KNOWLEDGE FOR THE FUTURE
Joint Assembly Gothenburg IAHS - IAPSO - IASPEI Sweden 22-26 July 2013
constituent Associations of International Union of Geodesy and Geophysics (IUGG)

Session HPS1S2: Advanced Statistical Methods for Hydrology, Oceanography and Seismology

Effect of time discretization and finite record length on continuous-time stochastic properties

Federico Lombardo, Elena Volpi
Department of Engineering,
"Roma Tre" University of Rome, Italy
(federico.lombardo@uniroma3.it)

Demetris Koutsoyiannis
Department of Water Resources and Environmental Engineering, School of Civil Engineering,
National Technical University of Athens, Greece

Outline

- The reason for fitting a statistical model to data is to make conclusions about some essential characteristics of the natural process which the data refer to.

- Such conclusions can be sensitive to the degree to which the datasets reflect the salient features of the process.

- Natural processes evolve in continuous time but their observation is inevitably made at discrete time, and it has finite length.

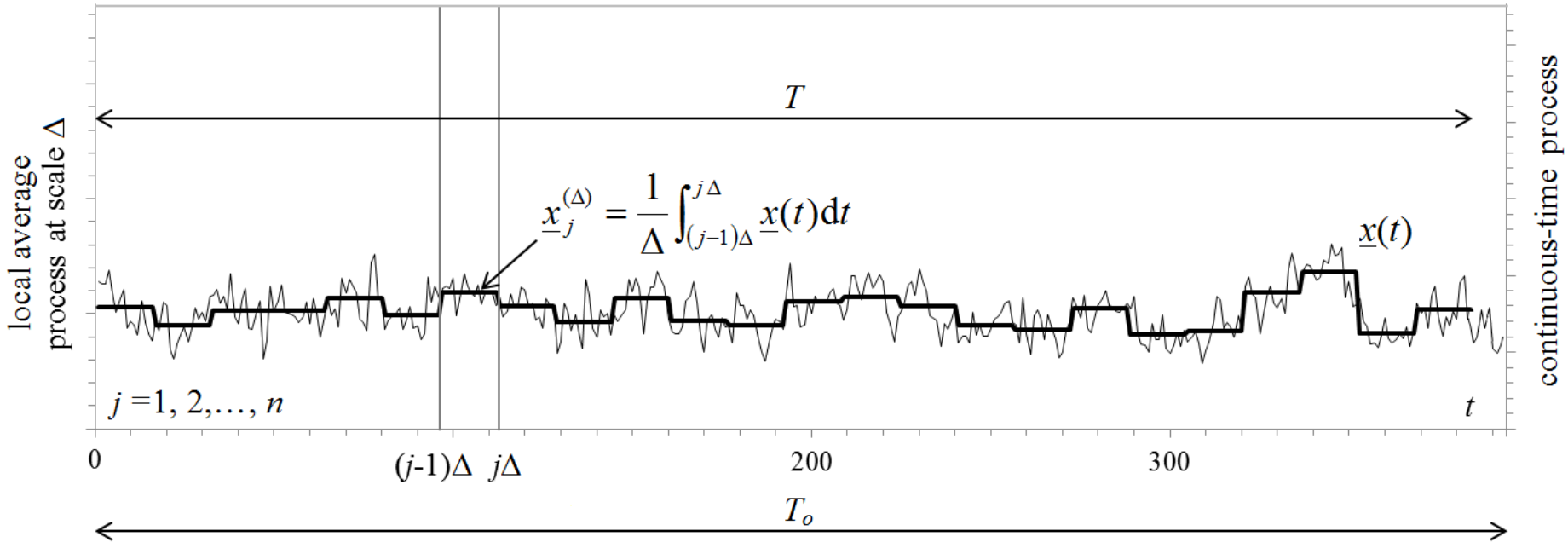
- Both time discretization and finite length may strongly affect the stochastic properties inferred from the data.

