Statistics for Data Science

Wintersemester 2024/25

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Twelfth lecture, January 13, 2025

Frequentist methods

The statistical methods that we have discussed so far are known as frequentist (or classical) methods. The frequentist point of view is based on the following postulates:

- F1 Probability refers to limiting relative frequencies. Probabilities are objective properties of the real world.
- F2 Parameters are fixed, unknown constants. Because they are not fluctuating, no useful probability statements can be made about parameters.
- F3 Statistical procedures should be designed to have well-defined long run frequency properties. For example, a 95 percent confidence interval should contain the true value of the parameter with limiting frequency at least 95 percent.

Bayesian methods

There is another approach to inference called Bayesian inference. The Bayesian approach is based on the following postulates:

- B1 Probability describes degree of belief, not limiting frequency. As such, we can make probability statements about lots of things, not just data which are subject to random variation. For example, "the probability that Albert Einstein drank a cup of tea on August 1, 1948" is 0.35. This does not refer to any limiting frequency; it reflects a subjective strength of belief that the proposition is true.
- B2 The parameters are modeled as random variables, not as fixed, unknown constants. We can make probability statements about the parameters.
- B3 We can make inferences about a parameter x by producing a probability distribution for x. Inferences, such as point estimates and interval estimates, may then be extracted from this distribution.

Bayesian inference embraces a subjective notion of probability. In general, Bayesian methods provide no guarantees on long run performance.

Notation / recap on conditional and marginal PDFs

Let x and y be random variables with values in $\mathbb{R}^d$ and $\mathbb{R}^k$, respectively. If the random variable (x, y) has a probability density $f_{x,y}$, i.e., if

$$\mathbb{P}(x \in A, y \in B) = \mathbb{P}((x, y) \in A \times B) = \int_{A \times B} f_{x,y}(u, v) \, \mathrm{d}u \, \mathrm{d}v$$

for all events $A \subset \mathbb{R}^d$ and $B \subset \mathbb{R}^k$, then $f_{x,y}$ is called the *joint probability density* of x and y. Here, $\mathbb{P}(x \in A, y \in B) := \mathbb{P}(x \in A \text{ and } y \in B)$. To simplify notation, we also write $f(x, y) = f_{x,y}(x, y)$.

Now, the *marginal probability density* f_x of x is defined by

$$f_x(u) = \int_{\mathbb{R}^k} f_{x,y}(u, v) \, \mathrm{d}v \quad \text{for all } u \in \mathbb{R}^d.$$

Analogously, the marginal density of y is

$$f_y(v) = \int_{\mathbb{R}^d} f_{x,y}(u, v) \, \mathrm{d}u \quad \text{for all } v \in \mathbb{R}^k.$$

The marginal density of x is indeed the probability density for x in the situation where we have no information about the random variable y, because

$$\begin{aligned}\mathbb{P}(x \in A) = \mathbb{P}(x \in A, y \in \mathbb{R}^k) &= \int_{A \times \mathbb{R}^k} f_{x,y}(u,v)\,\mathrm{d}u\,\mathrm{d}v \\ &= \int_A \left(\int_{\mathbb{R}^k} f_{x,y}(u,v)\,\mathrm{d}v \right) \mathrm{d}u = \int_A f_x(u)\,\mathrm{d}u\end{aligned}$$

for every event $A \subset \mathbb{R}^d$.

The random variables x and y are independent (denoted by $x \perp y$) if

$$\mathbb{P}(x \in A, y \in B) = \mathbb{P}(x \in A)\mathbb{P}(y \in B)$$

for all events $A \subset \mathbb{R}^d$ and $B \subset \mathbb{R}^k$ or, equivalently, if

$$f_{x,y}(u,v) = f_x(u) f_y(v) \quad \text{for all } u \in \mathbb{R}^d,\ v \in \mathbb{R}^k.$$

To simplify notation, we will also write $f(x) := f_x(x)$ and $f(y) := f_y(y)$.

Next, we consider the random variable x in the opposite situation where we know everything about the random variable y: we have observed it and know what value it has taken.

We say we consider the random variable x, *given* that we know the value y_0 taken by y, and denote this by $x|y = y_0$. For $y_0 \in \mathbb{R}^k$ with $f_y(y_0) > 0$, the *conditional probability density* of $x|y = y_0$, $f_{x|y=y_0}$, is then defined by

$$f_{x|y=y_0}(u) = \frac{f_{x,y}(u, y_0)}{f_y(y_0)}.$$

If x and y are independent and $f_y(y_0) > 0$, then

$$f_{x|y=y_0}(u) = f_x(u).$$

To simplify notation, we will also write $f(x|y) := f_{x|y}(x) := f_{x|y=y}(x)$.

Bayesian inference

Bayesian inference is usually carried out in the following way.

1. We choose a probability density $f(x)$ – called the prior distribution – that expresses our beliefs about a parameter x before we see any data.
2. We choose a statistical model $f(y|x)$ that reflects our beliefs about y given x.
3. After observing data $y_1, \ldots, y_n$, we update our beliefs and calculate the posterior distribution $f(x|y_1, \ldots, y_n)$.

In what follows, we will consider continuous $\mathbb{R}^d$*-valued random variables.*

Bayes' formula

Let (x, y) be a random variable with joint density $f(x, y)$ on $\mathbb{R}^d \times \mathbb{R}^k$. If $f(y) > 0$, then the conditional probability density of x, given y, equals

$$f(x|y) = \frac{f(x, y)}{\int_{\mathbb{R}^d} f(x, y)\,\mathrm{d}x}.$$

On the other hand, the conditional probability density of y in case we know the value of the unknown x, is the likelihood function

$$f(y|x) = \frac{f(x, y)}{f(x)}, \quad \text{if } f(x) > 0.$$

Since $f(x, y) = f(y|x)f(x)$, this leads to Bayes' formula

$$f(x|y) = \frac{f(y|x)f(x)}{Z(y)}, \quad Z(y) := \int_{\mathbb{R}^d} f(y|x)f(x)\,\mathrm{d}x.$$

If we have n i.i.d. observations $y_1, \dots, y_n$, then we replace $f(y|x)$ with

$$f(y_1, \dots, y_n|x) = \prod_{i=1}^{n} f(y_i|x) = \mathcal{L}_n(x).$$

Bayes' formula presents a way to express the conditional probability density of x, given y, assuming that the conditional density of y, given x, and the marginal density of x are known.

