A brief introduction to Bayesian statistics

H. Tyralis

Department of Water Resources and Environmental Engineering
National Technical University of Athens

1.1 Purpose of a statistical analysis

- The purpose of a statistical analysis is fundamentally an inversion purpose.

- It aims at retrieving the causes (reduced to the parameters θ) from the effects (summarized by the observations x)

- In other words, when observing a random phenomenon directed by θ, statistical methods allow to deduce from x an inference (that is a summary, a characterization) about θ.

nature x observations

statistical analysis

causes θ (parameters, deterministic)

reduction

nature φ (real parameters)

1.2 Purpose of a statistical analysis

- The purpose of a statistical analysis is fundamentally an inversion purpose.

- It aims at retrieving the causes (reduced to the parameters θ) from the effects (summarized by the observations x)

- In other words, when observing a random phenomenon directed by θ, statistical methods allow to deduce from x an inference (that is a summary, a characterization) about θ.

- Instead probabilistic modeling characterizes the behavior of the future x conditional on θ.

nature x observations

statistical analysis

causes θ (parameters) prediction of future x

probabilistic modeling

reduction

nature φ (real parameters)

1.3 Purpose of a statistical analysis

- The purpose of a statistical analysis is fundamentally an inversion purpose.

- It aims at retrieving the causes (reduced to the parameters θ) from the effects (summarized by the observations x)

- In other words, when observing a random phenomenon directed by θ, statistical methods allow to deduce from x an inference (that is a summary, a characterization) about θ.

- Instead probabilistic modeling characterizes the behavior of the future x conditional on θ.

- This inverting aspect of statistics is obvious in the notion of the likelihood function, since, formally, it is just the sample density rewritten in the proper order

$$l(\theta|x) = f(x|\theta)$$

i.e. as a function of θ, which is unknown, depending on the observed value x.

1.4 Purpose of a statistical analysis

Example of statistical analysis

nature x observations

maximize $l(\theta|x) = f(x|\theta)$

causes θ (parameters, deterministic)

$f(x|\theta)$

nature φ (real parameters)

2. Bayes's theorem

- A general description of the inversion of probabilities is given by Bayes's theorem

$$P(A|E) = \frac{P(E|A)P(A)}{P(E|A)P(A) + P(E|A^c)P(A^c)} = \frac{P(E|A)P(A)}{P(E)}$$

- This theorem is an actualization principle since it describes the updating of the likelihood of A from $P(A)$ to $P(A|E)$.

- Bayes (1764) proved a continuous version of this result, namely that given two random variables x and y, with conditional distribution $f(x|y)$ and marginal distribution $g(y)$, the conditional distribution of y given x is

$$g(y|x) = \frac{f(x|y)g(y)}{\int f(x|y)g(y)\mathrm{d}y}$$

3.1 Definition of the Bayesian statistical model

- Bayes and Laplace went further and considered that the uncertainty on the parameters θ of a model could be modeled through a probability distribution π on Θ, called prior distribution. The inference is then based on the distribution of θ conditional on x, $\pi(\theta|x)$, called posterior distribution and defined by

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta}$$

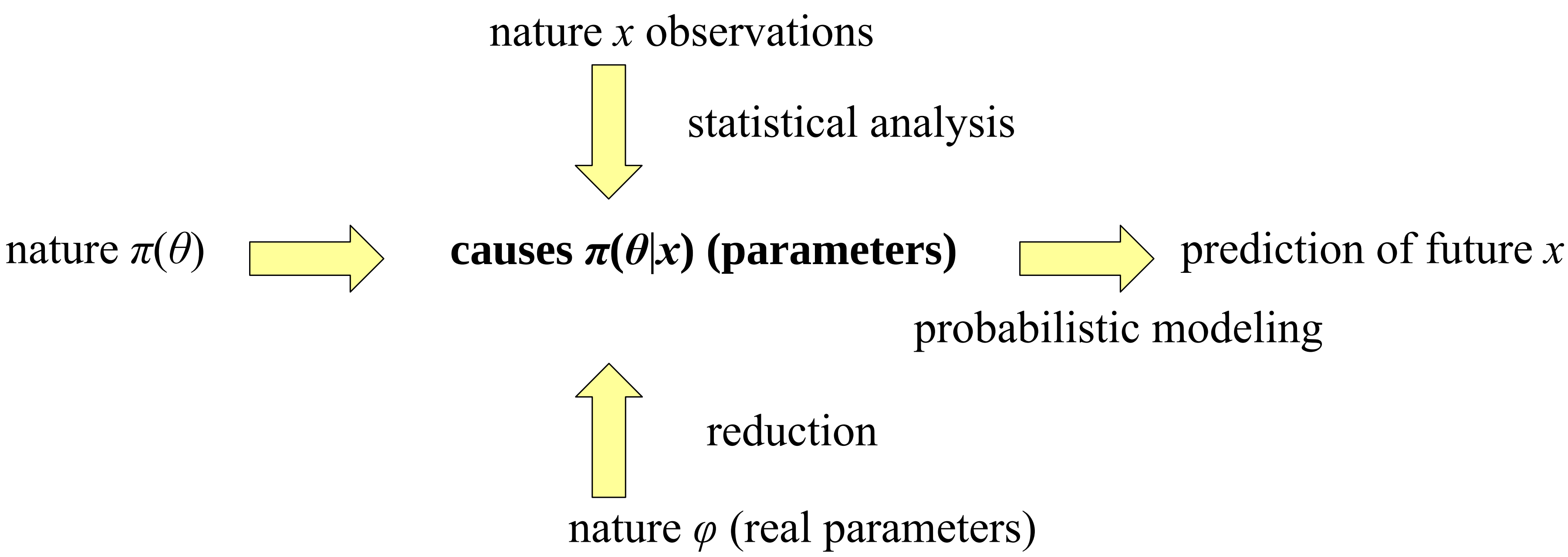
- The main addition brought by a Bayesian statistical model is thus to consider a probability distribution on the parameters

DEFINITION

A Bayesian statistical model is made of a parametric statistical model, $f(x|\theta)$ and a prior distribution on the parameters, $\pi(\theta)$.

