

Computational inverse problems

Nuutti Hyvönen, Jenni Heino,
Pauliina Hirvi

`nuutti.hyvonen@aalto.fi`, `pauliina.hirvi@aalto.fi`

Eighth lecture

General noise model

Assume that we have an observation model of the type $Y = f(X, E)$, where $X \in \mathbb{R}^n$ is the unknown, $Y \in \mathbb{R}^m$ is the measurement and $E \in \mathbb{R}^k$ is the noise/parameter vector. Since fixing X and E determines the value of Y , we may write

$$\pi(y | x, e) = \delta(y - f(x, e)).$$

In consequence,

$$\pi(y | x) = \int_{\mathbb{R}^k} \pi(y, e | x) de = \int_{\mathbb{R}^k} \delta(y - f(x, e)) \pi_{\text{noise}}(e | x) de.$$

Change of variables

Consider two random variables $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^n$ that are related via the formula

$$Y = f(X),$$

where f is continuously differentiable and injective (these conditions can be relaxed). Suppose we know the probability density of Y , namely π_Y .

Then, for a Borel set $B \subset \mathbb{R}^n$, it holds that

$$\begin{aligned} P\{X \in B\} &= P\{Y \in f(B)\} = \int_{f(B)} \pi_Y(y) dy \\ &= \int_B \pi_Y(f(x)) |\det Df(x)| dx \end{aligned}$$

where $Df(x) \in \mathbb{R}^{n \times n}$ is the *differential* or the *Jacobian matrix* of f . As a consequence,

$$\pi_X(x) = \pi_Y(f(x)) |\det Df(x)|.$$

Example: multiplicative noise

Consider an amplifier that takes in a signal $f(t) > 0$ and sends it out multiplied by a constant factor $\alpha > 1$. The ideal model for the output is thus

$$g(t) = \alpha f(t), \quad 0 \leq t \leq T.$$

Suppose that the amplification factor is not a constant but fluctuates slightly around a mean value $\alpha_0 > 0$ as a function of time. In order to write a likelihood model for the output, we first discretize the signal:

$$x_j = f(t_j), \quad y_j = g(t_j), \quad 0 = t_1 < t_2 < \cdots < t_n = T.$$

Let the amplification at $t = t_j$ be a_j , i.e.,

$$y_j = a_j x_j, \quad 1 \leq j \leq n,$$

and introduce the stochastic extension:

$$Y_j = A_j X_j, \quad 1 \leq j \leq n.$$

In vector notation, this reads

$$Y = A.X,$$

with the dot denoting componentwise multiplication of the vectors $A, X \in \mathbb{R}^n$; we also use a similar notation for componentwise division.

Assume that A is independent of X and has the probability density

$$A \sim \pi_{\text{noise}}(a).$$

To find the likelihood density of Y , conditioned on $X = x$ such that $x_j > 0$ for all $j = 1, \dots, n$, we write

$$A_j = \frac{Y_j}{x_j}, \quad 1 \leq j \leq n.$$

Thus, we obtain by the change of variables formula that

$$\pi(y | x) = \frac{1}{x_1 x_2 \cdots x_n} \pi_{\text{noise}} \left(\frac{y \cdot}{x} \right).$$

As an example, assume that the components of $A \in \mathbb{R}^n$ are mutually independent and *log-normally* distributed:

$$W_i := \log A_i \sim \mathcal{N}(w_0, \sigma^2), \quad w_0 = \log \alpha_0.$$

To find an explicit formula for the density of A , we note that if $w = \log a$, where the logarithm is applied componentwise, we have

$$dw = \frac{1}{a_1 a_2 \cdots a_n} da \quad \text{for } a_1, \dots, a_n > 0.$$

Thus, the probability density of A vanishes if any of the components of a is zero or negative, and otherwise it holds that

$$\pi_{\text{noise}}(a) \propto \frac{1}{a_1 a_2 \cdots a_n} \exp \left(-\frac{1}{2\sigma^2} \|\log(a/\alpha_0)\|^2 \right).$$

By substituting this formula in

$$\pi(y | x) = \frac{1}{x_1 x_2 \cdots x_n} \pi_{\text{noise}} \left(\frac{y \cdot}{x} \right),$$

we find that

$$\pi(y | x) \propto \frac{1}{y_1 y_2 \cdots y_n} \exp \left(-\frac{1}{2\sigma^2} \left\| \log \left(\frac{y \cdot}{\alpha_0 x} \right) \right\|^2 \right).$$

for $y \in \mathbb{R}^n$ such that $y_j > 0$ for all $j = 1, \dots, n$, and zero for other $y \in \mathbb{R}^n$. (Recall that it was assumed to begin with that the components of x are positive.)