Example

Consider the problem of estimating an unknown parameter $x \in \mathbb{R}^d$ from data $y \in \mathbb{R}^k$ that is connected to x via the model

$$y = F(x) + \varepsilon. \tag{1}$$

If

- A1 The noise ε has the probability density ν on $\mathbb{R}^k$;
- A2 The parameter x has the probability density f on $\mathbb{R}^d$;
- A3 The random variables x and ε are independent;

then the likelihood is

$$f(y|x) = \nu(y - F(x)).$$

This is because

$$\begin{aligned} f(y|x) = f_{y|x}(y) = f_{F(x)+\varepsilon|x}(y) = f_{\varepsilon|x}(y - F(x)) &= f_{\varepsilon}(y - F(x)) \\ &= \nu(y - F(x)) \end{aligned}$$

due to the assumptions $\varepsilon \perp x$ and $\varepsilon \sim \nu$.

Example

If assumptions A1–A3 hold and

$$Z(y) = \int_{\mathbb{R}^d} \nu(y - F(x)) f(x) \, \mathrm{d}x > 0,$$

then the posterior density corresponding to (1) is

$$f(x|y) = \frac{\nu(y - F(x)) f(x)}{Z(y)}.$$

Remarks.

- The condition that the marginal density $f(y)$ of the observed data y is positive means that the observed data is assumed to be consistent with the probabilistic assumptions A1–A3.
- An event cannot have positive probability under the posterior distribution if it does not have positive probability under the prior distribution.

Case study: source localization

Suppose that a particle with unit charge is located at some (unknown) point $x^* \in (0,1)$ and our goal is to locate it based on measurements of voltage at the interval end points $x = 0$ and $x = 1$. The mathematical model for the voltage at any point $x \in [0,1]$ is given by

$$y(x) = \frac{1}{|x^* - x|}.$$

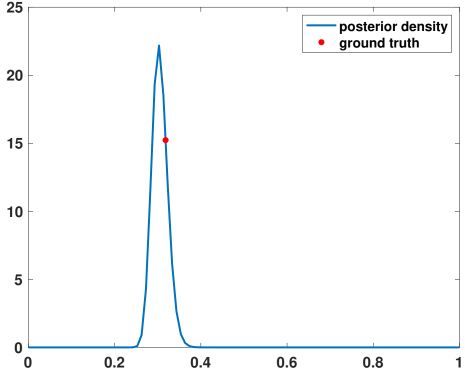
Our noisy measurements are modeled by $y_1 = \frac{1}{|x^*-0|} + \varepsilon_1$ and $y_2 = \frac{1}{|x^*-1|} + \varepsilon_2$, where ε_1 and ε_2 are i.i.d. realizations of $\mathcal{N}(0,\sigma^2)$. We take $x^* = 1/\pi$ (ground truth) and $\sigma = 0.2$ in the numerical experiments.

- The likelihood is given by $f(y|x) \propto \exp(-\frac{1}{2\sigma^2}\sum_{j=0}^{1}(y_{j+1} - \frac{1}{|x-j|})^2)$.
- We consider the prior $f(x) = \mathbf{1}_{(0,1)}(x) = \begin{cases} 1 & \text{if } x \in (0,1), \\ 0 & \text{otherwise.} \end{cases}$

Then the posterior density is given by Bayes' formula

$$f(x|y) \propto \mathbf{1}_{(0,1)}(x) \exp\left(-\frac{1}{2\sigma^2}\sum_{j=0}^{1}\left(y_{j+1} - \frac{1}{|x-j|}\right)^2\right).$$

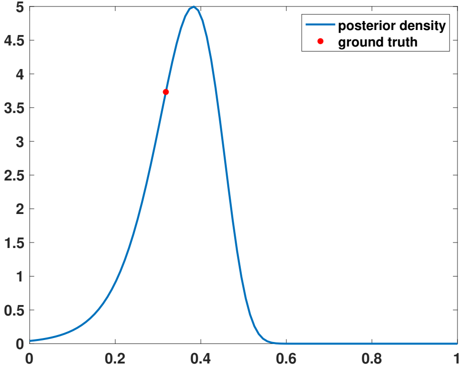
Let us visualize the posterior density against the ground truth solution. (See also the file `source.py` on the course website!)



We see that the posterior is localized around the true parameter value ("ground truth"). Note that in this case, the prior hardly plays any role.

We could take, e.g., the mean or mode of the posterior density as a point estimate for the unknown location of the point charge.

What if we modify the problem so that we have access to only one boundary measurement at $x = 1$?



The resulting posterior distribution carries substantially more uncertainty since we now have less measurement data!

Note that the posterior will generally be high-dimensional, meaning that it is usually not possible to visually inspect the posterior density.