3.2 Definition of the Bayesian statistical model

- In statistical terms, Bayes's theorem thus actualizes the information on θ contained in x.
- Its impact is based on the daring move that puts causes (observations) and effects (parameter) on the same conceptual level, since both of them have probability distributions.
- From a statistical viewpoint, there is thus little difference between observations and parameters, since conditional manipulations allow for an interplay of their respective roles.



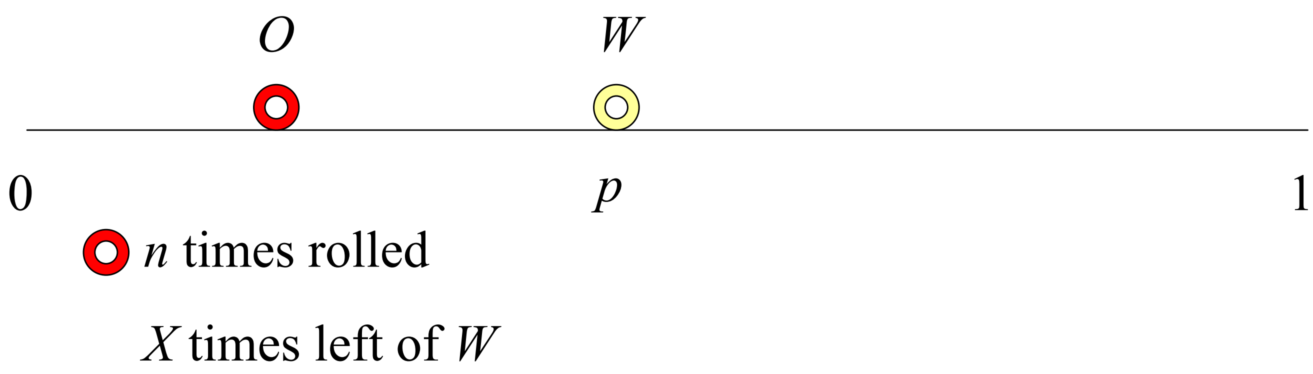
3.3 Definition of the Bayesian statistical model

- In statistical terms, Bayes's theorem thus actualizes the information on θ contained in x.
- Its impact is based on the daring move that puts causes (observations) and effects (parameter) on the same conceptual level, since both of them have probability distributions.
- From a statistical viewpoint, there is thus little difference between observations and parameters, since conditional manipulations allow for an interplay of their respective roles.
- Historically this perspective, that parameters directing random phenomena can also be perceived as random variables goes against the atheistic determinism of Laplace as well as the clerical position of Bayes, who was a nonconformist minister.
- The importance of the prior distribution in a Bayesian statistical analysis is not at all that the parameter of interest θ can (or cannot) be perceived as generated from π or even as a random variable, but rather that the use of a prior distribution is the best way to summarize the available information (or even the lack of information) about this parameter, as well as the residual uncertainty, thus allowing for incorporation of this imperfect information in the decision process.
- A more technical point is that the only way to construct a mathematically justified approach operating conditional upon the observations is to introduce a corresponding distribution on the parameters.

4.1 Bayes's example

Problem

- A billiard ball W is rolled on a line of length 1, with a uniform probability of stopping anywhere. It stops at p. A second ball O is then rolled n times under the same assumptions and X denotes the number of times the ball O stopped on the left of W. Given X, what inference can we make on p?



4.2 Bayes's example

Solution

- The problem is to derive the posterior distribution of p given X.
- The prior distribution of p is uniform on $[0,1]$, $\pi(p) = 1$.
- $X \sim B(n,p)$
- $P(X = x|p) = \binom{n}{x} p^x (1-p)^{n-x}$

- $P(a < p < b \text{ and } X = x) = \int_a^b \binom{n}{x} p^x (1-p)^{n-x} dp$

- $P(X = x) = \int_0^1 \binom{n}{x} p^x (1-p)^{n-x} dp$ and we derive that

- $$P(a < p < b|X = x) = \frac{\int_a^b \binom{n}{x} p^x (1-p)^{n-x} dp}{\int_0^1 \binom{n}{x} p^x (1-p)^{n-x} dp} = \frac{\int_a^b p^x (1-p)^{n-x} dp}{B(x+1, n-x+1)}$$

i.e. the distribution of p conditional on upon $X = x$ is a beta distribution $Be(x+1, n-x+1)$.

$Be(a,b)$: $f(x|a,b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)} I_{[0,1]}(x)$

5. Prior and posterior distributions

- Sample distribution: $f(x|\theta)$.

- Prior distribution on θ: $\pi(\theta)$.

- Joint distribution of (θ,x): $\varphi(\theta,x) = f(x|\theta)\pi(\theta)$.

