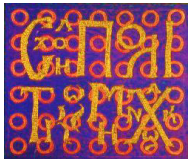
ROMA TRE
UNIVERSITÀ DEGLI STUDI

Lecture notes on stochastics – Part 1

Rome, 3-4 May 2018

A brief introduction to probability and stochastics



Demetris Koutsoyiannis
Department of Water Resources and Environmental Engineering
School of Civil Engineering
National Technical University of Athens, Greece
(dk@itia.ntua.gr, http://www.itia.ntua.gr/dk/)

Presentation available online: http://www.itia.ntua.gr/1835/

Senza probabilità (Without Probability): An example

- Problem: study the storage and outflow of a (toy) water supply reservoir in discrete time with ridiculously simple assumptions.
- Assumption 1 – **the ideal hydrological model**: The inflow I to the reservoir is constant for any time step, equal to 10 units.
- Assumption 2 – **the perfect socio-hydrological model**: If there is plenty of water in the reservoir, people consume more, while the consumption is reduced when the storage is low. We assume that this behaviour is expressed precisely by an exponential function: $Q = \varphi(S) = 0.2\ \mathrm{e}^{0.3\,S}$, where Q is the outflow and S the storage.
- Discrete time dynamics $Q_i = \varphi(S_{i-1})$, $S_i = S_{i-1} + I - Q_i$
- Question 1: Assume a specific initial storage S_0 in the interval (5, 15) and find S_1.
- Question 2: With the same initial condition, find S_{50}.
- Question 3: Is the system dynamics deterministic or stochastic?
- Question 4: Is the **system predictable (i.e., deterministic) or unpredictable (i.e., stochastic, random)?**

Impacts of the creation of a single electron somewhere at the limit of the known universe

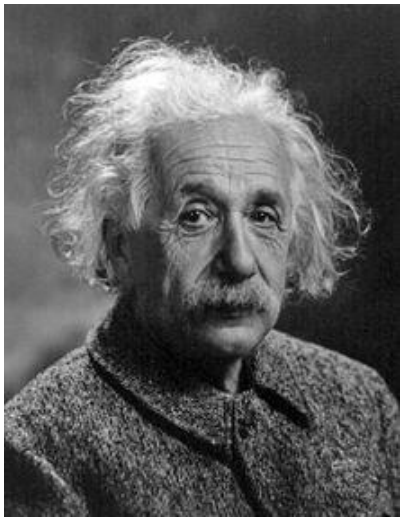
Perturbation scale	Cause	Effect	Time frame
1. Microscopic	An electron has been created at a distance of 10^{10} light years from Earth and has exerted gravitational attraction on the molecules of the atmosphere	After 50 collisions the trajectories of molecules would have changed (different molecules would collide)	10 ns (justifiably assuming $\sim 10^{10}$ collisions per second)
2. Macroscopic	Turbulence	Change in the fine structure of turbulence	1 min
3. Local	Turbulence	Change in the large (km) scale atmospheric turbulence (wind, clouds)	1 h – 1 d
4. Global		Change in the general circulation of the atmosphere (depressions, fronts→ a storm that would not occur without that electron)	1-2 weeks

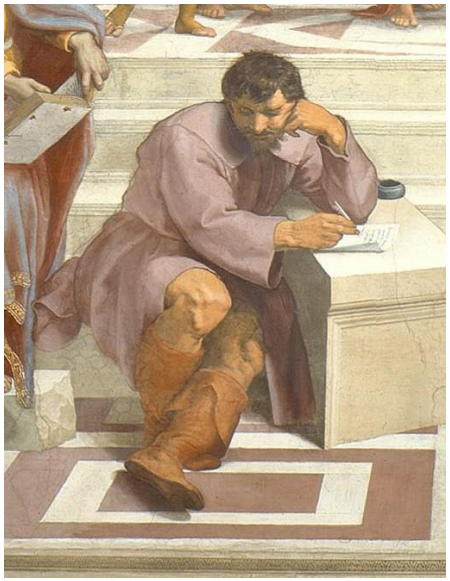
Adapted from Ruelle (1979, 1991, p. 75); based on Berry (1978) and some ideas of E. Borel και B. V. Chirikov.

So, who does not play dice?

Jedenfalls bin ich überzeugt, daß der nicht würfelt
I, at any rate, am convinced that He [God] does not throw dice

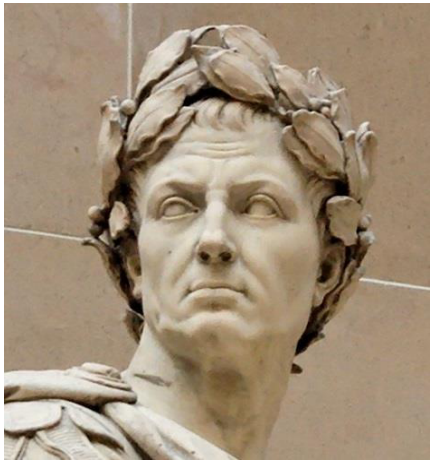
(Albert Einstein, in a letter to Max Born in 1926)





Αἰών παῖς ἐστι παίζων πεσσεύων
Time is a child playing, throwing dice

(Heraclitus; ca. 540-480 BC; Fragment 52)



Ἀνερρίφθω κύβος	Iacta alea est
Let the die have been cast	The die has been cast
[Plutarch's version, in Greek]	[Suetonius's version, in Latin]

(Julius Caesar, 49 BC, when crossing Rubicon River)

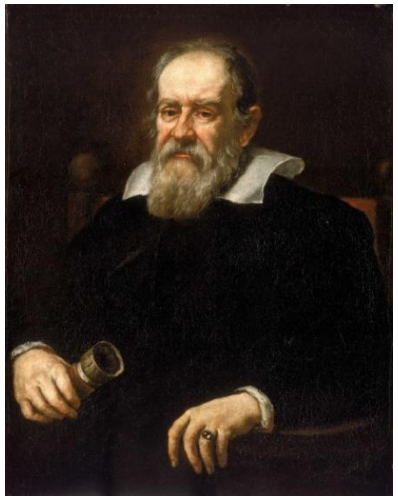
From old times dice games fascinated people—but perhaps not scientists and teachers



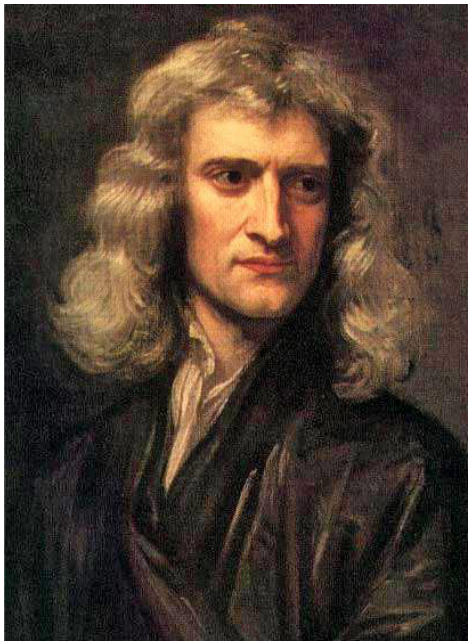
- All these dice are of the period 580-570 BC from Greek archaeological sites:
  - Left, Kerameikos Ancient Cemetery Museum, Athens, photo by author
  - Middle: Bronze die (1.6 cm), Greek National Archaeological Museum, www.namuseum.gr/object-month/2011/apr/7515.png
  - Right: Terracotta die (4 cm) from Sounion, Greek National Archaeological Museum, http://www.namuseum.gr/object-month/2011/dec/dies_b.png
- Much older dice (up to 5000 years old) have been found in Asia (Iran, India).

Modern determinism and the clockwise universe

- Johannes Kepler (1571-1630), Galileo Galilei (1564-1642) and René Descartes (1596-1650) introduced mathematical concepts to natural philosophy (science).
- They also introduced the idea of a clockwork universe, leading to the philosophical proposition of *determinism*, still widely accepted in science.
- Determinism was perfected by the French mathematician and astronomer Pierre-Simon Laplace (1749-1827; cf. *Laplace's demon,* a hypothetical entity that knows the precise location and momentum of every atom in the universe at present, and can deduce the future and the past using Newton's laws.).
- According to deterministic thinking, the roots of uncertainty about future are subjective, i.e. rely on the fact that we do not know exactly the present, or we do not have good enough methods and models. It is then a matter of time to eliminate uncertainty, with better data and better models.

Newton's awareness of the fragility of the universe (rejection of determinism)

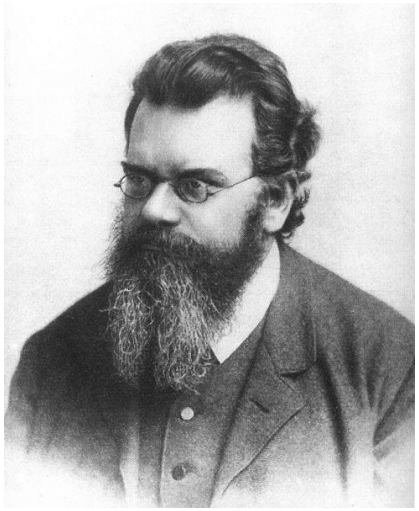


"*For while comets move in very eccentric orbs in all manner of positions, blind fate could never make all the planets move one and the same way in orbs concentric, some inconsiderable irregularities excepted which may have arisen from the mutual actions of comets and planets on one another, and which will be apt to increase, till this system wants a reformation*" (Newton, *Opticks*, Query 31).

- Newton regarded the complexity and fragility of the universe as proof of the existence of God.
- He rejected Leibniz' thesis that God would necessarily make a perfect world which requires no intervention from the creator.
- Newton simultaneously made an argument from design and for the necessity of intervention.

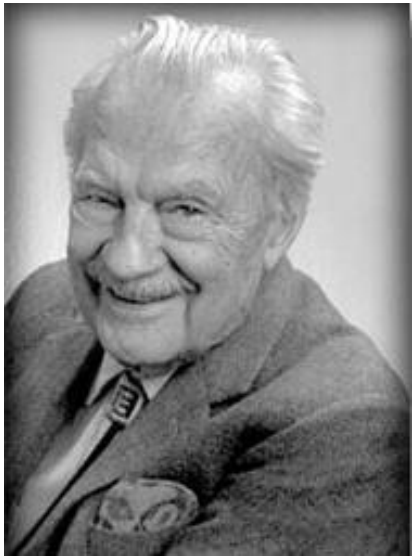
From the almighty determinism of the 17th century to the probabilistic world of the 20th century

- **Statistical physics** (cf. Boltzmann) used the probabilistic concept of entropy (which is nothing other than a quantified measure of uncertainty defined within the probability theory; see below) to explain fundamental physical laws (most notably the Second Law of Thermodynamics), thus leading to a new understanding of natural behaviours and to powerful predictions of macroscopic phenomena.
- **Dynamical systems** theory (cf. Poincare) has shown that uncertainty can emerge even from pure, simple and fully known deterministic (chaotic) dynamics, and cannot be eliminated.
- **Quantum theory** (cf. Heisenberg) has emphasized the intrinsic character of uncertainty and the necessity of probability in the description of nature.



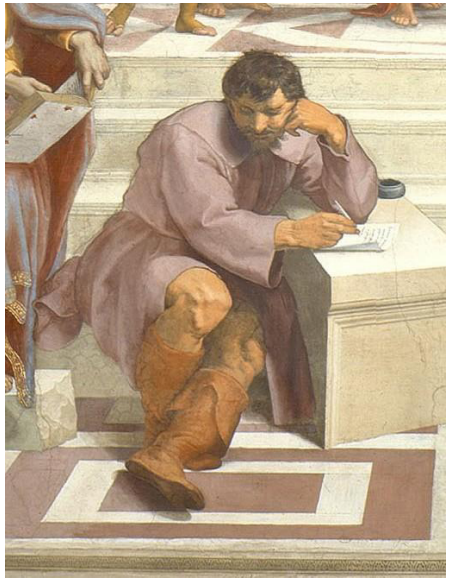
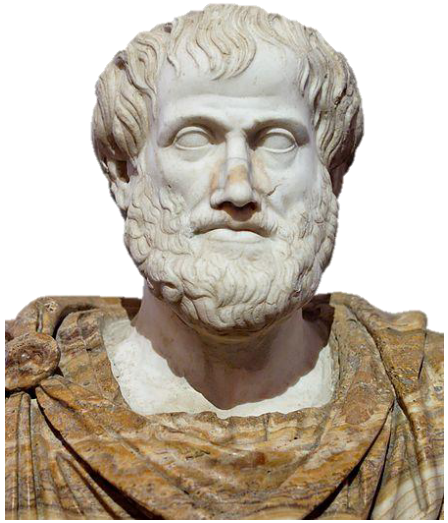
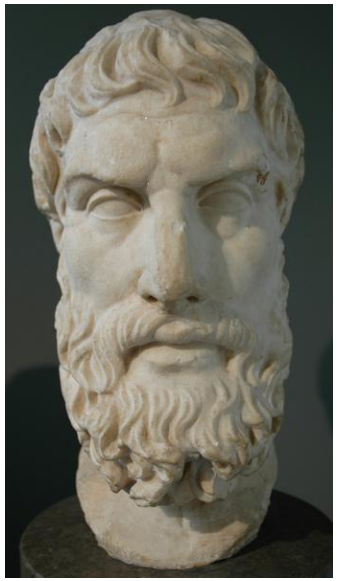
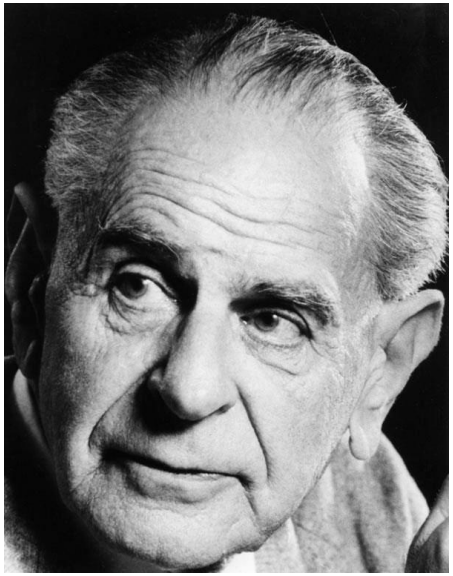
From the almighty determinism of the 17th century to the probabilistic world of the 20th century (2)

- Developments in **mathematical logic**, and particularly **Gödel's incompleteness theorem**, challenged the almightiness of deduction (inference by mathematical proof). Ironically, **Kurt Gödel** anticipated by one day (in 1930) **David Hilbert** who pronounced the opposite with his famous aphorism (also inscribed in his tombstone at Göttingen) "*Wir müssen wissen, wir werden wissen*" ("***We must know, we will know***").
- Developments in **numerical mathematics** (cf. **Nicholas Metropolis**) highlighted the effectiveness of stochastic methods in solving even purely deterministic problems, such as **numerical integration** in high-dimensional spaces and **global optimization** of non-convex functions (where stochastic techniques, e.g. evolutionary algorithms or simulated annealing, are in effect the only feasible solution in complex problems that involve many local optima).
- Advances in **evolutionary biology** emphasize the importance of stochasticity (e.g. in selection and mutation procedures and in environmental changes) as a driver of evolution.



Indeterminism vs. determinism

- In *indeterminism*, a philosophical belief contradictory to determinism, uncertainty may be a structural element of nature and thus cannot be eliminated.
- Indeterminism has its origin in the Greek philosophers **Heraclitus** (*ca*. 535–475 BC), **Aristotle** (384 – 322 BC) and **Epicurus** (341–270 BC).
- Its relationship with modern science was theorized by the Austrian-British philosopher **Karl Popper** (1902-1994).
- In science, indeterminism largely relies on the notion of ***probability***, which according to **Popper** is the extension (quantification) of the **Aristotelian** idea of ***potentia*** (***dynamis***). Practically, the idea is that several outcomes can be produced by a specified cause, while in deterministic thinking only one outcome is possible (albeit difficult to predict which one).