Incompletely known forward model

Consider having a noisy measurement with an incompletely known forward model: The deterministic model with additive noise is $y = A(v)x + e$, $y, e \in \mathbb{R}^m$, $x \in \mathbb{R}^n$ and $A(v) \in \mathbb{R}^{m \times n}$, where $A(v)$ depends on a parameter vector $v \in \mathbb{R}^k$.

The corresponding stochastic extension is

$$Y = A(V)X + E.$$

Assume that E , X and V are mutually independent. How to construct the likelihood model $\pi(y | x)$, assuming that the noise is distributed according to π_{noise} and the parameter according to π_{param} ?

To begin with, fix $X = x$ and $V = v$ in order to get the conditional density for Y :

$$\pi(y | x, v) = \pi_{\text{noise}}(y - A(v)x).$$

Subsequently, we marginalize with respect to the parameter V which is of secondary interest:

$$\begin{aligned}\pi(y | x) &= \int_{\mathbb{R}^k} \pi(y, v | x) dv = \int_{\mathbb{R}^k} \pi(y | x, v) \pi_{\text{param}}(v) dv \\ &= \int_{\mathbb{R}^k} \pi_{\text{noise}}(y - A(v)x) \pi_{\text{param}}(v) dv.\end{aligned}$$

On sampling

Before moving on to construction of priors, we touch the subject of how to draw a sample of realizations from a given probability distribution.

Why is such consideration relevant?

- visual inspection of priors,
- estimation of integrals of the type

$$I = \int f(x)\pi(x)dx$$

with the help of Markov chain Monte Carlo (MCMC) techniques.

In what follows, we assume to have random number generators for two elementary distributions at our disposal:

- Standard normal distribution

$$\pi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right);$$

in Matlab the command `randn`.

- Uniform distribution over the interval $[0, 1]$,

$$\pi(x) = \chi_{[0,1]}(x);$$

in Matlab the command `rand`.

Sampling from Gaussian distributions

Suppose that we want to create a sample of realizations for a multivariate Gaussian random variable $X \sim \mathcal{N}(x_0, \Gamma)$, with the probability density

$$\pi(x) = \left(\frac{1}{(2\pi)^n \det(\Gamma)} \right)^{1/2} \exp \left(-\frac{1}{2} (x - x_0)^T \Gamma^{-1} (x - x_0) \right).$$

Since Γ^{-1} is (by assumption) symmetric and positive definite, it has a Cholesky decomposition

$$\Gamma^{-1} = R^T R,$$

where R is an upper triangular matrix. Notice that the probability density of X can alternatively be written as

$$\pi(x) = \left(\frac{1}{(2\pi)^n \det(\Gamma)} \right)^{1/2} \exp \left(-\frac{1}{2} \|R(x - x_0)\|^2 \right).$$

Encouraged by this observation, we define a new random variable

$$W = R(X - x_0) \iff X = R^{-1}W + x_0,$$

which, in particular, means that

$$\pi_W(w) = \pi_X(R^{-1}w + x_0)|\det(R^{-1})| = \pi_X(R^{-1}w + x_0)|\det(R)|^{-1}.$$

Using the identity

$$\det(\Gamma)^{-1} = \det(\Gamma^{-1}) = \det(R^T) \det(R) = \det(R)^2,$$

leads finally to the formula

$$\pi(w) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\|w\|^2\right).$$

In consequence, W is *Gaussian white noise*, i.e.,

$$W \sim \mathcal{N}(0, I).$$

This transformation is called the *whitening* of X and the Cholesky factor R of the inverse of the covariance the *whitening matrix*.

