Quantifying uncertainty and correlation in complex systems

CO907 - MSc in Complexity Science

Stefan Grosskinsky

Warwick, 2013

Plan

- Basic probability
  random variables, PMF, PDF and CMF, independence, expectation, median, variance, standard deviation, characteristic functions, LLN, CLT;
  distributions: uniform, Bernoulli, Binomial, Poisson, Geometric, Exponential, Gaussian, power laws (Pareto, stable laws)
- Visualization of data
  Q-Q plot, probability plot, tail plot, histogram, kernel density estimate, Box plot
- Joint/multivariate RVs
  joint distributions, sum rule, product rule, Bayes rule, basic Bayesian inference, covariances, correlation, multi-variate Gaussians, scatter plot
- Basic statistics
  estimators, sample mean, variance, median, quantiles, bias, consistency, MLE, mean squared error, statistical significance
- Extreme value statistics
  order statistics for iid random variables, Frechet, Gumbel, Weibull distributions
- Basic time series analysis
  Gaussian regression, LSE, model selection, over-fitting and cross-validation, Markovian and auto-regressive models
- Stationary time series
  detrending and regression, autocorrelation function, spectral analysis, Fourier series and Fourier transform

We will only cover very basic Bayesian inference, and NO hypothesis testing, please refer to the module CO902 Probabilistic and statistical inference. I will produce a typed script as the course progresses which will be available for revision of the course material.

Contents

References

[B] C.M. Bishop, Pattern Recognition and Machine Learning, Springer 2006

[H] J.D. Hamilton, Time Series Analysis, Princeton University Press 1994

[BJR] G.E.P. Box, G.M. Jenkins and G.C. Reisel, Time Series Analysis: Forecasting and Control, 4th edition, Wiley 2008

[BP] J-P. Bouchaud, M. Potters, Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management, CUP 2000

[V] J. Voit, The Statistical Mechanics of Financial Markets, 3rd edition, Springer 2005

1 Basic probability

The basic ingredients for a probabilistic model are

$$\begin{array}{rl} \textbf{sample/state space} & S = \text{set of all possible values} \\ \textbf{outcome/state} & s \in S \quad \text{element of } S \\ \textbf{events} & A \subseteq S \quad \text{(certain) subsets of } S\ . \end{array} \tag{1.1}$$

S can be an abstract set (e.g. set of birds sitting on a rhinoceros), or a subset of $\mathbb{R}$ (or a higher dimensional set) which will be the most common case. While for discrete S, any subset $A \subseteq S$ is an event, this would lead to inconcistencies for continuous state spaces, and one has to restrict events to so-called σ-algebras. These still contain all interesting events for us anyway, so we will not go into details on this (which can be found in any standard textbook on probability or measure theory).

A **probability distribution** P on S is a function that assigns a number $P(A)$ to every event $A \subseteq S$ such that

- $P(A) \geq 0$ for all $A \subseteq S$ (positivity)
- $P(S) = 1$ (normalization)
- $P(\cup_i A_i) = \sum_i P(A_i)$ for all $A_i \subseteq S$ which are mutually disjoint. (additivity)

A **random variable** X on S with distribution P takes (random) values on S which are distributed according to P, i.e.

$$\mathbb{P}(X \in A) = P(A) \quad \text{for all events } A \subseteq S\ . \tag{1.2}$$

Two random variables X, Y are **independent**, if for all events $A, B \subseteq S$

$$\mathbb{P}(X \in A \text{ and } Y \in B) = \mathbb{P}(X \in A)\, \mathbb{P}(Y \in B)\ . \tag{1.3}$$

Independence, and the lack of independence corresponding to correlations among random variables, is one of the most important concepts and will be discussed in great detail. Note that in this course we understand the symbol $\mathbb{P}$ just intuitively as 'the probability that'.

1.1 Review of most important distributions and properties

For **discrete state space** $S = \{s_1, s_2, \ldots\}$ finite or countably infinite, distributions of a random variable X are characterized by a

$$\textbf{probability mass function (PMF)} \qquad p_s = \mathbb{P}(X = s)\ , \quad s \in S\ . \tag{1.4}$$

The simplest example is the

- **uniform distribution**, which can be defined on any finite $S = \{s_1, \ldots, s_n\}$ of size n, and is characterized by

$$p_s = 1/|S| = 1/n \quad \text{for all } s \in S\ , \quad \text{with shorthand} \quad X \sim U(S)\ . \tag{1.5}$$

If S is actually a subset of $\mathbb{R}$, the standard properties of a distribution are

$$\begin{aligned} \textbf{expectation} \quad & \mathbb{E}(X) = \sum_{s\in S} s\, p_s \\ \textbf{variance} \quad & \mathrm{Var}(X) = \mathbb{E}\big((X-\mathbb{E}(X))^2\big) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 \\ \textbf{standard deviation} \quad & \sigma(X) = \sqrt{\mathrm{Var}(X)}\ . \end{aligned} \tag{1.6}$$

Further characteristics of distrubion functions are

$$\begin{aligned} \textbf{cumulative distribution function (CDF)} \quad & F_X(s) = \mathbb{P}(X \le s) \\ \textbf{tail distribution function (TDF)} \quad & \bar{F}_X(s) = \mathbb{P}(X > s)\ . \end{aligned} \tag{1.7}$$

Deviding the CDF into regular intervals, we have the

$$\begin{aligned} \textbf{median} \quad & m \in \mathbb{R} \quad \text{such that} \quad \mathbb{P}(X \le m) \ge 1/2\ , \quad \mathbb{P}(X \ge m) \ge 1/2\ ; \\ \textbf{quantiles} \quad & Q_k^q \in \mathbb{R} \quad \text{such that} \quad \mathbb{P}(X \le Q_k^q) \ge k/q\ , \quad \mathbb{P}(X \ge Q_k^q) \ge 1 - k/q\ . \end{aligned} \tag{1.8}$$

Note that the median is simply $m = Q_1^2$, and in general there are $q-1$ q-quantiles. Other common special cases are quartiles $q = 4$ and percentiles $q = 100$. If F_X is not continuous (as is the case for discrete state space S), then quantiles are usually not well defined numbers but can lie in an interval, in which case one usually takes the midpoint of that interval.
In the following we discuss further common distributions and their interpretation.

- **Bernoulli** distribution with $S = \{0, 1\}$ and parameter $\theta \in [0, 1]$, writing $X \sim \mathrm{Be}(\theta)$, with PMF

  $$p_1 = 1 - p_0 = \theta\ . \tag{1.9}$$

  Models success (1) or failure (0) in single experiment with probability θ (e.g. coin throw).
  $\mathbb{E}(X) = \theta$, $\mathrm{Var}(X) = \theta(1-\theta)$.

- **Binomial** distribution with $S = \{0, \dots n\}$, parameters $n \in \mathbb{N}$ and $\theta \in [0, 1]$, writing $Y \sim \mathrm{Bin}(n, \theta)$ with PMF

  $$p_k = \mathbb{P}(Y = k) = \binom{n}{k} \theta^k\, (1-\theta)^{n-k}\ . \tag{1.10}$$

  Models the number of successes in n independent experiments with success probability θ. So if $X_i \sim \mathrm{Be}(\theta)$ are iid we have $Y = \sum_{i=1}^n X_i \sim \mathrm{Bin}(n, \theta)$.
  $\mathbb{E}(y) = n\theta$, $\mathrm{Var}(y) = n\theta(1-\theta)$ (both are additive for iid random variables).

- **geometric** distribution with $S = \{1, 2 \dots\}$ and parameter $\theta \in [0, 1]$, writing $X \sim \mathrm{Geo}(\theta)$ with PMF

  $$p_k = \mathbb{P}(X = k) = (1-\theta)^{k-1}\theta\ . \tag{1.11}$$

  Models the number of independent trials necessary until success.
  $\mathbb{E}(X) = 1/\theta$, $\mathrm{Var}(X) = (1-\theta)/\theta^2$.

The above distributions describe most interesting observables of successive, independent experiments/observations. The statistics of many large systems are described by particular scaling limits of those distributions.

- **Poisson** distribution with $S = \{0, 1, \ldots\}$, parameter $\lambda \geq 0$, writing $X \sim \mathrm{Poi}(\lambda)$ with PMF

  $$p_k = \mathbb{P}(X = k) = \frac{\lambda^k}{k!}\, e^{-\lambda}\ . \tag{1.12}$$

  Models the number of successes under vanishing success probability $\theta_n = \lambda/n$ in the limit of infinitely many trials, i.e.

  $$Y_n \sim \mathrm{Bin}(n, \lambda/n)\ \to\ Y \sim \mathrm{Poi}(\lambda) \quad \text{in distribution as } n \to \infty\ , \tag{1.13}$$

  since $\quad \mathbb{P}(X_n = k) = \frac{n\cdots(n-k+1)}{k!}\big(\frac{\lambda}{n}\big)^k\big(1 - \frac{\lambda}{n}\big)^{n-k} \to \frac{\lambda^k}{k!}\, e^{-\lambda}\ .$

For example, the degree distribution in an Erdös-Rényi random graph $G(n, p)$ is $\mathrm{Bin}(n-1, p)$ (edges are placed to $n-1$ neighbours independently with probability p. In the scaling limit $p = \lambda/n$ and $n \to \infty$, this converges to a $\mathrm{Poi}(\lambda)$ Poisson distribution. In general, the Poisson distribution models the statistics of rare events over a long period of time, such as the number of meteorites that have hit the earth in the last million years or so.
For further scaling limits we need to consider **continuous state spaces**, the most common choices being $S = [0, \infty)$ or $S = \mathbb{R}$. Distributions of random variables $X \in S$ are now described by a

probability density function (PDF) $\qquad f_X : S \to \mathbb{R} \quad$ such that $\quad \mathbb{P}(X \in A) = \int_A f_X(s)\, ds \qquad (1.14)$

Note that $f_X(s)$ is **not** equal to $\mathbb{P}(X = s)$, which is in general equal to 0. One can only associate non-vanishing probabilities to events A which contain at least one small intervall of non-zero length. For continuous S we have

$$\begin{aligned} \textbf{expectation} \qquad & \mathbb{E}(X) = \int_S s\, f_X(s)\, ds \\ \textbf{CDF} \qquad & F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f_X(s)\, ds\ , \end{aligned} \tag{1.15}$$

and the definition of all other quantities remains the same. The simplest example is again the

- **uniform** distribution with $S = [a, b)$, parameters $a < b \in \mathbb{R}$, writing $X \sim U([a, b))$ with PDF

  $$f_X(s) = \frac{1}{b-a}\, \mathbb{1}_{[a,b)} = \begin{cases} 1/(b-a) & ,\ s \in [a, b) \\ 0 & ,\ s \notin [a, b) \end{cases}\ . \tag{1.16}$$

Coming back to the scaling limit of rare successes, we may be interested in the time it takes to see such a rare success, which is given by the scaling limit of the geometric distribution.

- **exponential** distribution with $S = [0, \infty)$, parameter $\lambda > 0$, writing $X \sim \mathrm{Exp}(\lambda)$ with PMF

  $$f_X(s) = \lambda\, e^{-\lambda\, s}\ . \tag{1.17}$$

  It describes the time to the next success $X_n \sim \mathrm{Geo}(\lambda/n)$ in the limit $n \to \infty$ after rescaling, i.e. we have

  $$X_n/n \to X \sim \mathrm{Exp}(\lambda) \quad \text{since} \quad \mathbb{P}(X_n/n \geq s) = (1-\lambda/n)^{ns} \to e^{-\lambda s}\ . \tag{1.18}$$

  The tail has the simple form $\bar{F}_X(x) = e^{-\lambda s}$.
  $\mathbb{E}(X) = 1/\lambda$, $\mathrm{Var}(X) = 1/\lambda^2$ and λ can be interpreted as the success rate or intensity. Often the exponential distribution is also parametrized by the mean $\mu = 1/\lambda$ (e.g. in MATLAB). Scaling property: if $X \sim \mathrm{Exp}(\lambda)$ then $aX \sim \mathrm{Exp}(\lambda/a)$ for all $a > 0$.

The exponential distribution models the statistics of rare events which are purely driven by fluctuations, such as the lifetime distribution of light bulbs. To a good approximation they do not age (other than humans), but fail with small probability each time they are turned on. The statistics of times between rare events (such as meteorites, catastrophes in nuclear power plants,. . .) should therefore be exponentially distributed. If we are interested in the statistics of number of events S_t up to time t if events happen at rate λ this will be Poisson $S_t \sim \mathrm{Poi}(\lambda t)$, which follows directly from the scaling limit for the Poisson distribution.

1.2 Gaussian distribution and CLT

Let X be a real-valued random variable with PDF f_X. The **characteristic function** (CF) $\phi_X(t)$ is defined as the Fourier transform of the PDF, i.e.

$$\phi_X(t) = \mathbb{E}\big(e^{itX}\big) = \int_{-\infty}^{\infty} e^{itx} f_X(x)\, dx \quad \text{for all } t \in \mathbb{R}\ . \tag{1.19}$$

As the name suggests, ϕ_X uniquely determines (characterizes) the distribution of X and the usual inversion formula for Fourier transforms holds,

$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \phi_X(t)\, dt \quad \text{for all } x \in \mathbb{R}\ . \tag{1.20}$$

By normalization we have $\phi_X(0) = 1$, and moments can be recovered via

$$\tfrac{\partial^k}{\partial t^k} \phi_X(t) = (i)^k \mathbb{E}(X^k e^{itX}) \quad \Rightarrow \quad \mathbb{E}(X^k) = (i)^{-k} \tfrac{\partial^k}{\partial t^k} \phi_X(t)\, \big|_{t=0}\ . \tag{1.21}$$

Also, if we add independent random variables X and Y, their characteristic functions multiply,

$$\phi_{X+Y}(t) = \mathbb{E}\big(e^{it(X+Y)}\big) = \phi_X(t)\, \phi_Y(t)\ . \tag{1.22}$$

Furthermore, for a sequence $X_1, X_2, \ldots$ of real-valued random variables we have

$$X_n \to X \quad \text{in distribution, i.e.} \quad f_{X_n}(x) \to f_X(x)\ \forall x \in \mathbb{R} \quad \Leftrightarrow \quad \phi_{X_n}(t) \to \phi_X(t)\ \forall t \in \mathbb{R} \tag{1.23}$$

A real-valued random variable $X \sim N(\mu, \sigma^2)$ has **normal** or **Gaussian** distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \geq 0$ if its PDF is of the form

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2} \Big)\ . \tag{1.24}$$

Properties.

- The characteristic function of $X \sim N(\mu, \sigma^2)$ is given by

$$\phi_X(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2} + itx\Big)\, dx = \exp\Big(i\mu t - \frac{1}{2}\sigma^2 t^2\Big) \ . \quad (1.25)$$

  To see this (try it!), you have to complete the squares in the exponent to get

$$-\frac{1}{2\sigma^2}\big(x - (it\sigma^2 + \mu)\big)^2 - \frac{1}{2}t^2\sigma^2 + it\mu \ , \quad (1.26)$$

  and then use that the integral over x after re-centering is still normalized.

- This implies that linear combinations of independent Gaussians $X_1,\ X_2$ are Gaussian, i.e.

$$X_i \sim N(\mu_i, \sigma_i^2),\ a, b \in \mathbb{R} \quad \Rightarrow \quad aX_1 + bX_2 \sim N\big(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2\big) \quad (1.27)$$

For **discrete random variables** X taking values in $\mathbb{Z}$ with PMF $p_k = \mathbb{P}(X = k)$ we have

$$\phi_X(t) = \mathbb{E}\big(e^{itX}\big) = \sum_{k\in\mathbb{Z}} e^{itk} p_k \quad \text{for all } t \in \mathbb{R} \ . \quad (1.28)$$

So p_k is the inverse Fourier series of the function $\phi_X(t)$, the simplest example is

$$X \sim Be(p) \quad \Rightarrow \quad \phi_X(t) = pe^{it} + 1 - p \ . \quad (1.29)$$

Note that this is a 2π-periodic function in t, since only two coefficients are non-zero. We will come back to that later for time-series analysis.

Let $X_1, X_2, \ldots$ be a sequence of iidrv's with mean μ and variance σ^2 and set $S_n = X_1 + \ldots + X_n$. The following two important limit theorems are a direct consequence of the above.

Theorem 1.1 Weak law of large numbers (LLN)
Under the above assumptions we have

$$S_n/n \to \mu \quad \textit{in distribution as } n \to \infty \ . \quad (1.30)$$

There exists also a strong form of the LLN with almost sure convergence which is harder to prove.