Let $x \in \mathbb{R}^d, y \in \mathbb{R}^k$ be random variables (the unknown parameter and the measurement, respectively). Bayes' formula:

$$f(x|y) = \frac{f(y|x)f(x)}{Z(y)}, \quad Z(y) := \int_{\mathbb{R}^d} f(y|x)f(x)\,\mathrm{d}x > 0.$$

- The *prior model* $f(x)$ describes *a priori* information. It should assign high probability to objects x which are typical in light of *a priori* information, and low probability to unexpected x.
- The *likelihood model* $f(y|x)$ processes measurement information. It gives low probability to objects that produce simulated data which is very different from the measured data.
- The number $Z(y)$ can be treated as a normalization constant. It is often not of significant interest. If needed, we can recover it by computing the value of the integral $Z(y) = \int_{\mathbb{R}^d} f(y|x)f(x)\,\mathrm{d}x$.
- The *posterior distribution* $f(x|y)$ represents the updated knowledge about the parameter of interest x, given the evidence y.

Since the normalization constant $Z(y)$ is often not of interest, we write

$$f(x|y) \propto f(y|x)f(x),$$

where $\propto$ means equality up to a constant factor (not depending on x).

Bayesian estimators

The posterior distribution can be used to define estimators for the conditional random variable $x|y \sim f(x, y)$, where $y = (y_1, \ldots, y_n)$. In general, an estimator $\hat{x}$ is any function of the data y. The estimate $\hat{x} = \hat{x}(y)$ is itself an $\mathbb{R}^d$-valued random variable whose properties give information about the usefulness and quality of the estimator.

Bayesian estimators are those defined via the posterior distribution $f(x|y)$. We present the two most prominent ones. The conditional mean (CM) estimator is defined as the mean of the posterior distribution

$$\hat{x}_{\mathrm{CM}} = \mathbb{E}[x|y] = \int_{\mathbb{R}^d} x\, f(x|y)\, \mathrm{d}x$$

This is a high-dimensional integration problem.

The maximum a posteriori (MAP) estimator is defined as the mode

$$\hat{x}_{\mathrm{MAP}} = \operatorname*{arg\,max}_{x \in \mathbb{R}^d} f(x|y)$$

of the posterior distribution (if a unique mode exists). *This is a high-dimensional optimization problem.*

One way to estimate spread are Bayesian credible sets. A level $1-\alpha$ credible set $\mathcal{C}_\alpha$ with $\alpha \in (0,1)$ satisfies

$$\mathbb{P}(x \in \mathcal{C}_\alpha | y) = \int_{\mathcal{C}_\alpha} f(x|y)\,\mathrm{d}x = 1-\alpha.$$

For small α, it is a region that contains a large fraction of the posterior mass.

Example. Assume that $x \in \mathbb{R}$ and that the posterior density is given by

$$f(x|y) = \frac{c}{\sigma_1}\phi\left(\frac{x}{\sigma_1}\right) + \frac{1-c}{\sigma_2}\phi\left(\frac{x-1}{\sigma_2}\right),$$

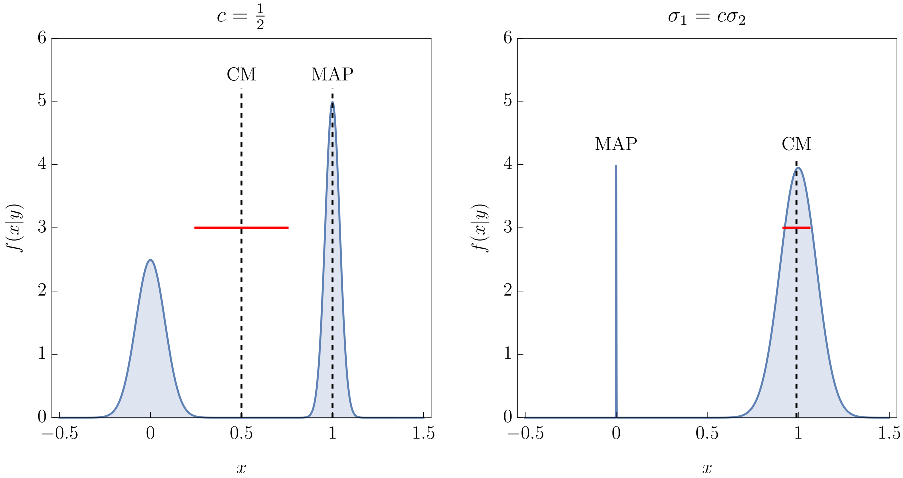
where $c \in (0,1)$, $\sigma_1, \sigma_2 > 0$, and ϕ is the density of the standard normal distribution, $\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$. In this case,

$$\hat{x}_{\mathrm{CM}} = 1 - c \quad \text{and} \quad \hat{x}_{\mathrm{MAP}} = \begin{cases} 0 & \text{if } c/\sigma_1 > (1-c)/\sigma_2, \\ 1 & \text{if } c/\sigma_1 < (1-c)/\sigma_2. \end{cases}$$

If $c = \frac{1}{2}$ and σ_1, σ_2 are small, the probability that x takes values near $\hat{x}_{\mathrm{CM}}$ is small. On the other hand, if $\sigma_1 = c\sigma_2$, then $c/\sigma_1 = 1/\sigma_2 > (1-c)/\sigma_2$, so that $\hat{x}_{\mathrm{MAP}} = 0$. If c is small, this is, however, a bad estimate for x, since the probability for x to take values near 0 is small. Last of all, we notice that when the conditional mean gives a poor estimate, this is reflected in a larger posterior variance

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \hat{x}_{\mathrm{CM}})^2 f(x|y)\,\mathrm{d}x.$$

We cannot say that one estimator is better than the other in all applications.



Left: the density with $\sigma_1 = 0.08$, $\sigma = 0.04$, and $c = \frac{1}{2}$. The CM estimate represents the distribution poorly. Notice that when the CM gives a poor estimate, this is reflected in wider variance (1 standard deviation is depicted as a red line). Right: the density with $\sigma_1 = 0.001$, $\sigma_2 = 0.1$, and $c = 0.01$. The MAP gives a poor estimate since it is in an unlikely part of the computational domain.