- We propose a new modelling strategy, in which the stochastic model is by definition a continuous-time process and the distortion due to discretization and finite-period observation is explicitly taken into account in model calibration.

Local averages

- Natural processes $\underline{x}(t)$ typically evolve in continuous time t, but their analysis is usually carried out in discrete-time j = 1, 2, …; practical interest often revolves around local average $\underline{x}_j^{(\Delta)}$ or aggregates of RVs over a time span Δ.

- Indeed, available data series generally consist of $n = T/\Delta$ time steps of $\underline{x}_j^{(\Delta)}$ in a finite observation period T_o, and $T = \lfloor T_o/\Delta \rfloor \Delta$ is the observation period rounded off to an integer multiple of Δ (i.e., 1/Δ is the sampling frequency).



Methodology

- We put the emphasis on autocorrelations and spectra, because they are the most extensively used concepts in the applications of stochastic processes (Papoulis, 1991).

- These concepts involve only second-order moments. Specifically, we focus on the power spectrum as well as the climacogram (Koutsoyiannis, 2010); the two are fully dependent on each other.

- The climacogram corresponds to the variance of the local average process $\underline{x}_j^{(\Delta)}$ as a function of the averaging scale Δ, which can be calculated from the autocovariance $c(\tau)$ of the continuous-time process (Vanmarke, 1983)

$$\mathrm{Var}\left[\underline{x}_j^{(\Delta)}\right] = \gamma(\Delta) = \frac{2}{\Delta^2}\int_0^{\Delta}(\Delta-\tau)c(\tau)d\tau$$

- Similarly, the power spectrum of a real continuous-time process is calculated from autocovariance function $c(\tau)$ as: $s(w) = 4\int_0^{\infty} c(\tau)\cos(2\pi w\tau)d\tau$
While for the local average process we have:
where $\omega = w\Delta \in [0, 1/2]$ $s_d^{(\Delta)}(\omega) = 2\gamma(\Delta) + 4\sum_{j=1}^{\infty} c_j^{(\Delta)}\cos(2\pi\omega j)$

Methodology

- To define uncertainty in statistical properties inferred from the data we need to specify a model for the underlying stochastic process.

- Since the statistics of a standard normal process are completely determined just in terms of its climacogram, we restrict ourselves to a discussion of a stationary, standard Gaussian stochastic process defined by a Cauchy-type climacogram:

$$\gamma(\Delta) = \lambda\left(1+(\Delta/\alpha)^{\kappa}\right)^{\frac{2H-2}{\kappa}}$$

  where we have four parameters: units of α and λ are [*time*] and $[x]^2$, respectively, while H and κ are dimensionless.

- This model was derived by modifying one proposed by Gneiting and Schlather (2004), and its important feature is that it provides power-law correlations asymptotically. Hence, it allows explicit control of both asymptotic logarithmic slopes of the climacogram $\gamma^{\#}(\Delta)$ and the power spectrum $s^{\#}(w)$:

$$s^{\#}(\infty) = -\kappa - 1; \quad \gamma^{\#}(0) = 0$$

$$s^{\#}(0) = 1 - 2H; \quad \gamma^{\#}(\infty) = 2H - 2$$

A synthetic experiment

- Knowing asymptotic stochastic properties of processes is crucial for the quantification of future uncertainty, for planning and design purposes.

- Our primary concern is to study how these properties can be better estimated from data.

- To accomplish this aim, we generate a time series of 1024 values from the known Cauchy-type process, assuming the following parameters: λ=1, α=10, H=0.8, κ=1. Hence, we have: $s^{\#}(\infty) = -\kappa - 1 = -2;$ $\gamma^{\#}(0) = 0$
  $s^{\#}(0) = 1 - 2H = -0.6;$ $\gamma^{\#}(\infty) = 2H - 2 = -0.4$

- Then, we compare the empirical power spectrum and the empirical climacogram with their known theoretical counterparts.

- Finally, we investigate the pros and cons of the climacogram-based pseudospectrum (CBPS, explained below) as a substitute of the power spectrum.