- Marginal distribution of x: $m(x) = \int\varphi(\theta,x)d\theta = \int f(x|\theta)\pi(\theta)d\theta$.

- Posterior distribution of θ: $\pi(\theta|x) = \dfrac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta} = \dfrac{f(x|\theta)\pi(\theta)}{m(x)}$.

- Predictive distribution of y when $y \sim g(y|\theta,x)$: $g(y|x) = \int g(y|\theta,x)\pi(\theta|x)d\theta$.

6.1 Improper prior distributions

- When the parameter θ can be treated as a random variable with known probability distribution π(θ), Bayes's theorem is the basis of Bayesian inference, since it leads to the posterior distribution.

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta}$$

- In many cases, however, the prior distribution is determined on a subjective or theoretical basis that provides a measure π, such that

$$\int_{\Theta}\pi(\theta)d\theta = \infty$$

- In such cases the prior distribution is said to be improper (or generalized)

Example

- Consider a distribution $f(x-\theta)$ where the location parameter $\theta \in R$. If no prior distribution is available on the parameter θ, it is quite acceptable to consider that the likelihood of an interval $[a,b]$ is proportional to its length $b-a$, therefore that the prior is proportional to the Lebesque measure on R. Therefore $\int_R \pi(\theta)d\theta = \infty$.

6.2 Improper prior distributions

Be careful

- When using an improper prior distribution always check that

$$\int f(x|\theta)\pi(\theta)d\theta < \infty$$

- This leads to a proper posterior distribution

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta}$$

7.1 Another example

- Consider $x \sim N(\theta,1)$.

- Sample distribution: $f(x|\theta) = (1/\sqrt{2\pi})\ \exp[-(1/2)(x-\theta)^2]$.

- Prior distribution: $\pi(\theta) = c,\ \theta \in R$.

- Posterior distribution of θ: $\pi(\theta|x) = \dfrac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta}$

- $f(x|\theta)\pi(\theta) = c(1/\sqrt{2\pi})\ \exp[-(1/2)(x-\theta)^2]$

- $\int_R f(x|\theta)\pi(\theta)d\theta = \int_R c(1/\sqrt{2\pi})\ \exp[-(1/2)(x-\theta)^2]d\theta = m(x)$

- This leads to $\pi(\theta|x) = \dfrac{c(1/\sqrt{2\pi})\ \exp[(1/2)(x-\theta)^2]}{m(x)} \propto \exp[-(1/2)(x-\theta)^2] = \exp[-(1/2)(\theta-x)^2]$

- In fact $\pi(\theta|x) = h(x)\exp[-(1/2)(\theta-x)^2]$

- Now look at the equivalence:

$y \sim N(\mu,1)$: $f(y|\mu) = (1/\sqrt{2\pi})\ \exp[-(1/2)(y-\mu)^2] = h(\mu)\exp[-(1/2)(y-\mu)^2] \propto \exp[-(1/2)(y-\mu)^2]$