If the whitening matrix is known, a random draw from a general Gaussian density can be generated as follows:

1. Draw $w \in \mathbb{R}^n$ from the Gaussian white noise density.
2. Compute the sought for realization $x \in \mathbb{R}^n$ by solving the linear system

$$w = R(x - x_0),$$

which is almost trivial since R is triangular.

Random draws from non-Gaussian densities using direct sampling

Let us next consider how to draw a random sample directly from the actual distribution in one dimension.

Let X be a real valued random variable with probability density $\pi(x)$ such that $\pi(x) = 0$ only at isolated points (this assumption can be relaxed). Define the cumulative distribution function via

$$\Phi(z) = \int_{-\infty}^z \pi(x) dx.$$

Due to the assumptions on π , it follows from the fundamental theorem of calculus that Φ is strictly increasing. In particular, $\Phi : \mathbb{R} \rightarrow (0, 1)$ has an inverse $\Phi^{-1} : (0, 1) \rightarrow \mathbb{R}$.

Define a new random variable,

$$T = \Phi(X).$$

Lemma. $T \sim \text{Uniform}([0, 1])$.

Proof. Observe first that,

$$P\{T < a\} = P\{\Phi(X) < a\} = P\{X < \Phi^{-1}(a)\}, \quad 0 < a < 1.$$

On the other hand, due to the definition of a probability density,

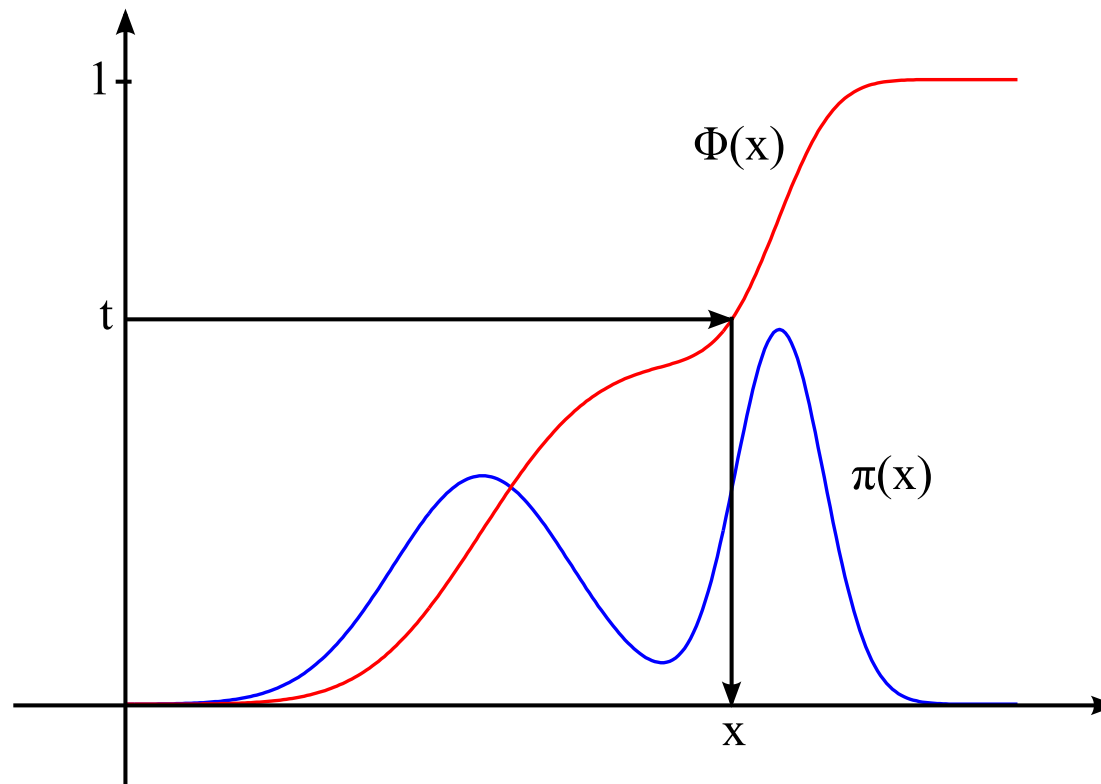
$$\begin{aligned} P\{X < \Phi^{-1}(a)\} &= \int_{-\infty}^{\Phi^{-1}(a)} \pi(x) dx = \int_{-\infty}^{\Phi^{-1}(a)} \Phi'(x) dx \\ &= \Phi(\Phi^{-1}(a)) - \lim_{x \rightarrow -\infty} \Phi(x) = a - 0 = a, \end{aligned}$$

which just means that T is distributed uniformly over the interval $[0, 1]$.