The meaning of probability (by examples)

(1) A fair coin has a probability of 0.5 of heads, and likewise 0.5 of tails; so the probability of tossing two heads in a row is 0.25.

(2) There is a 10% probability of rain tomorrow.

(3) There is a 10% probability of rain tomorrow according to the weather forecast.

(4) Fortunately there is only a 5% probability that her tumor is malignant, but this will not be known for certain until the surgery is done next week.

(5) Smith has a greater probability of winning the election than does Jones.

(6) I believe that there is a 75% probability that she will want to go out for dinner tonight.

(7) I left my umbrella at home today because the forecast called for only a 1% probability of rain.

(8) Among 100 patients in a clinical trial given drug *A,* 83 recovered, whereas among 100 other patients given drug *B,* only 11 recovered; so new patients will have a higher probability of recovery if treated with drug A.

Source of examples: Gauch (2003).

The utility of probability

- Commonly, probability is regarded to be a branch of applied mathematics that provides tools for data analysis (and also theorizes games of chance).
- Historically, as modern science was initiated from deterministic views of the world, probability had a marginal role for peculiar unpredictable phenomena.
- Nonetheless, probability is a more general concept that helps shape a consistent, realistic and powerful view of the world.
- Probability has provided grounds for philosophical concepts such as indeterminism and causality, as well as for extending the typical mathematical logic, offering the mathematical foundation of induction.
- In typical scientific and technological applications, probability provides the tools to quantify uncertainty, rationalize decisions under uncertainty, and make predictions of future events under uncertainty, in lieu of unsuccessful deterministic predictions.

See more details in Koutsoyiannis (2008).

Deduction and induction

- In mathematical logic, determinism can be paralleled to the premise that all truth can be revealed by *deductive reasoning* or *deduction* (the Aristotelian *apodeixis*). This type of reasoning consists of repeated application of strong syllogisms such as:

  If A is true, then B is true;
  A is true;
  Therefore, B is true.

  If A is true, then B is true;
  B is false;
  Therefore, A is false.

- Deduction uses a set of axioms to prove propositions known as theorems, which, given the axioms, are irrefutable, absolutely true statements. It is also irrefutable that deduction is the preferred route to truth; the question is, however, whether or not it has any limits.
- David Hilbert's belief "*Wir müssen wissen, wir werden wissen*", more formally known as *completeness*, according to which any mathematical statement could be proved or disproved by deduction from axioms, has been proved to be invalid.
- In everyday life, however, we use weaker syllogisms of the type:

  If A is true, then B is true;
  B is true;
  Therefore, A becomes more plausible.

  If A is true, then B is true;
  A is false;
  Therefore, B becomes less plausible.

- The latter type of syllogism is called *induction* (the Aristotelian *epagoge*). It does not offer a proof that a proposition is true or false and may lead to errors. However, it is very useful in decision making, when deduction is not possible.
- An important achievement of probability is that it quantifies (expresses in the form of a number between 0 and 1) the degree of plausibility of a certain proposition or statement. The formal probability framework uses both deduction, for proving theorems, and induction, for inference with incomplete information or data.

Definition of probability

- According to Kolmogorov's (1933) axiomatization, probability theory is based on three fundamental concepts and four axioms.
- The concepts, i.e., the triplet (Ω, Σ, P) called *probability space*, are:
  1. A non-empty set Ω, sometimes called the *basic set*, *sample space* or the *certain event* whose elements ω are known as *outcomes* or *states*.
  2. A set Σ known as *σ-algebra* or *σ-field* whose elements E are subsets of Ω, known as *events*. Ω and $\emptyset$ are both members of Σ, and, in addition, (a) if E is in Σ then the complement $\Omega - E$ is in Σ; (b) the union of countably many sets in Σ is also in Σ.
  3. A function P called *probability* that maps events to real numbers, assigning each event E (member of Σ) a number between 0 and 1.
- The four axioms, which define the properties of P, are:
  I. **Non-negativity**: For any event A, $P(A) \geq 0$.
  II. **Normalization**: $P(\Omega) = 1$.
  III. **Additivity**: For any events A, B with $AB = \emptyset$, $P(A + B) = P(A) + P(B)$.
  IV. **Continuity at zero**: If $A_1 \supset A_2 \supset \ldots \supset A_n \supset \ldots$ is a decreasing sequence of events, with $A_1A_2\ldots A_n\ldots = \emptyset$, then $\lim_{n\to\infty} P(A_n) = 0$.

  [Note: In the case that Σ is finite, axiom IV follows from axioms I-III; in the general case, however, it should be put as an independent axiom.]

The concept of a random variable

- A random variable $\underline{x}$ is a function that maps outcomes to numbers, i.e. quantifies the sample space Ω.
- More formally, a real single-valued function $\underline{x}(\omega)$, defined on the basic set Ω, is called a *random variable* if for each choice of a real number a the set $\{\underline{x} < a\}$ for all ω for which the inequality $\underline{x}(\omega) < \alpha$ holds true, belongs to Σ.
- With the notion of the random variable we can conveniently express events using basic mathematics. In most cases this is done almost automatically. For instance a random variable $\underline{x}$ that takes values 1 to 6 is intuitively assumed when we deal with a die through.
- We must be attentive that a random variable is not a number but a function. Intuitively, we could think of a random variable as an object that represents simultaneously all possible outcomes and only them.
- A particular value that a random variable may take in a random experiment, else known as a *realization* of the variable, is a number.
- We can denote a random variable by an underlined letter, e.g. $\underline{x}$ and its realization with a non-underlined letter x (another convention is to use an upper case letter, e.g. X, for the random variable and a lower case letter, e.g. x, for its realization. In any case, random variables and values thereof two should not be confused).

Probability distribution function

- *Distribution function* is a function of the real variable x defined by

  $$F(x) := P\{\underline{x} \le x\}$$

  where $\underline{x}$ is a random variable.
- The random variable with which this function is associated is not an argument of the function. If there risk of confusion (e.g. there are many random variables), the random variable is usually denoted as a subscript (e.g. $F_{\underline{x}}(x)$). Typically $F(x)$ has a mathematical expression depending on some parameters. The domain of $F(x)$ is not identical to the range of the random variable $\underline{x}$; rather it is always the set of real numbers.
- The distribution function is a non-decreasing function obeying the relationship

  $$0 = F(-\infty) \le F(x) \le F(+\infty) = 1$$
- For its non-decreasing attitude, in the English literature the distribution function is also known as *cumulative distribution function* (cdf) – though "cumulative" is not necessary. In practical applications the distribution function is also known as *non-exceedence probability*. Likewise, the non-increasing function

  $$\overline{F}(x) = P\{\underline{x} > x\} = 1 - F(x)$$

  is known as *exceedence probability* (or survival function, survivor function, tail function).
- The distribution function is always continuous on the right; however, if the basic set Ω is finite or countable, $F(x)$ is discontinuous on the left at all points x_i that correspond to outcomes ω_i, and it is constant between them (staircase-like). Such random variable is called *discrete*. If $F(x)$ is a continuous function, then the random variable is called *continuous*. A *mixed* case is also possible; in this the distribution function has some discontinuities on the left, but is not staircase-like.
- For continuous random variables, the inverse function $F^{-1}(\,)$ of $F(\,)$ exists. Consequently, the equation $u = F(x)$ has a unique solution for x, called *u-quantile* of the variable $\underline{x}$, that is:

  $$x_u = F^{-1}(u)$$

Probability density (or mass) function

- In continuous variables any particular value x has zero probability to occur. However, we can still tell which of two outcomes is more probable by examining the ratio of the two probabilities. As this is a $0/0$ expression, having in mind l'Hôpital's rule, we need to examine the ratio of derivatives of probabilities.
- The derivative of the distribution function is called the *probability density function*:

$$f(x) \coloneqq \frac{\mathrm{d}F(x)}{\mathrm{d}x}$$

- The basic properties of $f(x)$ are

$$f(x) \geq 0, \quad \int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$$

- Obviously, the probability density function does not represent a probability; therefore it can take values higher than 1. Its relationship with probability is described by the following equation:

$$f(x) = \lim_{\Delta x \to 0} \frac{P\{x \leq \underline{x} \leq x + \Delta x\}}{\Delta x}$$

- The distribution function can be calculated from the density function by

$$F(x) = \int_{-\infty}^{x} f(y)\mathrm{d}y$$

- In discrete random variables, the density is a sequence of Dirac δ functions. It is thus more convenient to use the so-called *probability mass function* $P_j \equiv P(x_j) = P\{\underline{x} = x_j\}$, $j = 1,\ldots,w$, where w is the number of possible outcomes (which can be infinite).

Some common distributions

Name	**Probability density function**	**Distribution function**
Uniform in [0, 1]	$f(x) = \begin{cases} 1 & \text{for } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$	$F(x) = \max(0, \min(x, 1))$
Exponential	$f(x) = \begin{cases} e^{-x/\mu} / \mu & \text{for } x \ge 0 \\ 0 & \text{for } x < 0 \end{cases}$	$F(x) = \begin{cases} 1 - e^{-x/\mu} & \text{for } x \ge 0 \\ 0 & \text{for } x < 0 \end{cases}$
Normal	$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} \exp\left(-\frac{(u-\mu)^2}{2\sigma^2}\right) du$

Independent and dependent events, conditional probability

- Two events A and B are called *independent* (or *stochastically independent*), if

$$P(AB) = P(A)P(B)$$

- Otherwise A and B are called (*stochastically*) *dependent*.
- The definition can be extended to many events. Thus, the events A_1, A_2, …, are *independent* if for any finite set of distinct indices i_1, i_2, …, i_n:

$$P\left(A_{i_1}A_{i_2}\dots A_{i_n}\right) = P\left(A_{i_1}\right)P(A_{i_2})\dots P(A_{i_n})$$

- The handling of probabilities of independent events is thus easy. However, this is a special case because usually natural events are dependent. In the handling of dependent events the notion of *conditional probability* is vital.
- By definition (Kolmogorov, 1933), conditional probability of the event A given B (i.e. under the condition that the event B has occurred) is the quotient

$$P(A|B) := \frac{P(AB)}{P(B)}$$

- Obviously, if $P(B) = 0$, this conditional probability cannot be defined, while for independent A and B, $P(A|B) = P(A)$. It follows that

$$P(AB) = P(A|B)P(B) = P(B|A)P(A)$$

- From this it follows the *Bayes theorem*:

$$P(B|A) = P(B)\frac{P(A|B)}{P(A)}$$

Random number generation

- **Sequence of random numbers** is a sequence of numbers x_i whose every one statistical property is consistent with that of a sample from a sequence of independent identically distributed random variables $\underline{x}_i$ (adapted from Papoulis, 1990).
- **Random number generator** is a device (typically computer algorithm) which generates a sequence of random numbers x_i with given distribution $F(x)$. As most algorithms are purely deterministic, sometimes the numbers are called pseudorandom—but this in not necessary.
- Random number generation is also known as **Monte Carlo** sampling.
- The basis of practically all random generators is the uniform distribution in [0,1]. A typical procedure is the following:
  - We generate a sequence of integers q_i from the recursive algorithm

    $$q_i = (k\, q_{i-1} + c) \bmod m$$

    where k, c and m are appropriate integers (e.g. $k = 69\,069$, $c = 1$, $m = 2^{32} = 4\,294\,967\,296$ or $k = 7^5 = 16\,807$, $c = 0$, $m = 2^{31} - 1 = 2\,147\,483\,647$; Ripley, 1987, p. 39).
  - We calculate the sequence of random numbers u_i with uniform distribution in [0,1] by

    $$u_i = q_i / m$$

- A more recent and better algorithm is the so-called *Mersenne twister* (en.wikipedia.org/wiki/Mersenne_twister). It is available in most languages and software packages. For example, for Excel (which by default includes the function rand) the Mersenne twister algorithm, called NtRand, can be found in www.ntrand.com/download/.
- A direct (but sometimes time demanding) algorithm to produce random numbers x_i from *any* $F(x)$ given random numbers u_i with uniform distribution in [0,1] is provided by:

  $$x_i = F^{-1}(u_i)$$

Exercise 1

Let $\underline{x}$ and $\underline{y}$ represent the outcomes of each of two dice. What is the probability of the following cases?

- $\{\underline{x} < \underline{y}\}$
- $\{\underline{x} < y\}$
- $\{x < \underline{y}\}$
- $\{x < y\}$

Verify the results by Monte Carlo simulations.

Exercise 2

- Assume that in a certain place on earth (specifically in the United Kingdom) and a certain period of the year a dry and a wet day are equiprobable and that in the different days the states (wet/dry) are independent. What is the probability that two consecutive days are wet under the following conditions?
  - Unconditionally.
  - If we know that the first day is wet.
  - If we know that the second day is wet.
  - If we know that one of the two days is wet.
  - If we know that one of the two days is dry.
- Verify the results by Monte Carlo simulations.
- Plot the distribution function of one day's state (wet/dry) (after introducing an appropriate random variable).
- Assuming that in a wet day the probability density function of the rainfall depth $\underline{x}$ (expressed in mm) is $f(x|\text{wet}) = e^{-x}$, plot the probability distribution function $F(x)$.

Exercise 3

- Three engineers A, B and C are biding for a 1 000 000 € project and the evaluation committee, in order to make the fairest possible selection, decided to throw a die, instead of evaluating the proposal, the experience of engineers, etc.. If the outcome is 1 or 2 the projects goes to A, if it is 3 or 4, then B wins and if it is 5 or 6, then C wins. The dice is cast, but the announcement of the winner is going to be done the next day by the minister.
- Engineer A approaches the chairman of the committee and offers him 1000 € to accept his following request: "I know you are not allowed to tell me who wins; however, two of the three will lose. Therefore, B or C or both will lose. Please tell me just one of these two will lose". The committee member accepts and says that C will lose. Then engineer A offers another 1000 € to swap him with B.
- Prove that the strategy of engineer A is consistent with awareness of probability.
- Compare this strategy with another one, in which engineer A offers the same amount to convince the chairman to re-decide on A and B by tossing a coin.
- Verify your result with Monte Carlo simulation.

Note: A different utterance of this problem is known as the "three prisoners problem" (http://en.wikipedia.org/wiki/Three_Prisoners_problem), which has puzzled many. For example, Ben-Naim, 2008, devotes several pages in his book about entropy (including a whole appendix) to solve this problem. However, its solution can be done in two lines.

Expectation

- For a discrete random variable $\underline{x}$, taking on the values $x_1, x_2, \ldots, x_w$ (where w could be ∞) with probability mass function $P_j \equiv P(x_j) = P\{\underline{x} = x_j\}$, if $g(\underline{x})$ is an arbitrary function of $\underline{x}$ (so that $g(\underline{x})$ is a random variable per se), we define the *expectation* or *expected value* or *mean* of $g(\underline{x})$ as

$$\mathrm{E}\left[g(\underline{x})\right] := \sum_{j=1}^{w} g(x_j) P(x_j)$$

- Likewise, for a continuous random variable $\underline{x}$ with density $f(x)$, the expectation is

$$\mathrm{E}\left[g(\underline{x})\right] := \int_{-\infty}^{\infty} g(x) f(x) \mathrm{d}x$$

- For certain types of functions $g(\underline{x})$ we get very commonly used statistical parameters, as specified below:

1. For $g(\underline{x}) = \underline{x}^r$, where $r = 0, 1, 2, \ldots$, the quantity $\mu'_r := \mathrm{E}[\underline{x}^r]$ is called the r*th moment* (or the r*th moment about the origin*) of $\underline{x}$. For $r = 0$, obviously the moment is 1.
2. For $g(\underline{x}) = \underline{x}$, the quantity $\mu := \mu'_1 = \mathrm{E}[\underline{x}]$ (that is, the first moment) is called the *mean* of $\underline{x}$.
3. For $g(\underline{x}) = (\underline{x} - \mu)^r$ where $r = 0, 1, 2, \ldots$, the quantity $\mu_r := \mathrm{E}[(\underline{x} - \mu)^r]$ is called the r*th central moment* of $\underline{x}$. For $r = 0$ and 1 the central moments are respectively 1 and 0. For
4. For $g(\underline{x}) = (\underline{x} - \mu)^2$ the quantity $\gamma \equiv \sigma^2 := \mu_2 = \mathrm{E}[(\underline{x} - \mu)^2]$ is called the *variance* of $\underline{x}$ (also denoted as $\mathrm{var}[\underline{x}]$); its square root σ (also denoted as $\mathrm{std}[\underline{x}]$ is called the standard deviation of $\underline{x}$.