θ: $\pi(\theta|x) \propto \exp[-(1/2)(\theta-x)^2]$

Thus $\theta|x \sim N(x,1)$

7.2 Another example

parameter θ

Prior knowledge
(or lack) on θ

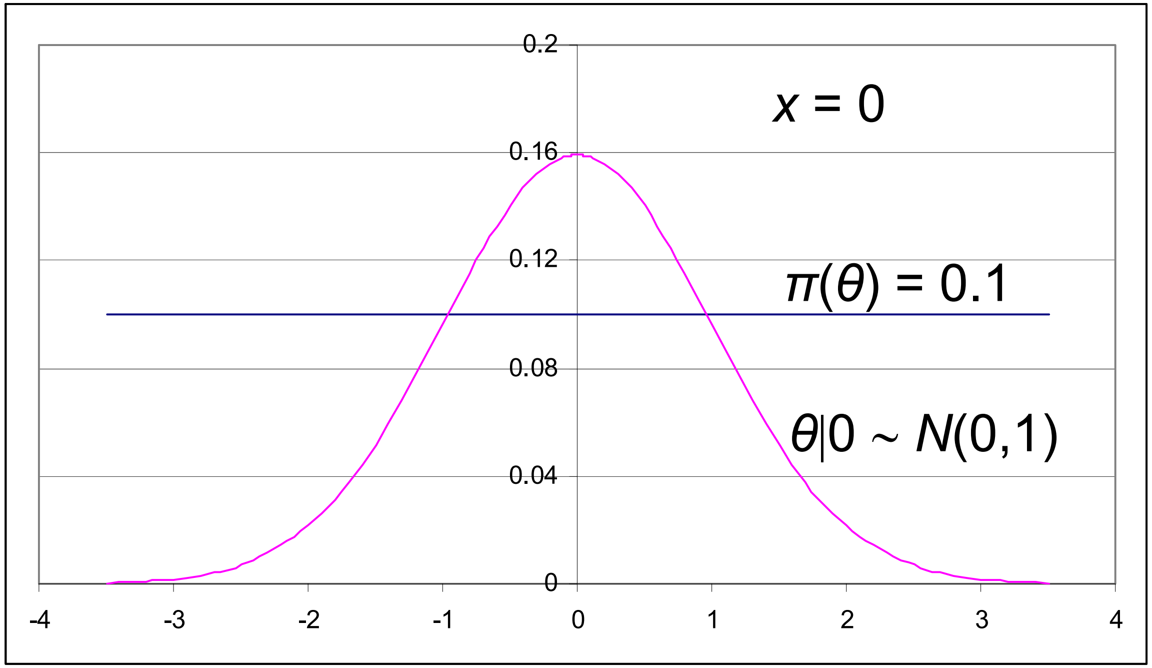
prior distribution $\pi(\theta)$

Observations x

Determination of sample density $f(x|\theta)$

posterior distribution $\pi(\theta|x)$

inference on θ



8.1 Hurst-Kolmogorov stochastic process

- Parameter: $\boldsymbol{\theta} = (\mu, \sigma^2, H)$

- Sample distribution: $f(\boldsymbol{x}_n|\boldsymbol{\theta}) = \dfrac{1}{(2\pi)^{n/2}}\ [\det(\sigma^2\, \boldsymbol{R})]^{-1/2} \exp[-1/(2\sigma^2)\ (\boldsymbol{x}_n - \mu\, \boldsymbol{e})^{\mathrm{T}}\, \boldsymbol{R}^{-1}\, (\boldsymbol{x}_n - \mu\, \boldsymbol{e})]$

- Prior distribution on $\boldsymbol{\theta}$: $\pi(\boldsymbol{\theta}) \propto 1/\sigma^2$

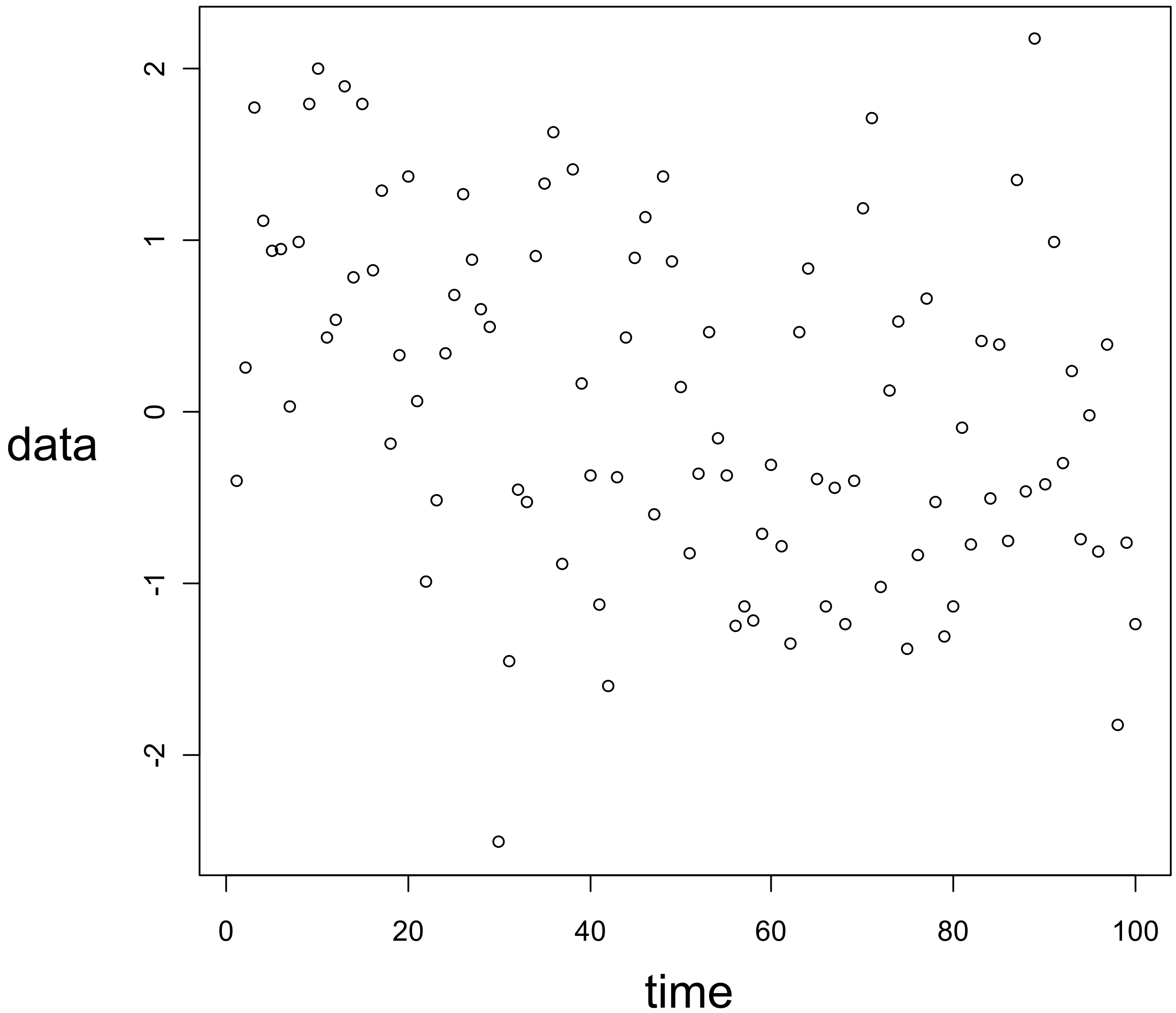
- Posterior distribution of $\boldsymbol{\theta}$:

$$\pi(H|\boldsymbol{x}_n) \propto (1/\sqrt{|\boldsymbol{R}|})\ [\boldsymbol{e}^{\mathrm{T}}\, \boldsymbol{R}^{-1}\, \boldsymbol{e}\, \boldsymbol{x}_n^{\mathrm{T}}\, \boldsymbol{R}^{-1}\, \boldsymbol{x}_n - (\boldsymbol{x}_n^{\mathrm{T}}\, \boldsymbol{R}^{-1}\, \boldsymbol{e})^2]^{-(n-1)/2}\ (\boldsymbol{e}^{\mathrm{T}}\, \boldsymbol{R}^{-1}\, \boldsymbol{e})^{n/2-1}$$