Theorem 1.2 Central limit theorem (CLT)
Under the above assumptions we have

$$\frac{S_n - \mu n}{\sigma\sqrt{n}} \to N(0,1) \quad \textit{in distribution as } n \to \infty \ . \quad (1.31)$$

The LLN and CLT imply that for $n \to \infty$, $\quad S_n \simeq \mu n + \sigma\sqrt{n}\,\xi \quad$ with $\quad \xi \sim N(0,1)$.

Proof. With $\phi(t) = \mathbb{E}\big(e^{itX_i}\big)$ we have from (1.22)

$$\phi_n(t) := \mathbb{E}\big(e^{itS_n/n}\big) = \big(\phi(t/n)\big)^n .$$

(1.21) implies the following Taylor expansion of ϕ around 0:

$$\phi(t/n) = 1 + i\mu\frac{t}{n} - \frac{\sigma^2}{2}\frac{t^2}{n^2} + o(t^2/n^2) ,$$

of which we only have to use the first order to see that

$$\phi_n(t) = \Big(1 + i\mu\frac{t}{n} + o(t/n)\Big)^n \to e^{it\mu} \quad \text{as } n \to \infty .$$

By (1.23) and uniqueness of characteristic functions this implies the LLN.

To show the CLT, set $\quad Y_i = \dfrac{X_i - \mu}{\sigma} \quad$ and write $\quad \tilde{S}_n = \displaystyle\sum_{i=1}^n Y_i = \frac{S_n - \mu n}{\sigma}$.

Then, since $\mathbb{E}(Y_i) = 0$, the corresponding Taylor expansion (now to second order) leads to

$$\phi_n(t) := \mathbb{E}\big(e^{it\tilde{S}_n/\sqrt{n}}\big) = \Big(1 - \frac{t^2}{2n} + o(t^2/n)\Big)^n \to e^{-t^2/2} \quad \text{as } n \to \infty ,$$

which implies the CLT. □

2 Less basic probability

2.1 Power laws and generalized CLT

A positive random variable with CDF F is said to have a **power-law tail** with exponent $\alpha > 0$, if

$$\bar{F}(x)x^\alpha \to C \in (0,\infty) \quad \text{i.e.} \quad \bar{F}(x) \propto x^{-\alpha} \quad \text{as } x \to \infty . \tag{2.1}$$

The simplest example is the

- **Pareto** distribution with $S = [x_m, \infty)$, $x_m > 0$ and parameter $\alpha > 0$, writing $X \sim \text{Pareto}(x_m, \alpha)$ with PDF

  $$f_X(x) = \alpha x_m^\alpha / x^{\alpha+1} \quad \text{for } x \geq x_m . \tag{2.2}$$

  The tail has the simple form $\bar{F}_X(x) = (x_m/x)^\alpha$.
  $\mathbb{E}(X) = \alpha x_m/(\alpha - 1)$ if $\alpha > 1$, otherwise ∞, $\text{Var}(X) = \frac{x_m^2\alpha}{(\alpha-1)^2(\alpha-2)}$ if $\alpha > 2$, otherwise ∞.

Power-law tails occur frequently in statistics of every-day life quantities, social sciences and finance. Power laws are also called scale-free distributions, due to the following:

$$X \sim \text{Pareto}(x_m, \alpha) \quad \text{then} \quad aX \sim \text{Pareto}(ax_m, \alpha) \quad \text{for } a > 0 . \tag{2.3}$$

So the power law exponent does not change under scaling, only the range does. Except for the lower cut-off at x_m, Pareto distributed phenomena look the same on all scales, and the system does not have a characteristic length scale (such as $1/\lambda$ for exponential or μ for Gaussian distributions). This is relevant in critical phenomena in statistical mechanics, where systems exhibit scale free distributions at points of phase transitions. Power law degree distributions in complex networks can emerge from preferential attachment-type dynamics, which is often used as an explanation for the abundance of power-law distributed observables in social or other types of networks.

For heavy-tailed distributions with diverging mean and/or variance the LLN and CLT have to be modified. The Gaussian has to be replaced by a generalized class of stable limit laws, the **α-stable Lévy distributions**. They are most easily characterized by their characteristic function, which for symmetric distributions is simply given by

$$\chi_\alpha(t) = e^{-|ct|^\alpha} \quad \text{where the scale} \quad c > 0 \quad \text{determines the width} \,. \tag{2.4}$$

Note that for $\alpha = 2$ this corresponds to the centred Gaussian, and for $\alpha = 1$ it is known as the **Cauchy-Lorentz distribution** with PDF

$$f_1(x) = \frac{1}{\pi} \frac{c}{c^2 + x^2} \,. \tag{2.5}$$

Asymptotically, symmetric Lévy distributions behave as

$$f_\alpha(x) \propto \frac{\alpha c}{|x|^{1+\alpha}} \quad \text{as} \quad |x| \to \infty \,, \tag{2.6}$$

i.e. they exhibit a power-law tail with exponent α. For the general asymmetric form of these distributions and more details on generalized LLN and CLT see e.g. [V], Chapter 5, or [BP], Chapter 1.

Theorem 2.1 Generalized LLN and CLT
Let $X_1, X_2 \ldots$ be iid random variables with symmetric power-law tail $\mathbb{P}(|X_i| \geq x) \propto x^{-\alpha}$ with $\alpha \in (0, 2)$. For $S_n = \sum_{i=1}^n X_i$ we have for $\alpha \in (0, 1)$

$$\frac{1}{n} S_n \quad \textit{does not converge, but} \quad \frac{1}{n^{1/\alpha}} S_n \quad \textit{does} \qquad \textit{(modified LLN)} \,. \tag{2.7}$$

If $\alpha \in (1, 2)$, $\mathbb{E}(|X_i|) < \infty$ and we have

$$\begin{aligned} \frac{1}{n} S_n \;&\to\; \mu = \mathbb{E}(X_i) \qquad \textit{(usual LLN)} \,, \\ \frac{1}{n^{1/\alpha}}(S_n - n\mu) \;&\to\; \alpha\textit{-stable Lévy} \qquad \textit{(generalized CLT)} \,. \end{aligned} \tag{2.8}$$

The proof follows the same idea as the usual CLT with $|t|^\alpha$ being the leading order term in the expansion of the characteristic function.

2.2 Extreme value statistics

Consider a sequence of iid random variables $X_1, X_2, \ldots$ with CDF F, and let

$$M_n = \max\{X_1, \ldots, X_n\} \tag{2.9}$$

be the maximum of the first n variables. Analogous to the CLT, the distribution of the maximum will converge to a universal limit distribution.

Theorem 2.2 Extreme value theorem (Fisher-Tippet Gnedenko)
If there exist normalizing sequences such that $\mathbb{P}\big(\frac{M_n-b_n}{a_n}\leq x\big)$ converges to a non-degenerate CDF $G(x)$ as $n\to\infty$, then G is of the form

$$G(x)=\exp\bigg(-\Big(1+k\Big(\frac{x-\mu}{\sigma}\Big)\Big)^{-1/k}\bigg) \tag{2.10}$$

*with parameters for location $\mu\in\mathbb{R}$, scale $\sigma>0$ and shape $k\in\mathbb{R}$, and is called the **generalized extreme value distribution**.*

The normalizing sequences and the parameter values of G are related to the tail of the distribution F in a rather complicated way (see e.g. [BP], Chapter 1 for details). Depending on the shape parameter k, one typically distinguishes the following 3 classes of extreme value distributions:

- **Gumbel** (Type I): $k=0$ and $G(x)=\exp\Big(-e^{-(x-\mu)/\sigma}\Big)$
  limit if $\bar{F}$ has exponential tail (including actual xponential or Gaussian rv's)
- **Fréchet** (Type II): $k=1/\alpha>0$ and $G(x)=\begin{cases} 0 & , x\leq\mu \\ \exp\Big(-\big(\frac{x-\mu}{\sigma}\big)^{\alpha}\Big) & , x>\mu \end{cases}$
  limit if $\bar{F}$ has power-law tail (including e.g. Pareto rv's)
- **Weibull** (Type III): $k=-1/\alpha<0$ and $G(x)=\begin{cases} \exp\Big(-\big(-\frac{x-\mu}{\sigma}\big)^{\alpha}\Big) & , x<\mu \\ 1 & , x\geq\mu \end{cases}$
  limit if $\bar{F}$ has light tail (including bounded support such as uniform rv's)

Note that the rescaled shape parameter α for Types II and III is usually taken to be positive, and the different types are compatible with (2.10) taking the limit $k\to 0$, and with $\sigma=k\mu$ for $k\neq 0$.

The typical scaling (location) s_n of $\mathbb{E}(M_n)$ as a function of n can be determined relatively easily. Note that for iid random variables

$$\mathbb{P}(M_n\leq x)=\mathbb{P}(X_1\leq x,\ldots,X_n\leq x)=\mathbb{P}(X_1\leq x)^n=(F(x))^n\ . \tag{2.11}$$

Now s_n is determined by requiring that $(F(s_n))^n$ has a non-degenerate limit in $(0,1)$ as $n\to\infty$, so that

$$(F(s_n))^n=\big(1-\bar{F}(s_n)\big)^n\to e^{-c}\ ,\ c>0\quad\text{which implies}\quad \bar{F}(s_n)\simeq c/n\ . \tag{2.12}$$

The proof of Theorem 2.2 uses the same idea, and it turns out that one can further parametrize all possible limit distributions according to (2.10), which is technical and we omit here.

For exponential $\text{Exp}(\lambda)$ random variables with tail $\bar{F}(s_n)=e^{-\lambda s_n}$ this leads to

$$s_n\simeq(\log n-\log c)/\lambda\quad\text{which implies}\quad M_n=\log n/\lambda+O(1) \tag{2.13}$$

where $O(1)$ is a random variable that does not scale with n. This implies that we may choose $b_n=\log n/\lambda$ and $a_n=1$ in Theorem 2.2 as normalizing sequences with convergence to Gumbel. For Pareto random variables $\text{Pareto}(x_m,\alpha)$ with power-law tail $\bar{F}(s_n)=\big(\frac{x_m}{s_n}\big)^{\alpha}$ we get

$$s_n\simeq x_m(n/c)^{1/\alpha}\ ,\quad\text{so that}\quad M_n=x_m n^{1/\alpha}\,O(1) \tag{2.14}$$

with multiplicative randomness, implying $b_n=0$ and $a_n=x_m n^{1/\alpha}$ as a valid normalization with convergence to Fréchet.

Note that for $\alpha \in (0, 1)$ where the power-law has infinite mean, this implies $\mathbb{E}(M_n) \ll 1$ scales faster than the number of variables n, and is of the same order as the total sum S_n. So the sum is therefore dominated by the largest contributions, whereas for $\alpha > 1$ we have $\mathbb{E}(M_n) \ll n \propto S_n$. For any $\alpha > 0$ iid variables with power-law tails exhibit a hierarchical **order statistics**, i.e. for the ordered sequence of random variables $X_{(1)} \leq \ldots \leq X_{(n)}$ we have

$$\mathbb{E}(X_{(k)}) \propto \Big(\frac{n}{n-k+1}\Big)^{1/\alpha} \quad \text{for all } k = 1, \ldots, n\ . \tag{2.15}$$

As a last example, for uniform $U([0, 1))$ random variables we expect $M_n \to 1$ as $n \to \infty$, and with $\bar{F}(x) = 1 - x$ we get $s_n \simeq 1 - c/n$ so that we can choose $b_n = 1$ and $a_n = 1/n$ with convergence to Weibull.

3 Joint distributions

3.1 Basic definitions and results

In many applications outcomes take values in higher dimensional spaces such as $S = \mathbb{R}^d$ or $\mathbb{N}^d$, and if $d > 1$ such random variables are called **multivariate**. They are described by **joint distributions**, and in the simplest case for $d = 2$ a multivariate discrete random variable (X, Y) this is given by mass function

$$p_{(X,Y)}(x, y) = \mathbb{P}(X = x, Y = y) \quad \text{for all } x \in S_x \quad \text{and} \quad y \in S_Y\ , \tag{3.1}$$

where S_X and S_Y are the sample space for X and Y, respectively. As before, $p_{(X,Y)}$ is normalized and non-negative.

Each component of a multivariate random variable is a random variable itself, and its distribution, the **marginal probability**, is given by the **sum rule**

$$p_X(x) = \mathbb{P}(X = x) = \sum_{y \in S_Y} \mathbb{P}(X = x, Y = y) = \sum_{y \in S_Y} p(x, y)\ . \tag{3.2}$$

The **conditional probability** distribution for X given that $Y = y$ takes a particular value is defined as

$$p_{X|Y=y}(x) = \mathbb{P}(X = x|Y = y) = \frac{P(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{p_{(X,Y)}(x, y)}{p_Y(y)}\ . \tag{3.3}$$

Note that from the sum rule (3.2) we get that

$$\sum_{x \in S_X} \mathbb{P}(X = x|Y = y) = \frac{1}{\mathbb{P}(Y = y)} \sum_{x \in S_X} \mathbb{P}(X = x, Y = y) = \frac{\mathbb{P}(Y = y)}{\mathbb{P}(Y = y)} = 1\ , \tag{3.4}$$

so for each $y \in S_y$ the conditional probability of X is normalized. The definition of conditional probabilities can be re-written as the **product rule**

$$p_{(X,Y)}(x, y) = \mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x|Y = y)\, \mathbb{P}(Y = y)\ . \tag{3.5}$$

By definition, X and Y are independent if

$$p_{(X,Y)}(x, y) = p_X(x)\, p_Y(y) \quad \text{for all } x \in S_X \quad \text{and} \quad y \in S_Y\ . \tag{3.6}$$

This implies that the conditional distribution

$$\mathbb{P}(X = x|Y = y) = \frac{p_X(x)\, p_Y(y)}{p_Y(y)} = p_X(x) \tag{3.7}$$

is actually independent of the value y, coinciding with the intuitive meaning of independence.

Simple examples.

- Draw from a deck of cards with suit $X \in S_X = (H, S, C, D)$ and rank $Y \in S_Y = (A, 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K)$. There are in total $4 \times 13 = 52$ outcomes for (X, Y) (cards), and the joint distribution can be represented in a table with entries $p_{(X,Y)}(x, y)$, which should all be equal to $1/4 \times 1/13 = 1/52$ for the first card from a well-mixed deck (suit and rank are independent).
  The second card will be uniform among the 51 remaining, and so on. Are suit and rank still independent?
- In medical applications X could signify a particular treatment, and Y characterize the outcome of the treatment; or X indicates if a person smokes, and Y if the person gets lung cancer.