Entropy

- For a **discrete random variable** $\underline{x}$, taking on the values $x_1, x_2, \ldots, x_w$ (where w could be ∞) with probability mass function $P_j \equiv P(x_j) = P\{\underline{x} = x_j\}$, the *entropy* is defined as the expectation of the minus logarithm of probability (Shannon, 1948), i.e.:

$$\Phi[\underline{x}] := \mathrm{E}[-\ln P(\underline{x})] = -\sum_{j=1}^{w} P_j \ln P_j$$

- Extension of the above definition for the case of a **continuous random variable** $\underline{x}$ with probability density function $f(x)$, is possible, although not contained in Shannon's (1948) original work. This extension involves a (so-called) '*background measure*' with density $h(x)$, which can be any probability density, proper (with integral equal to 1) or improper (meaning that its integral does not converge); typically it is an (improper) Lebesgue density, i.e. a constant with dimensions $[h(x)] = [f(x)] = [x^{-1}]$, so that the argument of the logarithm function that follows be dimensionless. Thus, the entropy of a continuous variable $\underline{x}$ is (see e.g. Jaynes, 2003, p. 375):

$$\Phi[\underline{x}] := \mathrm{E}\left[-\ln \frac{f(\underline{x})}{h(\underline{x})}\right] = -\int_{-\infty}^{\infty} \ln \frac{f(x)}{h(x)} f(x)\,\mathrm{d}x$$

- It is easily seen that for both discrete and continuous variables the entropy $\Phi[\underline{x}]$ is a *dimensionless* quantity.
- The importance of the entropy concepts relies in the **principle of maximum entropy** (Jaynes, 1957); it postulates that the entropy of a random variable $\underline{x}$ should be at maximum, under some conditions, formulated as constraints, which incorporate the information that is given about this variable.
- This principle can be used for **logical inference** as well as for **modelling physical systems**; for example, the tendency of entropy to become maximal (Second Law of thermodynamics) can result from this principle.

Exercise 4

- Find the mean, variance and entropy of the variable $\underline{x}$ representing the outcome of a fair die. Show that the entropy of a fair die is greater than in any loaded die.
- Find the mean, variance and entropy of a variable $\underline{x}$ with uniform distribution in [0,1]. Show that this entropy is the maximum possible among all distributions in [0,1].
- Find the mean, variance and entropy of a variable $\underline{x}$ with exponential distribution. Show that this entropy is the maximum possible among all distributions in $[0,\infty)$ which have specified mean.
- Find the mean, variance and entropy of a variable $\underline{x}$ with normal distribution. Show that this entropy is the maximum possible among all distributions in $(-\infty,\infty)$ which have specified mean and variance.

Two variables: joint distribution and joint moments

- Here we provide definitions referring to a pair of two random variables $(\underline{x}, \underline{y})$.
- **Joint probability distribution function**: $F_{xy}(x,y) := P\{\underline{x} \le x, \underline{y} \le y\}$
- **Joint probability density function** : $f_{xy}(x,y) := \dfrac{\partial^2 F_{xy}(x,y)}{\partial x\, \partial y}$
- **Marginal probability distribution functions** : $F_x(x) := P\{\underline{x} \le x\}$, $F_y(y) := P\{\underline{y} \le y\}$
- **Joint raw moment** of order $p + q$: $\mu'_{pq} := \mathrm{E}[\underline{x}^p\, \underline{y}^q] = \int_{-\infty}^{\infty} x^p y^q\, f_{xy}(x,y)\, \mathrm{d}x\, \mathrm{d}y$
- **Marginal first moments** (means): $\mu_x := \mu'_{10}, \mu_y := \mu'_{01}$
- **Joint raw moment** of order $p + q$:

$$\mu_{pq} := \mathrm{E}\left[(\underline{x} - \mu_x)^p \left(\underline{y} - \mu_y\right)^q\right] = \int_{-\infty}^{\infty} (\underline{x} - \mu_x)^p \left(\underline{y} - \mu_y\right)^q f_{xy}(x,y)\mathrm{d}x\, \mathrm{d}y$$

- **Variances**: $\mathrm{var}[\underline{x}] := \mathrm{E}\left[(\underline{x} - \mu_x)^2\right] \equiv \mu_{20} \equiv \gamma_x \equiv \sigma_x^2$; $\mathrm{var}\left[\underline{y}\right] := \mu_{02} \equiv \gamma_y \equiv \sigma_y^2$
- **Covariance**: $\mathrm{cov}\left[\underline{x}, \underline{y}\right] := \mathrm{E}\left[(\underline{x} - \mu_x)\left(\underline{y} - \mu_y\right)\right] \equiv \mu_{11} \equiv \sigma_{xy} = \mathrm{E}\left[\underline{x}\, \underline{y}\right] - \mathrm{E}[\underline{x}]\mathrm{E}\left[\underline{y}\right]$
- **Correlation coefficient**: $r_{xy} := \dfrac{\sigma_{xy}}{\sigma_x \sigma_y}$
- **Independent variables**: $F_{xy}(x,y) = F_x(x)\, F_y(y)$; $f_{xy}(x,y) = f_x(x)\, f_y(y)$
- **Uncorrelated variables**: $\sigma_{xy} = 0$, $r_{xy} = 0$, $\mathrm{E}[\underline{x}\, \underline{y}] = \mathrm{E}[\underline{x}]\, \mathrm{E}[\underline{y}]$

Correlation and climacogram

- **Linear combinations of random variables**: $\mathrm{E}\left[a_1\underline{x}_1 + a_2\underline{x}_2\right] = a_1\mathrm{E}\left[\underline{x}_1\right] + a_2\mathrm{E}\left[\underline{x}_2\right]$, $\mathrm{var}\left[a_1\underline{x}_1 + a_2\underline{x}_2\right] = a_1^2\mathrm{var}\left[\underline{x}_1\right] + a_2^2\mathrm{var}\left[\underline{x}_2\right] + 2a_1a_2\mathrm{cov}\left[\underline{x}_1, \underline{x}_2\right]$
- It follows that: $\mathrm{Var}\left[\frac{1}{2}\left(\frac{\underline{x}_1}{\sigma_1} + \frac{\underline{x}_2}{\sigma_2}\right)\right] = \frac{1}{4}\mathrm{E}\left[\left(\frac{\underline{x}_1-\mu_1}{\sigma_1} + \frac{\underline{x}_2-\mu_2}{\sigma_2}\right)^2\right] = \frac{1}{2} + \frac{1}{2}\mathrm{Cov}\left[\frac{\underline{x}_1}{\sigma_1}, \frac{\underline{x}_2}{\sigma_2}\right]$
- Likewise: $\mathrm{Var}\left[\frac{1}{2}\left(\frac{\underline{x}_1}{\sigma_1} - \frac{\underline{x}_2}{\sigma_2}\right)\right] = \frac{1}{4}\mathrm{E}\left[\left(\frac{\underline{x}_1-\mu_1}{\sigma_1} - \frac{\underline{x}_2-\mu_2}{\sigma_2}\right)^2\right] = \frac{1}{2} - \frac{1}{2}\mathrm{Cov}\left[\frac{\underline{x}_1}{\sigma_1}, \frac{\underline{x}_2}{\sigma_2}\right]$
- Thus, $r_{12} = \frac{\mathrm{Cov}[\underline{x}_1,\underline{x}_2]}{\sigma_1\sigma_2} = \mathrm{cov}\left[\frac{\underline{x}_1}{\sigma_1}, \frac{\underline{x}_2}{\sigma_2}\right] = 2\ \mathrm{var}\left[\frac{1}{2}\left(\frac{\underline{x}_1}{\sigma_1} + \frac{\underline{x}_2}{\sigma_2}\right)\right] - 1 = 1 - 2\ \mathrm{var}\left[\frac{\underline{x}_1}{\sigma_1} - \frac{\underline{x}_2}{\sigma_2}\right]$
- As the variance is by definition non-negative, it follows that $-1 \le r_{12} \le 1$; the value r_{12} =0 corresponds to uncorrelated variables, while positive or negative r_{12} corresponds to positively or negatively correlated variables, respectively.
- The same information as in r_{12} is provided by the quantity $\rho_{12} \coloneqq \mathrm{var}\left[\frac{1}{2}\left(\frac{\underline{x}_1}{\sigma_1} + \frac{\underline{x}_2}{\sigma_2}\right)\right]$, for which it is easily seen that $0 \le \rho_{12} \le 1$; the value ρ_{12} =1/2 corresponds to uncorrelated variables, while values of ρ_{12} greater or less than ½ correspond to positively or negatively correlated variables, respectively.
- The notion of ρ_{12} could be readily expanded to many variables. Assuming that all variables are identically distributed and multiplying by the common variance σ^2, we define the so-called **climacogram**, $\gamma_k \coloneqq \mathrm{var}\left[\underline{X}_k/k\right]$, where $\underline{X}_k \coloneqq \underline{x}_1 + \cdots + \underline{x}_k$ and $0 \le \gamma_k \le \sigma^2$.

Many variables and stochastic processes

- A **stochastic process** is a family of infinitely many random variables indexed by a (regular) variable, which takes values from an index set T, typically representing time. We distinguish between:
  - A **continuous-time stochastic process** $\underline{x}(t)$, when time is continuous, e.g. $T = [0, \infty)$.
  - A **discrete-time stochastic process** $\underline{x}_i$, when time is discrete, e.g., $T = \{0, 1, 2, \ldots\}$.
- **Time series or sample function**: a realization, x_i, of a stochastic process, $\underline{x}_i$ or $\underline{x}(t)$, at a finite set of discrete time instances i (or t_i). (**Caution**: A stochastic process is a family of random variables, infinitely many for discrete time processes and uncountably infinitely many for continuous time processes. On the other hand, a time series is a finite sequence of numbers).
- **First order distribution function** of the process: $F(x; t) := P\{\underline{x}(t) \le x\}$
- **Second order distribution function** : $F(x_1, x_2; t_1, t_2) := P\{\underline{x}(t_1) \le x_1, \underline{x}(t_2) \le x_2\}$
- ***n*th order distribution function**: $F(x_1, \ldots, x_n; t_1, \ldots, t_n) := P\{\underline{x}(t_1) \le x_1, \ldots, \underline{x}(t_n) \le x_n\}$
- **Mean**: $\mu(t) := \mathrm{E}[\underline{x}(t)]$
- **Autocovariance**: $c(t; h) := \mathrm{Cov}[\underline{x}(t), \underline{x}(t+h)] = \mathrm{E}[(\underline{x}(t) - \mu(t))\,(\underline{x}(t+h) - \mu(t+h))]$
- **Cross-covariance of two processes** $\underline{x}(t)$ and $\underline{y}(t)$: $c_{xy}(t; h) := \mathrm{cov}[\underline{x}(t), \underline{y}(t+h)]$

Stationarity

- Central to the notion of a stochastic process are the concepts of *stationarity* and *nonstationarity*, two widely misunderstood and misused concepts (see Koutsoyiannis and Montanari, 2014), whose definitions apply only to stochastic processes (thus, e.g., a time series cannot be stationary, nor nonstationary).
- A process is called **(strict-sense) stationary** if its statistical properties are invariant to a shift of time origin, i.e. the processes $\underline{x}(t)$ and $\underline{x}(t')$ have the same statistics for any t and t' (see further details in Papoulis, 1991; see also further explanations in Koutsoyiannis, 2006, 2011 and Koutsoyiannis and Montanari, 2015). Conversely, a process is nonstationary if some of its statistics are changing through time and their change is described as a deterministic function of time.
- A stochastic process is called **wide-sense stationary** if its mean is constant and its autocovariance depends on time difference only, i.e.

$$\mathrm{E}[\underline{x}(t)] = \mu = \text{constant}, \quad \mathrm{E}[(\underline{x}(t) - \mu)\,(\underline{x}(t+\tau) - \mu)] = c(\tau)$$

- Convenient tools for a stationary process, which can replace auto- and cross-covariance, are the following:
  - **Climacogram**: $\gamma(k) := \mathrm{var}[\underline{X}(k)/k]$, where $\underline{X}(k) := \int_0^k \underline{x}(t)\mathrm{d}t$.
  - **Cross-climacogram** of two stationary processes $\underline{x}(t)$ and $\underline{y}(t)$:

$$\gamma_{xy}^{\eta}(k) := \sigma_x \sigma_y \,\mathrm{var}\left[\frac{\underline{X}(k)}{k\sigma_x} + \frac{\underline{Y}((\eta+1)k) - \underline{Y}(\eta k)}{k\sigma_y}\right], \text{ where } \underline{Y}(k) := \int_0^k \underline{y}(t)\mathrm{d}t \text{ and } \eta \text{ is lag.}$$

Ergodicity

- Stationarity is also related to *ergodicity*, which in turn is a prerequisite to make inference from data, that is, induction. Without ergodicity inference from data would not be possible. Ironically, several studies use time series data to estimate statistical properties, as if the process were ergodic, while at the same time what they (cursorily) estimate may falsify the ergodicity hypothesis (see example on p. 22).
- While ergodicity is originally defined in dynamical systems (e.g. Mackey, 1992, p. 48), the ergodic theorem (e.g. Mackey, 1992 p. 54) allows redefining ergodicity within the stochastic processes domain (Papoulis 1991 p. 427; Koutsoyiannis 2010) in the following manner: A stochastic process $\underline{x}(t)$ is ergodic if the time average of any (integrable) function $g(\underline{x}(t))$, as time tends to infinity, equals the true (ensemble) expectation $\mathrm{E}[g(\underline{x}(t))]$, i.e., $\lim_{T\to\infty} \frac{1}{T}\int_0^T g\left(\underline{x}(t)\right) dt = \mathrm{E}[g(\underline{x}(t))]$.
- If the system that is modelled in a stochastic framework has deterministic dynamics (meaning that a system input will give a single system response, as happens for example in most hydrological models) then a theorem applies (Mackey 1992, p. 52), according to which a dynamical system has a stationary probability density *if and only if* it is ergodic. Therefore, a stationary system is also ergodic and vice versa, and a nonstationary system is also non-ergodic and vice versa.
- If the system dynamics is stochastic (a single input could result in multiple outputs), then ergodicity and stationarity do not necessarily coincide. However, recalling that a stochastic process is a model and not part of the real world, we can always conveniently device a stochastic process that is ergodic (see example in Koutsoyiannis and Montanari, 2015).
- In conclusion, from a practical point of view ergodicity can always be assumed when there is stationarity.

A note on statistical estimation

- Models are human inventions and not part of the real world. They are characterized by their mathematical structure and their parameters. The field of stochastics allows both testing the model structure and estimating the parameters, based on observation data. This is induction in practice and it is made possible by virtue of the ergodic theorem.
- We should be aware of the differences between **three concepts** related to a single parameter θ:
  - The **true** but unknown **value** θ (often called "population" parameter) .
  - The **estimator** $\underline{\hat{\theta}}$, which is a random variable depending on the stochastic process of interest $\underline{x}(t)$. $\underline{\hat{\theta}}$ is a model per se, not a number.
  - The **estimate** $\hat{\theta}$ which is a number calculated by using the observations and the estimator.
- Characteristic statistics of the estimator $\underline{\hat{\theta}}$ are its **bias**, $\mathrm{E}\left[\underline{\hat{\theta}}\right] - \theta$, and its **variance** $\mathrm{var}\left[\underline{\hat{\theta}}\right]$. When $\mathrm{E}\left[\underline{\hat{\theta}}\right] = \theta$ the estimator is called unbiased.
- As an example, the standard estimator of the **mean** from a finite set of random variables $\underline{x}_i$ (sample of size n), taken from a stochastic process $\underline{x}(t)$ at discrete time instances i, is $\underline{\hat{\mu}} \coloneqq \frac{1}{n}\sum_{i=1}^{n} \underline{x}_i$; it is easy to show that it is **unbiased**.
- However, the the standard estimator of the **variance** from the same set of random variables $\underline{x}_i$ is $\underline{\hat{\gamma}} \coloneqq \frac{1}{n-1}\sum_{i=1}^{n}\left(\underline{x}_i - \underline{\hat{\mu}}\right)^2$; even though it is often called unbiased, it is **biased**, unless $\underline{x}_i$ are independent, which is rarely the case in geophysics (see Koutsoyiannis, 2016).

