Inverse Problems

Sommersemester 2023

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Tenth lecture, June 19, 2023

Today's lecture

- Sampling from multivariate Gaussian distributions, inverse transform sampling
- Prior modeling
- The linear Gaussian setting
- Numerical example

Change of variables

Consider two random variables $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$ which are related via the formula

$$y = f(x),$$

where f is continuously differentiable and one-to-one (these conditions can be relaxed).

Then, for any $B \in \mathcal{B}(\mathbb{R}^n)$, it holds that

$$\mathbb{P}(x \in B) = \mathbb{P}(y \in f(B)) = \int_{f(B)} \pi_y(y)\,\mathrm{d}y = \int_B \pi_y(f(x))|\det Df(x)|\,\mathrm{d}x,$$

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

$$\pi_x(x) = \pi_y(f(x))|\det Df(x)|.$$

Sampling from Gaussian distributions

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

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

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

$$C^{-1} = R^{\mathrm{T}} R,$$

where R is an upper triangular matrix. The probability density of x can be alternatively written as

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

Let us define a new random variable $w = R(x - x_0) \Leftrightarrow x = R^{-1}w + x_0$.

On the last slide, we defined $w = R(x - x_0) \Leftrightarrow x = R^{-1}w + x_0$, where $x \sim \mathcal{N}(x_0, C)$. The change of variables formula yields

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

Noting that

$$\frac{1}{\det C} = \det(C^{-1}) = \det R^{\mathrm{T}} \det R = \det(R)^2,$$

we obtain

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

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

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

Coloring transformation

Let $x \sim \mathcal{N}(x_0, C)$ (C p.d.) and $w \sim \mathcal{N}(0, I)$, with $C^{-1} = R^{\mathrm{T}} R$. Then it holds that

$$w = R(x - x_0) \quad \Leftrightarrow \quad x = R^{-1} w + x_0.$$

This is the basis of the *coloring transformation*:

1. Draw $w \in \mathbb{R}^n$ from $\mathcal{N}(0, I)$.
2. A realization of $x \in \mathbb{R}^n$ from $\mathcal{N}(x_0, C)$ can be obtained via

$$x = R^{-1} w + x_0.$$

Remark: The function `mvnrnd` can be used to draw from multivariate normal distributions in MATLAB (`numpy.random.multivariate_normal` in Python, respectively). However, if the Cholesky factor R (or its inverse) has been precomputed, it may be slightly more efficient to apply the coloring transformation, e.g., as

```
x = R \ randn(n,1) + x0;                                  (in MATLAB)
x = numpy.linalg.solve(R,numpy.random.normal(size=(n,1)))\
    + x0                                                  (in Python)
```

Sampling from general univariate distributions

In order to sample a real-valued random variable x directly, we can use its inverse distribution function. Let us assume that the probability density $\pi(x)$ of x is almost surely positive (this condition can be relaxed). Then, the *cumulative distribution function* $\Phi\colon \mathbb{R} \to (0,1)$ of x is defined by

$$\Phi(t) = \mathbb{P}(x < t) = \int_{-\infty}^{t} \pi(x)\,\mathrm{d}x.$$

In other words, Φ is the antiderivative of π. It follows from the fundamental theorem of calculus that Φ is strictly increasing. In particular, its inverse $\Phi^{-1}\colon (0,1) \to \mathbb{R}$ exists.

Now, we define a new random variable $u = \Phi(x)$. First, we observe that

$$\mathbb{P}(u < t) = \mathbb{P}(\Phi(x) < t) = \mathbb{P}(x < \Phi^{-1}(t))$$

for all $t \in (0, 1)$. However, by definition of the cumulative distribution function,

$$\begin{aligned}\mathbb{P}(x < \Phi^{-1}(t)) &= \int_{-\infty}^{\Phi^{-1}(t)} \pi(x)\,\mathrm{d}x = \int_{-\infty}^{\Phi^{-1}(t)} \Phi'(x)\,\mathrm{d}x \\ &= \Phi(\Phi^{-1}(t)) - \lim_{x \to -\infty} \Phi(x) = t.\end{aligned}$$

Hence $\mathbb{P}(u < t) = t$, meaning that $u \sim \mathcal{U}(0, 1)$ is distributed uniformly on the interval $[0, 1]$. On the other hand, if $u \sim \mathcal{U}(0, 1)$ is given, then we obtain a random variable x with density π by setting $x = \Phi^{-1}(u)$. This reduces drawing a sample from the distribution π to drawing a sample from a uniform distribution, which can for example be performed in MATLAB using the `rand` command (`numpy.random.uniform` in Python).