The maximum likelihood estimate

$$\hat{x}_{\mathrm{ML}} = \arg\max_{x \in \mathbb{R}^d} f(y|x)$$

answers the question: “which value of the unknown is the most likely to produce the measured data?”

The ML estimate is a non-Bayesian estimate, and if the sample size is not large, it is not considered very useful by Bayesian statisticians.

Prior modeling

The prior density should reflect our beliefs on the unknown variable of interest before taking the measurements into account.

Often, the prior knowledge is qualitative in nature, and transferring the information into quantitative form expressed through a prior density can be challenging.

The prior probability distribution should be concentrated on those values of x we expect to see and assign a clearly higher probability to them than to the unexpected ones.

Gaussian priors

Gaussian densities

$$f(x) = \frac{1}{(2\pi)^{d/2}\sqrt{\det C}} \exp\left(-\frac{1}{2}(x-m)C^{-1}(x-m)\right)$$

are easy to construct and form a versatile class of distributions. They also often lead to explicit estimators.

Random samples from a standard normal distribution $\mathcal{N}(0, I)$ can be generated directly, for example via `numpy.random.normal` in Python. Samples from a general normal distribution $\mathcal{N}(m, C)$ and from a wide class of other distributions can then be derived from those, so that it is often not necessary to employ the inverse transform method.

Case study: signal recovery

Suppose that we want to estimate a one-dimensional signal $g\colon [0,1] \to \mathbb{R}$ from indirect observations. We discretize the interval $[0,1]$ by points $t_j = j/d$, $j \in \{1,\ldots,d\}$, and write $x_j = g(t_j)$. In what follows, we consider some priors we could place for the unknown signal $x \in \mathbb{R}^d$.

Gaussian priors with covariance $C = \alpha^2 I$, $\alpha > 0$, are often called (Gaussian) white noise priors. The variance α^2 controls the magnitude of the realizations, but the values at $t_1, \ldots, t_d$ are independent.

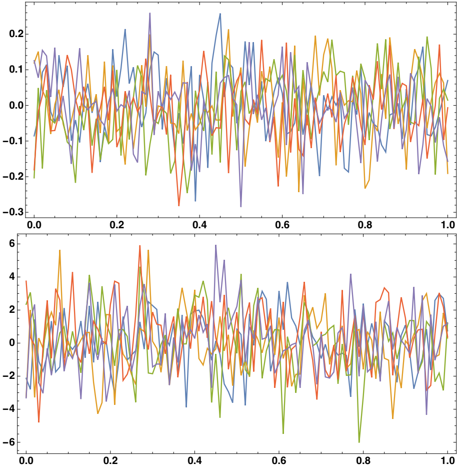


Figure: Top: 5 realizations of the Gaussian white noise prior with $\alpha = 0.1$. Bottom: 5 realizations of the Gaussian white noise prior with $\alpha = 2$.

Gaussian priors with covariance $C = \alpha^2(L^{\mathrm{T}}L)^{-1}$, where $\alpha > 0$ and $L = \mathrm{tridiag}(-1, 2, -1)$ (we will discuss the construction of this prior next week), are often called (Gaussian) smoothness priors. The parameter α^2 controls the variability of the realizations. Note that this prior enforces the Dirichlet boundary condition $g(0) = g(1) = 0$.

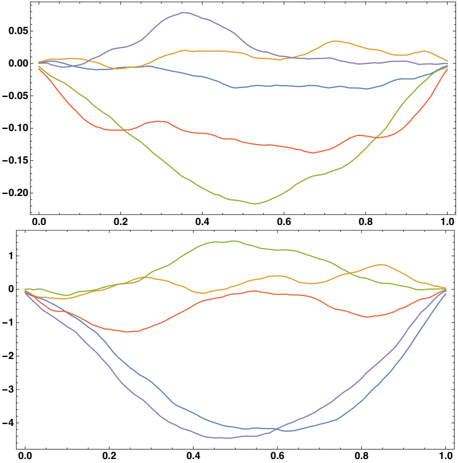


Figure: Top: 5 realizations of the Gaussian smoothness prior with $\alpha = 0.001$. Bottom: 5 realizations of the Gaussian smoothness prior with $\alpha = 0.02$.

Impulse priors

Assume that our prior information is that the signal contains small and well localized features in an almost constant background.

In such a case we could assume an impulse prior density, which means that it gives a low average amplitude but allows outliers. The tail of such a prior distribution is long, although the expected value is small.

Let $x \in \mathbb{R}^d$ represent the signal, where the component $x_j = f(t_j)$ is the values at the j^{th} coordinate. In what follows, x_i and x_j are assumed to be independent for $i \neq j$.

One example of an impulse prior is the ℓ^1 prior. It has the density

$$f(x) = \left(\frac{\alpha}{2}\right)^d \exp(-\alpha\|x\|_1)$$

with $\alpha > 0$, where the ℓ^1-norm is defined as

$$\|x\|_1 = \sum_{j=1}^{d} |x_j|.$$

The impulse effect can be enhanced by choosing an even smaller power $p \in (0,1)$ of the components of x, that is, using $\sum_{j=1}^{d} |x_j|^p$ instead of the ℓ^1-norm.

Another choice that produces images with few distinctly different function values and a low-amplitude background is the Cauchy density

$$f(x) = \left(\frac{\alpha}{\pi}\right)^d \prod_{j=1}^{d} \frac{1}{1+\alpha^2 x_j^2}$$

with $\alpha > 0$.

Since we assumed that each component is independent of the others, random draws can be performed componentwise using, e.g., inverse transform sampling.