References

- Ben-Naim, A., *A Farewell to Entropy: Statistical Thermodynamics Based on Information*, World Scientific Pub., Singapore, 384 pp., 2008.
- Berry, M., Regular and irregular motion, in *Topics in nonlinear dynamics: A tribute to Sir Edward Bullard*, edited by S. Jorna, American Institute of Physics, New York, 1978 (pp. 16-120)
- Gauch, H.G., Jr., *Scientific Method in Practice*, Cambridge University Press, Cambridge, 2003.
- Jaynes, E.T. Information theory and statistical mechanics, *Physical Review*, 106 (4), 620-630, 1957.
- Jaynes, E.T. *Probability Theory: The Logic of Science*, Cambridge Univ. Press, Cambridge, 728 pp., 2003.
- Kolmogorov, A. N., Grundbegrijfe der Wahrscheinlichkeitsrechnung, *Ergebnisse der Math.* (2), Berlin, 1933; 2nd English Edition: Foundations of the Theory of Probability, 84 pp. Chelsea Publishing Company, New York, 1956.
- Koutsoyiannis, D., Nonstationarity versus scaling in hydrology, *Journal of Hydrology*, 324, 239–254, 2006.
- Koutsoyiannis, D., A random walk on water, *Hydrology and Earth System Sciences*, 14, 585–601, 2010.
- Koutsoyiannis, D., Hurst-Kolmogorov dynamics and uncertainty, *Journal of the American Water Resources Association*, 47 (3), 481–495, 2011.
- Koutsoyiannis, D., *Probability and statistics for geophysical processes*, National Technical University of Athens, Athens, 2008 (itia.ntua.gr/1322/).
- Koutsoyiannis, D., Generic and parsimonious stochastic modelling for hydrology and beyond, *Hydrological Sciences Journal*, 61 (2), 225–244, doi: 10.1080/02626667.2015.1016950, 2016.
- Koutsoyiannis, D.. and Montanari, A., Negligent killing of scientific concepts: the stationarity case, *Hydrological Sciences Journal*, 60 (7-8), 1174–1183, doi:10.1080/02626667.2014.959959, 2015.
- Mackey, M.C., *Time's Arrow: The Origins of Thermodynamic Behavior*, Dover, Mineola, NY, USA, 175 pp., 2003.
- Papoulis, A., *Probability and Statistics*, Prentice-Hall, New Jersey, 1990.
- Ripley, B. D., *Stochastic Simulation*, Wiley, New York, 1987.
- Ruelle, D., Microscopic fluctuations and turbulence, *Phys. Letters*, 72A, 81-82, 1979.
- Ruelle, D., Chance and chaos, Princeton University Press, 1991.
- Shannon, C.E. The mathematical theory of communication, *Bell System Technical Journal*, 27 (3), 379-423, 1948.

ROMA TRE
UNIVERSITÀ DEGLI STUDI

Lecture notes on stochastics –Part 2

Rome, 3-4 May 2018

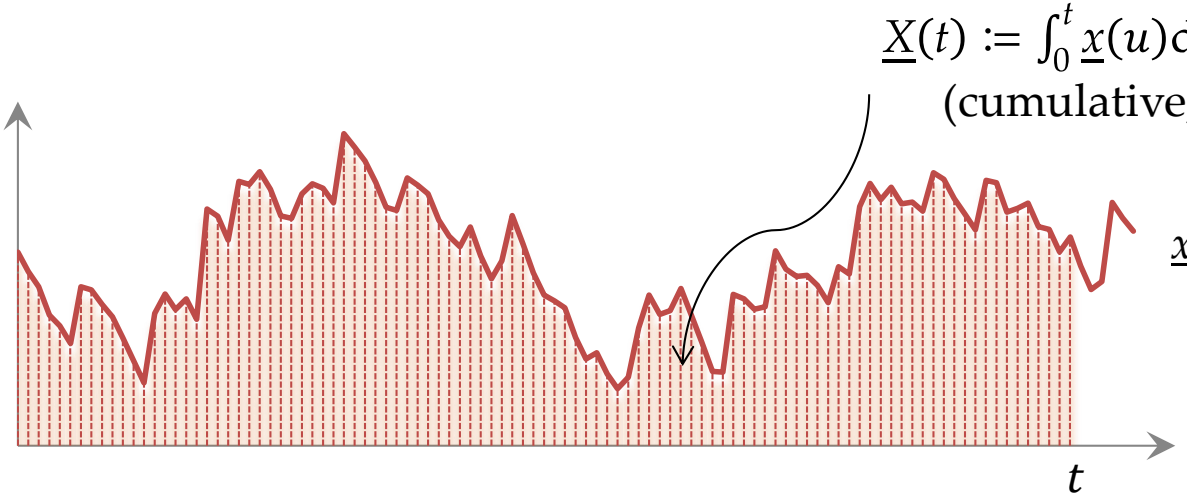
Aspects of stochastics

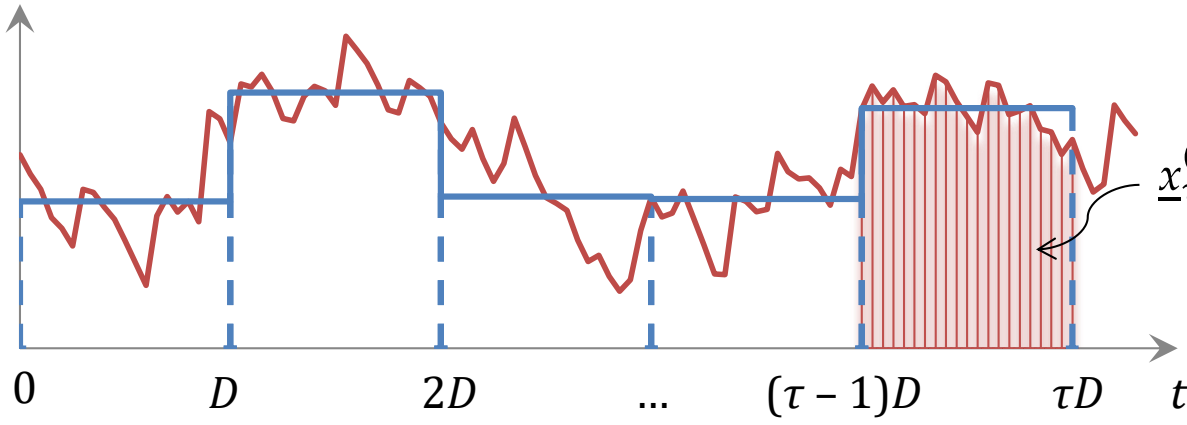
Entropy production, scaling, climacogram, climacospectrum, generic simulation

Demetris Koutsoyiannis
Department of Water Resources and Environmental Engineering
School of Civil Engineering
National Technical University of Athens, Greece
(dk@itia.ntua.gr, http://www.itia.ntua.gr/dk/)

Presentation available online: http://www.itia.ntua.gr/1835/

A stochastic process in continuous and discrete time





Note that the graphs display a realization of the process (it is impossible to display the process as such) while the notation is for the process per se.

Definitions and notation – continuous time

Name of quantity or characteristic	Symbol and definition	Remarks	Ref.
Stochastic process of interest	$\underline{x}(t)$	Assumed stationary	
Time, continuous	t	Dimensional quantity	
Cumulative process	$\underline{X}(t) \coloneqq \int_0^t \underline{x}(\xi)\mathrm{d}\xi$	Nonstationary	(1)
Variance, instantaneous	$\gamma_0 \coloneqq \mathrm{Var}[\underline{x}(t)]$	Constant (not a function of t)	(2)
Cumulative climacogram	$\Gamma(t) \coloneqq \mathrm{Var}[\underline{X}(t)]$	A function of t, $\Gamma(0) = 0$	(3)
Climacogram	$\gamma(k) \coloneqq \mathrm{Var}[(1/k)(\underline{X}(t+k) - \underline{X}(t))]$ $= \mathrm{Var}[\underline{X}(k)/k] = \Gamma(k)/k^2$	Not a function of t, $\gamma(0) = \gamma_0$	(4)
Time scale, continuous	k	Units of time	
Autocovariance function	$c(h) \coloneqq \mathrm{Cov}[\underline{x}(t), \underline{x}(t+h)]$	$c(0) = \gamma_0$	(5)
Time lag, continuous	h	Units of time	
Structure function (or semivariogram or variogram)	$v(h) \coloneqq \frac{1}{2}\mathrm{Var}\left[\underline{x}(t) - \underline{x}(t+h)\right]$		(6)
Climacostructure function	$\xi(k) \coloneqq \gamma_0 - \gamma(k)$		(7)
Power spectrum (or spectral density)	$s(w) \coloneqq 4\int_0^\infty c(h)\cos(2\pi w h)\,\mathrm{d}h$		(8)
Frequency, continuous	$w = 1/k$	Units of inverse time	(9)

Definitions and notation – discrete time

Name of quantity or characteristic	Symbol and definition	Remarks	Ref.
Stochastic process, discrete time	$\underline{x}_\tau^{(D)} := \frac{1}{D}\int_{(\tau-1)D}^{\tau D} \underline{x}(u)\mathrm{d}u = \frac{1}{D}\left(\underline{X}(\tau D) - \underline{X}((\tau-1)D)\right)$		(10)
Time unit = discretization time step	D	Length of time window of averaging	
Time, discrete	$\tau := t/D$	Dimensionless quantity, integer	(11)
Characteristic variance	$\mathrm{Var}[\underline{x}_\tau^{(D)}] = \gamma(D)$		(12)
Climacogram	$\gamma_\kappa^{(D)} = \gamma(\kappa D) = \frac{\Gamma(\kappa D)}{(\kappa D)^2}$	$\gamma_1^{(D)} = \gamma(D)$	(13)
Time scale, discrete	$\kappa = k/D$	Dimensionless quantity	(14)
Autocovariance function	$c_\eta^{(D)} := \mathrm{Cov}[\underline{x}_\tau^{(D)}, \underline{x}_{\tau+\eta}^{(D)}]$	$c_0^{(D)} = \gamma(D)$	
Time lag, discrete	$\eta = h/D$	Dimensionless quantity	(15)
Structure function	$v_\eta^{(D)} = \gamma(D) - c_\eta^{(D)}$		(16)
Power spectrum	$s_\mathrm{d}^{(D)}(\omega) = \frac{1}{D}\sum_{j=-\infty}^{\infty} s\left(\frac{\omega+j}{D}\right)\ \mathrm{sinc}^2(\pi(\omega+j))$		(17)
Frequency, discrete	$\omega = wD = 1/\kappa$	Dimensionless quantity	(18)

Note: In time-related quantities, Latin letters denote dimensional quantities and Greek letters dimensionless ones. The Latin *i, j, l* may also be used as integers to denote quantities τ, η, κ, depending on the context.

Relationships between characteristics of a process in continuous and discrete time

Related characteristics	Symbol and definition	Inverse relationship	Ref.
$\gamma(k) \leftrightarrow c(h)$	$\gamma(k) = 2\int_0^1 (1-\chi)c(\chi k)\mathrm{d}\chi$	$c(h) = \frac{1}{2}\frac{\mathrm{d}^2\left(h^2\gamma(h)\right)}{\mathrm{d}h^2}$	(19)
$s(w) \leftrightarrow c(h)$	$s(w) \coloneqq 4\int_0^\infty c(h)\cos(2\pi wh)\,\mathrm{d}h$	$c(h) = \int_0^\infty s(w)\cos(2\pi wh)\,\mathrm{d}w$	(20)
$\gamma(k) \leftrightarrow s(w)$	$\gamma(k) = \int_0^\infty s(w)\ \mathrm{sinc}^2(\pi wk)\,\mathrm{d}w$	$s(w) \coloneqq 2\int_0^\infty \frac{\mathrm{d}^2\left(h^2\gamma(h)\right)}{\mathrm{d}h^2}\cos(2\pi wh)\,\mathrm{d}h$	(21)
$v(h) \leftrightarrow c(h)$	$v(h) = \gamma_0 - c(h)$	$c(h) = v(\infty) - v(h)\ (v(\infty) = \gamma_0)$	(22)
$\xi(k) \leftrightarrow \gamma(k)$	$\xi(k) \coloneqq \gamma_0 - \gamma(k)$	$\gamma(k) = \xi(\infty) - \xi(k)\ \ (\xi(\infty) = \gamma_0)$	(23)
$\xi(k) \leftrightarrow v(h)$	$\xi(k) = 2\int_0^1 (1-\chi)v(\chi k)\mathrm{d}\chi$	$v(h) = \frac{1}{2}\frac{\mathrm{d}^2\left(h^2\xi(h)\right)}{\mathrm{d}h^2}$	(24)
$\gamma_\kappa^{(D)} \equiv \gamma(\kappa D) \leftrightarrow c_\eta^{(D)}$	$\gamma_\kappa^{(D)} = \frac{1}{\kappa}\left(c_0^{(D)} + 2\sum_{\eta=1}^{\kappa-1}\left(1-\frac{\eta}{\kappa}\right)c_\eta^{(D)}\right)$ Alternatively, $\gamma_\kappa^{(D)} = \frac{\Gamma(\kappa D)}{(\kappa D)^2}$ where, in recursive mode, $\Gamma(\kappa D) = 2\Gamma\left((\kappa-1)D\right) - \Gamma\left((\kappa-2)D\right) + 2c_{j-1}^{(D)}D^2$ with $\Gamma(0) = 0, \Gamma(D) = c_0^{(D)}D^2$	$c_\eta^{(D)} = \frac{1}{D^2}\left(\frac{\Gamma(\lvert\eta+1\rvert D)+\Gamma\left(\lvert\eta-1\rvert D\right)}{2} - \Gamma(\lvert\eta\rvert D)\right)$	(25)
$c_\eta^{(D)} \leftrightarrow s_\mathrm{d}^{(D)}(\omega)$	$s_\mathrm{d}^{(D)}(\omega) = 2c_0^{(D)} + 4\sum_{\eta=1}^\infty c_\eta^{(D)}\cos(2\pi\eta\omega)$	$c_\eta^{(D)} = \int_0^{1/2} s_\mathrm{d}^{(D)}(\omega)\cos(2\pi\omega\eta)\,\mathrm{d}\omega$	(26)
$v_\eta^{(D)} \leftrightarrow c_\eta^{(D)}$	$v_\eta^{(D)} = \gamma(D) - c_\eta^{(D)}$	$c_\eta^{(D)} \coloneqq \gamma(D) - v_\eta^{(D)}$	(27)

Asymptotic power laws and the log-log derivative

It is quite common that functions $f(x)$ defined in $[0, \infty)$, whose limits at 0 and ∞ exist, are associated with asymptotic power laws as $x \to 0$ and ∞ (Koutsoyiannis, 2014b).

Power laws are functions of the form

$$f(x) \propto x^b \tag{28}$$

A power law is visualized in a graph of $f(x)$ plotted in logarithmic axis vs. the logarithm of x, so that the plot forms a straight line with slope b. Formally, the slope b is expressed by the **log-log derivative** (LLD):

$$f^{\#}(x) := \frac{\mathrm{d}(\ln f(x))}{\mathrm{d}\,(\ln x)} = \frac{x f'(x)}{f(x)} \tag{29}$$

If the power law holds for the entire domain, then $f^{\#}(x) = b =$ constant. Most often, however, $f^{\#}(x)$ is not constant. Of particular interest are the **asymptotic values** for $x \to 0$ and ∞, symbolically $f^{\#}(0)$ and $f^{\#}(\infty)$, which **define two asymptotic power laws**.

