A 依赖于 λ_0 和 λ_1 的常数。模型在 (μ, Δ) 空间所能表示的范围由图 3a 的灰色部分标出。模型可以通过调整参数覆盖所有观察到的范围。但是,这可能是个假象: 比如人类活动的 $P(\tau)$ 是胖尾的,不同于模型。这表明 Δ and μ 只提供了阵发性原因的第一阶近似,而我们需要在比较系统的时候考虑其它的度量, 比如 $P(\tau)$ 的函数形态。这也表明现在还没有能阵发性背后机制的细节的模型,未来还有很多工作可以做。

参考文献

[1] S. N. Dorogovtsev and J. F. F. Mendes, *Evolution of Networks* (Oxford University Press, Oxford, 2002); M. E. J. Newman, SIAM Rev. **45**, 167 (2003); R. Pastor-Satorras and A. Vespignani, *Structure and evolution of the Internet* (Cambridge University Press, Cambridge, 2003); S. Boccaletti, *et al.*, Phys. Rep. **424**, 175 (2006); M.

E. J. Newman, D. J. Watts, and A.-L. Barabási (eds.), *Structure and Dynamics of Complex Networks* (Princeton University Press, Princeton, 2006).

[2] J. P. Eckmann, E. Moses, and D. Sergi, Proc. Natl. Acad. Sci. U.S.A. **101**, 14333 (2004).

[3] A.-L. Barabási, Nature (London) **207**, 435 (2005); A. Vázquez, Phys. Rev. Lett. **95**, 248701 (2005).

[4] A. Vázquez, *et al.*, Phys. Rev. E **73**, 036127 (2006).

[5] Z. Dezső, *et al.*, Phys. Rev. E **73**, 066132 (2006).

[6] I. Golding, *et al.*, Cell **123**, 1025 (2005); J. R. Chubb, *et al.*, Curr. Biol. **16**, 1018 (2006).

[7] P. Bak, *et al.*, Phys. Rev. Lett. **88**, 178501 (2002).

[8] A. Corral, Phys. Rev. E **68**, 035102(R) (2003).

[9] A. Bunde, *et al.*, Phys. Rev. Lett. **94**, 048701 (2005).

[10] V. N. Livina, S. Havlin, and A. Bunde, Phys. Rev. Lett. **95**, 208501 (2005).

[11] W. E. Leland, *et al.*, IEEE/ACM Trans. Networking **2**, 1 (1994).

[12] S. Thurner, M. C. Feurstein, and M. C. Teich, Phys. Rev. Lett. **80**, 1544 (1998).

[13] As an alternative, σ_τ/m_τ can also be used intead of Δ to meausre burstiness [A. Vázquez, private communications].

[14] J. Laherrère and D. Sornette, Eur. Phys. J. B **2**, 525 (1998).

[15] A. Saichev and D. Sornette, Phys. Rev. Lett. **97**, 078501 (2006).

[16] Project Gutenberg, http://gutenberg.org.

[17] Japan University Network Earthquake Catalog, http://wwweic.eri.u-tokyo.ac.jp/CATALOG/junec/.

[18] National Resources Conservation Service, http://www.nm.nrcs.usda.gov/snow/data/historic.htm.

[19] U. Harder and M. Paczuski, Physica A **361**, 329 (2006).

[20] I. Guedj and A. Mandelbaum, http://iew3.technion.ac.il/serveng/callcenterdata/.

[21] PhysioBank, http:// www.physionet.org/ physiobank/.

[22] J. G. Oliveira and A.-L. Barabási, Nature (London) **437**, 1251 (2005).

[23] J. Kleinberg, Proc. ACM SIGKDD '02, pp. 91 (2002).