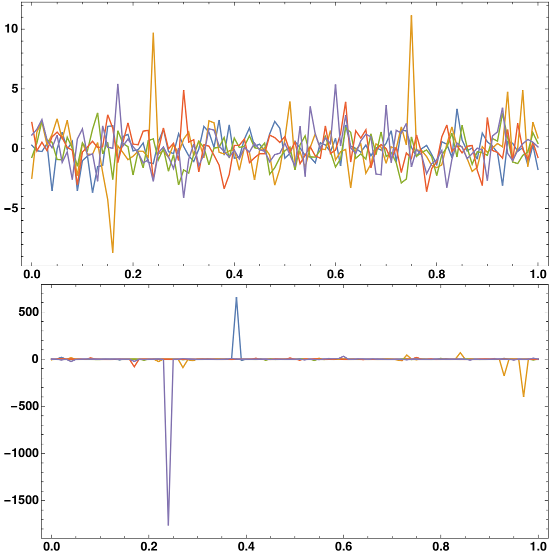


Figure: Top: 5 realizations of the ℓ^1 prior with $\alpha = 1$. Bottom: 5 realizations of the Cauchy prior with $\alpha = 1$.

Discontinuities

Assume still that we want to estimate a one-dimensional signal $g\colon [0,1] \to \mathbb{R}$ with $g(0) = 0$ from indirect observations. Our prior knowledge is that the signal is usually relatively stable but can have large jumps every now and then. We may also have information on the size of the jumps or the rate of their occurrence.

We obtain one possible prior by taking the finite difference approximation of the derivative of g and assigning an impulsive noise distribution to it. Let us discretize the interval $[0,1]$ by points $t_j = j/d$ and write $x_j = g(t_j)$. Consider the density

$$f(x) = \left(\frac{\alpha}{\pi}\right)^d \prod_{j=1}^{d} \frac{1}{1+\alpha^2(x_j - x_{j-1})^2}. \tag{2}$$

To draw samples from the above distribution we define new random variables for the jumps

$$u_j = x_j - x_{j-1}, \quad j = 1, \ldots, d.$$

These each have the density

$$f(u) = \left(\frac{\alpha}{\pi}\right)^d \prod_{j=1}^{d} \frac{1}{1+\alpha^2 u_j^2}.$$

In particular, the u_j are independent from each other, so that they can be drawn from a one-dimensional Cauchy density. Also note that $x = (x_1, \ldots, x_d)^{\mathrm{T}} \in \mathbb{R}^d$ satisfies $x = Lu$, where $L \in \mathbb{R}^{d\times d}$ is a lower triangular matrix with $L_{ij} = 1$ for $i \geq j$.[†] Generalizing the idea behind the above prior leads, e.g., to total variation priors.

[†] Note that in Python, it is more efficient to implement this as `x = numpy.cumsum(u)`.

Example: drawing realizations from the prior (2)

```
import numpy as np
import matplotlib.pyplot as plt

d = 1200
t = np.arange(1,d+1)/d
alpha = 1
quantile = lambda t: 1/alpha * np.tan(np.pi * (t-1/2))
unif = np.random.uniform(size=d)
draw = quantile(unif)
y = np.cumsum(draw)
plt.plot(t,y)
plt.xlabel('$t$',fontsize=14)
plt.ylabel('$g(t)$',fontsize=14)
plt.show()
```

Example: drawing realizations from the prior (2)

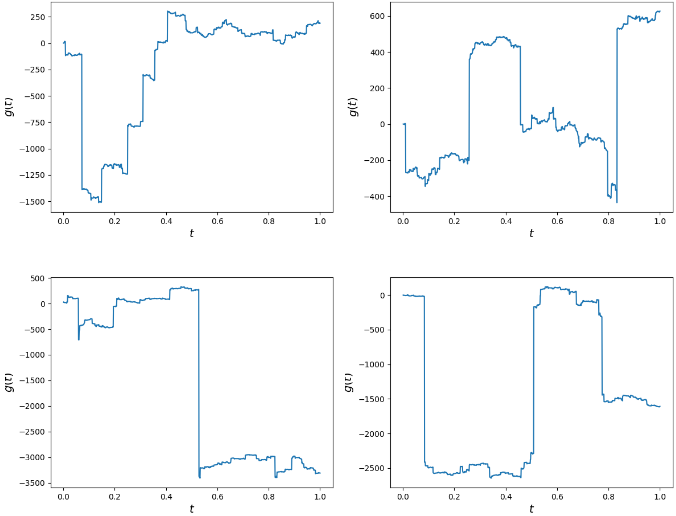


Figure: Four realizations drawn from the prior (2)

Hierarchical models

The prior density may depend on some parameter, such as variance or mean. So far we have assumed that these parameters are known. However, we often do not know how to choose them. If a parameter is not known, it can be estimated as a part of the statistical inference problem on the data. This leads to hierarchical models that include hypermodels for the parameters defining the prior density.

Assume that the prior distribution depends on a parameter α, which is assumed to be unknown. We then write the prior as a conditional density

$$f(x|\alpha).$$

We model the unknown α with a hyperprior $f_{\mathrm{h}}(\alpha)$ and write the joint distribution of x and α as

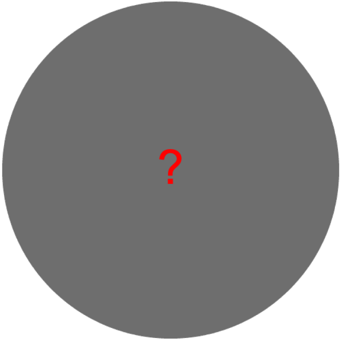
$$f(x,\alpha) = f(x|\alpha)f_{\mathrm{h}}(\alpha).$$

Assuming we have a likelihood model $f(y|x)$ for the measurement y, we get the posterior density for x and α, given y, using Bayes' formula

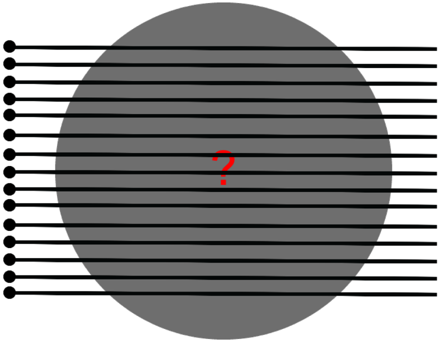
$$f(x,\alpha|y) \propto f(y|x,\alpha)f(x,\alpha) = f(y|x,\alpha)f(x|\alpha)f_{\mathrm{h}}(\alpha).$$

The hyperprior density f_{h} may again depend on some hyperparameter α_0. The main reason for the use of a hyperprior model is that the construction of the posterior is considered to be more robust with respect to fixing a value for the hyperparameter α_0 than fixing a value for α.