Analogous formulations hold continuous random variables with joint PDF $f_{(X,Y)}$:

$$\textbf{sum rule} \qquad f_X(x) = \int_{S_Y} f_{(X,Y)}(x, y)\, dy \tag{3.8}$$

$$\textbf{product rule} \qquad f_{(X,Y)}(x, y) = f_{X|Y=y}(x)\, f_Y(y)\ , \tag{3.9}$$

where f_X is the PDF of the first marginal of (X, Y) and $f_{X|Y=y}$ is the PDF of the conditional law of X given $Y = y$. Note that the conditional distribution of X given Y has to be defined through a limit

$$F_{X|Y=y}(x) = \lim_{\epsilon \searrow 0} \mathbb{P}\big(X \le x \big| Y \in [y, y + \epsilon)\big)\ , \tag{3.10}$$

and if this exists and has a well-defined derivative, the conditional density is given by

$$f_{X|Y=y}(x) = \frac{d}{dx} F_{X|Y=y}(x)\ . \tag{3.11}$$

The standard and most common example for which all this can be worked out in detail is the **multivariate Gaussian**. $\mathbf{X} = (X_1, \ldots, X_d) \in \mathbb{R}^d$ is a d-dimensional multivariate Gaussian, $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$ if it has PDF

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} \exp\Big(-\frac{1}{2} \big\langle (\mathbf{x} - \boldsymbol{\mu})\, |\Sigma^{-1}|\, (\mathbf{x} - \boldsymbol{\mu}) \big\rangle \Big) \quad \text{with} \quad \mathbf{x} = (x_1, \ldots, x_d), \tag{3.12}$$

with **mean** $\mu = (\mu_1, \ldots, \mu_d) \in \mathbb{R}^d$ and **covariance matrix** $\Sigma = (\sigma_{ij} : i, j = 1, \ldots, d) \in \mathbb{R}^{d \times d}$ with entries

$$\sigma_{ij} = \mathrm{Cov}(X_i, X_j) = \mathbb{E}\big((X_i - \mu_i)(X_j - \mu_j)\big)\ . \tag{3.13}$$

We use the notation $\langle .|.\rangle$ for a scalar product of a row vector $\langle .|$ with a column vector $|.\rangle$. The characteristic function of $\mathbf{X}$ is given by

$$\phi_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}\big(e^{i\langle \mathbf{t}|\mathbf{X}\rangle}\big) = \exp\Big(i\langle \mathbf{t}|\boldsymbol{\mu}\rangle - \frac{1}{2}\langle \mathbf{t}|\Sigma|\mathbf{t}\rangle \Big)\ , \quad \mathbf{t} \in \mathbb{R}^d\ .$$

Note that the PDF (3.12) factorizes if and only if the covariance matrix is diagonal, i.e. $\sigma_{ij} = \sigma_i^2 \delta_{ij}$ where $\sigma_i^2 = \mathrm{Var}(X_i)$. So Gaussian random variables are independent if and only if they are **uncorrelated**, i.e.

$$\sigma_{ij} = \mathrm{Cov}(X_i, X_j) = 0 \quad \text{for all } i \neq j \ . \tag{3.14}$$

The covariance matrix as defined in (3.13) can be used for general distributions. It is symmetric by definition, it is positive semi-definite, since

$$\langle \mathbf{v}|\Sigma|\mathbf{v}\rangle = \mathrm{Var}(\langle \mathbf{v}|\mathbf{X}\rangle) \geq 0 \quad \text{for all } \mathbf{v} \in \mathbb{R}^d \ , \tag{3.15}$$

and therefore has d real eigenvalues $\lambda_1, \ldots, \lambda_d \geq 0$ which are non-negative. If all eigenvalues are positive (i.e. $\mathrm{Var}(X_i) > 0$ for all $i = 1, \ldots, d$) its inverse Σ^{-1} that appears in the multivariate Gaussian PDF is well defined. The inverse is also called **concentration** or **precision matrix**, and its entries can be interpreted in terms of conditional correlations, i.e. fixing the values of all other coordinates with conditional expectations

$$\mathbb{E}\big((X_i - \mu_i)(X_j - \mu_j)\big|X_k, k \neq i, j\big) \ , \tag{3.16}$$

as compared to the full (or marginal) expectations in the covariance matrix. Since Σ is symmetric, the eigenvectors $|\mathbf{v}_1\rangle, \ldots, |\mathbf{v}_d\rangle$ can be chosen orthogonal, i.e. $\langle \mathbf{v}_i|\mathbf{v}_j\rangle = \delta_{ij}$, and form a basis of $\mathbb{R}^d$. The inverse is then given by

$$\Sigma^{-1} = \sum_{i=1}^{d} \lambda_i^{-1} |\mathbf{v}_i\rangle\langle \mathbf{v}_i| \ , \tag{3.17}$$

where $|\mathbf{v}_i\rangle\langle \mathbf{v}_i|$ is the projector matrix on the eigenspace of eigenvalue λ_i and Σ itself can be written as $\Sigma = \sum_{i=1}^d \lambda_i |\mathbf{v}_i\rangle\langle \mathbf{v}_i|$. Then we have

$$\Sigma^{-1}\Sigma = \sum_{i,j=1}^{d} \lambda_i^{-1}\lambda_j |\mathbf{v}_i\rangle \underbrace{\langle \mathbf{v}_i|\mathbf{v}_j\rangle}_{=\delta_{ij}} \langle \mathbf{v}_j| = \sum_{i=1}^{d} |\mathbf{v}_i\rangle\langle \mathbf{v}_i| = \mathrm{Id} \ . \tag{3.18}$$

The **correlation coefficient** $\rho_{X,Y}$ for two random variables X, Y with means μ_X, μ_Y and standard deviations σ_X, σ_Y is defined as

$$\rho_{X,Y} = \mathrm{corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X\, \sigma_Y} = \frac{\mathbb{E}\big((X - \mu_X)(Y - \mu_Y)\big)}{\sigma_X\, \sigma_Y} \ , \tag{3.19}$$

and takes values in $[-1, 1]$. If Y is a linear function of X, $|\rho_{X,Y}| = 1$ and is $+1$ for an increasing function, and -1 for a decreasing function (anticorrelation). Again, for independent variables $\rho_{X,Y} = 0$, but the converse is not true in general (but holds for Gaussians). Correlations can be well detected by eye in so-called **scatter plots**, where many realizations of $(X, Y) \in \mathbb{R}$ are shown in a 2-dimensional plot.

3.2 Bayes' rule and simple hypothesis testing

In statistics the (dangerously) compact notation $P(X) = \mathbb{P}(X = x)$ is often used to represent the distribution of a random variable X. We will do the same in the following to get some exercise in using and understanding it. The sum and product rule can be written in the simple form

$$P(X) = \sum_Y P(X, Y) \quad \text{and} \quad P(X, Y) = P(X|Y)\, P(Y) \ . \tag{3.20}$$

As a direct consequence of the product rule and its symmetric left-hand side we get

$$P(X|Y)\,P(Y) = P(Y|X)\,P(X)\,, \tag{3.21}$$

which can be rewritten as the famous

$$\textbf{Bayes' rule} \qquad P(X|Y) = \frac{P(Y|X)\,P(X)}{P(Y)}\,. \tag{3.22}$$

The main interpretation is

- $P(X)$ is the **prior distribution**, summarizing the initial knowledge on X
- $P(X|Y)$ is the **posterior distribution**, summarizing the updated knowledge on X knowing about Y, which is usually an observation or outcome of an experiment (data)
- $P(Y|X)$ is the **likelihood** of data Y, given X (which is often parameters of a model or a hypothesis)

The ratio $P(Y|X)/P(Y)$ represents the support that Y provides for X. Usually $P(Y)$ is not easily accessible directly, but can just be found as a normalization of the right-hand side $P(Y) = \sum_X P(Y|X)\,P(X)$. Bayes' rule is therefore often written as

$$P(X|Y) \propto P(Y|X)\,P(X)\,, \tag{3.23}$$

which contains all crucial components listed above.

Example. Suppose the secret service of some country monitors mobile phone conversations and can have devised a clever test that can identify terrorists with 99% accuracy. Suppose 1 in 10000 people are terrorists. If the police kick in your door on the basis of one of one of your intercepted phone calls, what is the probability that you are a terrorist?
Let X be a random variable on $S = \{0, 1\}$ indicating whether I am a terrorist, and Y a variable on S indicating positivity of the terrorist test. Without further knowledge on me our prior distribution is

$$P(X) = \begin{cases} \frac{1}{10000} & , X = 1 \\ \frac{9999}{10000} & , X = 0 \end{cases} \tag{3.24}$$

For the likelihood we know from the test description

$$\begin{aligned} P(Y|X=1) \;&=\; \begin{cases} \frac{99}{100} & , Y = 1 \\ \frac{1}{100} & , Y = 0 \end{cases} \qquad 99\% \;\; \textbf{sensitivity} \\ P(Y|X=0) \;&=\; \begin{cases} \frac{1}{100} & , Y = 1 \\ \frac{99}{100} & , Y = 0 \end{cases} \qquad 99\% \;\; \textbf{specificity}\,. \end{aligned} \tag{3.25}$$

In general, sensitivity (related to false negatives) and specificity (related to false positives) could have different values. By Bayes' rule we get for the posterior distribution

$$P(X|Y) = \frac{P(Y|X)\,P(X)}{P(Y|X=1)\,P(X=1) + P(Y|X=0)\,P(X=0)} \tag{3.26}$$

and in particular:

$$
\begin{aligned}
P(X=1|Y=1) \;&=\; \frac{P(Y=1|X=1)\,P(X=1)}{P(Y=1|X=1)\,P(X=1)+P(Y=1|X=0)\,P(X=0)} = \\
&= \frac{\frac{99}{100}\frac{1}{10000}}{\frac{99}{100}\frac{1}{10000}+\frac{1}{100}\frac{9999}{10000}} = \frac{99}{10098} \approx 0.01 \qquad (3.27)
\end{aligned}
$$

which is actually of the same order as the probability that the test result is incorrect.
In general, if $a \in [0,1]$ is the accuracy of the test (specificity and sensitivity), and $f \in [0,1]$ is the fraction of terrorists in the population, we get

$$
P(X=1|Y=1) = \frac{a\,f}{a\,f+(1-a)(1-f)} \approx \frac{a\,f}{(1-a)(1-f)} \approx \frac{f}{1-a} \ll 1 \qquad (3.28)
$$

as long as $f \ll 1-a \ll 1$. In that case the probability of false positives

$$
P(X=0|Y=1) \approx 1 - \frac{f}{1-a} \approx 1 \qquad (3.29)
$$

is very close to 1 and results have to be interpreted with care.
On the other hand, the probability of false negatives

$$
P(X=1|Y=0) = \frac{(1-a)\,f}{(1-a)\,f+a(1-f)} \approx \frac{(1-a)\,f}{a(1-f)} \approx \frac{f}{a} \ll 1 \qquad (3.30)
$$

is low and those are usually not problematic.

Q: How accurate would the classifier have to be to achieve $b = 99\%$ accuracy over the entire population (as opposed to only over the terrorist population)?

A: Since false positives are the crucial problem, we have to solve

$$
P(X=1|Y=1) = \frac{a\,f}{a\,f+(1-a)(1-f)} = b \qquad (3.31)
$$

for a, which yields after some steps (still assuming $f \ll 1-a \propto 1-b \ll 1$)

$$
a \approx 1 - f\frac{1-b}{b} = 1 - f\frac{1}{99} \gg 1-f\,. \qquad (3.32)
$$

So not surprisingly, the error rate $1-a$ has to be significantly less than the fraction f of terrorists.

Often a problem: Posterior probabilities depend significantly on the prior distribution, which is in general not determined uniquely and should be chosen as 'informatively' as possible.
Using the prior $P(X=1) = 1/10000$ above, we assumed implicitly that we were kicking the door of a guy who was as likely to be a terrorist as anyone else before the test result came up. But if one can narrow down the list of prior suspects, this could significantly improve results. In general, choosing an informative prior is one of the central problems in Bayesian inference.

Q: Think about genetic paternity tests. Even if they come with 99.99% accuracy, there are about 3 billion possible fathers of the child. How can these tests give reliable answers?

4 Basic statistics

4.1 Most common statistics

Let $\mathbf{X} = (X_1, \ldots, X_N) \in \mathbb{R}^N$ be a **sample** of N real valued datapoints. We assume in this section that the data are purely distributional and are independent samples from a single distribution.
A **statistic** $\hat{\theta}(\mathbf{X})$ is a function of the data, and therefore directly measurable as opposed to the true parameters θ of the underlying distribution. The most common examples are

- **sample mean** $\hat{\mu} = \hat{\mu}(\mathbf{X}) := \frac{1}{N} \sum_{i=1}^N X_i$
- **sample variance** $\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \hat{\mu})^2$

- **order statistics** $X_{(1)} \leq \ldots \leq X_{(N)}$ is the ordered data sample, where $X_{(1)} = \min\{X_1, \ldots, X_N\}$ and $X_{(N)} = \max\{X_1, \ldots, X_N\}$
- **empirical distribution:** the empirical CDF is given by the monotone increasing step function

$$\hat{F}_{\mathbf{X}}(x) = \frac{1}{N} \sum_{i=1}^N \mathbb{1}_{[X_i, \infty)}(x) , \tag{4.1}$$

and the empirical tail is given by $\hat{\bar{F}}_{\mathbf{X}}(x) = 1 - \hat{F}_{\mathbf{X}}(x) = \dfrac{1}{N} \displaystyle\sum_{i=1}^N \mathbb{1}_{(-\infty, X_i)}(x)$.

Formally, the derivative of the step function $\mathbb{1}_{[X_i,\infty)}(x)$ is given by the delta function $\delta_{X_i}(x)$, which leads to an expression for the **empirical density**

$$\hat{f}_{\mathbf{X}}(x) = \frac{1}{N} \sum_{i=1}^N \delta_{X_i}(x) , \tag{4.2}$$

which mathematically is a random point measure on $\mathbb{R}$ (not a function). Since it is purely atomic, for practical purposes it is usually smoothed with a filter, e.g. the

$$\textbf{Gaussian kernel} \quad K(x) = \frac{1}{\sqrt{2\pi s^2}} \, e^{-x^2/(2s^2)} \tag{4.3}$$

with a filter parameter s^2. This has nothing to do with the actual distribution of the data which does not have to be Gaussian. The kernel density estimate of $\hat{f}$ is then given by the convolution

$$\begin{aligned} \hat{f}_K(x) \ &= \ (\hat{f} \star K)(x) := \int_{\mathbb{R}} \hat{f}(y)\, K(x-y)\, dy = \frac{1}{N} \int_{\mathbb{R}} \sum_{i=1}^N \delta_{X_i}(y) K(x-y)\, dy = \\ &= \ \frac{1}{N} \sum_{i=1}^N K(x - X_i) = \frac{1}{N} \sum_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \, e^{-(x-X_i)^2/(2\sigma^2)} \ . \end{aligned} \tag{4.4}$$

A **histogram** consists of a partition of $\mathbb{R}$ given by intervals $I_k, k = 1, \ldots, M$ where $I_k = [v_{k-1}, v_k)$, where one possible convention at the boundaries is to choose $v_0 = -\infty$ and

$v_M = \infty$. The values associated to each interval reflect the proportion of data points in that interval, so that

$$\text{hist}(\mathbf{X}) = \frac{1}{N}\Big(\sum_{i=1}^{N} \mathbb{1}_{I_1}(X_i), \ldots, \sum_{i=1}^{N} \mathbb{1}_{I_M}(X_i)\Big) \, . \tag{4.5}$$

- **Quantiles** remember that the k-th q-quantile Q_k^q of a distribution is any $x \in \mathbb{R}$ such that

  $$\mathbb{P}(X \leq x) \geq k/q \quad \text{and} \quad \mathbb{P}(X \geq x) \geq 1 - k/q \, . \tag{4.6}$$

  As an estimator, one usually picks the datapoint with the closest rank in the order statistics, i.e.

  $$\hat{Q}_k^q = X_{(\lceil kN/q \rceil)} \qquad \text{rounding up if } kN/q \text{ is not an integer ,} \tag{4.7}$$

  and averaging data points

  $$\hat{Q}_k^q = \frac{1}{2}\big(X_{(kN/q)} + X_{(kN/q+1)}\big) \qquad \text{if } kN/q \text{ is an integer .} \tag{4.8}$$

  This leads e.g. to a consistent definition of the median as $\hat{Q}_2^2$ for samples of odd and even length.

4.2 Statistical models and estimators

Generative statistical models are mathematical modesl that generate data which is statistically identical to the sample. For distributional data (iid samples) which we focus on here, these are simply probability distributions, for time series data studied in later sections the models become more involved.

The simplest approach is to use the empirical distribution as a statistical model. This is **non-parametric**, i.e. there are no parameter values to be fitted, and summarizes all the statistical information about the data available from the sample. This is often used to resample datasets to get realistic error bars or confidence intervals on estimates of statistics or hypothesis tests. This method is called **bootstrap**. In terms of informative mathematical modelling the empirical distribution does not provide any further understanding and has in a sense as many parameters as the sample has datapoints. Modelling is usually associated with a reduction of the underlying mechanism to a lower dimensional space, which is achieved by fitting parameters of **parametric models**, i.e. probability distributions. The idea is that those models have usually far less parameters than the sample has data points, but can still give a good representation of their statistical properties.

An **estimator** $\hat{\theta} = \hat{\theta}(\mathbf{X})$ is a statistic which provides an estimate for the parameter (or parameter vector) θ of a model (distribution) F_θ, such as $\theta = \lambda$ for $\text{Exp}(\lambda)$ or $\text{Poi}(\lambda)$, or $\theta = (\mu, \sigma^2)$ for $N(\mu, \sigma^2)$.

As before, the **likelihood** of the sample, given the model F_θ is

$$\mathcal{L}(\theta) := P(\mathbf{X}|\theta) \tag{4.9}$$

where P denotes a PDF for continuous distributions and a PMF for discrete. Note that $\mathcal{L}(\theta)$ is not a distribution for θ, in general it is not normalized. The **maximum likelihood estimator (MLE)** is then defined as

$$\hat{\theta} := \arg\max \mathcal{L}(\theta) \tag{4.10}$$

the set of parameter values that maximize the likelihood. I most examples maximization of the likelihood has a unique solution and the MLE is well defined. Since models of iid samples are usually product distributions, it is often simpler to maximize the **log-likelihood** $\log \mathcal{L}(\theta)$ instead, which gives the same MLE since the logarithm is a monotone increasing function.