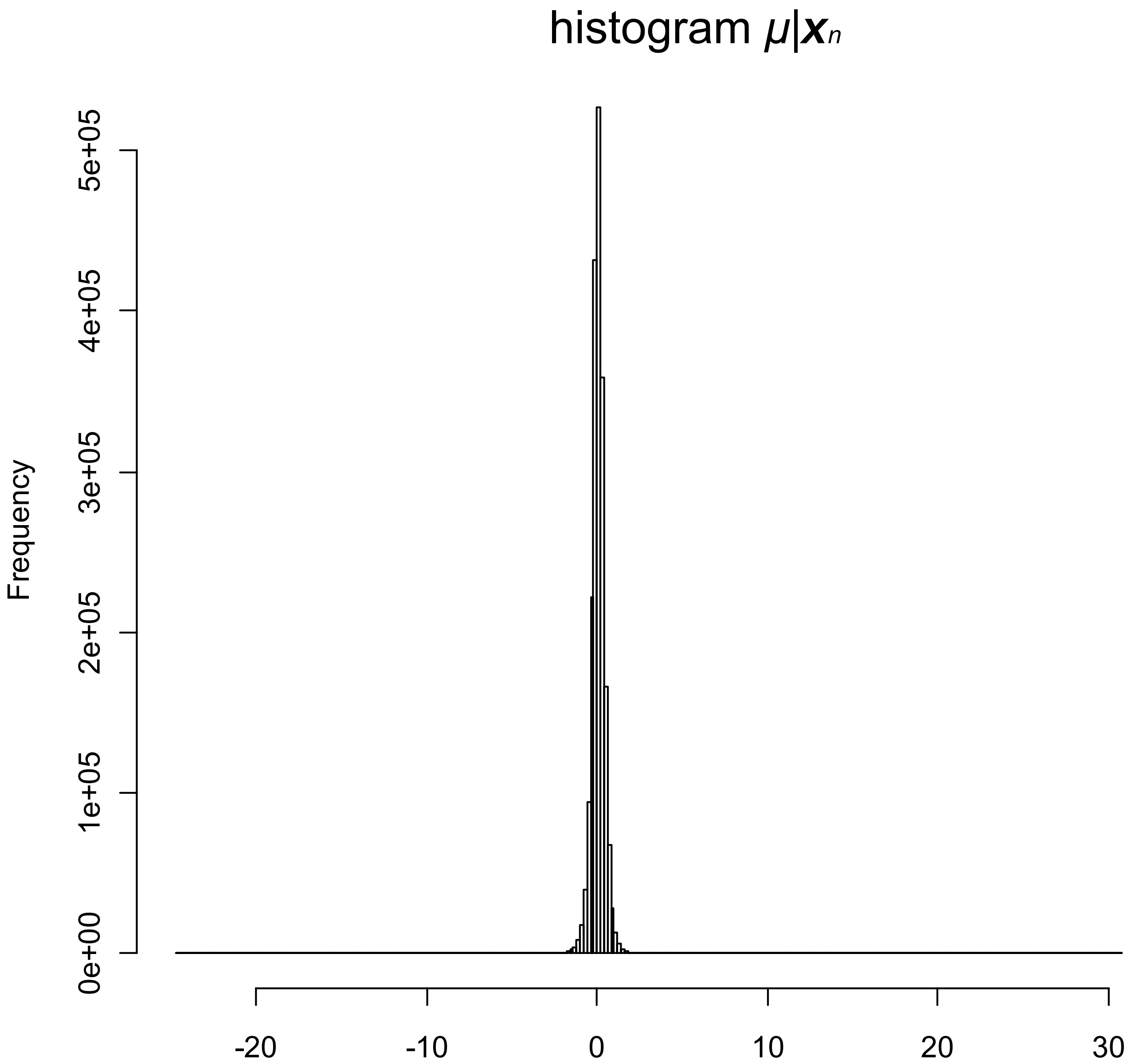
$$\sigma^2|H,\boldsymbol{x}_n \sim \text{inv-gamma}\left(\frac{n-1}{2}, \frac{\boldsymbol{e}^{\mathrm{T}}\, \boldsymbol{R}^{-1}\, \boldsymbol{e}\, \boldsymbol{x}_n^{\mathrm{T}}\, \boldsymbol{R}^{-1}\, \boldsymbol{x}_n - (\boldsymbol{x}_n^{\mathrm{T}}\, \boldsymbol{R}^{-1}\, \boldsymbol{e})^2}{2\, \boldsymbol{e}^{\mathrm{T}}\, \boldsymbol{R}^{-1}\, \boldsymbol{e}}\right)$$

$$\mu|\sigma^2,H,\boldsymbol{x}_n \sim N\left(\frac{\boldsymbol{x}_n^{\mathrm{T}}\, \boldsymbol{R}^{-1}\, \boldsymbol{e}}{\boldsymbol{e}^{\mathrm{T}}\, \boldsymbol{R}^{-1}\, \boldsymbol{e}}, \frac{\sigma^2}{\boldsymbol{e}^{\mathrm{T}}\, \boldsymbol{R}^{-1}\, \boldsymbol{e}}\right)$$