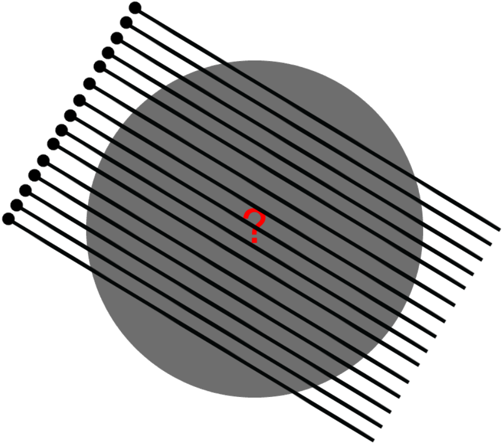
Case study: parallel-beam X-ray tomography



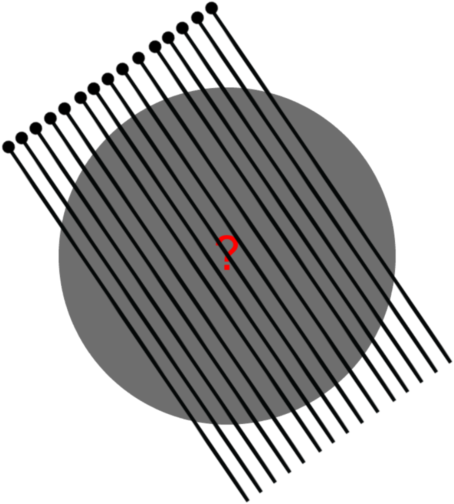
Case study: parallel-beam X-ray tomography



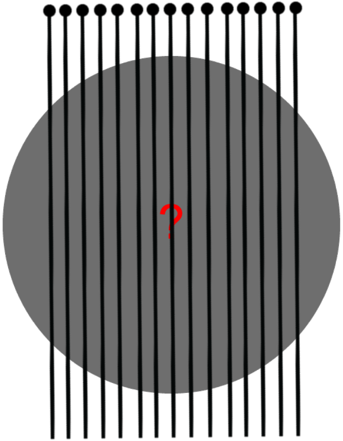
Case study: parallel-beam X-ray tomography



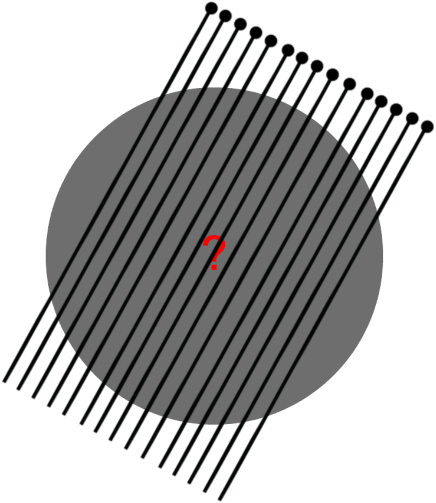
Case study: parallel-beam X-ray tomography



Case study: parallel-beam X-ray tomography

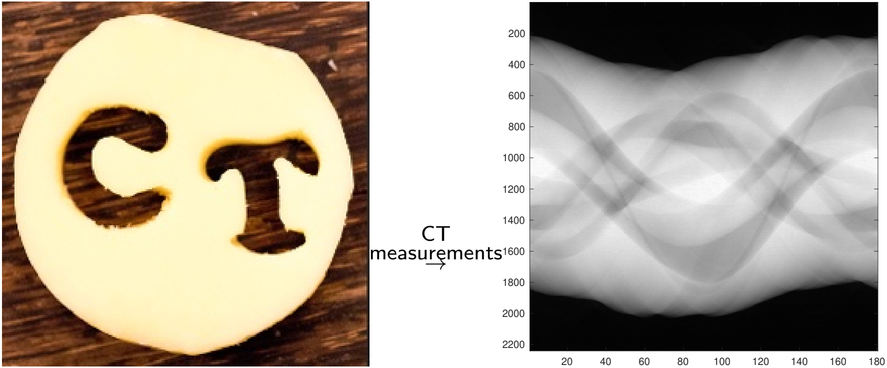


Case study: parallel-beam X-ray tomography



An object (left) is illuminated using a beam front of X-rays and the intensities of the X-rays are recorded after passing through the object. The measurements can be represented as a sinogram (right). In this case, the beam front consists of 2 240 parallel X-rays (arranged as rows) taken at 180 equally spaced angles at 1° increments (arranged as columns).

The goal is to reconstruct the interior density of the object based on the sinogram measurements.



Formation of a CT sinogram (by Samuli Siltanen): `https://www.youtube.com/watch?v=q7Rt_OY_7tU`

Let us consider the inference problem of recovering the attenuation coefficient (density) of an object given a set of X-ray measurements. The mathematical model can be expressed as

$$y = Ax,$$

where $y \in \mathbb{R}^Q$ denotes the (noisy) measurements for Q X-rays, $A \in \mathbb{R}^{Q \times n^2}$ is the projection matrix subject to an $n \times n$ pixel discretization of the computational domain, and $x \in \mathbb{R}^{n^2}$ denotes the (piecewise constant) discretization of the unknown attenuation inside the object of interest.