An algorithm for drawing from the density π :

1. Draw $t \sim \text{Uniform}([0, 1])$,
2. Calculate $x = \Phi^{-1}(t)$.

This technique is sometimes referred to as the *Golden Rule*.



Example: Gaussian distribution with a bound constraint

Consider a one-dimensional normal distribution with a bound constraint,

$$\pi(x) \propto \pi_c(x) \exp\left(-\frac{1}{2}x^2\right),$$

where

$$\pi_c(x) = \begin{cases} 1 & \text{if } x > c, \\ 0 & \text{if } x \leq c \end{cases}$$

for some $c \in \mathbb{R}$. Our aim is to generate a sample from this distribution.

In this case, the cumulative distribution function is

$$\Phi(z) = C \int_c^z e^{-x^2/2} dx, \quad C = \left(\int_c^\infty e^{-x^2/2} dx \right)^{-1},$$

where $C > 0$ is the normalizing constant of the corresponding probability density.

The function Φ has to be calculated numerically. Fortunately, there are routines available to do the needed integration: In Matlab, the built-in *error function*, `erf`, is defined as

$$\text{erf}(t) = \frac{2}{\sqrt{\pi}} \int_0^t e^{-s^2} ds.$$

We observe that

$$\begin{aligned} \Phi(z) &= C \left(\int_0^z - \int_0^c \right) e^{-x^2/2} dx = \sqrt{2}C \left(\int_0^{z/\sqrt{2}} - \int_0^{c/\sqrt{2}} \right) e^{-s^2} ds \\ &= \sqrt{\frac{\pi}{2}} C \left(\text{erf}(z/\sqrt{2}) - \text{erf}(c/\sqrt{2}) \right). \end{aligned}$$

Since $\text{erf}(t) \rightarrow 1$ as $t \rightarrow \infty$, the same logic also shows that

$$C = \left(\sqrt{\frac{\pi}{2}} \left(1 - \text{erf}(c/\sqrt{2}) \right) \right)^{-1}.$$

Altogether we have

$$\Phi(z) = \frac{\operatorname{erf}(z/\sqrt{2}) - \operatorname{erf}(c/\sqrt{2})}{1 - \operatorname{erf}(c/\sqrt{2})}.$$

How about the inverse then?

Setting

$$\Phi(z) = t \iff z = \Phi^{-1}(t),$$

we find through a straightforward algebraic manipulation that

$$\operatorname{erf}(z/\sqrt{2}) = t(1 - \operatorname{erf}(c/\sqrt{2})) + \operatorname{erf}(c/\sqrt{2}),$$

or in other words (see `erfinv` in Matlab)

$$\Phi^{-1}(t) = \sqrt{2} \operatorname{erf}^{-1}\left(t(1 - \operatorname{erf}(c/\sqrt{2})) + \operatorname{erf}(c/\sqrt{2})\right).$$

The generation of random draws in Matlab is then very simple:

```
a = erf(c/sqrt(2));  
t = rand;  
z = sqrt(2)*erfinv(t*(1-a)+a);
```

Note: If the bound c is large, the above program does not work because the error function saturates quickly to unity. To be more precise, e.g. for $c=10$, Matlab interprets that a in the above code is exactly one, which means that the value of z is Inf independently of the random draw t . In consequence, it is actually more advisable to use the *complementary error function* $\text{erfc} = 1 - \text{erf}$:

```
a = erfc(c/sqrt(2));  
t = rand;  
z = sqrt(2)*erfcinv((1-t)*a);
```

Prior models

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.