Power spectrum estimation

- We use the classical non-parametric approach (periodogram) because it explicitly estimates the power spectrum of the process without assuming that the process has any particular structure.

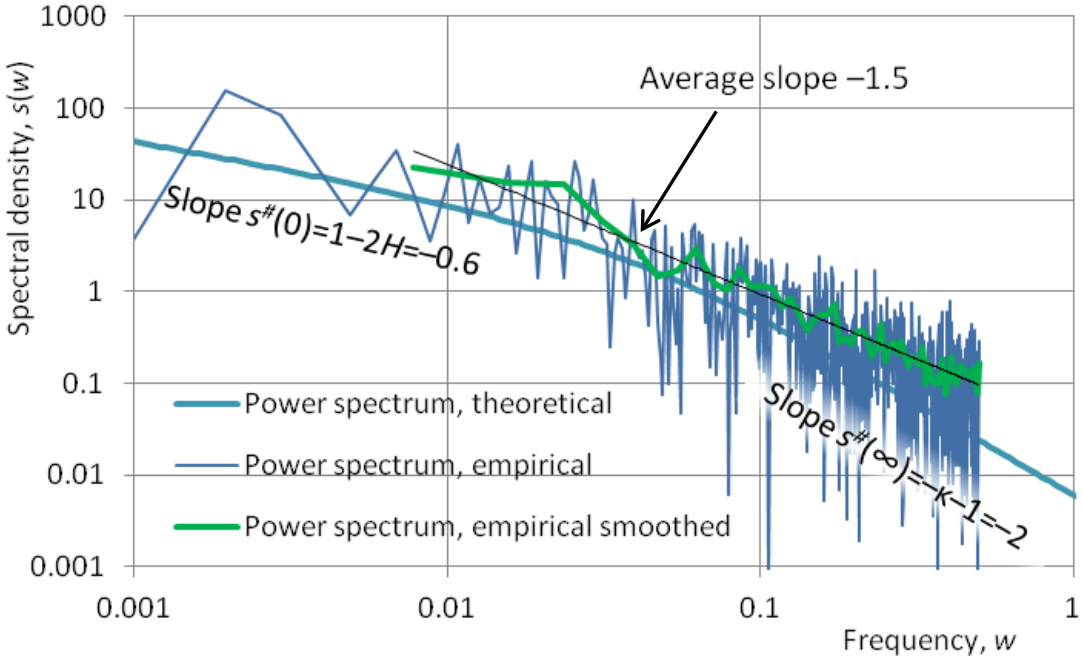
- We compute the periodogram from a finite-length digital sequence using the fast Fourier transform (FFT), that is why we chose $n=2^{10}=1024$.

- The raw periodogram is an unbiased estimator of the power spectrum only asymptotically (i.e. shorter samples cause higher bias, even when windowing the data), and it has extremely poor variance characteristics which are not affected by the length of data used (Papoulis, 1991).

- The variance problem can be reduced by smoothing the periodogram. Here we show results for the Bartlett's method, which provides estimate of the spectrum at a given frequency by averaging the estimates from the periodograms (at the same frequency) derived from a non-overlapping portions (segments) of the original series.

Power spectrum estimation

- It can be shown that we can control the power-spectrum estimator variance by averaging more segments, but shorter segments mean larger bias; so for a fixed sample size, there is a basic trade-off between segment length, which controls the bias, and the number of segments, which controls the variance.
- Both bias and uncertainty in estimation may cause problems in estimating either asymptotic slopes or statistically significant peaks. In particular, the bias depends on frequency and this distorts the estimated slopes (e.g. too steep slopes, $s^{\#}(0)<-1$: unfeasible, Koutsoyiannis 2013)



- Also, time discretization causes folding (i.e. symmetry of empirical power spectrum about the Nyquist frequency $\omega=1/2$); therefore the calculated slope $s^{\#}(½)=0$, and it does not equal the actual asymptotic slope.