The data y can be reshaped into an $n \times n$ array, which is a graphical representation of the X-ray measurements (sinogram). The unknown can be reshaped into an $n \times n$ image of the density of the imaged object.

We use the FIPS open dataset of carved cheese available at
https://doi.org/10.5281/zenodo.1254210

The files `DataFull_128x15.mat` and `DataLimited_128x15.mat` contain sparse angle and limited angle tomography measurements, respectively. The data has been collected using 15 projections spanning either the full 360° circle in the first dataset, and 15 projections spanning a limited 90° angle of view in the second dataset. The computational domain is a 128×128 pixel grid in both cases. Each file contains a projection matrix `A` and a sinogram measurement matrix `m`.

By defining `y = m.reshape((m.size,1))`, our naïve maximum likelihood (ML) reconstruction of the unknown `x` is precisely the least squares solution of the problem

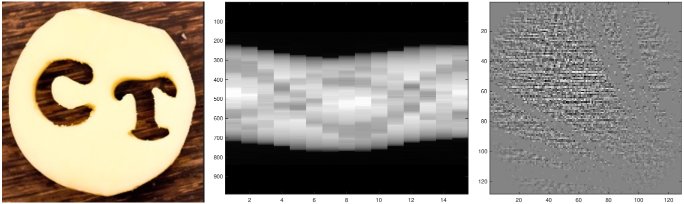
$$\mathtt{y} = \mathtt{Ax}.$$

The reconstruction is the image `x.reshape((128,128))`.

In addition, we also consider the MAP estimators of the unknown `x` corresponding to a Gaussian white noise prior and a total variation prior (which gives a high probability to piecewise constant signals).

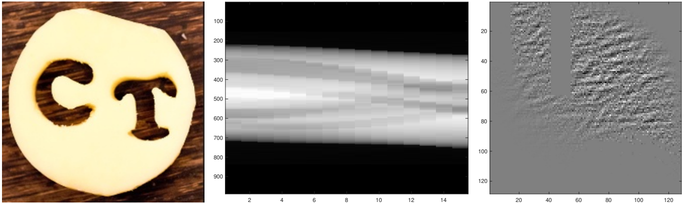
Maximum likelihood (ML) estimator

Sparse angle tomography data:



Left: the actual object. Middle: sinogram data for sparse angle tomography. Right: ML estimator of the unknown density.

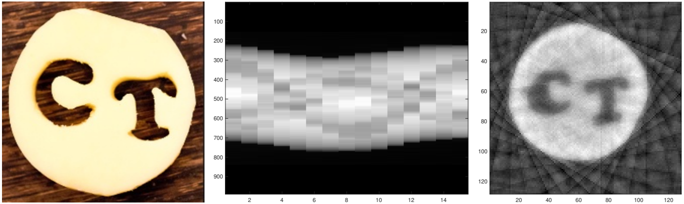
Limited angle tomography data:



Left: the actual object. Middle: sinogram data for limited angle tomography. Right: ML estimator of the unknown density.

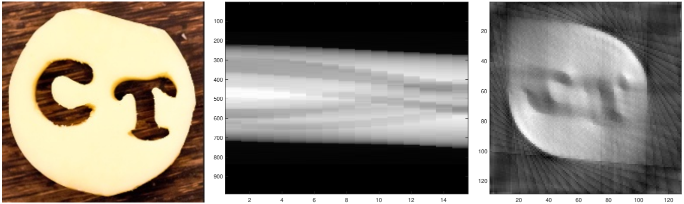
MAP estimator (Gaussian prior)

Sparse angle tomography data:


100
200
300
400
500
600
700
800
900
2
4
6
8
10
12
14
20
40
60
80
100
120
20
40
60
80
100
120

Left: the actual object. Middle: sinogram data for sparse angle tomography. Right: MAP estimator with a Gaussian prior.

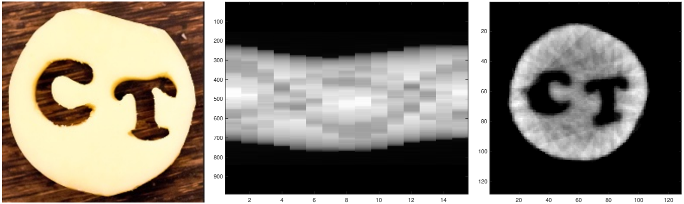
Limited angle tomography data:


100
200
300
400
500
600
700
800
900
2
4
6
8
10
12
14
20
40
60
80
100
120
20
40
60
80
100
120

Left: the actual object. Middle: sinogram data for limited angle tomography. Right: MAP estimator with a Gaussian prior.

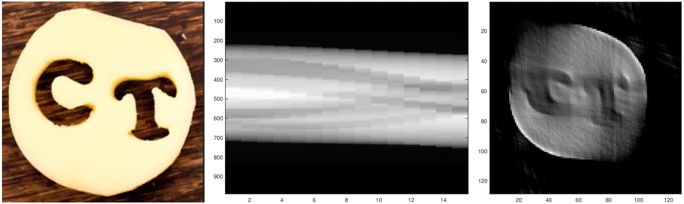
MAP estimator (total variation prior)

Sparse angle tomography data:


100
200
300
400
500
600
700
800
900
2
4
6
8
10
12
14
20
40
60
80
100
120
20
40
60
80
100
120

Left: the actual object. Middle: sinogram data for sparse angle tomography. Right: MAP estimator with a total variation prior.

Limited angle tomography data:


100
200
300
400
500
600
700
800
900
2
4
6
8
10
12
14
20
40
60
80
100
120
20
40
60
80
100
120

Left: the actual object. Middle: sinogram data for limited angle tomography. Right: MAP estimator with a total variation prior.