Definition and importance of entropy

Historically entropy was introduced in thermodynamics but later it was given a rigorous definition within probability theory (owing to Boltzmann, Gibbs and Shannon). Thermodynamic and probabilistic entropy are essentially the same thing (Koutsoyiannis, 2013, 2014a; but others have different opinion).

Entropy is a dimensionless measure of uncertainty defined as follows:

For a **discrete random variable** $\underline{z}$ with probability mass function $P_j := P\{\underline{z} = z_j\}$

$$\Phi[\underline{z}] := \mathrm{E}[-\ln P(\underline{z})] = -\sum_{j=1}^{w} P_j \ln P_j \tag{30}$$

For a **continuous random variable** $\underline{z}$ with probability density function $f(z)$:

$$\Phi[\underline{z}] := \mathrm{E}\left[-\ln \frac{f(\underline{z})}{m(\underline{z})}\right] = -\int_{-\infty}^{\infty} \ln \frac{f(z)}{m(z)} f(z)\,\mathrm{d}z \tag{31}$$

where $m(z)$ is the density of a background measure (usually $m(z) = 1[z^{-1}]$).

Entropy acquires its importance from the **principle of maximum entropy** (Jaynes, 1957), which postulates that the entropy of a random variable should be at maximum, under some conditions, formulated as constraints, which incorporate the information that is given about this variable.

Its physical counterpart, the tendency of **entropy to become maximal (2nd Law of thermodynamics) is the driving force of natural change**.

Entropy production in stochastic processes

In a stochastic process the change of uncertainty in time can be quantified by the **entropy production,** i.e. the time derivative (Koutsoyiannis, 2011):

$$\Phi'[\underline{X}(t)] := \mathrm{d}\Phi[\underline{X}(t)]/\mathrm{d}t \tag{32}$$

A more convenient (and dimensionless) measure is the **entropy production in logarithmic time** (EPLT):

$$\varphi(t) \equiv \varphi[\underline{X}(t)] := \Phi'[\underline{X}(t)]\, t \equiv \mathrm{d}\Phi[\underline{X}(t)] \,/\, \mathrm{d}(\ln t) \tag{33}$$

For a Gaussian process, the entropy depends on its variance $\Gamma(t)$ only and is given as (cf. Papoulis, 1991):

$$\Phi[\underline{X}(t)] = (1/2)\ln(2\pi \mathrm{e}\, \Gamma(t)/m^2) \tag{34}$$

The EPLT of a Gaussian process is thus easily shown to be:

$$\varphi(t) = \Gamma'(t)\, t \,/\, 2\Gamma(t) = 1 + \gamma'(t)\, t \,/\, 2\gamma(t) = \tfrac{1}{2}\, \Gamma^{\#}(t) = 1 + \tfrac{1}{2}\, \gamma^{\#}(t) \tag{35}$$

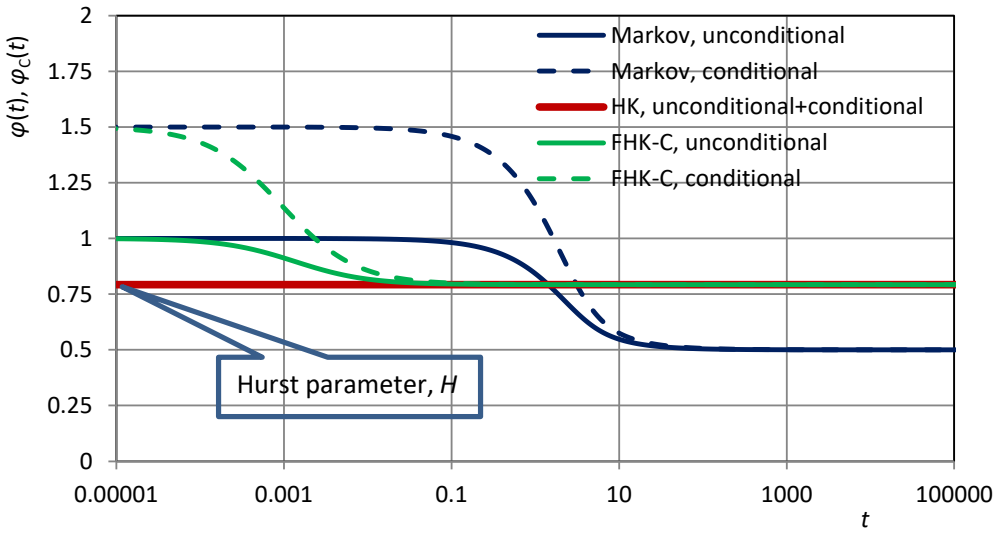
That is, **EPLT** is visualized and estimated by the **slope of a log-log plot of the climacogram.** When the past and the present are observed, instead of the unconditional variance $\gamma(t)$ we should use a variance $\gamma_C(t)$ conditional on the known past and present. This turns out to equal a **differenced climacogram** (Koutsoyiannis, 2017):

$$\gamma_C(k) = \varepsilon\big(\gamma(k) - \gamma(2k)\big), \qquad \varepsilon = \frac{1}{1 - 2^{\gamma^{\#}(\infty)}} \tag{36}$$

The **conditional entropy production in logarithmic time (CEPLT)** becomes:

$$\varphi_C(t) = 1 + \tfrac{1}{2}\gamma_C^{\#}(t) \tag{37}$$

Examples of stochastic processes and their entropy production



All three processes have same:
variance $\gamma(1) = 1$;
autocovariance for lag 1, $c_1^{(1)} = 0.5$;
fractal parameter $M = 0.5$

The HK and FHK processes have Hurst parameter $H = 0.7925$.

Markov process, maximizing entropy production for small times ($t \to 0$) but minimizing it for large times ($t \to \infty$):

$$c(h) = \lambda e^{-h/\alpha},\ \gamma(k) = \frac{2\lambda}{k/\alpha}\left(1 - \frac{1-e^{-k/\alpha}}{k/\alpha}\right) \quad (38)$$

Hurst-Kolmogorov (HK) process, maximizing entropy production for large times ($t \to \infty$) but minimizing it for small times ($t \to 0$):

$$\gamma(k) = \lambda(\alpha/k)^{2-2H} \quad (39)$$

Filtered Hurst-Kolmogorov process with a generalized Cauchy-type climacogram (FHK-C), maximizing entropy production for large ($t \to \infty$) and small times ($t \to 0$):

$$\gamma(k) = \lambda(1 + (k/\alpha)^{2M})^{\frac{H-1}{M}} \quad (40)$$

The parameters a and λ are scale parameters. The parameter H is the Hurst parameter and determines the global properties of the process with the notable property $H = \varphi(\infty) = \varphi_C(\infty)$. The parameter M (for Mandelbrot) is the fractal parameter. Both M and H are dimensionless parameters varying in the interval (0, 1] with $M < ½$ or $> ½$ indicating a rough or a smooth process, respectively, and with $H < ½$ or $> ½$ indicating an antipersistent or a persistent process, respectively (see also the graph in p. 12).

The climacospectrum

By slightly modifying the differenced climacogram (in order to make it integrable in $(0, \infty)$), i.e. by multiplying with k, we can obtain an additional tool, which resembles the power spectrum and thus is referred to as the **climacospectrum**:

$$\zeta(k) := \frac{k\big(\gamma(k)-\gamma(2k)\big)}{\ln 2} \tag{41}$$

The climacospectrum is also written in an alternative manner in terms of frequency $w = 1/k$:

$$\tilde{\zeta}(w) := \zeta(1/w) = \frac{\gamma(1/w)-\gamma(2/w)}{(\ln 2)w} \tag{42}$$

The inverse transformation, i.e., that giving the climacogram $\gamma(k)$ once the climacospectrum $\zeta(k)$ is known, is

$$\gamma(k) = \ln 2 \sum_{i=0}^{\infty} \frac{\zeta(2^i k)}{2^i k} = \gamma(0) - \ln 2 \sum_{i=1}^{\infty} \frac{\zeta(2^{-i} k)}{2^{-i} k} \tag{43}$$

As also happens with the power spectrum, the entire area under the curve $\tilde{\zeta}(w)$ is precisely equal to the variance $\gamma(0)$ of the instantaneous process. The climacospectrum has also the same asymptotic behaviour with the power spectrum, i.e.,

$$\tilde{\zeta}^{\#}(0) = -\zeta^{\#}(\infty) = s^{\#}(0), \qquad \tilde{\zeta}^{\#}(\infty) = -\zeta^{\#}(0) = s^{\#}(\infty) \tag{44}$$

This property holds almost always, with the exception of the cases where $\zeta^{\#}(0)$ is a specific integer ($\zeta^{\#}(\infty) = -1$ or $\zeta^{\#}(0) = 3$).

The climacospectrum is also connected with the CEPLT trough:

$$\varphi_C(k) = ½\left(1 + \zeta^{\#}(k)\right) = ½\left(1 - \tilde{\zeta}^{\#}(1/k)\right) \tag{45}$$

The climacogram and the climacogram-based metrics compared to more standard metrics

- In stochastic processes, almost all classical statistical estimators are biased and uncertain; in processes with LTP bias and uncertainty are very high.
- In the climacogram (variance), bias and uncertainty are easy to control as they can be calculated analytically (and a priori known; see Koutsoyiannis, 2016).
- The autocovariance function is the second derivative of the climacogram.
  - Estimation of the second derivative from data is too uncertain and makes a very rough graph.
  - Estimation of autocovariance is too biased in processes with LTP.
- The power spectrum is the Fourier transform of the autocovariance and entails an even rougher shape and more uncertain estimation than in the autocovariance (see also Dimitriadis and Koutsoyiannis, 2015).
- An additional advantage of the climacogram is its close relationship with entropy production.
- A further advantage is its expandability to high-order moments (see part 3 of the Lecture Notes).

Asymptotic scaling of second order properties

EPLT and the CEPLT are related to LLDs (slopes of log-log plots) of second order tools such as climacogram, climacospectrum, power spectrum, etc. With a few exceptions, these slopes are nonzero asymptotically, hence entailing **asymptotic scaling** or **asymptotic power laws** with the **LLDs being the scaling exponents**. It is intuitive to expect that an emerging asymptotic scaling law would provide a good approximation of the true law for a range of scales.

If the scaling law was appropriate for the entire range of scales, then we would have a simple scaling law. Such simple scaling sounds attractive from a mathematical point of view, but it turns out to be **impossible in physical processes** (Koutsoyiannis, 2017; see also the graph in p. 12).

It is thus physically more realistic to expect **two different types of asymptotic scaling** laws, one in each of the ends of the continuum of scales. The respective scaling exponents are the following:

Local scaling or **smoothness** or **fractal behaviour**, when $k \to 0$ or $w \to \infty$:

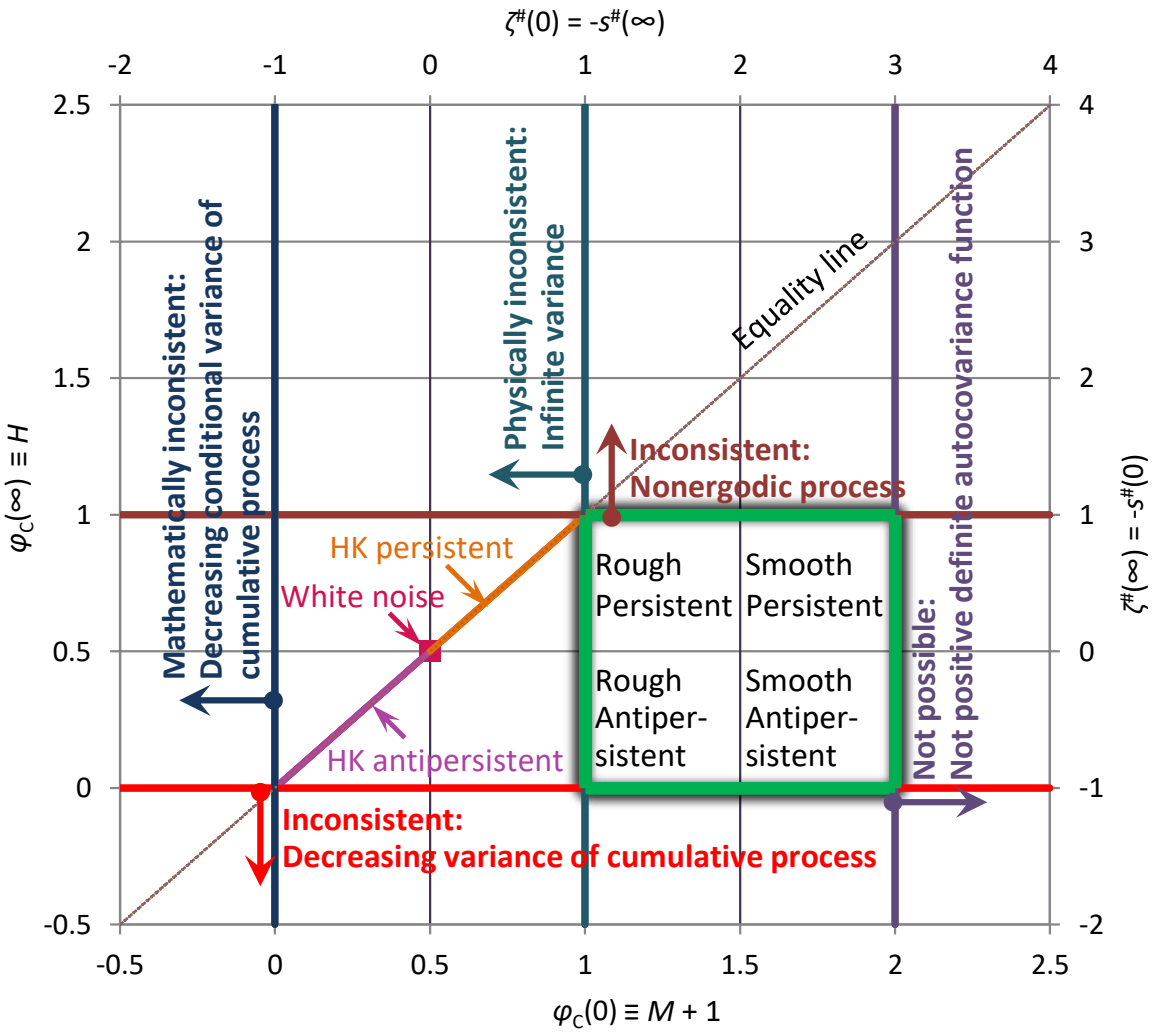
$$\gamma_C^{\#}(0) = \xi^{\#}(0) = \nu^{\#}(0) = \zeta^{\#}(0) - 1 = 2\varphi_C(0) - 2 = -s^{\#}(\infty) - 1 = 2M \tag{46}$$

Global scaling or **persistence** or **Hurst- Kolmogorov behaviour**, when $k \to \infty$ or $w \to 0$**:**

$$\gamma_C^{\#}(\infty) = \gamma^{\#}(\infty) = c^{\#}(\infty) = \zeta^{\#}(\infty) - 1 = 2\varphi_C(\infty) - 2 = -s^{\#}(0) - 1 = 2H - 2 \tag{47}$$

Here, the emergence of scaling has been related to maximum entropy considerations, and this may provide the theoretical background in modelling complex natural processes by such scaling laws. Generally, scaling laws are a mathematical necessity and could be constructed for virtually any continuous function defined in (0, ∞). In other words, there is no magic in power laws, except that they are, logically and mathematically, a necessity.

Bounds of scaling



Bounds of asymptotic values of CEPLT, $\varphi_C(0)$ and $\varphi_C(\infty)$, and corresponding bounds of the log-log slopes of power spectrum and climacospectrum.
The "green square" represents the admissible region (note that $s^{\#}$ can, by exception, take on values out of the square when $\varphi_C(0) = 2$ or $\varphi_C(\infty) = 0$). The reasons why a process out of the square would be impossible or inconsistent are also marked. The lines $\varphi_C(0) = 3/2$ and $\varphi_C(\infty) = 1/2$ define "neutrality" (which is represented by a Markov process) and support the classification of stochastic processes into the indicated four categories (smaller squares within the "green square").