Example. Let $\mathbf{X} = (X_1, \ldots, X_N)$ be a sample of iid coin tosses, i.e. $X_i \in \{0, 1\}$. The obvious model is iid $\mathrm{Be}(\theta)$ random variables with $\theta \in [0, 1]$. The likelihood is given by

$$\mathcal{L}(\theta) = P(\mathbf{X}|\theta) = \prod_{i=1}^{N} \theta^{X_i}(1-\theta)^{1-X_i} , \tag{4.11}$$

and the log-likelihood

$$\log \mathcal{L}(\theta) = \sum_{i=1}^{N} X_i \log \theta + (1 - X_i) \log(1 - \theta) . \tag{4.12}$$

Maximizing the latter yields

$$\frac{d}{d\theta} \log \mathcal{L}(\theta) = \sum_{i=1}^{N} \frac{X_i}{\theta} - \frac{1 - X_i}{1 - \theta} = \sum_{i=1}^{N} \frac{X_i - \theta}{\theta(1 - \theta)} = 0 \quad \text{for } \theta = \hat{\theta} , \tag{4.13}$$

which implies that $\hat{\theta} = \frac{1}{N} \sum_{i=1}^{N} X_i$.

This isintuitively reasonable, since $\theta = \mathbb{E}(X_i)$ is in fact the expected value of a Bernoulli random variable.

A more quantitative analysis of the value of an estimator is given by the followint two concepts, for which we assume that the data $\mathbf{X}$ are indeed a sample of the model F_θ, which turns the estimator $\hat{\theta}(\mathbf{X})$ into a random variable w.r.t. the model.

- **Bias.** an estimator $\hat{\theta}$ is **unbiased** if it's expectation is equal to θ, i.e. the **bias**

$$\mathbb{E}[\hat{\theta}|\theta] - \theta = 0 . \tag{4.14}$$

- **Consistency.** An estimator $\hat{\theta}_N$ from a sample of size N is **consistent**, if $\hat{\theta}_N \to \theta$ as $N \to \infty$ in probability, which means that

$$\forall \epsilon > 0 \quad \mathbb{P}(|\hat{\theta} - \theta| > \epsilon) \to 0 \quad \text{as } N \to \infty . \tag{4.15}$$

  Intuitively this means that the distribution of the estimator concentrates around the true value θ for larger and larger sample size. Two sufficient conditions to assure consistency and which are simpler to check are

$$\begin{aligned} \mathbb{E}[\hat{\theta}_N|\theta] &\to \theta \quad \text{i.e. the estimator is } \textbf{asymptotically unbiased} , \\ \mathrm{Var}[\hat{\theta}_N|\theta] &\to 0 \quad \text{as } N \to \infty . \end{aligned} \tag{4.16}$$

In particular, for unbiased estimators only the second condition on variances needs to be checked.

Example. The MLE $\hat{\theta} = \frac{1}{N}\sum_{i=1}^{N} X_i$ for iid coin tosses is unbiased and consistent, since

$$\begin{aligned} \mathbb{E}[\hat{\theta}|\theta] &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[X_i|\theta] = \theta \;, \\ \mathrm{Var}[\hat{\theta}|\theta] &= \sum_{i=1}^{N}\mathrm{Var}\Big[\frac{X_i}{N}\Big|\theta\Big] = \frac{N}{N^2}\mathrm{Var}[X_1|\theta] = \frac{\theta(1-\theta)}{N} \to 0 \;. \end{aligned} \tag{4.17}$$

In general, the exact value of the estimator variance is not important for consitency, it should simply vanish in the limit $N \to \infty$.

4.3 MLE for Gaussians

Let $\mathbf{X} = (X_1, \ldots, X_N)$ be an iid sample of real-valued random variables, with a Gaussian model $N(\mu, \sigma^2)$. The likelihood is given by the PDF for the sample

$$\mathcal{L}(\mu,\sigma^2) = \prod i = 1^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(X_i-\mu)^2}{2\sigma^2}\Big) \tag{4.18}$$

and the log-likelihood by

$$\begin{aligned} \log\mathcal{L}(\mu,\sigma^2) &= \sum_{i=1}^{N}\Big(-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma^2 - \frac{(X_i-\mu)^2}{2\sigma^2}\Big) = \\ &= -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(X_i-\mu)^2 \;. \end{aligned} \tag{4.19}$$

To find the MLE we have to find the roots of two partial derivatives. The first one

$$\frac{\partial}{\partial\mu}\log\mathcal{L}(\mu,\sigma^2) = -\frac{-2}{2\sigma^2}\sum_{i=1}^{N}(X_i-\mu) = 0 \tag{4.20}$$

implies that, as expected, the MLE for μ is given by the sample mean,

$$\sum_{i=1}^{N} X_i - N\hat{\mu} = 0 \quad \Rightarrow \quad \hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i \;. \tag{4.21}$$

The second derivative

$$\frac{\partial}{\partial\sigma^2}\mathcal{L}(\mu,\sigma^2) = -\frac{N}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{N}(X_i-\mu)^2 = 0 \tag{4.22}$$

leads to the sample variance as MLE for σ^2,

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i-\hat{\mu})^2 \;. \tag{4.23}$$

Analogously to (4.17) $\hat{\mu}$ is unbiased, and for $\hat{\sigma}^2$ we get

$$\mathbb{E}[\hat{\sigma}^2|\mu,\sigma^2] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\big[(X_i-\hat{\mu})^2\big|\mu,\sigma^2\big] = \mathbb{E}\big[(X_1-\hat{\mu})^2\big|\mu,\sigma^2\big] \tag{4.24}$$

using that the X_i are iid under our model. This leads to

$$\begin{aligned}\mathbb{E}[\hat{\sigma}^2|\mu,\sigma^2] &= \mathbb{E}[X_1^2-2\hat{\mu}X_1+\hat{\mu}^2|\mu,\sigma^2] = \\ &= \sigma^2+\mu^2-2\mathbb{E}\Big[\frac{1}{N}\sum_{i=1}^{N}X_iX_1\Big|\mu,\sigma^2\Big]+\mathbb{E}\Big[\frac{1}{N^2}\sum_{i,j=1}^{N}X_iX_j\Big|\mu,\sigma^2\Big] = \\ &= \sigma^2+\mu^2-\tfrac{2}{N}\big[(\sigma^2+\mu^2)+(N-1)\mu^2\big]+\tfrac{1}{N^2}\big[N(\sigma^2+\mu^2)+N(N-1))\mu^2\big] \\ &= \sigma^2\frac{N-1}{N}<\sigma^2\ ,\end{aligned} \tag{4.25}$$

so the sample variance is a biased estimator for the variance. Form the above computation it is clear that

$$\hat{\sigma}^2 := \frac{1}{N-1}\sum_{i=1}^{N}(X_i-\hat{\mu})^2 \tag{4.26}$$

is an unbiased estimator σ^2. Since we have not used the fact that we have a Gaussian distribution to compute the bias, this holds in fact for estimators for mean and variance for all iid distributed samples.

For consistency of $\hat{\mu}$ we get analogous to the Bernoulli case

$$\mathrm{Var}[\hat{\mu}|\mu] = \frac{1}{N}\mathrm{Var}[X_1|\mu] = \frac{\sigma^2}{N}\to 0 \quad \text{as } N\to\infty\ . \tag{4.27}$$

For the variance estimator we get, using that the X_i are iid,

$$\mathrm{Var}[\hat{\sigma}^2|\mu,\sigma^2] = \frac{1}{N-1}\mathrm{Var}[(X_1-\hat{\mu})^2|\mu,\sigma^2] = \frac{C}{N-1}\to 0\ , \tag{4.28}$$

where the constant C can in principle be obtained from a cumbersome computation, but is not really relevant to check for consistency.

4.4 Confidence intervals

Recall that the sample mean

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N}X_i \quad \text{has variance} \quad \mathrm{Var}[\hat{\mu}|\mu\sigma^2] = \frac{\sigma^2}{N}\ .$$

The **standard error (SE)** of the mean is a statistic that estimates the standard deviation of $\hat{\mu}$,

$$\mathrm{SE}(\mathbf{X}) := \frac{\hat{\sigma}}{\sqrt{N}} \quad \text{where} \quad \hat{\sigma} = \sqrt{\hat{\sigma}^2}\ . \tag{4.29}$$

By the CLT $\hat{\mu}$ is asymptotically Gaussian, i.e.

$$\hat{\mu}\sim N\big(\mu,\sigma^2/N\big) \quad \text{for large } N\ . \tag{4.30}$$

Therefore one can use the standard **confidence intervals** for the Gaussian, the boundaries of which are determined by percentiles of the normal centered CDF, e.g. for 95% we have

$$\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{x} e^{-y^2/(2\sigma^2)}\,dy = 0.025\ . \tag{4.31}$$

and a symmetric condition for the tail. Since $\Phi^{-1}(0.025) \approx -1.96\sigma$ the 95% confidence interval for the estimate of the mean is that

$$\mu \in \big[\hat{\mu} - 1.96\,\mathrm{SE}, \hat{\mu} + 1.96\,\mathrm{SE}\big] \quad \text{with probability } 0.95\ . \tag{4.32}$$

Ususally one uses simply ±2SE for the confidence interval or error bars for the estimator of the mean. Analogous intervals can be computed for other estimators, explicit formulas for the standard error can become quite complicated.

Note that the Gaussian approximation only holds for large enough N. For small or moderate sample sizes the confidence interval is larger (since confidence is smaller). For Gaussian data it is given by the 0.025-percentile of the well-known **Student's t-distribution** with N degrees of freedom, which has PDF

$$f(x) = C_N\Big(1 + \frac{x^2}{N}\Big)^{-(N+1)/2} \text{with a normalizing constant } C_N\ . \tag{4.33}$$

Note that for finite N it has a power-law tail, so fluctuations are more likely as for Gaussians, and as $N \to \infty$ it converges to the Gaussian PDF. In general, an α**-confidence interval** I_α for any distribution with PDF f is usually defined symmetrically around the mean μ, i.e. $I_\alpha = [\mu - a, \mu + a]$ such that

$$\int_{I_\alpha} f(x)\,dx = \alpha\ . \tag{4.34}$$

The CDF of the standard Gaussian $N(0,1)$ is also called the **error function**

$$\Phi(x) := \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-y^2/2}\,dy\ . \tag{4.35}$$

There is no analytic formula for this function, but it has a slightly modified Gaussian tail, i.e. to leading order we have

$$1 - \Phi(x) = e^{-x^2}/(x\sqrt{\pi})\big(1 + o(1)\big) \quad \text{as } x \to \infty\ . \tag{4.36}$$

This follows from re-arranging the derivative $e^{-x^2} = -(2x)^{-1}(e^{-x^2})'$ and integrating by parts.

5 Time series

5.1 Mathematical models for time series data

Time series data are given in the form $\mathbf{X} = (X_1, \ldots, x_N)$ or $\mathbf{X} = \{(t_i, X_i) : i = 1, \ldots, N\}$.
If the X_i are iid the data are actually distributional, and the time series would look like a scatter plot since X_t and t are actually independent, any permutation of the data would give a statistically equivalent time series.

Standard models.

- **trend/signal plus noise:** $X_t = f(t) + \xi_t$ with ξ_t iid noise, usually Gaussian, and the signal is given by a deterministic function $f(t)$. This is a typical model to describe measurement errors, and the X_t are independent but not identical random variables. To detect the independence structure in correlation functions, the time series has to be det-trended first, which we will discuss in detail.

- **moving average process MA(q):**

$$X_t = \mu + \xi_t + \theta_1 \xi_{t-1} + \ldots + \theta_q \xi_{t-q} \ , \tag{5.1}$$

  where the ξ_t are iid noise and $\mu, \theta_1, \ldots, \theta_q \in \mathbb{R}$, $q \in \mathbb{N}$ are paremeters. This corresponds to integrated or averaged noise and is the simplest model for a stationary time series with $\mathbb{E}[X_t] = \mu$ for all t. Note that X_t and X_{t+q+1} are independent and the largest correlation length in the system is q. Such processes also occur as (truncated) solutions for the following auto regressive models.

- **autoregressive process AR(q):**

$$X_t = c + \phi_1 X_{t-1} + \ldots + \phi_q X_{t-q} + \xi_t \ , \tag{5.2}$$

  where ξ_t is iid noise and $c, \phi_1, \ldots, \phi_q \in \mathbb{R}$, $q \in \mathbb{N}$ are parameters. This recursive definition of the process (X_t) can admit stationary solutions, depending on the parameter values, but solutions could also diverge. The recursion can generate long range dependences which only decay asymptotically, and infinite moving average processes can be solutions to the recursion, which often can be truncated at a finite, large value. AR(p) and MA(q) can be combined to general **ARMA(p,q)** models, with recursion depth p and average depth q. In MATLAB this class of processes is implemented under the name `arima`.

- Any time series can in general be viewed as a realization or sample of a **stochastic process**. The above models are all examples of processes with linear/additive noise, and the ARMA(1,0) process of depth 1 is actually a **Markov process**, i.e. the future value X_{t+1} only depends on the present X_t and noise ξ_t.
  In some examples multiplicative noise is more appropriate, e.g. geometric Brownian motion given by the recursion

$$X_{t+1} = X_t(\mu + \xi_t) \quad \text{with } \xi_t \text{ iid noise and } \mu \in \mathbb{R} \tag{5.3}$$

  is a common discrete-time model for stock prices. We will not consider these models further, often they can be turned into additive noise models by taking logarithms.

5.2 Correlations and stationarity

The most general model of a timeseries is a **stochastic process** $(X_t : t \in \mathbb{T})$, where we denote by $\mathbb{T}$ a general time index set which could be discrete (e.g. $\mathbb{N}$, $\mathbb{Z}$) or continuous (e.g. $[0, \infty)$, $\mathbb{R}$).

- If the X_t are iid, $(X_t : t \in \mathbb{T})$ is a pure noise process, and if they are standard Gaussian $X_t \sim N(0,1)$ this process is called **white noise**. This process exists for discrete and continuous time.
- If the X_t are independent but not identically distributed, signal plus noise is often a good model.
- If the X_t are identical but not independent, $(X_t : t \in \mathbb{T})$ could be a stationary process, which is a very important class of timeseries models, as we will discuss in the following.

A stochastic process $(X_t : t \in \mathbb{T})$ is called **stationary**, if its joint distributions are invariant under time shifts, i.e.

$$\big(X_{t_1+\tau}, \ldots, X_{t_k+\tau}\big) \sim (X_{t_1}, \ldots, X_{t_k}) \tag{5.4}$$

for all $k \in \mathbb{N}$, $t_i \in \mathbb{T}$ and $t_i + \tau \in \mathbb{T}$. The problem with this definition is that it cannot be checked from data, so one has to resort to a weaker version from a statistical point of view which is related to correlations.
Let $(X_t : t \in \mathbb{T})$ and $(Y_t : t \in \mathbb{T})$ be two processes with the same time index set. Then the **cross correlation** is defined as

$$R_{XY}(s,t) = \frac{\mathrm{Cov}(X_s, Y_t)}{\sigma(X_s)\,\sigma(Y_t)} = \frac{\mathbb{E}((X_s - \mathbb{E}X_s)(Y_t - \mathbb{E}Y_t))}{\sigma(X_s)\,\sigma(Y_t)} \,. \tag{5.5}$$

By the **Cauchy-Schwarz inequality** we have

$$\mathbb{E}\big((X_s - \mathbb{E}X_s)(Y_t - \mathbb{E}Y_t)\big)^2 \leq \mathbb{E}\big((X_s - \mathbb{E}X_s)^2\big)\mathbb{E}\big((Y_t - \mathbb{E}Y_t)^2\big)\big) \,, \tag{5.6}$$

and therefore, the cross correlation is bounded by $R_{XY}(s,t) \in [-1,1]$ for all s, t. The **auto correlation** of a process $(X_t : t \in \mathbb{T})$ is then given by

$$R(s,t) = R_{XX}(s,t) = \frac{\mathrm{Cov}(X_s, X_t)}{\sigma(X_s)\,\sigma(X_t)} = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}} \,, \tag{5.7}$$

where we use the notation $\gamma(s,t) = \mathrm{Cov}(X_s, X_t)$ for the self-covariances.
The process $(X_t : t \in \mathbb{T})$ is then called **weakly stationary** (or covariance stationary), if

$$\mathbb{E}[X_t] \equiv \mu \quad \text{and} \quad \gamma(t, t+\tau) \equiv \gamma(\tau) \quad \text{for all } t \in \mathbb{T} \,, \tag{5.8}$$

i.e. the covariances depend only on the time-lag τ. In this case we have

$$R(t, t+\tau) = R(\tau) = \frac{\gamma(\tau)}{\gamma(0)} \tag{5.9}$$

and $R(\tau)$ is a symmetric function by definition with $R(0) = 1$. Furthermore, $|\gamma(\tau)| \leq \gamma(0)$ for all t and $R(\tau) \in [-1,1]$.