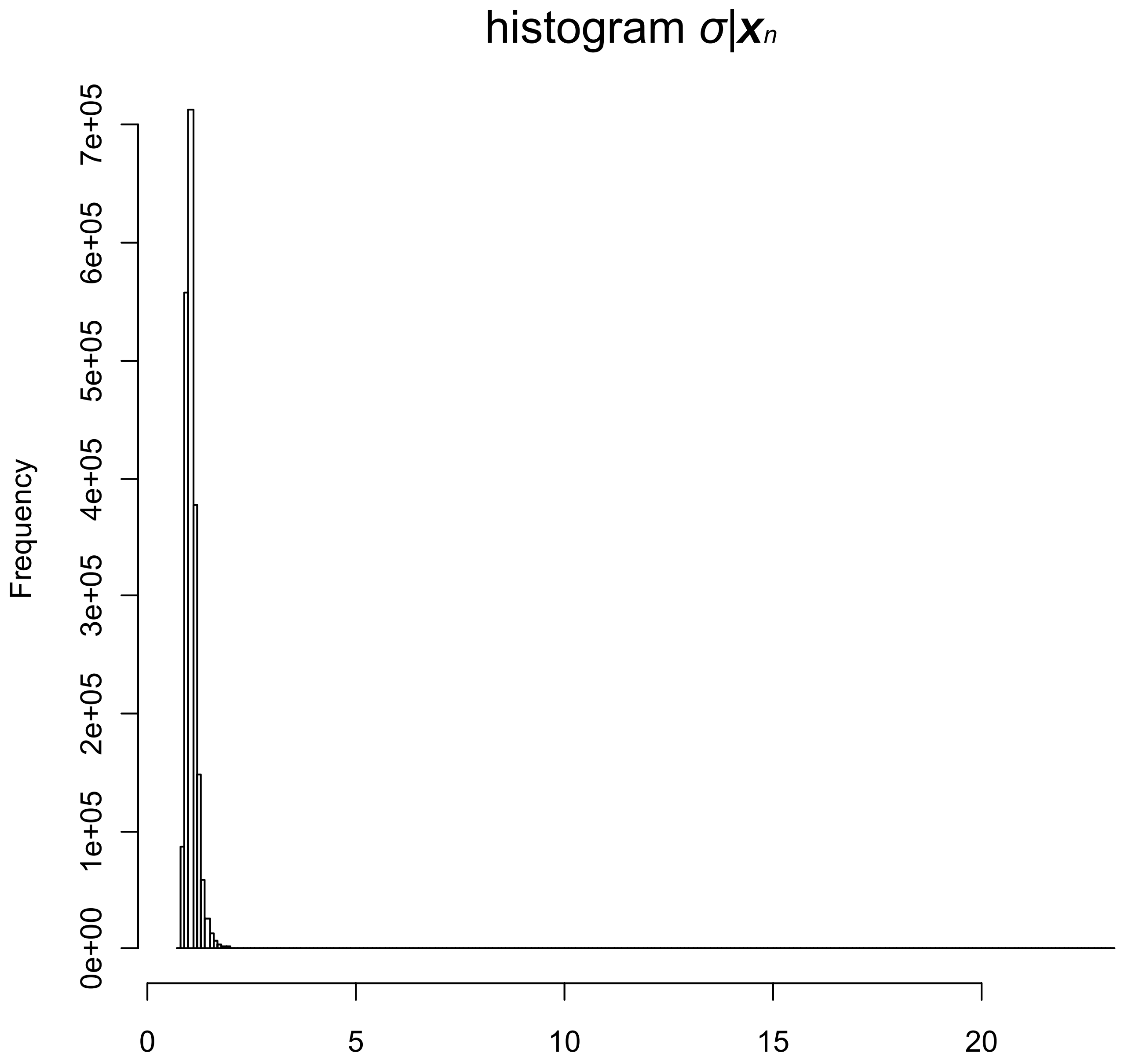
8.2 Hurst-Kolmogorov stochastic process



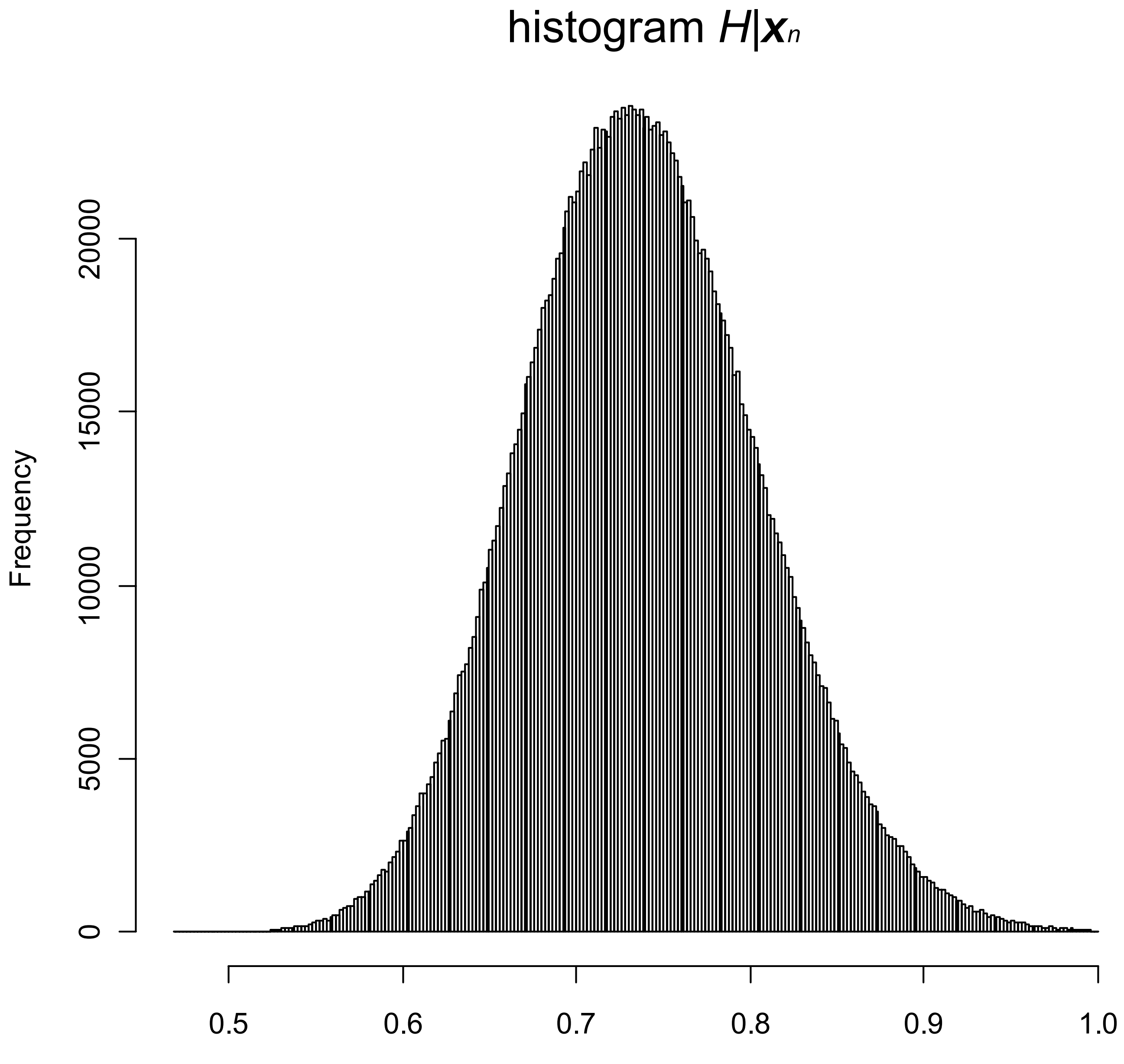
8.3 Hurst-Kolmogorov stochastic process



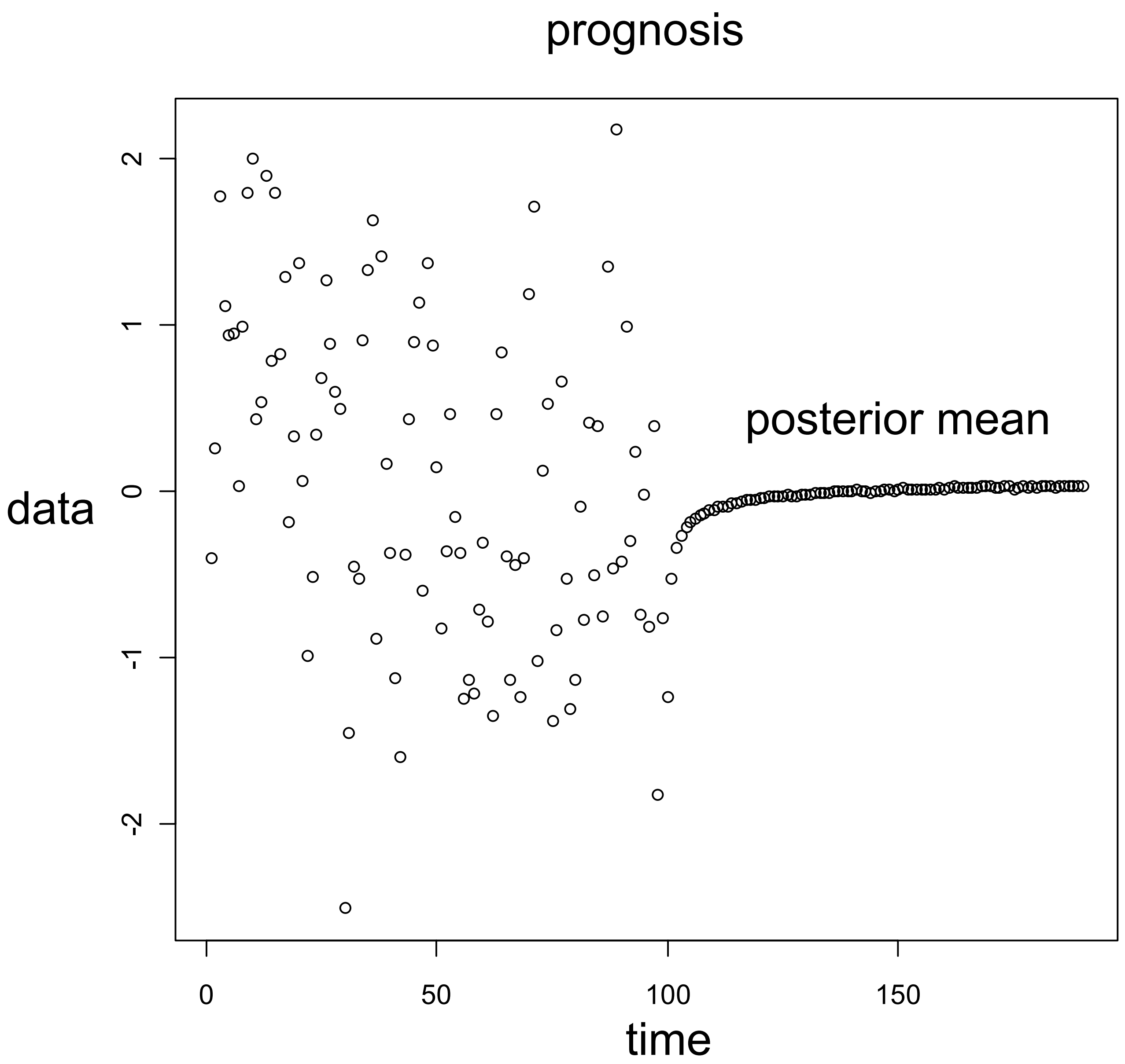
8.4 Hurst-Kolmogorov stochastic process



8.5 Hurst-Kolmogorov stochastic process