A good prior should have the following property: Denote by x a possible realization of a random variable $X \sim \pi_{\text{pr}}(x)$. If E is a collection of expected (i.e., something you would expect to see) vectors x and U is a collection of unexpected ones, then it should hold that

$$\pi_{\text{pr}}(x) \gg \pi_{\text{pr}}(x') \quad \text{when } x \in E, x' \in U,$$

i.e., the prior assigns a clearly higher probability to the realization that we expect to see.

Example: Impulse prior densities

Consider, e.g., an imaging problem where the unknown is the discretized distribution of a physical parameter, i.e., a pixel image.

Assume that our prior information is that the image contains small and well localized objects in almost constant background. In such a case, one may try impulse prior densities, which have low average amplitude but allow outliers. (The 'tail' of an impulse prior density is long, although the expected value is small.)

Examples of impulse prior densities: Let $x \in \mathbb{R}^n$ represent a pixel image, where the component x_j is the intensity of the j th pixel. (In all of the following examples, X_j and X_k are assumed to be independent for $j \neq k$.)

The ℓ_1 prior:

$$\pi_{\text{pr}}(x) = \left(\frac{\alpha}{2}\right)^n \exp(-\alpha\|x\|_1), \quad \alpha > 0.$$

where the ℓ_1 -norm is defined as

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

More enhanced impulse noise effect can be obtained by taking even smaller power of the components of x :

$$\pi_{\text{pr}}(x) \propto \exp\left(-\alpha \sum_{j=1}^n |x_j|^p\right), \quad 0 < p < 1, \quad \alpha > 0.$$

Another choice is the *Cauchy density* that is defined via

$$\pi_{\text{pr}}(x) = \left(\frac{\alpha}{\pi}\right)^n \prod_{j=1}^n \frac{1}{1 + \alpha^2 x_j^2}, \quad \alpha > 0.$$

The entropy of an image is defined as

$$\mathcal{E}(x) = - \sum_{j=1}^n x_j \log \frac{x_j}{x_0},$$

where it is assumed that $x_j > 0$, $j = 1, \dots, n$, and $x_0 > 0$ is a given constant. The *entropy density* is then of the form

$$\pi(x) \propto \exp(\alpha \mathcal{E}(x)), \quad \alpha > 0.$$

Log-normal density: The logarithm of a single pixel $x \in \mathbb{R}$ is normally distributed, i.e.,

$$w = \log x, \quad w \sim \mathcal{N}(w_0, \sigma^2).$$

The explicit density of x is then

$$\pi(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(\log x - w_0)^2\right), \quad x > 0.$$

Do these priors represent our beliefs? How do these priors look like?

To underline the interpretation as a pixel image, we add a positivity constraint to the above introduced priors, that is, we make the replacement

$$\pi_{\text{pr}}(x) \rightarrow C\pi_+(x)\pi_{\text{pr}}(x),$$

where $\pi_+(x)$ is one if all components of x are positive, and zero otherwise. Here, C is a normalizing constant: If $\pi_{\text{pr}}(x)$ is a probability density, the same does not typically apply to $\pi_+(x)\pi_{\text{pr}}(x)$ without appropriate scaling.

For visual inspection we make random draws of pixel images from the constrained densities. As all components are independent, drawing can be done componentwise.

To make the draws from one-dimensional densities, we calculate the cumulative distribution of the prior density and employ the Golden Rule, as presented at the previous lecture.

Example: Drawing from ℓ_1 prior

The one-dimensional cumulative distribution of the positively constrained ℓ_1 prior is