In general, if we have M samples of our timeseries $\mathbf{X}^1, \ldots, \mathbf{X}^M$, we can use the obvious estimator for the auto correlation function

$$R_M(s,t) := \frac{1}{M} \sum_{k=1}^{M} \frac{(X_t^k - \hat{\mu}_t)(X_s^k - \hat{\mu}_s)}{\hat{\sigma}_t \hat{\sigma}_s} \,, \tag{5.10}$$

where $\hat{\mu}_t = \frac{1}{M} \sum_{k=1}^{M} X_t^k$ and $\hat{\sigma}_t^2 = \frac{1}{M} \sum_{k=1}^{M} (X_t^k - \hat{\mu}_t)^2$ are the sample mean and variance at each time $t \in \mathbb{T}$. This situation is common if one can perform several realizations of an experiment, and the function $R_N(s,t)$ can then be used to test for stationarity. In many cases, however, there is only a single sample available of a time series, such as temperature records and other observational data. In this case, the only possibility to compute auto correlations is to use stationarity of the series and replace the sample average by a time average. So let us assume that we are given a single realization $\mathbf{X} = (X_1, \ldots, X_N)$ of a stationary timeseries. Then an estimator of the auto correlation is given by

$$R_N(\tau) := \frac{1}{N-\tau} \sum_{t=1}^{N-\tau} \frac{(X_t^k - \hat{\mu})(X_s^k - \hat{\mu})}{\hat{\sigma}^2} \tag{5.11}$$

where $\hat{\mu} = \frac{1}{N} \sum_{t=1}^{N} X_t$ and $\hat{\sigma}^2 = \frac{1}{N} \sum_{t=1}^{N} (X_t - \hat{\mu})^2$ are the estimates for the time-independent mean and variance of the sample. $R_N(0) = 1$ by defitition and is also a symmetric function in τ, which is consistent with the true auto correlation. If $\mathbf{X}$ actually consists of iid random variables with finite mean μ and variance σ^2, one can show that

$$\mathbb{E}[R_N(\tau)] \propto -\frac{1}{N} \quad \text{and} \quad \mathrm{Var}[R_n(\tau)] \propto \frac{1}{N} \quad \text{for all } \tau > 0 \,. \tag{5.12}$$

So in this case $R_N(\tau)$ is actually a asymptotically unbiased, consistent estimator for the true auto correlation. One can also show that for large N, $R_N(\tau)$ is approximately Gaussian distributed, so by the CLT the 95% confidence interval has width proportional to $1/\sqrt{N}$. So if $R_N(\tau)$ takes values in this interval, there is no statistically significant correlation at time lag τ since an iid model would lead to a similar value.

If a given timeseries is not stationary, the above analysis does not give reasonable results and the correlation function decays very slowly or does not decay at all even for independent random variables. In this case, the first thing to do is to de-trend the time series using techniques such as linear regression, which we will discuss next.

5.3 Gaussian processes

$(X_t : t \in \mathbb{T})$ is a **Gaussian process** if all joint distributions are Gaussian, i.e.

$$(X_{t_1}, \ldots, X_{t_k}) \sim N(\boldsymbol{\mu}, \Sigma) \tag{5.13}$$

with mean $\boldsymbol{\mu} = \big(\mu(t_1), \ldots, \mu(t_k)\big)$ and covariance matrix $\Sigma = \big(\sigma(t_i, t_j) : i, j = 1, \ldots, k\big)$. Such a process is uniquely determined by its

$$\textbf{mean} \quad \mu(t) := \mathbb{E}[X_t] \quad \text{and } \textbf{covariance function} \quad \sigma(s,t) := \mathrm{Cov}(X_s, X_t) \,. \tag{5.14}$$

A Gaussian process has auto correlation function

$$R(s,t)=\frac{\sigma(s,t)}{\sqrt{\sigma(s,s)\sigma(t,t)}} \tag{5.15}$$

and is weakly stationary if $\mu(t)\equiv\mu$ and $\sigma(t,t+\tau)=\sigma(\tau)$ for all $t\in\mathbb{T}$. As with independence and correlation, weak stationarity is actually equivalent to stationarity since the process is fully determined by mean and covariance functions.
Examples include the signal plus noise process in the case of Gaussian noise, where the signal $f(t)=\mu(t)$ is equal to the mean. Also **Brownian motion** is a Gaussian process with mean 0 and covariance $\sigma(t,t+\tau)=t$, which is increasing with t and the process is not stationary (even though it has constant mean). The formal derivative of Brownian motion is the **white noise process**, which is a stationary Gaussian process with

$$\text{mean}\quad \mu=0\quad\text{and covariance}\quad \sigma(t,t+\tau)=\sigma(\tau)=\delta_0(\tau)\ . \tag{5.16}$$

So the random variables $X_t\sim N(0,1)$ are iid Gaussians.
Due to the nice invariance properties of Gaussians, combinations of Gaussian processes are again Gaussian processes and they are used as a common class of models for timeseries data.

6 Linear regression

6.1 Least squares and MLE

Given a time series of the form $\big\{(t_i,X_i):i=1,\ldots,N\big\}$ we consider the model

$$X_i=f(t_i)+\xi_i\quad\text{with iid noise, and} \tag{6.1}$$

deterministic **trend** or **signal** $f:\mathbb{R}\to\mathbb{R}$. For **linear** regression, the trend is given by a linear combination of $M\in\mathbb{N}$ **basis functions** $\phi_0,\ldots,\phi_{M-1}$, parametrized as

$$f(t)=f(z|\mathbf{w})=\sum_{i=0}^{M}w_i\,\phi_i(t)=\langle\mathbf{w}|\boldsymbol{\phi}(t)\rangle\ , \tag{6.2}$$

with parameters $w_0,\ldots,w_{M-1}\in\mathbb{R}$ that can be inferred from the data. One usually chooses $\phi_0(t)=1$ to account for a constant shift in the data, the other functions are arbitrary but have to be chosen to form a basis of a linear space. That means that they have to be linearly independent and none of the basis functions can be written as a linear combination of others. The most common choice are

$$\textbf{polynomial basis functions}\quad \phi_i(t)=t^i\ ,\quad i=0,\ldots,M-1\ , \tag{6.3}$$

other choices include $\sin(\omega t)$ and $\cos(\omega t)$ e.g. for seasonal trends in climate data, or Gaussians or sigmoid basis functions.

The **least squares estimate (LSE)** $\hat{\mathbf{w}}$ for the parameters $\mathbf{w}$ is defined as the minimizer of the **least squares error function**

$$E(\mathbf{w})=\frac{1}{2}\sum_{i=1}^{N}\big(X_i-\langle\mathbf{w}|\boldsymbol{\phi}(t_i)\rangle\big)^2\ . \tag{6.4}$$

This is equivalent to the MLE under a Gaussian model

$$X_i = \langle \mathbf{w} | \boldsymbol{\phi}(t_i) \rangle + \xi_i \quad \text{with} \quad \xi_i \sim N(0, \sigma^2)\text{iid} \tag{6.5}$$

with mean 0 and variance σ^2. The log-likelihood of the model is given by

$$\log \mathcal{L}(\mathbf{w}, \sigma^2) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\big(X_i - \langle \mathbf{w}|\boldsymbol{\phi}(t_i)\rangle\big)^2 \,, \tag{6.6}$$

and maximizing this expression over $\mathbf{w}$ for given σ^2 is equivalent to minizing (6.4).

To compute the LSE/MLE, the partial derivative is given by

$$\frac{\partial}{\partial w_k}E(\mathbf{w}) = \sum_{i=1}^{N}\big(X_i - \langle \mathbf{w}|\boldsymbol{\phi}(t_i)\rangle\big)(-\phi_k(t_i)) \,, \tag{6.7}$$

which leads to the following condition for the gradient (written as a row vector)

$$\begin{aligned} \langle\nabla|E(\mathbf{w}) \;&=\; \sum_{i=1}^{N}\Big(\langle \mathbf{w}|\boldsymbol{\phi}(t_i)\rangle\langle\boldsymbol{\phi}(t_i)| - X_i\langle\boldsymbol{\phi}(t_i)|\Big) = \\ &=\; \Big\langle \mathbf{w}\Big|\sum_{i=1}^{N}|\boldsymbol{\phi}(t_i)\rangle\langle\boldsymbol{\phi}(t_i)| - \sum_{i=1}^{N}X_i\langle\boldsymbol{\phi}(t_i)| = \langle 0| \,. \end{aligned} \tag{6.8}$$

This can be written in a shorter way in terms of the **design matrix**

$$\Phi = \begin{pmatrix} \phi_0(t_1) & \dots & \phi_{M-1}(t_1) \\ \vdots & & \vdots \\ \phi_0(t_N) & \dots & \phi_{M-1}(t_N) \end{pmatrix} \in \mathbb{R}^{N\times M} \,, \tag{6.9}$$

which consists of the basis functions evaluated at the base points $t_1, \dots, t_N$. Then

$$\langle\nabla|E(\mathbf{w}) = \langle\mathbf{w}|\underbrace{\Phi^T\Phi}_{\in\mathbb{R}^{M\times M}} - \underbrace{\langle\mathbf{X}}_{\in\mathbb{R}^N}|\underbrace{\Phi}_{\mathbb{R}^{N\times M}} = \langle 0| \quad \in \mathbb{R}^M \,, \tag{6.10}$$

which has the solution

$$\langle\mathbf{w}| = \langle\mathbf{X}|\Phi(\Phi^T\Phi)^{-1} \quad \text{or} \quad |\mathbf{w}\rangle = (\Phi^T\Phi)^{-1}\Phi^T|\mathbf{X}\rangle \,. \tag{6.11}$$

$(\Phi^T\Phi)^{-1}\Phi^T$ is called the **Moore-Penrouse pseudo inverese** of the desing matrix Φ, and is well defined as long as all basis functions evaluated at $t_1, \dots, t_N$ (columns of Φ) are linearly independent, and the $\mathrm{rank}(\Phi) = M$ (full rank). This is only possible if $N \geq M$, since otherwise the rank of Φ is bounded by $N < M$ and $\Phi^T\Phi$ is not invertible. In practice numerical inversion becomes unstable long before M reaches N, and usually requires $M \ll N$ is required. This is also consistent with the main idea of regression, where the number of inferred parameters in the signal should be clearly less than the number of data points.

6.2 Goodness of fit and model selection

The quality of the fit $\hat{\mathbf{w}}$ is characterized by the **residual sum of squares (RSS)**

$$\mathrm{RSS} = 2E(\hat{\mathbf{w}}) = \sum_{i=1}^{N} \big(X_i - \langle \hat{\mathbf{w}} | \boldsymbol{\phi}(t_i) \rangle \big)^2 = N\hat{\sigma}^2 \, , \tag{6.12}$$

which is also proportional to the MLE for the variance σ^2 in the Gaussian model (6.5). For given parameter dimension M this quantity is minimzed, and it is instructive to compare its value it to the **total sum of squares (TSS)**

$$\mathrm{TSS} = \sum_{i=1}^{N} \big(X_i - \hat{\mu} \big)^2 \quad \text{with sample mean} \quad \hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} X_i \, . \tag{6.13}$$

This is equivalent to using $\mathbf{w} = (\hat{\mu}, 0, \ldots)$ instead of $\hat{\mathbf{w}}$, i.e. $\mathrm{TSS} = 2E\big((\hat{\mu}, 0, \ldots)\big)$ (which would be the LSE for a model with $M = 1$). If the data does not exhibit any particular trend in addition to a simple shift, TSS and RSS are of the same order, whereas for non-trival trends the latter should be much smaller. This is quantified in the **coefficient of determination**

$$R^2 := 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} \tag{6.14}$$

which is close to 1 if the fit leads to a good approximation.

In general, the RSS is a decreasing function of M and decays to 0 if $M = N$, i.e. there are as many parameters as data points. In this case, since the basis functions are linearly independent, the trend $f(t|\mathbf{w})$ can go through all the points. This is, however, not desirable, since the model has simply learned the noise of the data, and the data itself could be as good a model without any regression. The problem of fitting too many parameters in a regression is called **overfitting**, and has to be avoided in order for the mathematical model to be informative and separate the signal from noise effectively. In the following we discuss the most common systematic approaches to **model selection**, i.e. to avoid overfitting and determine an optimal number of parameters in a regression.

Cross validation. We partition the data $\big\{(t_i, X_i) : i = 1, \ldots, N\big\}$ into a **training set** $\mathbf{X}^A = \big\{(t_i, X_i) : i =\in A\big\}$ and a **test set** $\mathbf{X}^B = \big\{(t_i, X_i) : i \in B\big\}$ with $A \cup B = \{1, \ldots, N\}$ and $A \cap B = \emptyset$. We fit the model (6.1) through linear regression on the training set so that

$$\hat{\mathbf{w}} = \mathrm{argmin}_{\mathbf{w}} E^A(\mathbf{w}) = \mathrm{argmin}_{\mathbf{w}} \frac{1}{2} \sum_{i \in A} \big(X_i - \langle \mathbf{w} | \boldsymbol{\phi}(t_i) \rangle \big)^2 \, . \tag{6.15}$$

As we have discussed above, the **training error** $E^A(\hat{\mathbf{w}})$ is a decreasing function of the number M of parameters since the model learns the noise of the trainig set. The **test error**

$$E^B(\hat{\mathbf{w}}) = \frac{1}{2} \sum_{i \in B} \big(X_i - \langle \hat{\mathbf{w}} | \boldsymbol{\phi}(t_i) \rangle \big)^2 \tag{6.16}$$

however, typically increases only initially and then increases with M, and exhibits a minimum around the desired value for M. In order to learn all features of the model the training set should be typical, which is best achieved by picking the training set A at random. A natural size for training and test set is $N/2$, but as long as N is large enough other fractions can work well as well. A typical procedure is to compute an averaged test error $\bar{E}(M)$ for each M over about 10 independent choices of training and test set, and select the best model size M^* as the minimum of $\bar{E}(M)$.

Regularized least squares regression (ridge regression). Another approach is to adapt the cost/error function $E(\mathbf{w})$ to punish large values of the coefficients w_k, which usually come along with overfitting. This looks like

$$\tilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}\big(X_i - \langle \mathbf{w}|\boldsymbol{\phi}(t_i)\rangle\big)^2 + \frac{\lambda}{2}\langle \mathbf{w}|\mathbf{w}\rangle \ , \tag{6.17}$$

where $\lambda \geq 0$ is a parameter that multiplies the norm $\|\mathbf{w}\|_2^2 = \langle \mathbf{w}|\mathbf{w}\rangle$. Minimization of this function contains a new term

$$\frac{\partial}{\partial w_k}\Big(\sum_{j=1}^{M} w_j^2\Big) = 2w_k \ , \tag{6.18}$$

and is otherwise analogous to standard LS regression, leading to

$$\langle \nabla|\tilde{E}(\mathbf{w}) = \langle \mathbf{w}|\Phi^T\Phi - \langle \mathbf{X}|\Phi + \lambda\langle \mathbf{w}| = \langle 0| \tag{6.19}$$

with the solution

$$|\mathbf{w}\rangle = (\Phi^T\Phi + \lambda \mathrm{Id})^{-1}\Phi^T|\mathbf{X}\rangle \ . \tag{6.20}$$

In Bayesian context this is equivalent to having a Gaussian prior on $\mathbf{w}$ with mean 0 and variance $1/\lambda$, where the standard case corresponds to a uniform prior, i.e. no a-priori knowledge on the possible parameter values. This can be generalized to Gaussians with inverse covariance matrix Γ by adding $\frac{1}{2}\langle \mathbf{w}|\Gamma|\mathbf{w}\rangle$, which leads to

$$|\mathbf{w}\rangle = (\Phi^T\Phi + \Gamma^T\Gamma)^{-1}\Phi^T|\mathbf{X}\rangle \ . \tag{6.21}$$

The most important point is to choose λ 'correctly'. For a fixed large enough number M of parameter values, the test error as a function of λ should have a minimum at the optimal value λ^*, whereas the training error will exhibit a plateau around this value and then further decrease with decreasing λ. Typical values of λ^* are very small (around $10^{-3} - 10^{-6}$), since the coefficients usually can have quite different values which have to be equally likely under the prior.

Maximize adjusted coefficient of determination. The adjusted coefficient is defined as

$$\bar{R}^2 := 1 - (1 - R^2)\frac{N-1}{M-M} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}\frac{N-1}{M-M} = R^2 - (1-R^2)\frac{M}{N-M} \ . \tag{6.22}$$

From the last expression we see that compared to R^2, $\bar{R}^2$ has a negative component with a prefactor that is increasing with M and eventually leads to a decrasing evaluation of the fit for increasing M. However, since R^2 itself is M-dependent, $\bar{R}^2(M)$ as a function of M takes often a maximum at a particular value M^*, which can be taken as the optimal choice for the number of parameters in the regression.