Stochastic simulation

The so-called symmetric moving average (SMA) method (Koutsoyiannis, 2000) can directly generate time series with any arbitrary autocorrelation function provided that it is mathematically feasible. It consists of the following generation equation which transforms white noise $\underline{v}_i$ averaged in discrete time (and not necessarily Gaussian), to a process $\underline{x}_i$ with the specified autocorrelation:

$$\underline{x}_i = \sum_{l=-q}^{q} a_{|l|}\underline{v}_{i+l} \tag{48}$$

In theory, the limit q should be ∞ but in practice a truncation to a specific finite q is made (see Koutsoyiannis, 2016, for methods to handle the truncation error).

To calculate the series of coefficients a_l we first determine their Fourier transform $s_{\mathrm{d}}^{a}(\omega)$ from the power spectrum of the process, i.e.,

$$s_{\mathrm{d}}^{a}(\omega) = \sqrt{2 s_{\mathrm{d}}(\omega)} \tag{49}$$

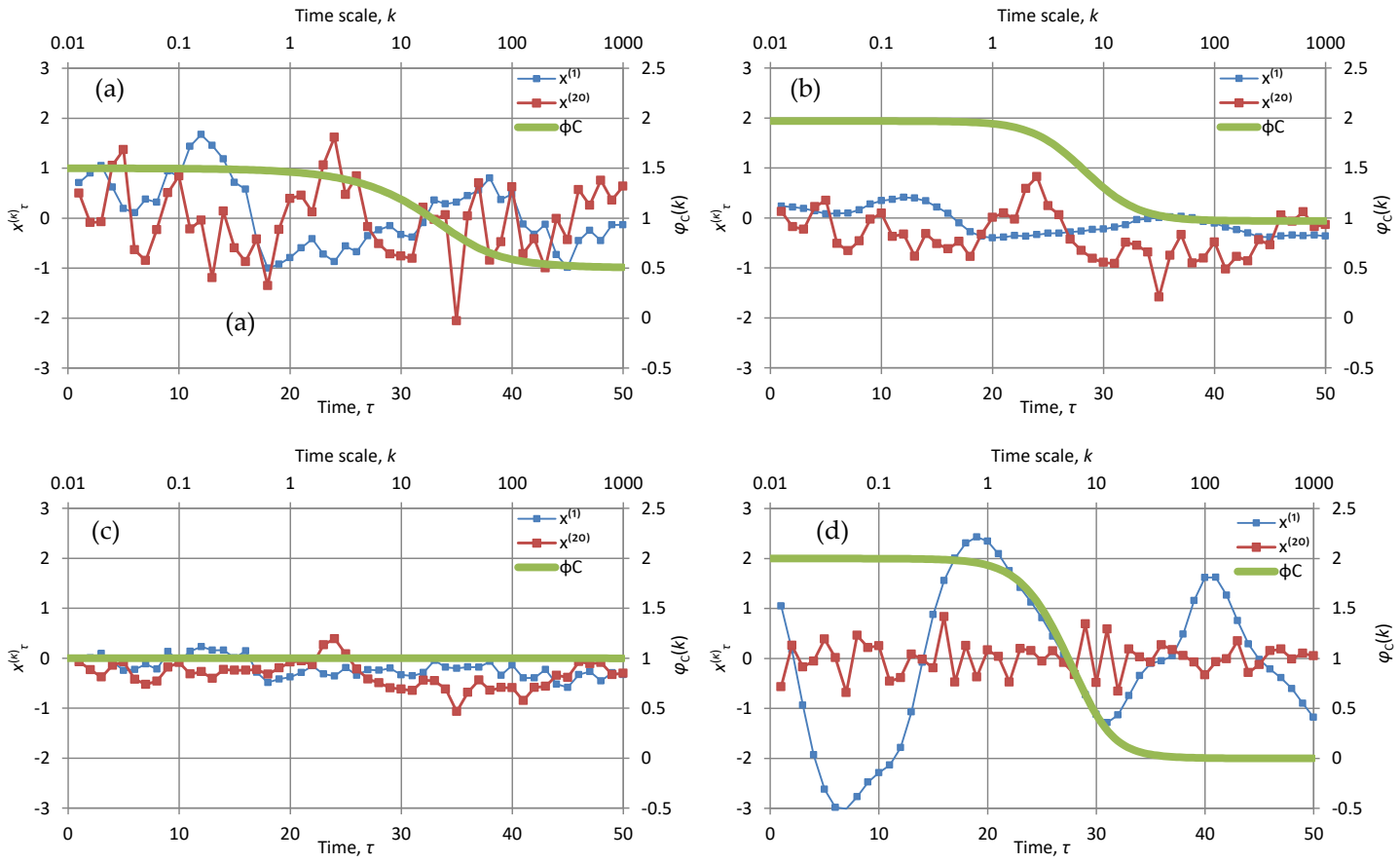
and then we inverse the transform and get the coefficients a_l. Note that the coefficients are internal constants of the model, not model parameters.

For the HK process with $H > 0.5$, there is an explicit analytical solution (Koutsoyiannis, 2016):

$$a_l = \sqrt{\frac{2\Gamma(2H+1)\sin(\pi H)\gamma(\Delta)}{\Gamma^2(H+3/2)\,(1+\sin(\pi H))}}\left(\frac{|l+1|^{H+0.5}+|l-1|^{H+0.5}}{2} - |l|^{H+0.5}\right) \tag{50}$$

By properly calculating the high-order moments of $\underline{v}_i$, we can preserve any moment of $\underline{x}_i$ that we wish (Dimitriadis and Koutsoyiannis, 2018). Thus, the scheme can handle any marginal distribution of $\underline{x}_i$.

Some results of simulations



(a) **Markov**;

(b) FHK, with **CEPLT** close to the **absolute maximum** ($H = M = 0.97$);

(c) FHK, close to "**red noise**", i.e., with CEPLT close to the absolute maximum for large scales ($H = 0.99$) and close to the absolute minimum for small scales ($M = 0.01$);

(d) process with the **blackbody** spectrum, i.e. with CEPLT equal to the absolute minimum (0) for large scales and to the absolute maximum (2) for small scales.

The first fifty terms of times series at time scales $k = 1$ and 20 of time series produced by various models, along with "stamps" of the models (green lines plotted with respect to the secondary axes) represented by the CEPLT, $\varphi_C(k)$. In all cases the discretization time scale is $D = 1$, the characteristic time scale $a = 10$, and the characteristic variance scale λ is chosen so that for time scale D, $\gamma(D) = 1$. The mean is 0 in all cases and the marginal distribution is normal (see details in Koutsoyiannis, 2017).

References

- Dimitriadis, P., and D. Koutsoyiannis, Climacogram versus autocovariance and power spectrum in stochastic modelling for Markovian and Hurst–Kolmogorov processes, *Stochastic Environmental Research & Risk Assessment*, 29 (6), 1649–1669, doi: 10.1007/s00477-015-1023-7, 2015.
- Dimitriadis, P., and D. Koutsoyiannis, Stochastic synthesis approximating any process dependence and distribution, Stochastic Environmental Research & Risk Assessment, doi:10.1007/s00477-018-1540-2, 2018.
- Jaynes, E.T., Information theory and statistical mechanics, *Physical Review*, 106 (4), 620-630, 1957.
- Koutsoyiannis, D., A generalized mathematical framework for stochastic simulation and forecast of hydrologic time series, *Water Resources Research*, 36 (6), 1519–1533, 2000.
- Koutsoyiannis, D., Hurst-Kolmogorov dynamics as a result of extremal entropy production, *Physica A*, 390 (8), 1424–1432, 2011.
- Koutsoyiannis, D., Physics of uncertainty, the Gibbs paradox and indistinguishable particles, *Studies in History and Philosophy of Modern Physics*, 44, 480–489, 2013.
- Koutsoyiannis, D., Entropy: from thermodynamics to hydrology, *Entropy*, 16 (3), 1287–1314, 2014a.
- Koutsoyiannis, D., Random musings on stochastics (Lorenz Lecture), *AGU 2014 Fall Meeting*, San Francisco, USA, American Geophysical Union, 2014b (http://www.itia.ntua.gr/en/docinfo/1500/).
- Koutsoyiannis, D., Generic and parsimonious stochastic modelling for hydrology and beyond, *Hydrological Sciences Journal*, 61 (2), 225–244, doi: 10.1080/02626667.2015.1016950, 2016.
- Koutsoyiannis, D., Entropy production in stochastics, Entropy, 19 (11), 581, doi:10.3390/e19110581, 2017.
- Papoulis, A., *Probability, Random Variables and Stochastic Processes*, 3rd edn., McGraw-Hill, New York, 1991.

ROMA TRE
UNIVERSITÀ DEGLI STUDI

Lecture notes on stochastics – Part 3

Rome, 3-4 May 2018

Knowable moments and K-climacogram

Demetris Koutsoyiannis
Department of Water Resources and Environmental Engineering
School of Civil Engineering
National Technical University of Athens, Greece
(dk@itia.ntua.gr, http://www.itia.ntua.gr/dk/)

Presentation available online: http://www.itia.ntua.gr/1835/

Introduction

Classical moments, raw or central, express important theoretical properties of probability distributions but cannot be estimated from typical samples for order beyond 2—cf. Lombardo et al. (2014): "*Just two moments!*".

L-moments are better estimated but they are all of first order in terms of the random variable of interest. They are good to characterize independent series or to infer the marginal distribution of stochastic processes, but they cannot characterize even second order dependence of processes.

Picking from both categories, we introduce K-moments, which combine advantages of both classical and L moments. They enable reliable estimation from samples (in some cases even more reliable than L moments) and effective description of high order statistics, useful for marginal and joint distributions of stochastic processes.

High-order joint statistics of stochastic properties involve multivariate functions expressing joint high-order moments. Here, by extending the notion of climacogram (Koutsoyiannis, 2010, 2016) and climacospectrum (Koutsoyiannis, 2017) we introduce the K-climacogram and the K-climacospectrum, which enable characterization of high-order properties of stochastic processes, as well as preservation thereof in simulations, in terms of univariate functions.

A note on classical moments

The classical definitions of raw and central moments of order p are:

$$\mu'_p := \mathrm{E}\left[\underline{x}^p\right], \qquad \mu_p := \mathrm{E}\left[\left(\underline{x} - \mu\right)^p\right] \tag{1}$$

respectively, where $\mu := \mu'_1 = \mathrm{E}\left[\underline{x}\right]$ is the mean of the random variable $\underline{x}$. Their standard estimators from a sample $\underline{x}_i$, $i = 1, \ldots, n$, are

$$\underline{\hat{\mu}}'_p = \frac{1}{n}\sum_{i=1}^{n} \underline{x}_i^p \,, \quad \underline{\hat{\mu}}_p = \frac{b(n,p)}{n}\sum_{i=1}^{n}\left(\underline{x}_i - \hat{\mu}\right)^p \tag{2}$$

where $a(n, p)$ is a bias correction factor (e.g. for the variance $\mu_2 =: \sigma^2$, $b(n, 2) = n/(n - 1)$). The estimators of the raw moments $\underline{\hat{\mu}}'_p$ are in theory unbiased, but it is practically impossible to use them in estimation if $p > 2$—cf. Lombardo et al. (2014), "Just two moments".

In fact, because for large p, it holds that $\left(\frac{1}{n}\sum_{i=1}^{n} x_i^p\right)^{1/p} \approx \max_{1 \le i \le n}(x_i)$ [*], we can conclude that, for an unbounded variable $\underline{x}$, asymptotically $\underline{\hat{\mu}}'_p$ **is not an estimator of $\boldsymbol{\mu'_p}$** but one of an extreme quantity, i.e., the nth order statistic raised to power p. Thus, unless p is very small, **$\boldsymbol{\mu'_p}$ is not a *knowable* quantity**: we cannot infer its value from a sample. **This is the case even if n is very large**!

[*] This is precise if x_i are positive; see also p. 5.

Definition of K-moments

To derive *knowable* moments for high orders p, in the expectation defining the pth moment we raise $\underline{x} - \mu$ to a lower power $q < p$ and for the remaining $p - q$ terms we replace $\underline{x} - \mu$ with $2F(\underline{x}) - 1$, where $F(x)$ is the distribution function. This leads to the following (central) *K-moment* definition:

$$K_{pq} \coloneqq (p - q + 1)\mathrm{E}\left[\left(2F\left(\underline{x}\right) - 1\right)^{p-q}(\underline{x} - \mu)^{q}\right] \tag{3}$$

Likewise, we define non-central K-moments as:

$$K'_{pq} \coloneqq (p - q + 1)\mathrm{E}\left[\left(F\left(\underline{x}\right)\right)^{p-q}\underline{x}^{q}\right] \tag{4}$$

The quantity $\left(2F\left(\underline{x}\right) - 1\right)^{p-q}$ is estimated from a sample without using powers of $\underline{x}$. Specifically, for the ith element of a sample $x_{(i)}$ of size n, sorted in ascending order, $F(x_{(i)})$, is estimated as $\hat{F}\left(x_{(i)}\right) = (i - 1)/(n - 1)$, thus taking values from 0 to 1 precisely and irrespective of the values $x_{(i)}$; likewise, $2F(x_{(i)}) - 1$ is estimated as $2\hat{F}\left(x_{(i)}\right) - 1 = (2i - n + 1)/(n - 1)$, taking values from –1 to 1 precisely and irrespective of the values $x_{(i)}$. Hence, the estimators are:

$$\underline{\hat{K}}'_{pq} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{i-1}{n-1}\right)^{p-q}\underline{x}_{(i)}^{q}\,,\quad \underline{\hat{K}}_{pq} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{2i-n+1}{n-1}\right)^{p-q}\left(\underline{x}_{(i)} - \hat{\mu}\right)^{q} \tag{5}$$

Rationale of the definition

1. Assuming that the distribution mean is close to the median, so that $F(\mu) \approx 1/2$ (this is precisely true for a symmetric distribution), the quantity whose expectation is taken in (3) is $A(x) \coloneqq \big(2F(\underline{x}) - 1\big)^{p-q}(\underline{x} - \mu)^q$ and its Taylor expansion is

$$A(x) = \big(2f(\mu)\big)^{p-q}(\underline{x} - \mu)^p + (p-q)\big(2f(\mu)\big)^{p-q-1} f'(\mu)(\underline{x} - \mu)^{p+1} + O((\underline{x} - \mu)^{p+2}) \tag{6}$$

where $f(x)$ is the probability density function of $\underline{x}$. Clearly then, K_{pq} depends on μ_p as well as classical moments of $\underline{x}$ of order higher than p. The independence of K_{pq} from classical moments of order < p makes it a good knowable surrogate of the unknowable μ_p.

2. As p becomes large, by virtue of the multiplicative term $(p - q + 1)$ in definition (3), K_{pq} shares similar asymptotic properties with $\hat{\mu}_p^{q/p}$ (the estimate, not the true $\mu_p^{q/p}$). To illustrate this for q = 1, we consider the variable $\underline{z} \coloneqq \max_{1 \le i \le p} \underline{x}_i$ and denote $f(\,)$ and $h(\,)$ the probability densities of $\underline{x}_i$ and $\underline{z}$, respectively. Then (Papoulis, 1990, p. 209):

$$h(z) = pf(z)\big(F(z)\big)^{p-1} \tag{7}$$

and thus, by virtue of (4),

$$\mathrm{E}[\underline{z}] = p\mathrm{E}\left[\big(F(\underline{x})\big)^{p-1}\,\underline{x}\right] = K'_{p1} \tag{8}$$

On the other hand, as seen in p. 2, for positive $\underline{x}$ and large $p \to n$,

$$E[\underline{\hat{\mu}}'_p{}^{1/p}] = E\left[\left(\frac{1}{n}\sum_{i=1}^{n} \underline{x}_i^p\right)^{1/p}\right] \approx E\left[\max_{1 \le i \le n} x_i\right] = \mathrm{E}[\underline{z}] = K'_{p1} \tag{9}$$

Note also that the multiplicative term $(p - q + 1)$ in definition (3) and (4) makes K-moments increasing functions of p.