$$\Phi(t) = \alpha \int_0^t e^{-\alpha s} ds = 1 - e^{-\alpha t}.$$

The inverse cumulative distribution is thus

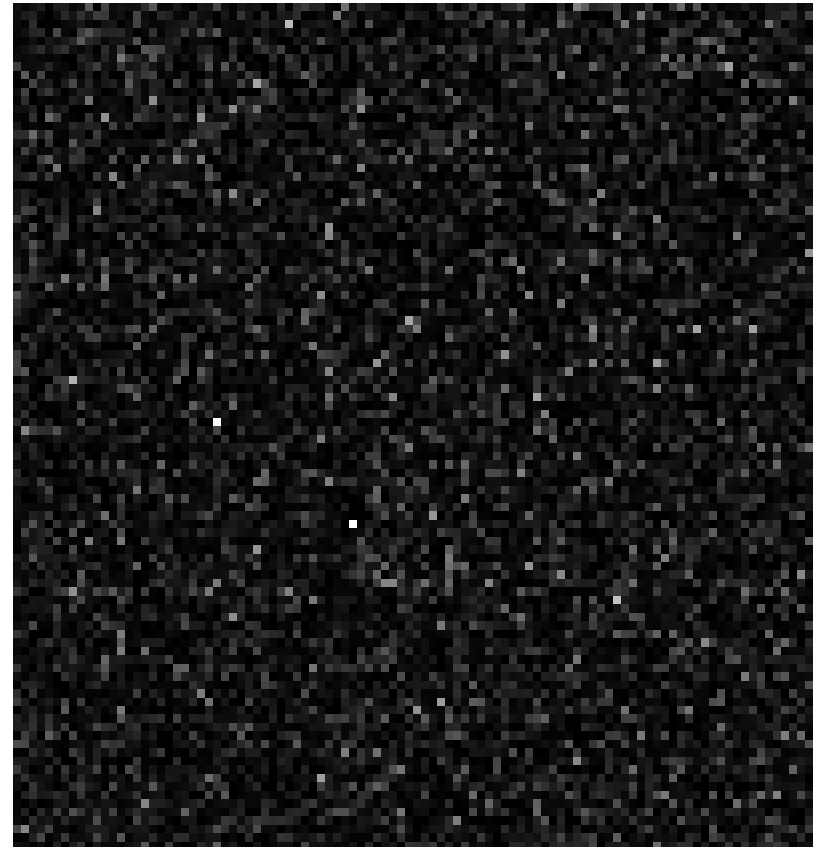
$$\Phi^{-1}(t) = -\frac{1}{\alpha} \log(1 - t).$$

For each pixel x_j , we draw t_j from the uniform distribution $\text{Uniform}([0, 1])$ and calculate $x_j = -1/\alpha \log(1 - t_j)$.

The Matlab code for doing this is very simple:

```
A=rand(100,100);  
alfa=1;  
AL1inv=-1/alfa*log(1-A);  
figure  
imagesc(AL1inv)  
colormap gray  
axis square
```

Two random draws of pixel images from a ℓ_1 -prior.



Example: Drawing from Cauchy prior

The one-dimensional cumulative distribution of the positively constrained Cauchy prior is

$$\Phi(t) = \frac{2\alpha}{\pi} \int_0^t \frac{1}{1 + \alpha^2 s^2} ds = \frac{2}{\pi} \arctan(\alpha t),$$

meaning that the inverse cumulative distribution is

$$\Phi^{-1}(t) = \frac{1}{\alpha} \tan \frac{\pi t}{2}.$$

As in the case of the ℓ_1 -prior, we draw t_j from the uniform distribution and then calculate $x_j = 1/\alpha \tan(\pi t/2)$.

Two random draws of pixel images from a Cauchy prior.



How do these priors compare to white noise?

Let us consider a Gaussian prior with a positivity constraint, i.e.,

$$\pi_{\text{pr}}(x) \propto \pi_+(x) \exp\left(-\frac{1}{2\alpha^2}\|x\|^2\right), \quad \alpha > 0.$$

Previously, we implemented drawing from a standard Gaussian distribution with a bound c . In particular, we were able to calculate the one-dimensional cumulative distribution function

$$\Phi^{-1}(t) = \sqrt{2} \operatorname{erf}^{-1}\left(t(1 - \operatorname{erf}(c/\sqrt{2})) + \operatorname{erf}(c/\sqrt{2})\right).$$

A similar derivation for $c = 0$ and the variance α^2 instead of 1 yields in the current case that

$$\Phi^{-1}(t) = \sqrt{2}\alpha \operatorname{erf}^{-1}(t).$$

L_1 prior



Cauchy prior



White noise prior