From the second expression in (6.22) we see that the motivation for the adjustment is a result of replacing the biased estimators RSS and TSS (both normalized by $1/N$) by their unbiased version, normalized by $1/(N-M)$ and $1/(N-1)$, respectively. The general concept behind this normalization is called **degrees of freedom**, where one such degree has been used to compute the sample mean for the TSS (hence $N-1$), and M degrees have been used to determine the coefficients $\mathbf{w}$ in RSS (hence $N-M$).

The adjusted $\bar{R}^2$-value is the simplest of the above approaches and often leads to reasonable results. There are also automated implementations of the regularized/ridge regression (e.g. `ridge`$(\mathbf{X}, \Phi, \lambda)$ in MATLAB), which make this a computationally viable method. The most flexible approach which can always be implemented and will deliver some controlled understanding of the optimum M^* (but requires some work) is cross validation. In general, it is also very important to use **common sense** in model selection. If the coefficients of the fitted polynomial decay very quickly and there are only $Q < M^*$ which are not essentially 0, choosing Q over M^* (by however method this was determined) can be well justified. In general the simplest way to argue is that the adjusted $\bar{R}^2(M)$ as a function of the number of parameters grows rapidly until Q and than more slowly and $\bar{R}^2(Q)$ and $\bar{R}^2(M^*)$ are very similar. This provides a good justification to choose Q. In general both choices could be compared by an hypothesis test, which is not part of this module (see CO902 in term 2).

6.3 Detrending

One of the most important uses of linear regression is detrending of a timeseries. Assume the following model

$$X_t = f(t) + Y_t \quad \text{where } Y_t \text{ is a stationary process (not necessarily iid)} . \tag{6.23}$$

The goal is to extract the deterministic signal $F(t)$ from the timeseries in order to study the correlation structure of the stationary part Y_t. A **linear trend** $f(t) = \beta_1 + \beta_2 t$ in the timeseries can simply be removed by **differencing**,

$$(\nabla X)_t := X_t - X_{t-1} = \beta_1 + \beta_2 t + Y_t - \beta_1 - \beta_2(t-1) - Y_{t-1} = \beta_2 + (\nabla Y)_t . \tag{6.24}$$

If Y_t is stationary, the difference process ∇Y is stationary with mean 0, so ∇X is stationary with mean β_2. The variance is given by

$$\mathrm{Var}\big((\nabla X)_t\big) = \mathrm{Var}\big((\nabla Y)_t\big) = 2\sigma^2 - 2\mathrm{Cov}(Y_t, Y_{t-1}) = 2\big(\gamma(0) - \gamma(1)\big) , \tag{6.25}$$

where $\sigma^2 = \mathrm{Var}(Y_t) = \gamma(0)$ denotes variance and covariance of the stationary process Y_t. This approach could be extended to higher order polynomial trends.

The most general and usual approach is detrending by **regression**. We assume the above model to be of the usual form

$$X_t = \langle \mathbf{w} | \boldsymbol{\phi}(t) \rangle + Y_t \quad \text{with basis functions} \quad \phi_0, \ldots, \phi_{M-1} . \tag{6.26}$$

After finding the LSE estimate $\hat{\mathbf{w}}$ the series

$$Y_t := X_t - \langle \hat{\mathbf{w}} | \boldsymbol{\phi}(t) \rangle \tag{6.27}$$

is a stationary sequence (hopefully, as far as can be determined from the data). Usual choices for basis functions are as mentioned before polynomials or $\sin$ and $\cos$ for seasonal trends in data or a combination of both.

The use of $\sin$ and $\cos$ is also common in **extracting a signal from noise**, which consists of identifying a carrier wave in a noisy signal. If the frequency ν of the wave is known, this can be achieved by linear regression and is possible even for very high noice levels. The associated model is

$$X_t = A\,\cos(2\pi\nu t + \phi) + \xi_t \quad \text{with noise } \xi_t \sim N(0\sigma^2) \text{ iid} , \tag{6.28}$$

amplitude $A > 0$ and **phase** $\phi \in [0, 2\pi)$. One often uses the notation $\omega = 2\pi\nu$ for the angular frequency. By trigonometric identities this can also be written in the form

$$X_t = B_1 \sin(\omega t) + B_2 \cos(\omega t) + \xi_t , \tag{6.29}$$

which is linear in the coefficients $B_1, B_2 \in \mathbb{R}$. These are functions of A and ϕ and can be estimated by linear regression to determine the latter.

7 Autoregressive models

7.1 Correlation functions and stationary solutions

Consider the autoregressive model AR(q) of degree $q \in \mathbb{N}$,

$$X_t = c + \phi_1 X_{t-1} + \ldots + \phi_q X_{t-q} + \xi_t , \tag{7.1}$$

where ξ_t is iid noise with $\mathbb{E}(\xi_t) = 0$ and $\text{Var}\xi_t = \sigma^2$. Under certain conditions on the paramters, this recursion relation admits a stationary solution, i.e. a stationary process X_t that fulfills (7.1). Assuming that X_t is stationary, we can get a condition on the mean $\mathbb{E}(X_t) = \mu$ from the recursion,

$$\mu = c + (\phi_1 + \ldots + \phi_q)\,\mu \quad \Rightarrow \quad \mu = \frac{c}{1 - \phi_1 - \ldots - \phi_q} . \tag{7.2}$$

This already provides a first condition on existence of a stationary process with finite mean, namely that $\phi_1 + \ldots + \phi_q \neq 1$ which we will explain later. Substracting the self consistent relation for the mean from (7.1) we get the homogeneous equation

$$X_t - \mu = \phi_1(X_{t-1} - \mu) + \ldots + \phi_q(X_{t-q} - \mu) + \xi_t . \tag{7.3}$$

Multiplying with $(X_0 - \mu)$ and taking expectation this leads to

$$\gamma(t) = \phi_1\gamma(t-1) + \ldots + \phi_q\gamma(t-q) , \tag{7.4}$$

where we have used the usual notation $\gamma(t) = \text{Cov}(X_t, X_0)$ and the fact that $\mathbb{E}\big(\xi_t(X_0 - \mu)\big) = 0$ since the noise is iid. This is a linear recursion/difference equation of degree q with constant coefficients, and can be solved by the **exponential ansatz** $\gamma(t) = \lambda^t$ (also called a mode). Pluggin into (7.1), this leads to the characteristic equation

$$\lambda^q - \phi_1\lambda^{q-1} - \ldots - \phi_q = 0 \tag{7.5}$$

for the parameter $\lambda \in \mathbb{C}$ after multiplication with λ^{q-t}. This polynomial equation has q complex roots $\lambda_1, \ldots, \lambda_q$, and assuming they are all different, by linearity the general solution of (7.4) is given by

$$\gamma(t) = A_1\lambda_1^t + \ldots + A_q\lambda_q^t \quad \text{with} \quad A_1, \ldots, A_q \in \mathbb{R} . \tag{7.6}$$

If roots coincide, this leads to polynomial corrections of this expressions which are well known but we do not consider here. Typically, for coefficients ϕ_k estimated from data roots will not coincide, and all coefficients A_k will take non-zero values which are distinct (unless they below to a complex conjugate pair of roots), so that no cancellations happen and the model is non-degenerate. We have

$$A_1 + \ldots + A_q = \gamma(0) = \mathrm{Var}(X_1) \tag{7.7}$$

and further conditions necessary to determine the parameters can be written in terms of $\gamma(t)$ for $t = 1, \ldots, q-1$, which can be inferred from the data. In general, they depend on the actual distribution of the noise, and in the Gaussian case can be computed also analytically in terms of known expressions for covariances.
We know that for stationary processes $|\gamma(t)| \leq \gamma(0)$ for all t, so the solution has to be a bounded function. This is possible if $|\lambda_k| < 1$ for all $k = 1, \ldots, q$, where strict inequality is required due to the presence of noise. If $|\lambda_k| = 1$ for some k, the noise would lead to a random walk-type, non-stationary process for $q \geq 2$ or a ballistic motion with random direction for $q = 1$. In fact, for a non-degenerate model the condition

$$|\lambda_k| < 1 \quad \text{for all} \quad k = 1, \ldots, q \tag{7.8}$$

is equivalent to existence of a stationary solution. In particular, this implies that

$$1 - \phi_1 - \ldots - \phi_q \neq 0 \tag{7.9}$$

since otherwise 1 would be a root of (7.5).
So the amplitutes of all modes decay exponentially to 0 since

$$\lambda_k^t = \mathrm{Re}(\lambda_k)^t \, e^{i\,\mathrm{phase}(\lambda_k)t} \,, \tag{7.10}$$

and $\mathrm{Re}(\lambda_k) < 1$ and $|e^{i\mathrm{phase}(\lambda_k)t}| = 1$. Nevertheless, typical realizations of the process show stationary oscillations or correlated fluctuations around the mean value, which are driven by the noise term.

Example. The simplest case is $q = 1$, for which we get $\mu = c/(1 - \phi_1)$ and need $\phi_1 = \phi \in (-1, 1)$. So if ϕ gets close to 1 the mean can become arbitrarily large, and if $\phi = 0$, X_t is just an iid noise process with mean c. The characteristic equation is simply $\lambda - \phi = 0$, so that $\lambda = \phi$ and the covariances of the process are

$$\gamma(t) = \frac{\sigma^2}{1 - \phi^2} \phi^t \,. \tag{7.11}$$

For $q = 1$ we can determine the constant of this solution from a self-consistent equation for the variance (analogous to the mean for the general case), from (7.1) we get from stationarity and independence of the noise

$$\mathrm{Var}(X_1) = \phi^2 \mathrm{Var}(X_1) = \sigma^2 \quad \Rightarrow \quad \gamma(0) = \mathrm{Var}(X_1) = \frac{\sigma^2}{1 - \phi^2} \,. \tag{7.12}$$

Note that this simple approach does not work for $q > 1$ since the variables in the recursion are not independent. Negative values of ϕ lead to anti-correlations and oscillations with period 1 around the mean with added noise, whereas for positive ϕ the noise is dominating sign changes, and the recursion is just damping the noise. Similar statements hold also for larger q, where negative real λ_k lead to period 1 oscillations, complex pairs lead to higher period oscillations, and real positive

λ_k to damping of noise without sign change around the mean.

In general, discrete-time stationary processes can be defined for all $t \in \mathbb{Z}$ since the distribution at time t is independent of t, and for a semi-infinite sequence $(X_0, X_1, \ldots)$ one can simply shift the value of the initial time to any negative value. So let's assume that $\mathbf{X} = (\ldots X_{-1}, X_0, X_1 \ldots)$ is a full stationary sequence which fulfilles (7.1). Then we can rewrite this as a vector valued equation

$$\mathbf{X} = \mathbf{c} + (\phi_1 L + \ldots + \phi_q L^q)\, \mathbf{X} + \boldsymbol{\xi} \,, \tag{7.13}$$

where we use the **left-shift operator** $L : \mathbb{R}^{\mathbb{Z}} \to \mathbb{R}^{\mathbb{Z}}$, defined by

$$(L\mathbf{X})_t = X_{t-1} \,, \tag{7.14}$$

and the obvious notation $\mathbf{c} = (\ldots c, c \ldots)$ and $\boldsymbol{\xi} = (\ldots \xi_{-1}, \xi_0, \xi_1 \ldots)$. We can also write an equation for the mean vector $\boldsymbol{\mu} = (\ldots, \mu, \mu \ldots)$

$$\mathbf{c} = (\mathrm{Id} - \phi_1 L - \ldots - \phi_q L^q)\, \boldsymbol{\mu} \,. \tag{7.15}$$

This leads to the following formal solution of the recursion (7.13)

$$\mathbf{X} = (\mathrm{Id} - \phi_1 L - \ldots - \phi_q L^q)^{-1} (\mathbf{c} + \boldsymbol{\xi}) = \boldsymbol{\mu} + (\mathrm{Id} - \phi_1 L - \ldots - \phi_q L^q)^{-1} \boldsymbol{\xi} \,. \tag{7.16}$$

The operator/matrix inverse is defined by the following series

$$(\mathrm{Id} - \phi_1 L - \ldots - \phi_q L^q)^{-1} = \sum_{j=0}^{\infty} (\phi_1 L + \ldots + \phi_q L^q)^j \tag{7.17}$$

which is analogous to the geometric sum formula. In the simplest case $q = 1$ this leads to

$$\mathbf{X} = \boldsymbol{\mu} + \xi_t + \phi_1 \xi_{t-1} + \phi_1^2 \xi_{t-2} + \phi_1^3 \xi_{t-3} + \ldots \tag{7.18}$$

which is equivalent to a **moving average process** of infinite degree MA(∞). This is also true for higher values of q with more complicated expressions for the coefficients, and ingeneral MA processes can be interpreted as truncated expansions of solutions to AR models. For stationary solutions this expansions actually converge since $|\lambda_k| < 1$, and MA processes are good approximations of stationary AR solutions. However, the solution formula (7.16) also holds in some non-stationary cases (modulo invertability of the operator), and can lead to exponentially diverging solutions if $|\lambda_k| > 1$ for some k, or linearly diverging solutions if $|\lambda_k| = 1$ for some k.

7.2 Linear regression for AR(1) models

In this subsection we consider an AR(1) model with iid $N(0, \sigma^2)$ Gaussian noise

$$X_t = c + \phi X_{t-1} + \xi_t \quad \text{which implies} \quad \mu = \frac{c}{1 - \phi} \quad \text{and} \quad \mathrm{Var}(X_1) = \frac{\sigma^2}{1 - \phi^2} \,. \tag{7.19}$$

From the solution expansion formula (7.18) we see that X_t is given by a combination of Gaussian variables $\xi_t, \xi_{t-1}, \ldots$, so that $X_t \sim N\big(c/(1 - \phi), \sigma^2/(1 - \phi^2)\big)$ is itself a Gaussian for all t. Moreover, $(X_t : t \in \mathbb{Z})$ is a stationary Gaussian process with covariances $\gamma(t)$ given in (7.11).
Let us denote a finite sample of this stationary process by $\mathbf{X}_N = (X_1, \ldots, X_N)$. Since we have no further information on X_0, the distribution of X_1 under our model is simply $N\big(c/(1-\phi), \sigma^2/(1-$

$\phi^2)\big)$. X_2 will be correlated with X_1 via the recursion, and in general the conditional distributions for further values are

$$X_2|_{X_1} \sim N(c + \phi X_1, \sigma^2)\ , \quad X_3|_{X_1,X_2} \sim N(c + \phi X_2, \sigma^2)\ , \ldots \tag{7.20}$$

Using the product rule, the joint PDF of $\mathbf{X}_N$ can be written as

$$f_{\mathbf{X}_N}(x_1, \ldots x_N) = f_{X_N|\mathbf{X}_{N-1}}(x_N|x_{N-1}, \ldots x_1)\, f_{\mathbf{X}_{N-1}}(x_1, \ldots x_{N-1})\ . \tag{7.21}$$

This holds in general for any PDF and can be iterated, for one-step recursions we have the additional simplification that X_N only depends on X_{N-1}, which leads to

$$f_{\mathbf{X}_N}(x_1, \ldots x_N) = f_{X-1}(x_1)\, f_{X_2|X_1}(x_2|x_1) \cdot f_{X_N|X_{N-1}}(x_N|x_{N-1})\ . \tag{7.22}$$

The log-likelihood for this joint PDF is then given by

$$\begin{aligned} \log \mathcal{L}(\mathbf{X}_N|c, \phi, \sigma^2) \;&=\; -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\frac{\sigma^2}{1-\phi^2} - \frac{(X_1 - c/(1-\phi))^2}{2\sigma^2/(1-\phi^2)} \\ &\quad -\frac{N-1}{2}\log(2\pi) - \frac{N-1}{2}\log\sigma^2 - \sum_{t=2}^{N}\frac{(X_t - c - \phi X_{t-1})^2}{2\sigma^2} \end{aligned} \tag{7.23}$$

This expression can also be determined by using the usual formula for the multivariate Gaussian $\mathbf{X}_N \sim N(\boldsymbol{\mu}_N, \Sigma)$ with constant mean vector $\boldsymbol{\mu}_N = c/(1-\phi)(1, \ldots, 1)$ and covariance matrix (from (7.11))

$$\Sigma = (\sigma_{ij}\ i,j = 1, \ldots, N)\ , \quad \sigma_{ij} = \mathrm{Cov}(X_i, X_j) = \frac{\sigma^2}{1-\phi^2}\,\phi^{|i-j|}\ . \tag{7.24}$$