Asymptotic properties of moment estimates

Generally, as p becomes large (approaching n), estimates of both classical and K moments, central or non-central, become estimates of expressions involving extremes such as $\left(\max_{1\le i\le p} x_i\right)^q$ or $\max_{1\le i\le p}(x_i-\mu)^q$. For negatively skewed distributions these quantities can also involve minimum, instead of maximum quantities.

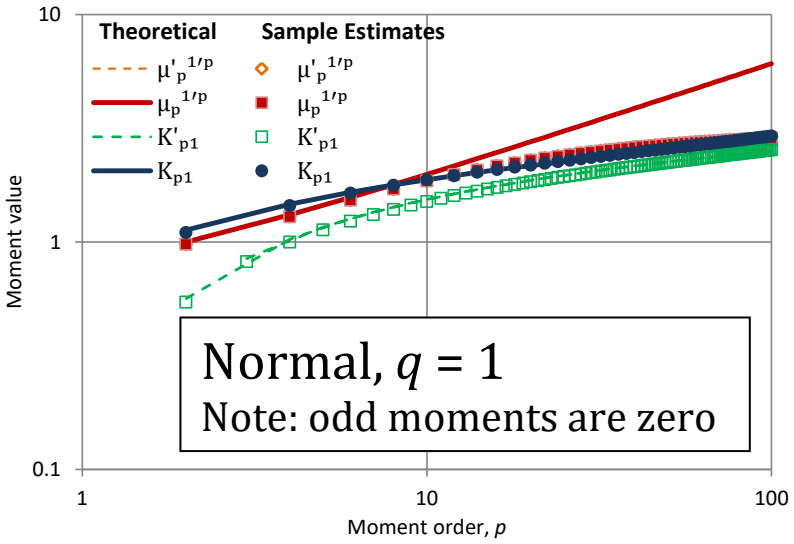
For the K-moments this is consistent with their theoretical definition. For the classical moments this is an inconsistency.

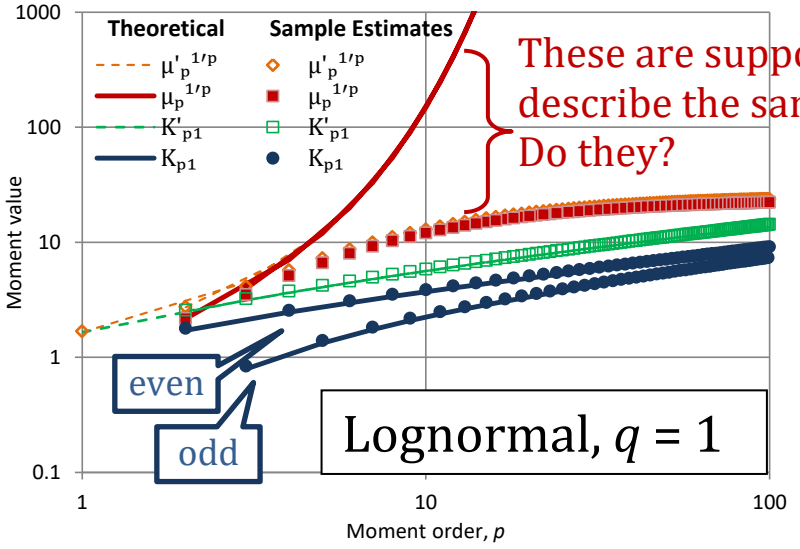
A common property of both classical and K moments is that symmetrical distributions have all their odd moments equal to zero.

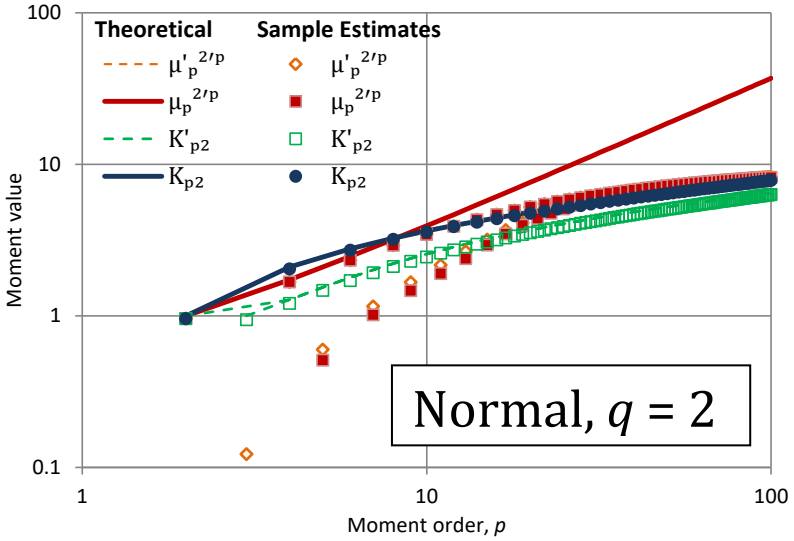
Both classical and K moments are non-decreasing functions of p, separately for odd and even p.

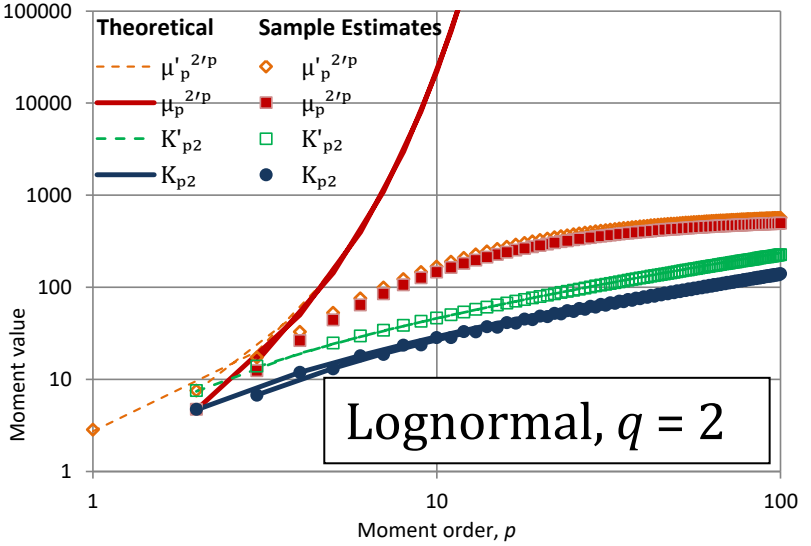
In geophysical processes we can justifiably assume that the variance $\mu_2 \equiv \sigma^2 \equiv K_{22}$ is finite (an infinite variance would presuppose infinite energy to materialize, which is absurd). Hence, high order K-moments K_{p2} will be finite too, even if classical moments μ_p diverge to infinity beyond a certain p (i.e., in heavy tailed distributions).

Justification of the notion of *unknowable* vs. *knowable*









Note: Sample sizes are ten times higher than the maximum p shown in graphs, i.e., 1000.

Relationship among different moment types

The classical moments can be recovered as a special case of K moments: $M_p \equiv K_{pp}$. In particular, in uniform distribution, classical and K moments are proportional to each other:

$$K'_{pq} \coloneqq (p-q+1)\mu'_p, \quad K_{pq} \coloneqq (p-q+1)\mu_p \tag{10}$$

The probability weighted moments (PWM), defined as $\beta_p \coloneqq \mathrm{E}\left[\underline{x}\left(F(\underline{x})\right)^p\right]$, are a special case of K-moments corresponding to $q = 1$:

$$K'_{p1} = p\beta_{p-1} \tag{11}$$

The L-moments defined as $\lambda_p \coloneqq \frac{1}{p}\sum_{k=0}^{p-1}(-1)^k \binom{p-1}{k} \mathrm{E}[\underline{x}_{(p-k):p}]$, where $\underline{x}_{k:p}$ denotes the kth order statistic in an independent sample of size p. L-moments are also related to PWM and through them to K moments. In particular, the relationships for the different types of moments for the first four orders are:

$$\begin{aligned}
K'_{11} &= \mu = \beta_0, & K_{11} &= 0 \\
K'_{21} &= 2\beta_1, & K_{21} &= 2(K'_{21} - \mu) = 4\beta_1 - 2\beta_0 = 2\lambda_2 \\
K'_{31} &= 3\beta_2, & K_{31} &= 4(K'_{31} - \mu) - 6(K'_{21} - \mu) = 12\beta_2 - 12\beta_1 + 2\beta_0 = 2\lambda_3 \\
K'_{41} &= 4\beta_3, & K_{41} &= 8(K'_{41} - \mu) - 16(K'_{31} - \mu) + 12(K'_{21} - \mu) \\
& & &= 32\beta_3 - 48\beta_2 + 24\beta_1 - 4\beta_0 = \tfrac{8}{5}\lambda_4 + \tfrac{12}{5}\lambda_2
\end{aligned} \tag{12}$$

Basic characteristics of marginal distribution

Within the framework of K-moments, we can (and should) use "Just two moments" in terms of the power of $\underline{x}$, i.e. $q = 1$ or 2, but we can obtain knowable statistical characteristics for much higher order p.

In this manner, for $p > 1$ we have two alternative options to define statistical characteristics related to moments of the distribution, as in the table below. (Which of the two is preferable depends on the statistical behaviour and in particular, mean, mode and variance of the estimator.)

Characteristic	**Order p**	**Option 1**	**Option 2**
Location	1	$K'_{11} = \mu$	
Variability	2	$K_{21} = 2(K'_{21} - \mu) = 2\lambda_2$	$K_{22} = \sigma^2$ (the classical variance)
Skewness (dimensionless)	3	$\frac{K_{31}}{K_{21}} = \frac{\lambda_3}{\lambda_2}$	$\frac{K_{32}}{K_{22}}$
Kurtosis (dimensionless)	4	$\frac{K_{41}}{K_{21}} = \frac{4}{5}\frac{\lambda_4}{\lambda_2} + \frac{6}{5}$	$\frac{K_{42}}{K_{22}}$

High order moments for stochastic processes: the K-climacogram and the K-climacospectrum

Second order properties of stochastic properties are typically expressed by the autocovariance function, $c(h) := \mathrm{cov}[\underline{x}(t), \underline{x}(t+h)]$. An equivalent description is by the power spectrum, which is the Fourier transform of the autocovariance, $s(w) := 4\int_0^\infty c(h)\cos(2\pi w h)\,\mathrm{d}h$.

Another fully equivalent description with many advantages (Dimitriadis and Koutsoyiannis 2015, Koutsoyiannis 2016) is through the climacogram, the variance of the averaged process, i.e., $\gamma(k) := \mathrm{var}[\underline{X}(k)/k]$, where $\underline{X}(t) := \int_0^t \underline{x}(\xi)\mathrm{d}\xi$. The climacogram is connected to autocovariance by $\gamma(k) = 2\int_0^1 (1-\chi)c(\chi k)\mathrm{d}\chi$ and $c(h) = \frac{1}{2}\frac{\mathrm{d}^2\left(h^2\gamma(h)\right)}{\mathrm{d}h^2}$. A surrogate of the power spectrum with several advantages over it is the climacospectrum (Koutsoyiannis, 2017) defined as $\zeta(k) := \frac{k\left(\gamma(k)-\gamma(2k)\right)}{\ln 2}$.

Full description of the third-order, fourth-order, etc., properties of a stochastic process requires functions of 2, 3, …, variables. For example, the third order properties are expressed in terms of $c_3(h_1, h_2) := \mathrm{E}[(\underline{x}(t)-\mu)\,(\underline{x}(t+h_1)-\mu)\,(\underline{x}(t+h_2)-\mu)]$.

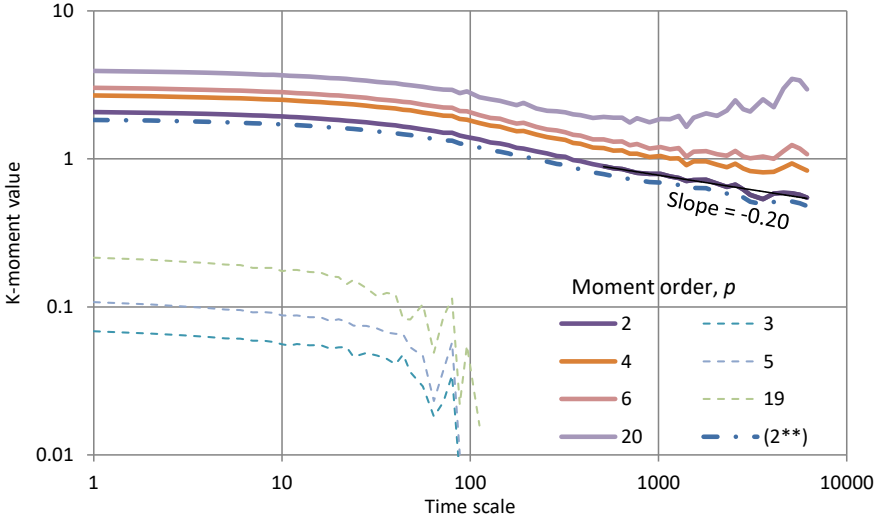
Such a description is not parsimonious and its accuracy holds only in theory, because sample estimates are not reliable. Therefore we introduce single-variable descriptions for any order p, expanding the idea of the climacogram and climacospectrum based on K-moments.

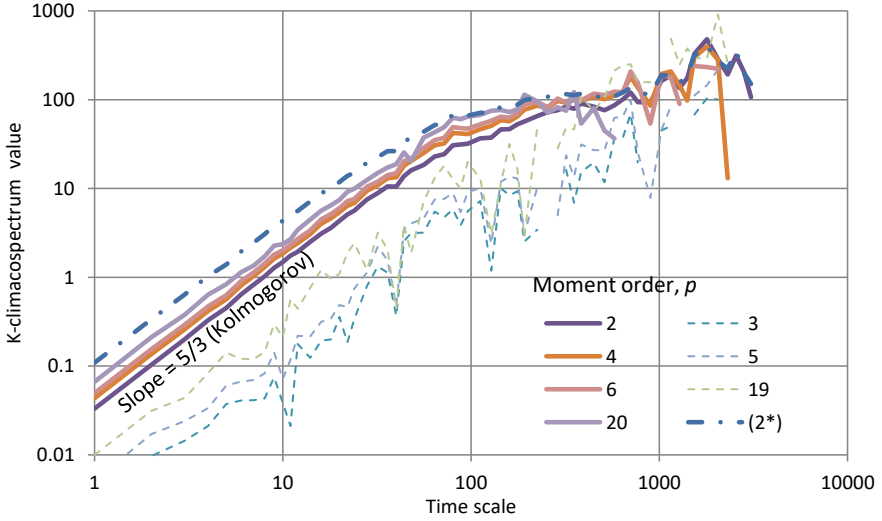
K-climacogram: $$\gamma_{pq}(k) = (p-q+1)\mathrm{E}\left[\left(2F\left(\underline{X}(k)/k\right)-1\right)^{p-q}\left(\underline{X}(k)/k-\mu\right)^q\right] \quad (13)$$

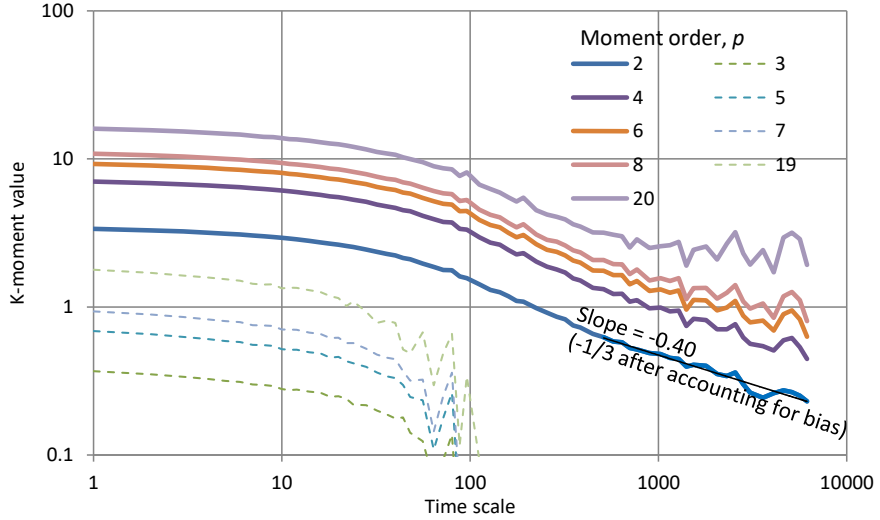
K-climacospectrum: $$\zeta_{pq}(k) = \frac{k\left(\gamma_{pq}(k)-\gamma_{pq}(2k)\right)}{\ln 2} \quad (14)$$

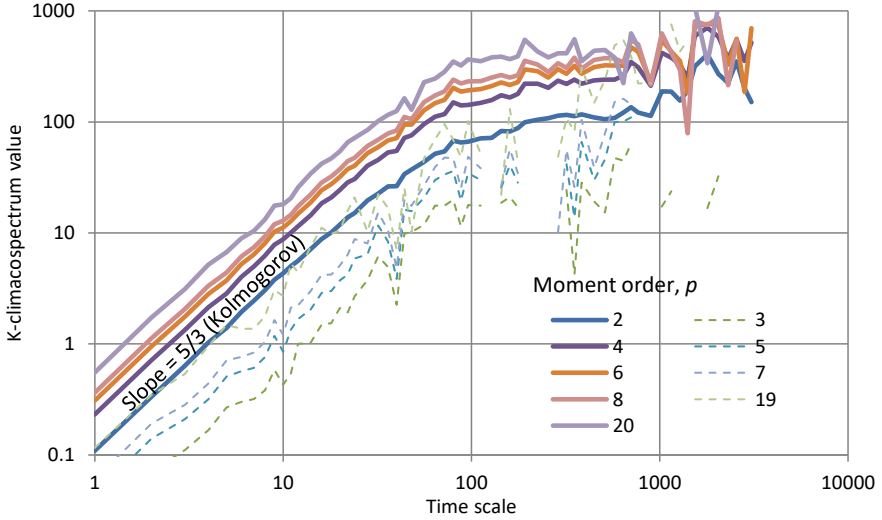
where $\gamma_{22}(k) \equiv \gamma(k)$ and $\zeta_{22}(k) \equiv \zeta(k)$. Even though the K-moment description is not equivalent to the multivariate high-order one, it suffices to define the marginal distribution at any scale k.