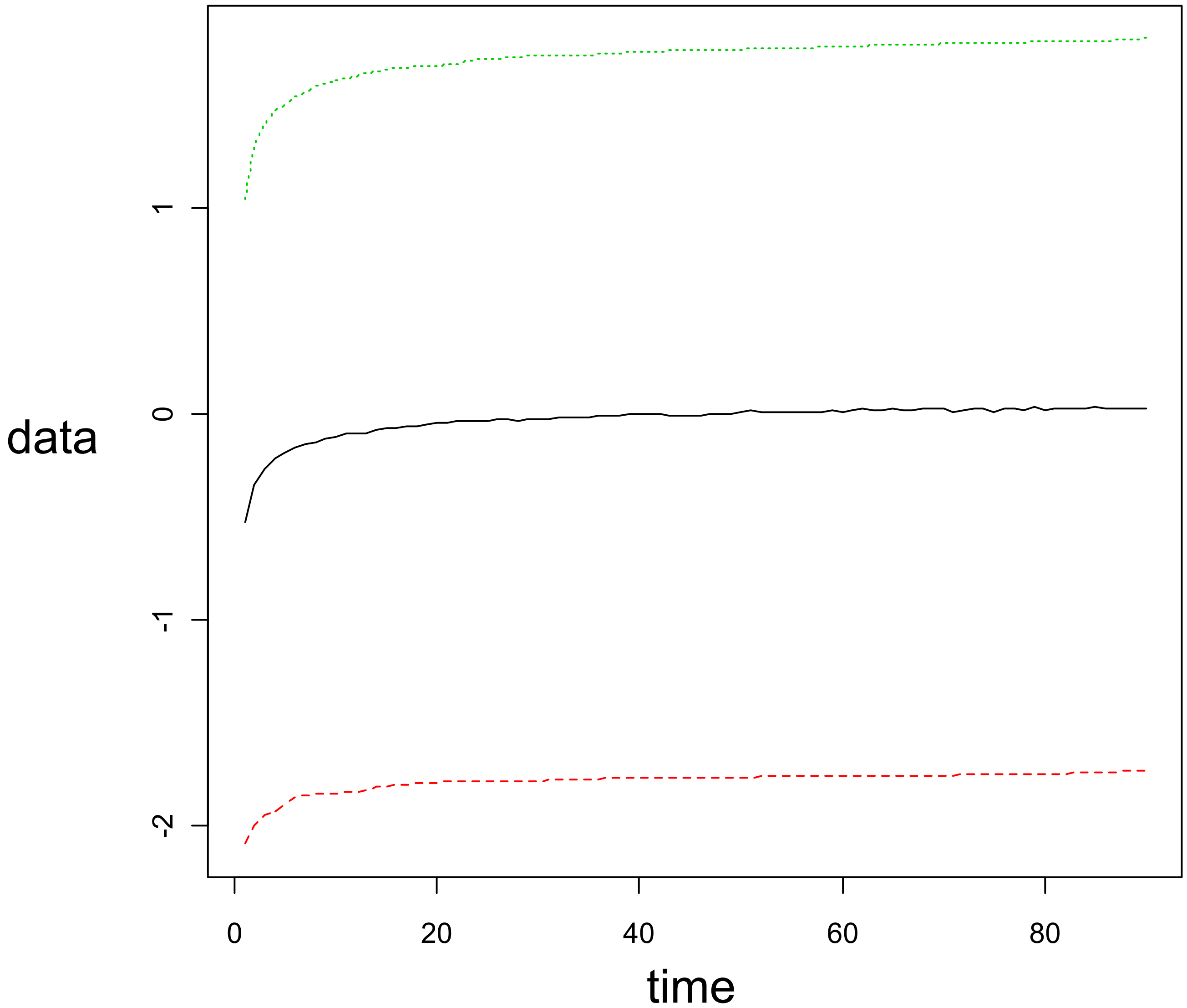


8.6 Hurst-Kolmogorov stochastic process



8.7 Hurst-Kolmogorov stochastic process

0,90 predictive percentile for $y_i|\boldsymbol{x}_n$



9. References

Beran J (1994) Statistics for Long-Memory Processes, Volume 61 of Monographs on Statistics and Applied Probability. Chapman and Hall, New York

Gelman A, Carlin J, Stern H, Rubin D (2004) Bayesian Data Analysis. Chapman & Hall/CRC

Robert C (2007) The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation. Springer ,New York