- In the previous example, using sparse or limited angle measurements means that the matrix system $y = Ax$ is underdetermined. Since we have very little data in the sparse or limited angle measurement settings, the ML estimator is useless in practice.
- In the Bayesian approach even a fairly weak Gaussian prior can produce a reconstruction. Using a more sophisticated prior such as a total variation prior improves the reconstruction quality even further (in this case, the density of the object can be well approximated using piecewise constant functions).
- The prior essentially compensates for the lack of data in the Bayesian approach to parameter recovery problems

Solution strategies

Bayes' formula produces an expression for the (in general high-dimensional) posterior distribution of the unknown parameter $x \in \mathbb{R}^d$, given the available data $y \in \mathbb{R}^k$. The main Bayesian estimators of the unknown parameter x are the MAP estimate $\hat{x}_{\mathrm{MAP}}$ (high-dimensional optimization problem) and the CM estimate $\hat{x}_{\mathrm{CM}}$ (high-dimensional integration problem). One may also be interested in quantifying the uncertainty in these estimates by computing the (co)variance of the posterior distribution or Bayesian credible sets (high-dimensional numerical integration problems). Typical solution strategies include the following.

- Conjugate inference: for a given likelihood, the prior is chosen such that the posterior is in the same probability distribution family as the prior (for example, if the likelihood and prior are both Gaussian, then the posterior is also Gaussian with known mean and covariance). In these cases, the MAP, CM, and (co)variance of the posterior have closed form solutions. This is an algebraic convenience, which avoids numerical difficulties otherwise associated with the computation of the MAP, CM, or other statistics of the posterior distribution.

- Numerical methods:
  - The computation of the MAP estimate is a high-dimensional optimization problem. It is often convenient to work with the *negative log-posterior*

    $$\hat{x}_{\mathrm{MAP}} = \arg\min_{x\in\mathbb{R}^d} (-\log f(x|y)).$$

    In some cases, the MAP estimator can be expressed as the solution to a Tikhonov functional. For example, consider the problem

    $$y = F(x) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I),$$

    where $x \in \mathbb{R}^d$ is the unknown parameter, $y \in \mathbb{R}^k$ is the data, and $\sigma > 0$ is the noise level. If we endow x with a Gaussian prior, e.g., $x \sim \mathcal{N}(x_0, \gamma^2 I)$, $\gamma > 0$, then the MAP estimator can be found as the minimizer of the Tikhonov functional

    $$\hat{x}_{\mathrm{MAP}} = \arg\min_{x\in\mathbb{R}^d} \left(\|y - F(x)\|^2 + \lambda^2\|x - x_0\|^2\right),$$

    where $\lambda = \frac{\sigma}{\gamma}$. If $F\colon \mathbb{R}^d \to \mathbb{R}^k$ is linear, i.e., $F(x) = Ax$ for some matrix $A \in \mathbb{R}^{k\times d}$, then we can solve $\hat{x}_{\mathrm{MAP}}$ from the (invertible) linear system

    $$(A^{\mathrm{T}}A + \lambda^2 I)\hat{x}_{\mathrm{MAP}} = A^{\mathrm{T}}y + \lambda^2 x_0. \qquad \text{(exercise)}$$

- Numerical methods:
  - The computation of the CM estimate is a high-dimensional numerical integration problem:

    $$\hat{x}_{\mathrm{CM}} = \int_{\mathbb{R}^d} x\, f(x|y)\,\mathrm{d}x. \tag{3}$$

    Typical solution strategies involve using high-dimensional cubatures or sampling-based methods. We will discuss the latter. Namely, if we are able to draw an i.i.d. sample $x_1, \ldots, x_n$ from the posterior $f(x|y)$, then we can in principle use the Monte Carlo method to approximate (3) as

    $$\hat{x}_{\mathrm{CM}} \approx \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}_n$$

    and likewise for the posterior variance $\mathrm{Var}(x|y) \approx \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2$.

    The difficulty with this approach lies in drawing a sample from a high-dimensional posterior distribution. To this end, we will discuss Markov Chain Monte Carlo (MCMC), which is an algorithm that can be used to draw a sample from a high-dimensional distribution with a known (unnormalized) density function.

    Another approach is to use, e.g., importance sampling to obtain a (biased) estimate of the integral (3).

Appendix: Remark on Bayesian hypothesis testing

Remark on Bayesian hypothesis testing

It is possible to perform statistical hypothesis testing from a Bayesian point of view. We will only give a brief sketch of the main idea here.

The Bayesian approach to testing involves putting a prior on H_0 and on the parameter x and then computing $\mathbb{P}(H_0|y)$. Consider the case where x is a vector and we are testing

$$H_0\colon x = x_0 \quad \text{versus} \quad H_1\colon x \neq x_0.$$

It is usually reasonable to use the prior $\mathbb{P}(H_0) = \mathbb{P}(H_1) = 1/2$ (although this is not essential in what follows). Under H_1, we need a prior for x; let us denote this prior density by $f(x)$. From Bayes' theorem,

$$\begin{aligned}\mathbb{P}(H_0|y) &= \frac{f(y|H_0)\mathbb{P}(H_0)}{f(y|H_0)\mathbb{P}(H_0) + f(y|H_1)\mathbb{P}(H_1)} = \frac{\frac{1}{2}f(y|x_0)}{\frac{1}{2}f(y|x_0) + \frac{1}{2}f(y|H_1)} \\ &= \frac{f(y|x_0)}{f(y|x_0) + \int_{\mathbb{R}^d} f(y|x)f(x)\,\mathrm{d}x} = \frac{\mathcal{L}(x_0)}{\mathcal{L}(x_0) + \int_{\mathbb{R}^d} \mathcal{L}(x)f(x)\,\mathrm{d}x}.\end{aligned}$$