Example 1: Turbulent velocity



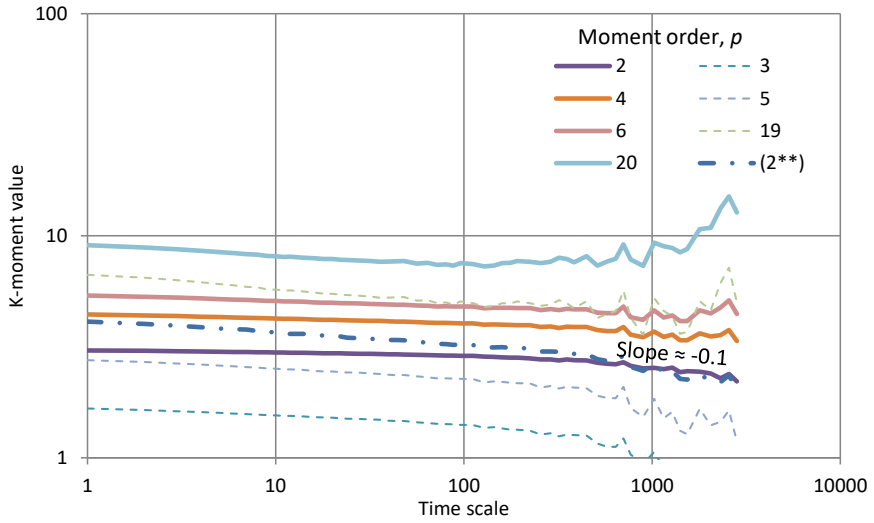


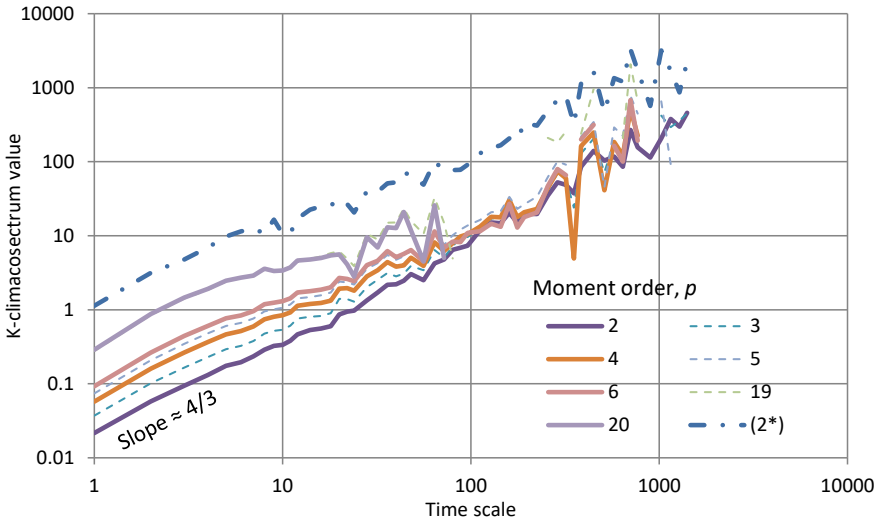


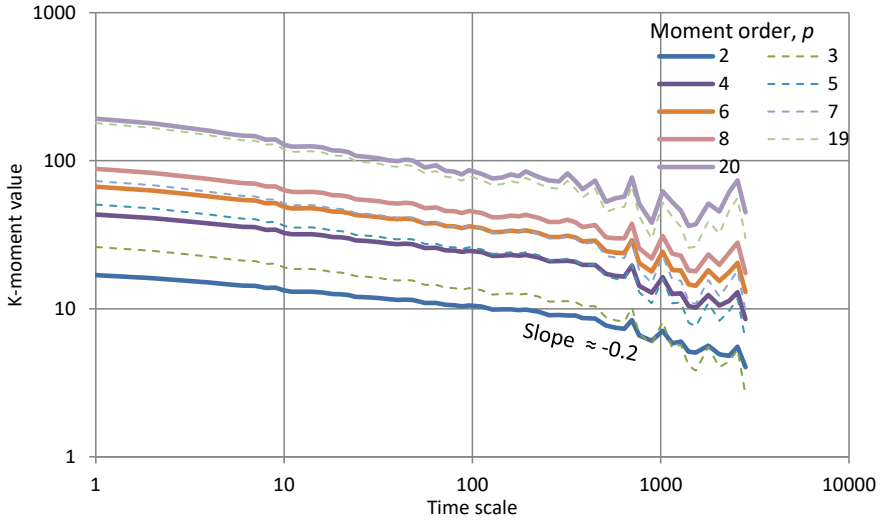


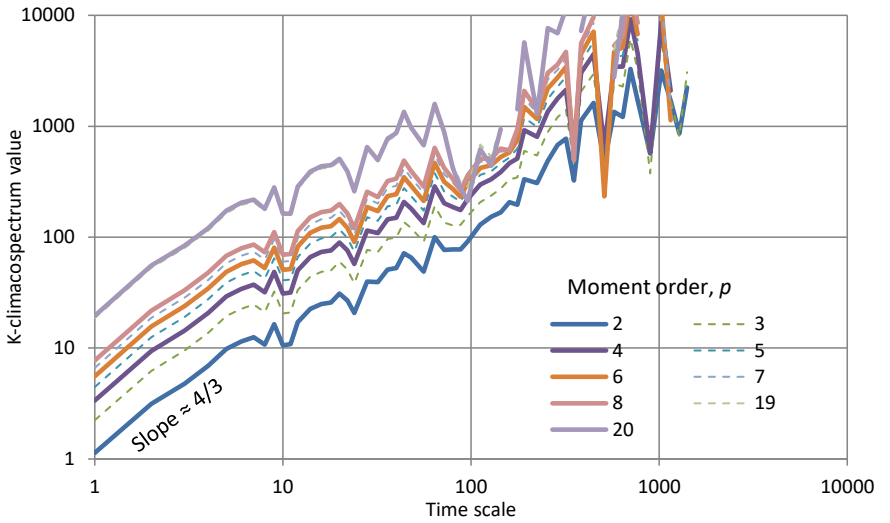
Data: 60 000 values of turbulent velocity along the flow direction (Kang, 2003; Koutsoyiannis 2017, Dimitriadis and Koutsoyiannis, 2018); the original series was averaged so that time scale 1 corresponds to 0.5 s.
Note: Plot (2*) is constructed from the variance and (2**) corresponds to standard deviation.

Example 2: Rainfall rate at Iowa measured every 10 s



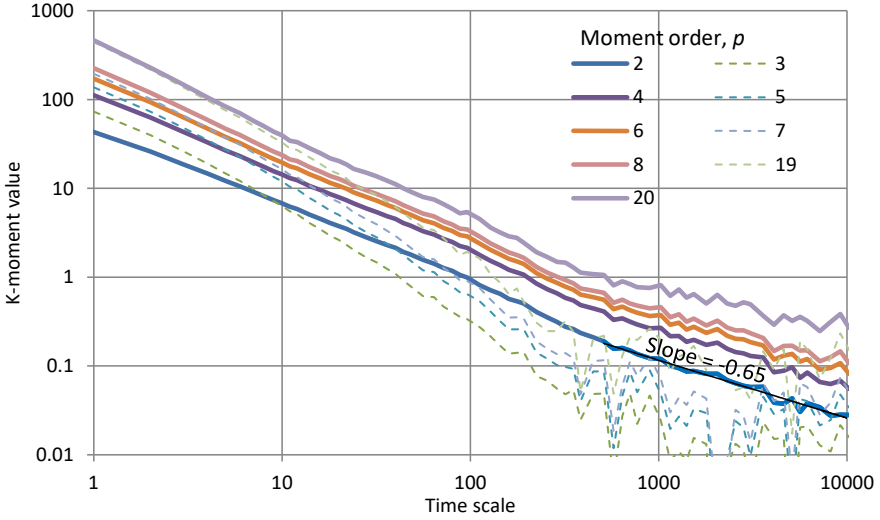


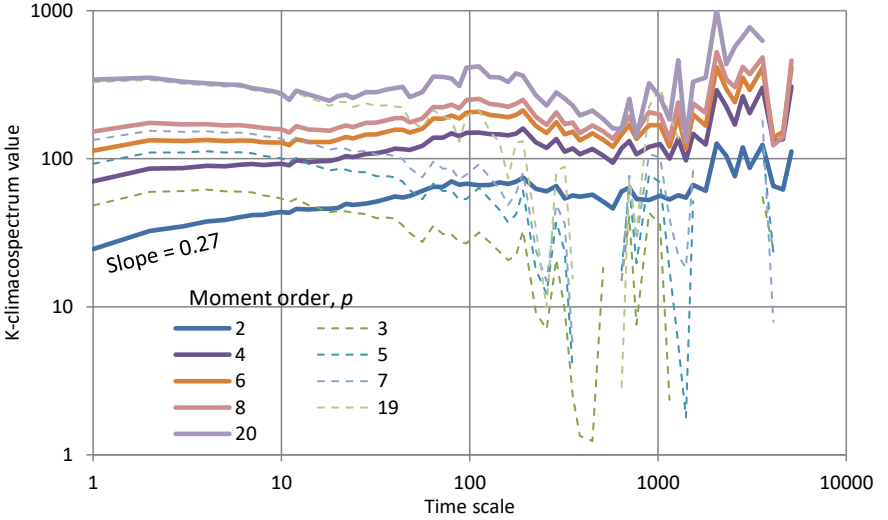


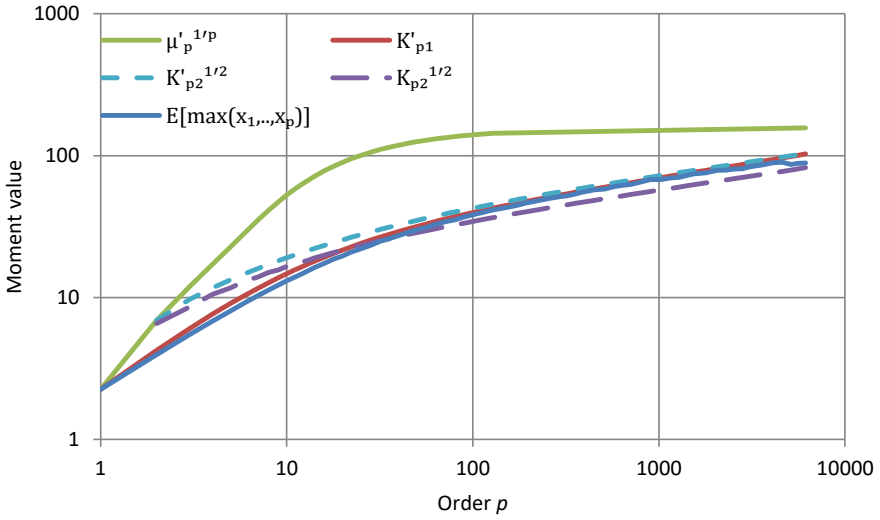


Data: 29542 values of rainfall at Iowa measured at temporal resolution of 10 s (merger of seven events from Georgakakos *et al.* 1994; see also Lombardo et al. 2012). Plot (2*) is constructed from the variance and (2**) corresponds to standard deviation.

Example 3: Daily rainfall at Padova







Data: 100 442 values of daily rainfall at Padova (the longest rainfall record existing worldwide; Marani and Zanetti, 2015).

Note about the graph on the left: Notice that moments are plotted against order p and thus approximately represent maxima for a time window of length p. For independent processes $\mathrm{E}\left[\max\left(\underline{x}_1, \ldots, \underline{x}_p\right)\right]$ should be equal to K'_{p1} , but when there is dependence the two quantities slightly differ; the former reflects the joint distribution and the latter the marginal one.

References

Dimitriadis, P., and Koutsoyiannis, D. (2015), Climacogram versus autocovariance and power spectrum in stochastic modelling for Markovian and Hurst–Kolmogorov processes, *Stochastic Environmental Research & Risk Assessment*, 29 (6), 1649–1669, doi:10.1007/s00477-015-1023-7.

Dimitriadis, P., and Koutsoyiannis, D. (2018), Stochastic synthesis approximating any process dependence and distribution, *Stochastic Environmental Research & Risk Assessment*, doi:10.1007/s00477-018-1540-2.

Georgakakos, K. P., Cârsteanu, A. A., Sturdevant, P. L. and Cramer, J. A. (1994), Observation and Analysis of Midwestern Rain Rates, *Journal of Applied Meteorology*, 33, 1433-1444.

Kang, H.S., Chester, S., and Meneveau, C. (2003). Decaying turbulence in an active-grid-generated flow and comparisons with large-eddy simulation, *Journal of Fluid Mechanics*, 480, 129–160.

Koutsoyiannis, D. (2010), A random walk on water, *Hydrology and Earth System Sciences*, 14, 585–601, doi:10.5194/hess-14-585-2010.

Koutsoyiannis, D. (2016), Generic and parsimonious stochastic modelling for hydrology and beyond, *Hydrological Sciences Journal*, 61 (2), 225–244, doi:10.1080/02626667.2015.1016950.

Koutsoyiannis, D. (2017), Entropy production in stochastics, *Entropy*, 19 (11), 581, doi:10.3390/e19110581.

Lombardo, F., Volpi, E., and Koutsoyiannis, D. (2012), Rainfall downscaling in time: Theoretical and empirical comparison between multifractal and Hurst-Kolmogorov discrete random cascades, *Hydrological Sciences Journal*, 57 (6), 1052–1066.

Lombardo, F., Volpi, E., Koutsoyiannis, D., and Papalexiou, S.M. (2014), Just two moments! A cautionary note against use of high-order moments in multifractal models in hydrology, *Hydrology and Earth System Sciences*, 18, 243–255, doi:10.5194/hess-18-243-2014.

Marani, M., and Zanetti, S. (2015), Long-term oscillations in rainfall extremes in a 268 year daily time series, *Water Resources Research*, 51, 639–647.