Climacogram estimation

- Consider the synthetic time series x_1, …, x_n generated by our experiment, where n = 1024. Its sample variance is given by:

$$\hat{\gamma}(1) = \frac{(x_1 - x_{av})^2 + (x_2 - x_{av})^2 + \cdots + (x_{1024} - x_{av})^2}{1023}$$

where the argument (1) stands for Δ=1 and x_{av}:=(x_1+x_2+…+x_{1024})/1024 is the sample average.

- Next, we average the given time series locally over the moving time interval of size Δ=2, and find its variance (the same procedure is repeated with Δ>2)

$$x_1^{(2)} := \frac{x_1 + x_2}{2}, x_2^{(2)} = \frac{x_3 + x_4}{2}, \ldots, x_{512}^{(2)} := \frac{x_{1023} + x_{1024}}{2} \to \hat{\gamma}(2)$$

- Finite length n implies that we need to stop at Δ=$\lfloor n/10 \rfloor$=102, so that sample variance can be estimated from at least 10 data values (Koutsoyiannis, 2003)

$$x_1^{(102)} := \frac{x_1 + \cdots + x_{102}}{102}, x_2^{(102)} = \frac{x_{103} + \cdots + x_{204}}{102}, \ldots, x_{10}^{(102)} := \frac{x_{919} + x_{1020}}{102} \to \hat{\gamma}(102)$$

- The empirical climacogram (Koutsoyiannis, 2010) is the log-log plot of the sample variance versus Δ.

Climacogram estimation

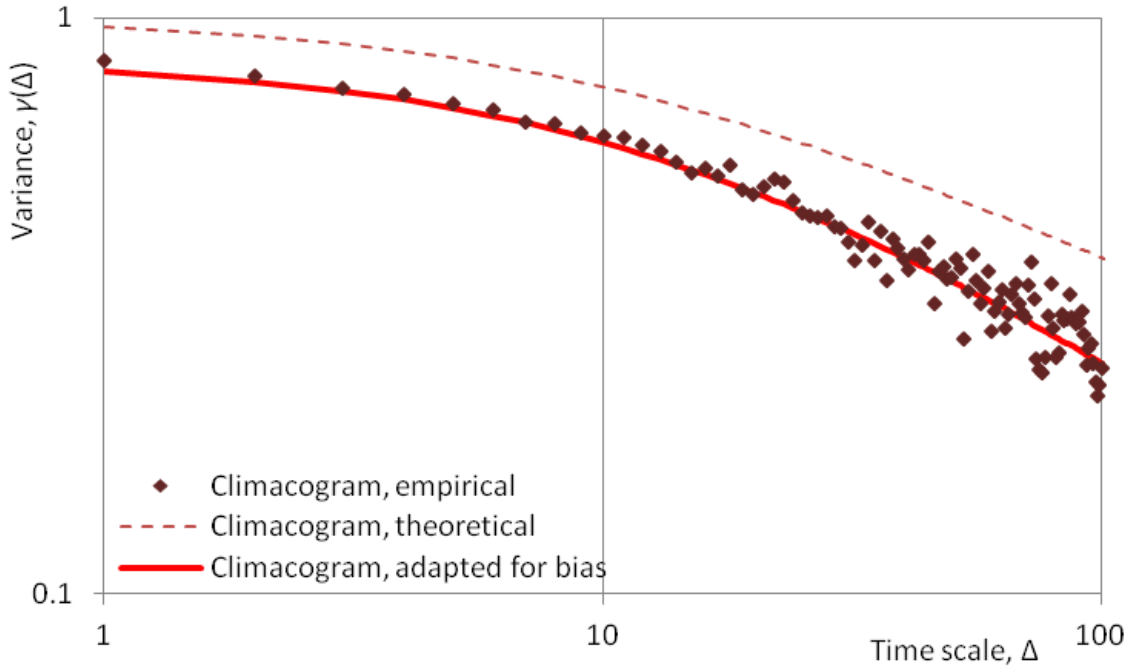
- The standard (but biased except for white noise) estimator of the variance $\gamma(\Delta)$ of averaged process $\underline{x}_j^{(\Delta)}$ is:

$$\hat{\underline{\gamma}}(\Delta) := \frac{1}{n-1}\sum_{j=1}^{n}\left(\underline{x}_j^{(\Delta)} - \underline{x}_{\text{av}}^{(\Delta)}\right)^2$$