The inverse of this matrix, which enters the joint PDF, is given by

$$\Sigma^{-1} = \frac{1}{\sigma^2}\begin{pmatrix} 1 & -\phi & & & \\ -\phi & 1+\phi^2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & 1+\phi^2 & -\phi \\ & & & -\phi & 1 \end{pmatrix}\ . \tag{7.25}$$

This is tri-diagonal and simpler than Σ itself, since it describes the conditional correlations between variables, which are given by a one-step recursion in our case leading to only one non-zero off-diagonal. The joint Gaussian PDF is then

$$f_{\mathbf{X}_N}(\mathbf{x}_N) = \frac{1}{((2\pi)^{N/2}\det(\Sigma)}\exp\Big(-\frac{1}{2}\big\langle \mathbf{x}_N - \boldsymbol{\mu}_N\big|\Sigma^{-1}\big|\mathbf{x}_N - \boldsymbol{\mu}_N\big\rangle\Big)\ , \tag{7.26}$$

which leads to the same expression as above, since one can compute

$$\det(\Sigma)^{-1} = \det(\Sigma^{-1}) = \frac{1-\phi^2}{(\sigma^2)^N} \tag{7.27}$$

for the normalization terms to match. For the quadratic term in (7.23) in the second line one gets

$$\begin{aligned} (X_t - c - \phi X_{t-1})^2 \;&=\; \big(X_t - \tfrac{c}{1-\phi} + \tfrac{c\phi}{1-\phi} - \phi X_{t-1}\big)^2 = \\ &=\; \big(X_t - \tfrac{c}{1-\phi}\big)^2 - 2\phi\big(X_t - \tfrac{c}{1-\phi}\big)\big(X_{t-1} - \tfrac{c}{1-\phi}\big) + \phi^2\big(X_{t-1} - \tfrac{c}{1-\phi}\big)^2 , \end{aligned} \tag{7.28}$$

which together with the term corresponding to X_1 leads to the desired expression $\frac{1}{2}\big\langle \mathbf{x}_N - \boldsymbol{\mu}_N \big| \Sigma^{-1} \big| \mathbf{x}_N - \boldsymbol{\mu}_N \big\rangle$.
Note that the log-likelihood (7.23) can be minimized w.r.t. c and ϕ without considering σ^2 (analogous to previous cases) and the error function to be minimized is

$$E(c,\phi) = \frac{1}{2}\Big(X_1 - \frac{c}{1-\phi}\Big)(1-\phi^2) - \log(1-\phi^2) + \frac{1}{2}\sum_{t=2}^{N}\big(X_t - c - \phi X_{t-1}\big)^2 \,. \quad (7.29)$$

This, however, does not have the nice symmetric form as for usual Gaussian linear regression, and we cannot apply the formalism with the design matrix developed in the previous section.

LS regression with fixed X_1.
In the following we therefore consider a simpler model, where we take the first data point X_1 to be deterministally fixed to the observed value, i.e. the likelihood of this observation is simply 1 and the first two terms in the error function associated to X_1 drop out. We are left with minimizing

$$\tilde{E}(c,\phi) = \frac{1}{2}\sum_{t=2}^{N}\big(X_t - c - \phi X_{t-1}\big)^2 \,. \quad (7.30)$$

Writing it in the standard form $\langle c,\phi|\psi_0(t),\psi_1(t)\rangle$ we can identify the basis functions

$$\psi_0(t) = (1,\ldots,1) \quad \text{and} \quad \psi_1(t) = (X_1,\ldots,X_{N-1}) \,. \quad (7.31)$$

Note that ψ_1 depends in fact on the data $\mathbf{X}_N$ in this case. The design matrix is then given gy

$$\Phi = \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_{N-1} \end{pmatrix} \quad \text{and} \quad \Phi^T\Phi = \begin{pmatrix} N-1 & \sum_{t=1}^{N-1} X_t \\ \sum_{t=1}^{N-1} X_t & \sum_{t=1}^{N-1} X_t^2 \end{pmatrix} \in \mathbb{R}^{2\times 2} \,. \quad (7.32)$$

Writing out $\Phi^T|\mathbf{X}_N\rangle = \Big|\sum_{t=2}^{N} X_t, \sum_{t=2}^{N} X_t X_{t-1}\Big\rangle$, this leads to the solution

$$|\hat{c},\hat{\phi}\rangle = (\Phi^T\Phi)^{-1}\Phi^T|\mathbf{X}_N\rangle = \begin{pmatrix} N-1 & \sum_{t=1}^{N-1} X_t \\ \sum_{t=1}^{N-1} X_t & \sum_{t=1}^{N-1} X_t^2 \end{pmatrix}^{-1} \Big|\sum_{t=2}^{N} X_t, \sum_{t=2}^{N} X_t X_{t-1}\Big\rangle \,, \quad (7.33)$$

which can easily be implemented numerically. As before, the estimate for σ^2 is then simply given by the residual sum of squares (RSS),

$$\hat{\sigma}^2 = \frac{1}{N-1}\,\mathrm{RSS} = \frac{1}{N-1}\sum_{t=2}^{N}\big(X_t - \hat{c} - \hat{\phi}X_{t-1}\big)^2 \,. \quad (7.34)$$

8 Spectral analysis

8.1 Fourier series

A function $f:\mathbb{R}\to\mathbb{R}$ is **periodic** with **period** T if $f(t+T)=f(t)$ for all $t\in\mathbb{R}$. A periodic function is completely determined by its values on a single period, which we take to be $[-T/2,T/2)$. Simple examples are

$$\begin{aligned}
&\text{trigonometric functions} \qquad f(t)=\sin\big(2\pi nt/T\big)\ ,\quad n\in\mathbb{Z}\ ,\\
&\text{square wave} \qquad f(t)=\left\{\begin{array}{rl} 1 & ,\ -T/2\le t<0\\ -1 & ,\ 0\le t<T/2\end{array}\right.
\end{aligned}\tag{8.1}$$

Fourier series. Let f be a periodic real or complex valued function with period T. Then it can be written as a series of periodic exponentials

$$f(t)=\sum_{n\in\mathbb{Z}}A_n\,e^{2\pi int/T}\ ,\tag{8.2}$$

where the **Fourier coefficients** are unique, and given by

$$A_n=\frac{1}{T}\int_{-T/2}^{T/2}f(t)\,e^{-2\pi int/T}\,dt\quad\in\ \mathbb{C}\ .\tag{8.3}$$

To show this, we use the **orthogonality relations** for complex exponentials,

$$\int_{-T/2}^{T/2}e^{2\pi int/T}\,e^{-2\pi imt/T}\,dt=T\,\delta_{n,m}\tag{8.4}$$

since for $n=m$ we simply integrate 1 and for $n\neq m$ we have

$$\int_{-T/2}^{T/2}e^{2\pi int/T}\,e^{-2\pi imt/T}\,dt=\frac{T}{2\pi i(n-m)}\big(e^{2\pi i(n-m)/2}-e^{-2\pi i(n-m)/2}\big)=0\ .\tag{8.5}$$

With this, using the representation (8.2) we get

$$\begin{aligned}
\frac{1}{T}\int_{-T/2}^{T/2}f(t)\,e^{-2\pi imt/T}\,dt\ &=\ \frac{1}{T}\sum_{n\in\mathbb{Z}}A_n\int_{-T/2}^{T/2}e^{2\pi int/T}\,e^{-2\pi imt/T}\,dt\\
&=\ \frac{1}{T}\sum_{n\in\mathbb{Z}}A_nT\,\delta_{n,m}=A_m\ ,
\end{aligned}\tag{8.6}$$

which confirms (8.3).

Simple examples.

$$\sin(2\pi t/T)=\frac{1}{2i}\big(e^{2\pi it/T}-e^{-2\pi it/T}\big)\quad\Rightarrow\quad A_1=-A_{-1}=\frac{1}{2i}\ ,\tag{8.7}$$

and all other coefficients vanish. For $f(t)=1$ we have $A_n=\delta_{n,0}$.

In general, if $f(t)$ is a real-valued function then we have

$$\begin{aligned} f^*(t) = f(t) \quad &\Rightarrow \quad \sum_{n\in\mathbb{Z}} A_n^*\, e^{+2\pi int/T} = \sum_{n\in\mathbb{Z}} A_n\, e^{-2\pi int/T} \\ &\Rightarrow \quad \sum_{n\in\mathbb{Z}} A_{-n}^*\, e^{-2\pi int/T} = \sum_{n\in\mathbb{Z}} A_n\, e^{-2\pi int/T} \\ &\Rightarrow \quad \sum_{n\in\mathbb{Z}} (A_{-n}^* - A_n)\, e^{-2\pi int/T} = 0\ . \end{aligned} \tag{8.8}$$

Since the Fourier coefficients are unique (in this case for the simple function $f(t) = 0$), this implies the following symmetry for the coefficients,

$$f(t) \text{ real} \quad \Leftrightarrow \quad A_{-n} = A_n^* \quad \text{for all } n \in \mathbb{Z}\ . \tag{8.9}$$

In particular, $A_0^* = A_0 \in \mathbb{R}$. Further symmetries that can be derived in the same way are

$$\begin{aligned} f(t) \text{ even, i.e. } f(-t) = f(t) \quad &\Leftrightarrow \quad A_{-n} = A_n \quad \text{for all } n \in \mathbb{Z}\ , \\ f(t) \text{ odd, i.e. } f(-t) = -f(t) \quad &\Leftrightarrow \quad A_{-n} = -A_n \quad \text{for all } n \in \mathbb{Z}\ . \end{aligned} \tag{8.10}$$

In particular, for odd functions $A_0 = 0$ and if f is also real, the coefficients are purely imaginary ($\in i\mathbb{R}$) and for even functions they are real ($\in \mathbb{R}$). For real-valued functions, the Fourier series can also be expressed in terms of $\sin$ and $\cos$. Using (8.9) we have

$$\begin{aligned} f(t) &= \sum_{n\in\mathbb{Z}} A_n\, e^{2\pi int/T} = \sum_{n\in\mathbb{Z}} A_n \big(\cos(2\pi nt/T) + i\sin(2\pi nt/T)\big) \\ &= A_0 + \sum_{n\geq 1} \underbrace{(A_n + A_{-n})}_{=A_n + A_n^*} \cos(2\pi nt/T) + \sum_{n\geq 1} i\, \underbrace{(A_n - A_{-n})}_{=A_n - A_n^*} \sin(2\pi nt/T) \\ &= \underbrace{A_0}_{\in\mathbb{R}} + \sum_{n\geq 1} \underbrace{2\mathrm{Re}(A_n)}_{:=a_n\in\mathbb{R}} \cos(2\pi nt/T) + \sum_{n\geq 1} i\, \underbrace{2\mathrm{Im}(A_n)}_{:=b_n\in\mathbb{R}} \sin(2\pi nt/T)\ . \end{aligned} \tag{8.11}$$

So for even, real functions we have $b_n = 0$ and for odd, real functions $A_0, a_n = 0$.

Example. Consider the square wave $f(t) = \left\{ \begin{array}{rl} 1 & , \ -T/2 \leq t < 0 \\ -1 & , \ 0 \leq t < T/2 \end{array} \right.$.

This is an odd function so $A_0 = 0$, and for all $n \neq 0$ we have

$$\begin{aligned} A_n &= \frac{1}{T}\int_{-T/2}^{0} e^{-2\pi int/T}\, dt - \frac{1}{T}\int_0^{T/2} e^{-2\pi int/T}\, dt \\ &= -\frac{1}{2\pi in}\big(1 - e^{\pi in}\big) + \frac{1}{2\pi in}\big(e^{-\pi in} - 1\big) = \frac{i}{2\pi n}\big(2 - e^{\pi in} - e^{-\pi in}\big) \\ &= \frac{i}{\pi n}\big(1 - (-1)^n\big) = \left\{ \begin{array}{ll} 0 & , \ n \text{ even} \\ \frac{2i}{\pi n} & , \ n \text{ odd} \end{array} \right. . \end{aligned} \tag{8.12}$$

Thus $a_n = 0$ and $\quad b_n = 2\mathrm{Im}(A_n) = \left\{ \begin{array}{ll} 0 & , \ n \text{ even} \\ \frac{4}{\pi n} & , \ n \text{ odd} \end{array} \right.$ and we have

$$f(t) = \sum_{k=0}^{\infty} \frac{4}{(2k+1)\pi} \sin\Big(\frac{2\pi(2k+1)}{T}\, t\Big)\ . \tag{8.13}$$

A mathematical subtlety: Fourier series can be proven to converge pointwise at every ponit of continuity of f and to the average of the left and right limits at a point of discontinuity of f. However, convergence is not absolute, and if we define the partial sum

$$S_N(t) := \sum_{|n|\leq N} A_n \, e^{2\pi int/T} \tag{8.14}$$

for the square wave example above, the limits $N \to \infty$ and $t \to 0$ do not commute, since 0 is a point of discontinuity,

$$\lim_{t\to 0} \lim_{N\to\infty} S_N(t) \neq \lim_{N\to\infty} \lim_{t\to 0} S_N(t) = 0 \, . \tag{8.15}$$

This results in 'ringing artifacts' at points of discontinuity, which is also called **Gibbs phenomenon**.

Stationary sequences.
Let $(X_n : n \in \mathbb{Z})$ be a stationary, discrete-time process with autocorrelation function $R(n)$, $n \in \mathbb{Z}$. The corresponding Fourier series

$$D(\omega) := \frac{1}{2\pi} \sum_{n\in\mathbb{Z}} e^{i\omega n} R(n) \; , \quad \omega \in [-\pi, \pi) \tag{8.16}$$

is called **power spectral density**. Slightly different from above, this is defined as a normalized, 2π-periodic function on the interval $[-\pi, \pi)$ in terms of the angular frequency ω. The formula for Fourier coefficients gives the representation

$$R(n) = \int_{-\pi}^{\pi} D(\omega) \, e^{-i\omega n} \, d\omega \; . \tag{8.17}$$

$D(\omega)$ provides a spectral decomposition of the covariance structure of the process. The simplest example is an **iid sequence** with $R(n) = \delta_{n,0}$, which leads to

$$D(\omega) = \frac{1}{2\pi} \sum_{n\in\mathbb{Z}} e^{i\omega n} \delta_{n,0} = \frac{1}{2\pi} \; . \tag{8.18}$$

There is no particular structure in this process and the flat spectrum corresponds to a discrete version of white noise.
For a stationary, autoregressive **AR(1) model** $X_n = c + \phi X_{n-1} + \xi_n$ we have derived above that $R(n) = \phi^{|n|}$ for all $n \in \mathbb{Z}$. This leads to

$$\begin{aligned} D(\omega) &= \frac{1}{2\pi} \sum_{n\in\mathbb{Z}} e^{i\omega n} \phi^{|n|} = \frac{1}{2\pi} \Big(\sum_{n\geq 0} (e^{i\omega}\phi)^n + \sum_{n\geq 0} (e^{-i\omega}\phi)^n - 1 \Big) \\ &= \frac{1}{2\pi} \Big(\frac{1}{1 - e^{i\omega}\phi} + \frac{1}{1 - e^{-i\omega}\phi} - 1 \Big) = \frac{1}{2\pi} \Big(\frac{2 - 2\phi \cos\omega}{1 - 2\phi\cos\omega + \phi^2} - 1 \Big) \\ &= \frac{1}{2\pi} \frac{1 - \phi^2}{1 - 2\phi\cos\omega + \phi^2} \; . \end{aligned} \tag{8.19}$$

This is a symmetric/even function on $[-\pi, \pi)$ since $R(n) = R(-n)$, with a maximum $D(0) = \frac{1-\phi^2}{2\pi(1-\phi)^2}$, and reduces to the flat case for $\phi = 0$ as the AR(1) model is then iid noise.