- The following equation is used to estimate its bias (Koutsoyiannis, 2013):

$$\mathrm{E}\left[\hat{\underline{\gamma}}(\Delta)\right] = \frac{1}{1-1/n}\left(\mathrm{Var}\left[\underline{x}_j^{(\Delta)}\right] - \mathrm{Var}\left[\underline{x}_{\text{av}}^{(\Delta)}\right]\right)$$

- Being $n = T/\Delta$ and $\underline{x}_{\text{av}}^{(\Delta)} = \underline{x}_1^{(T)}$, then: $\mathrm{E}\left[\hat{\underline{\gamma}}(\Delta)\right] = \frac{1}{1-\Delta/T}\left(\gamma(\Delta) - \gamma(T)\right) = \eta(\Delta,T)\gamma(\Delta)$



- So, the bias correction coefficient η is given by:

$$\eta(\Delta,T) = \frac{1-\gamma(T)/\gamma(\Delta)}{1-\Delta/T}$$

- From the above, we see that direct estimation of $\gamma(\Delta)$ from the data is not possible; we need to know ratio $\gamma(\Delta)/\gamma(T)$: we should assume a model.

The climacogram-based pseudospectrum (CBPS)

- Once a stochastic model for the climacogram is assumed and its parameters estimated based on the data, we can estimate the variance for any timescale, including that of the instantaneous process $\gamma(0) = c(0)$.

- Therefore, the important advantage of the climacogram over other common statistical tools is that its bias can be determined analytically (usually in a closed form) and included in the estimation problem.

- The concept of climacogram can be used also in the frequency domain to find a substitute of the power spectrum, which has similar properties: The climacogram-based pseudospectrum (CBPS) defined as (Koutsoyiannis 2013)

$$\psi(w) := \frac{2\gamma(1/w)}{w}\left(1 - \frac{\gamma(1/w)}{\gamma(0)}\right)$$

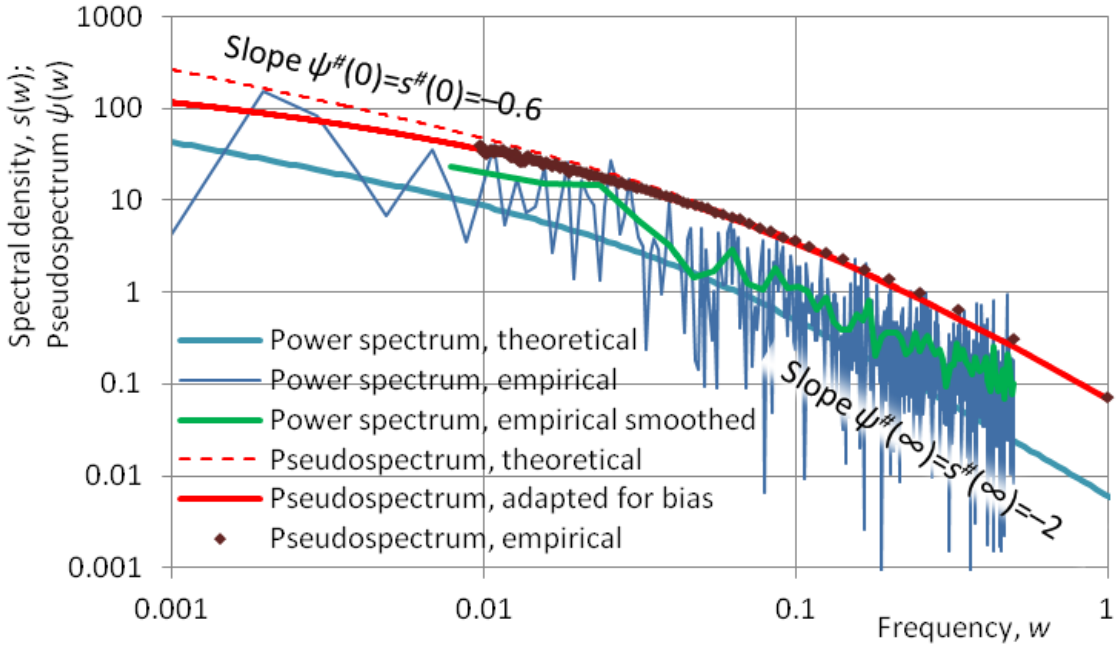
- In processes with infinite variance ($\gamma(0) = c(0) = \infty$) the CBPS simplifies to:

$$\psi(w) := \frac{2\gamma(1/w)}{w}$$

The climacogram-based pseudospectrum (CBPS)

- It can be shown that the CBPS value of at $w = 0$ equals that of the power spectrum, therefore $\psi(0) = s(0)$.

- Furthermore, the asymptotic logarithmic slopes $\psi^{\#}(w)$ of CBPS at frequencies $w \to 0$ and ∞ follow those of the power spectrum $s^{\#}(w)$, and in most processes the asymptotic slopes are precisely equal to each other. In our case we have indeed: $\psi^{\#}(0) = s^{\#}(0) = 1 - 2H$; $\psi^{\#}(\infty) = s^{\#}(\infty) = -\kappa - 1$



- At frequencies where the power spectrum has peaks, the CBPS has troughs (negative peaks).

- In contrast to the empirical periodogram, the empirical $\psi(w)$ is pretty smooth.

Conclusions

- Natural processes typically evolve in continuous time, but we observe and study them at discrete time.

- In order to make reliable inferences about the stochastic properties of natural processes, we should always be aware of the effect of time discretization and finite record length on classical statistical estimators.

- In particular, time discretization distorts the stochastic properties at small time scales, while the finite length affects the properties at large time scales.

- We mainly focus on second-order moments, and specifically on climacograms and power spectra.

- Moreover, we analyse a possible substitute of the power spectrum, which is based on the concept of climacogram.

Conclusions

- The power spectrum is very powerful in identifying strong periodicities in time series. However, it has some problems in identifying scaling laws and weak periodicities.

- Specifically, time discretization and finite length of data alter asymptotic slopes of periodograms by introducing biases and uncertainties that are uncontrollable.

- The climacogram-based pseudospectrum has an asymptotic behaviour similar to that of the power spectrum and offers some advantages.

- Indeed, we showed that when the power spectrum and pseudospectrum are estimated from data, the latter is much smoother and its bias is a priori known, thus enabling a more direct and accurate estimation of slopes and fitting on a model.

References

- Gneiting, T. and Schlather, M., *Stochastic models that separate fractal dimension and the Hurst effect*, SIAM review, 46 (2), 269-282, 2004.

- Koutsoyiannis, D., *Climate change, the Hurst phenomenon, and hydrological statistics*, Hydrological Sciences Journal, 48 (1), 3–24, 2003.

- Koutsoyiannis, D., *A random walk on water*, Hydrology and Earth System Sciences, 14, 585–601, 2010.

- Koutsoyiannis, D., *Encolpion of stochastics: Fundamentals of stochastic processes*, Department of Water Resources and Environmental Engineering – National Technical University of Athens, Athens, 2013 (itia.ntua.gr/1317/).

- Papoulis, A., *Probability, Random Variables and Stochastic Processes*, 3rd edition, McGraw Hill, 666 pp., 1991.

- Vanmarcke, E., *Random fields: Analysis and synthesis*, MIT Press, Cambridge, MA., 382 pp., 1983.