8.2 Fourier transform and power spectra

The **Fourier transform** extends the notion of a Fourier series to non-periodic functions, by taking the limit $T \to \infty$. For each term in the series (8.2) we introduce the the **angular frequency** variable

$$\omega_n := \frac{2\pi n}{T} \quad \text{with spacings} \quad \Delta\omega = \omega_n - \omega_{n-1}\frac{2\pi}{T} \to 0 \quad \text{as } T \to \infty \; . \tag{8.20}$$

We rewrite the Fourier series as

$$f(t) = \frac{1}{2\pi} \sum_{n\in\mathbb{Z}} \frac{2\pi}{T} \hat{f}(\omega_n)\, e^{i\omega_n t} = \frac{1}{2\pi} \sum_{n\in\mathbb{Z}} \hat{f}(\omega_n)\, e^{i\omega_n t} \Delta\omega \; , \tag{8.21}$$

where

$$\hat{f}(\omega_n) = \int_{-T/2}^{T/2} f(t)\, e^{-i\omega_n t}\, dt \quad \in \mathbb{C} \; . \tag{8.22}$$

The series is a Riemann sum approximation to an integral over ω which becomes exact in the limit $T \to \infty$ and we get

$$\begin{aligned} f(t) &= \frac{1}{2\pi} \int_{\mathbb{R}} \hat{f}(\omega)\, e^{i\omega t} \quad \text{and} \\ \hat{f}(\omega) &= \int_{\mathbb{R}} f(t)\, e^{-i\omega t}\, dt \; . \end{aligned} \tag{8.23}$$

$\hat{f}$ is the **Fourier transform** of f and the first line is the **inverse Fourier transform**. Note that one could use different conventions for the normalization,

$$\begin{aligned} f(t) &= \sqrt{\frac{|b|}{(2\pi)^{1+a}}} \int_{\mathbb{R}} \hat{f}(\omega)\, e^{ib\omega t} \quad \text{and} \\ \hat{f}(\omega) &= \sqrt{\frac{|b|}{(2\pi)^{1-a}}} \int_{\mathbb{R}} f(t)\, e^{-bi\omega t}\, dt \end{aligned} \tag{8.24}$$

are equally good for any choice of constants $a, b \in \mathbb{R}$, and we use $a = 1$, $b = 1$. The Fourier transform allows us to decompose any function or signal into its constituent frequencies. The quantity

$$D(\omega) := \hat{f}(\omega)\, \hat{f}^*(\omega) \quad \in \mathbb{R} \tag{8.25}$$

is called the **power spectrum** of $f(t)$ and is the squared amplitude (energy) of the frequency ω.

Properties of Fourier transforms.

- **Translation.** If $h(t) := f(t + \tau)$ we have

  $$\frac{1}{2\pi} \int_{\mathbb{R}} \hat{h}(\omega)\, e^{i\omega t}\, d\omega = \frac{1}{2\pi} \int_{\mathbb{R}} \hat{f}(\omega)\, e^{i\omega\tau}\, e^{i\omega t}\, d\omega \quad \Rightarrow \quad \hat{h}(\omega) = e^{i\omega\tau}\, \hat{f}(\omega) \; . \tag{8.26}$$

  So a shift in the time domain corresponds to a multiplicative phase factor for the Fourier transform.
  Note that for a T-periodic function f this implies

  $$\hat{f}(\omega) = e^{i\omega T}\, \hat{f}(\omega) \; . \tag{8.27}$$

This can only hold if $\hat{f}(\omega) = 0$, or

$$e^{i\omega T} = 1 \quad \Rightarrow \quad \omega = 2\pi n/T \quad \text{with } n \in \mathbb{Z}\,. \tag{8.28}$$

If A_n are the coefficients of the Fourier series for the periodic function f, the Fourier transform is then given by

$$\hat{f}(\omega) = \sum_{n\in\mathbb{Z}} A_n \delta(\omega - 2\pi n/T) \tag{8.29}$$

which is consistent with (8.2) and is basically the Fourier series coefficients embedded in the real line.

- **Convolution theorem.** Consider smoothing of f by a kernel K,

$$\begin{aligned}
(f * k)(t) \;&:=\; \int_{\mathbb{R}} ds\, f(t-s)\, k(s) \\
&= \int_{\mathbb{R}} ds \frac{1}{2\pi} \int_{\mathbb{R}} d\omega_1\, \hat{f}(\omega_1) e^{i\omega_1 (t-s)} \frac{1}{2\pi} \int_{\mathbb{R}} d\omega_2\, \hat{K}(\omega_2) e^{i\omega_2 s} \\
&= \frac{1}{2\pi} \int_{\mathbb{R}} \int_{\mathbb{R}} d\omega_1 d\omega_2\, \hat{f}(\omega_1) \hat{K}(\omega_2) e^{i\omega_1 t} \frac{1}{2\pi} \int_{\mathbb{R}} ds\, e^{i(\omega_2 - \omega_1)s} \\
&= \frac{1}{2\pi} \int_{\mathbb{R}} \int_{\mathbb{R}} d\omega_1 d\omega_2\, \hat{f}(\omega_1) \hat{K}(\omega_2) \delta(\omega_2 - \omega_1)\, e^{-i\omega_1 t} \\
&= \frac{1}{2\pi} \int_{\mathbb{R}} d\omega_1\, \hat{f}(\omega_1) \hat{K}(\omega_1)\, e^{i\omega_1 t} = \mathcal{F}^{-1}\big[\mathcal{F}[f]\, \mathcal{F}[K]\big]
\end{aligned} \tag{8.30}$$

This is useful, since Fourier transforms (denoted by $\mathcal{F}$) can be implemented more efficiently $O(N \ln N)$ than convolution products which take $O(N^2)$ steps (see also below).

- **Fourier transform of a Gaussian.** Let $f(t) = e^{-at^2}$. Then

$$\begin{aligned}
\hat{f}(\omega) \;&=\; \int_{\mathbb{R}} e^{-at^2} e^{-i\omega t} dt = \int_{\mathbb{R}} \exp\Big[-a\Big(t^2 + \frac{i\omega}{a} t - (\frac{i\omega}{2a})^2 + (\frac{i\omega}{2a})^2\Big)\Big] dt \\
&= e^{-a\frac{\omega^2}{4a^2}} \int_{\mathbb{R}} e^{-a\left(t + \frac{i\omega}{a}\right)^2} dt = \frac{1}{\sqrt{a}} e^{-\frac{\omega^2}{4a}} \int_{\mathbb{R}} e^{z^2} dz = \sqrt{\frac{\pi}{a}} e^{-\frac{\omega^2}{4a}}\;,
\end{aligned} \tag{8.31}$$

so the Gaussian is invariant under Fourier transformation. Note that if a is large, i.e. the pulse narrow/localized in t-space, $1/a$ is small, i.e. it is broad in ω-space. This is a general property of Fourier representations which is related to the uncertainty principle in quantum mechanics.

Let $(X_t : t \in \mathbb{R})$ be a continuous-time weakly stationary process with positive variance. Then, if it is continuous at $\tau = 0$, the autocorrelation function

$$R(\tau) = (\mathcal{F}[D])(\tau) = \int_{\mathbb{R}} e^{i\omega\tau}\, D(\omega)\, d\omega \tag{8.32}$$

can be written as the Fourier transform of the **power spectral density**

$$D(\omega) := (\mathcal{F}[R])(\omega) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-i\omega\tau}\, R(\tau)\, d\tau\;. \tag{8.33}$$

Note that $R(\tau)$ is a real and even function, so that the same holds for $D(\omega)$. $D(\omega)$ is not necessarily a function, and can contain atomic parts in terms of δ functions. In general we have for some index set I,

$$D(\omega) = \sum_{k \in I} \delta_{\lambda_k}(\omega) + \text{ smooth part } , \tag{8.34}$$

and the set $\{\lambda_k : k \in I\}$ is called the **point spectrum** of the process.
The simplest example is **white noise**, a Gaussian process with mean 0 and (degenerate) covariance $\mathrm{Cov}(X_t, X_s) = \delta(t-s)$. Then we get the smooth spectral density

$$R(t) = \delta(t) \quad \text{and} \quad D(\omega) = \frac{1}{2\pi} \int_{\mathbb{R}} \delta(t) e^{i\omega t} dt = \frac{1}{2\pi} \,, \tag{8.35}$$

analogous to the iid discrite-time noise process.

8.3 Discrete Fourier transform

Consider a timeseries or a discrete function $\{(t_i, f_i) : i = 1, \ldots, N\}$, with evenly spaced sampling points $t_i = iT/N$ over the time interval $(0, T]$, where we take N even for simplicity. The necessarily finite range of observation and the finite sampling rate lead to two effects for the Fourier transform, which are a-priori defined only for functions $f(t),\ T \in \mathbb{R}$.

- **Discrete spectrum.** The usual approach is to simply assume the function extends periodically to $\pm\infty$, which leads to a discrete Fourier transform as explained above with $\Delta\omega = 2\pi/T$. The angular frequencies are therefore

$$\omega_n = n\Delta\omega = 2\pi n/T \,, \quad n \in \mathbb{Z} \,. \tag{8.36}$$

- **Aliasing.** Due to the finite sampling rate with $\Delta t = T/N$, the shortest period that can be detected from the timeseries is $2\Delta t$, which corresponds to the so-called

$$\textbf{Nyquist frequency} \quad \omega_{N/2} = 2\pi N/(2T) = \pi N/T \,. \tag{8.37}$$

Higher frequencies ω in the Fourier spectrum are indistinguishable from lower ones such as $\omega/2$ from the timeseries, and therefore one usually considers the spectrum only for

$$\omega_n \quad \text{with} \quad n \in \{-N/2+1, \ldots, N/2\} \,. \tag{8.38}$$

We represent the timeseries as a function $f : \mathbb{R} \to \mathbb{R}$ as

$$f(t) = \Delta t \sum_{k=1}^{N} f_k \, \delta_{t_k}(t) \,. \tag{8.39}$$

This is more convenient than a piecewise constant interpolation, and the normalizing factor $\Delta t = T/N$ is chosen so that the integral over the function has the same value. Now we can simply use the usual formula for the Fourer transform (8.23) to get

$$\hat{f}(\omega_n) = \Delta t \int_{\mathbb{R}} dt \sum_{k=1}^{N} f_k \, \delta_{t_k}(t) \, e^{-i\omega_n t} = \Delta t \sum_{k=1}^{N} f_k \, e^{-i\omega_n t_k} = \Delta t \hat{f}_n \tag{8.40}$$

where $\quad \hat{f}_n := \sum_{k=1}^{N} f_k \, e^{-2\pi i n k/N} \ , \quad n \in \{-N/2+1, \ldots, N/2\}$

is called the **discrete Fourier transform** of the timeseries. The normalization is usually not included in implementations, but sometimes is ($\rightarrow$ important to check the documentation!). Remember that for real f_k we know that $\hat{f}_{-n} = f_n^*$. For the inverse formula we get

$$f_k = \sum_{n=-N/2+1}^{N/2} \hat{f}_n \, e^{2\pi i n k/N} \ . \tag{8.41}$$

The direct computation time for the discrete FT is $O(N^2)$, but **Fast Fourier Transform (FFT)** developed by Cooley and Tukey in 1965 provides an $O(N \log N)$ algorithm.

The discrete FT of a real-valued timeseries provides an estimate for the power spectral density

$$D(\omega_n) = \hat{f}_n^* \hat{f}_n = |f_n|^2 \ , \quad n \in \{0, \ldots, \mathbb{N}/2\} \ , \tag{8.42}$$

which is non-negative and symmetric, and therefore usually only considered for non-negative ω_n. There are several methods for stationary timeseries $\mathbf{X} = \{(t_i, X_i) : i = 1, \ldots, N\}$ based on different ways to extend the series/signal to infinity.

- **Periodogram.** This is based on a strictly periodic extension of the timeseries as used above, and since the latter can contain a significant amount of noise the periodogram is usually rather noisy itself. It is most appropriate for signals which are deterministic or have little noise (such as measurement errors).
(Implemented in MATLAB as `periodogram`.)

- **Autoregressive power spectral density estimate.** This is most appropriate for a noisy timeseries which can be fitted well to an autoregressive model, and provides the FT of the autocorrelation function $R(n)$ which is a symmetric, real-valued function. Modulo normalization, this is given in (8.16), and provides a smoother version of the spectral density as the periodogram.
(Implemented in MATLAB as `pmcov`, needs parameter q for the AR(q) model.)

Note that the timeseries itself can be seen as a noisy, unnormalized version of the autocorrelation function and therefore the periodogram is a noisy version of the autoregressive estimate of the spectral density.

The discrete FT is also often used as a fast method for smoothing and filtering timeseries. This is usually done by convolution products (cf. kernel density estimates discussed before), which require $O(N^2)$ operations. Using the convolution formula for Fourier transforms (8.30) and FFT this can be done in $O(N \log N)$.

Possible viva questions

1. Basic probability

- What is a probability distribution and a random variable?
  Explain mentioning concepts of state space, outcome and event.
- What is independence of random variables, and how is it related to correlations?
- Define expectation, variance and standard deviation, CDF, TDF, PMF, median, quantiles.
  Explain PDF, CDF and expectation for continuous rv's through integrals.
- Give state space (support) and PDF and/or CDF and/or tail of the
  uniform, Bernoulli, binomial, geometric, Poisson, exponential and Gaussian distribution.
  Be able to compute (or know) mean and variance, give typical examples where distributions show up and how they are related, including Poisson as scaling limit of binomial, exponential from geometric.

2. Less basic probability

- Define heavy tail, Pareto distribution and characterize Lévy distribution.
- Scaling properties of Gaussians, exponentials and Pareto variables.
- Define characteristic functions and state their basic properties.
- State the weak LLN and the CLT (with assumptions), also the generalized version for heavy tails.
  If you want $80+$, be able to prove Gaussian case using characteristic functions.
- State the extreme value theorem, define the 3 types of extreme value distributions by their CDF and for which tails they apply.
- Explain how to compute typical values and fluctuations of the maximum of iidrv's.

3. Joint distributions

- Define the joint PMF, marginal and conditional probability, give sum rule and product rule.
  Corresponding versions for continuous rv's, with special care for conditional probabilities.
- give PDF of multivariate Gaussian, explain mean, covariance matrix, correlations and independence.
- Definition and interpretation of the concentration/precision matrix, and correlation coefficient.
- Give Bayes' rule, prior, posterior and likelihood. Be ready to do an example.
  Explain the problem of false positives when testing for rare events, related choice of prior.

4. Basic statistics

- Give the definition of sample mean, variance, order statistics and quantiles.
- Give the definition of empirical density, CDF and tail, histogram and kernel density estimate.
- Empirical distribution as simplest non-parametric model, explain bootstrap.
- For parametric models explain likelihood, log-likelihood, MLE and be ready to do an example computation.
- Explain bias and consistency, compute for simple examples.
  Explain unbiased variance estimator and degrees of freedom.
- Define standard error, and explain confidence intervals base on the Gaussian distribution.

5. Time series

- Give standard models for timeseries data: signal plus noise, MA(q), AR(q), Markov process.
- Define stationarity and weak stationarity, how are the two related?
  Define cross correlation and auto correlation function. Be ready to compute it for examples (iid noise, AR or AM models).
- Give an estimator for the auto correlation function for single or multiple datasets, explain the difference.
- Define a Gaussian process, and give a simple example (e.g. white noise).

6. Linear regression

- Write down the basic model for linear regression, define LSE and LS error function.
  How is this related to the MLE?
- Define the design matrix and be ready to show (or know) how the LSE can be written in terms of the data X. What is the Moore-Penrose pseudo inverse?
- Define the RSS and TSS and the R^2 coefficient of determination. How can this be interpreted to measure the goodness of fit?
- Explain the problem of overfitting and standard approaches to model selection: cross validation, regularized LS regression, adjusted R^2 coefficient.
- Explain how to detrend data using differencing and regression.
  Explain how to detect a periodic signal in noise.

7. Autoregressive models

- Compute the mean of a stationary AR(q) model and explain how to compute the covariances and auto correlation. Be able to do an explicit computation for AR(1).

- Explain how to write the stationary solution of an AR(1) model in terms of shift operators and a series expansion over the noise.
- Compute mean and variance of an AR(1) model and be able to write the most important terms of the log likelihood based on the recursion.
  Give the LS error function for fixed initial value and show how to derive the MLE using the design matrix. Explain the last part for the AR(2) model.

8. Spectral analysis

- Define the Fourier series for periodic functions and use orthogonality of basis functions to show inverse formula. Be ready to compute a simple example.
- Explain the symmetry relations of the coefficients for real/even and odd functions.
- Explain the Gibbs phenomenon and how it is related to convergence of Fourier series.
- Define the power spectral density for general stationary processes, and compute it for an AR(1) model or another model with given auto correlation function.
- Explain how to derive the Fourer transform in the limit of period $T \to \infty$.
  What is the power spectrum of a function f?
- State basic properties of FTs (translation, convolution theorem, FT of Gaussian and uncertainty)
- Explain the two basic issues of discrete FT: finite range and periodic extension leads to discrete spectrum, finite sampling rate leads to aliasing and Nyquist frequency
- Give formulas for discrete FT and estimator for the power spectral density. Explain the difference between a periodogram and an AR power spectral density estimate.
- Explain how to use spectral analysis to extract a periodic signal from noise.

General questions.

- How would you go about systematically analyzing a timeseries?
  What preprocessing is necessary to do what?
- How can you check if your data are iid or a timeseries?
- What is a scatter plot, box plot?
- How can you plot distributions most appropriately (log/lin etc)?
- Explain practical problems that can occur in model selection and how to use common sense (e.g. set small parameter values to 0, when is that justified?)