Inverse transform sampling ("Golden rule")

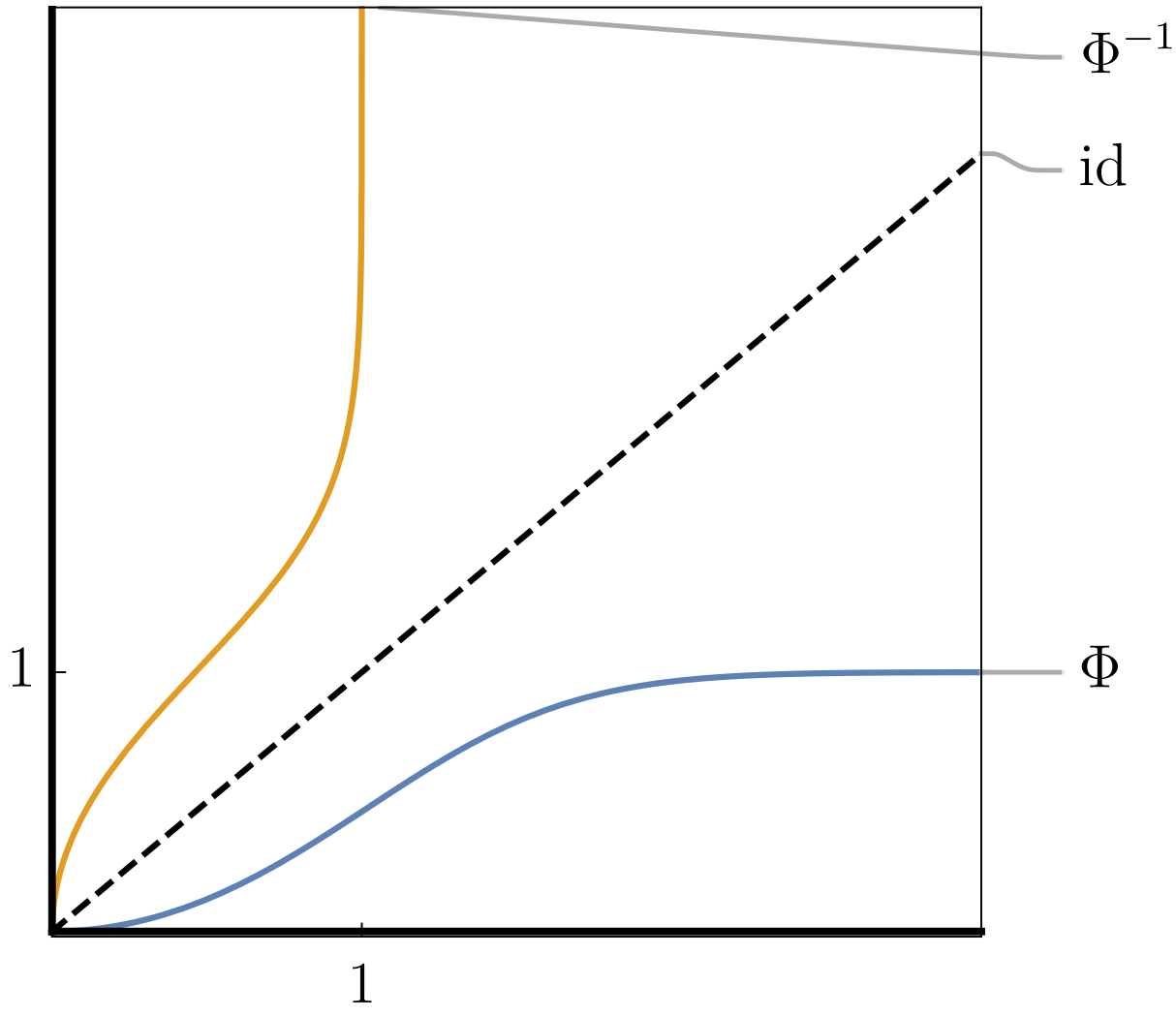
An algorithm for drawing from the density π with CDF Φ:

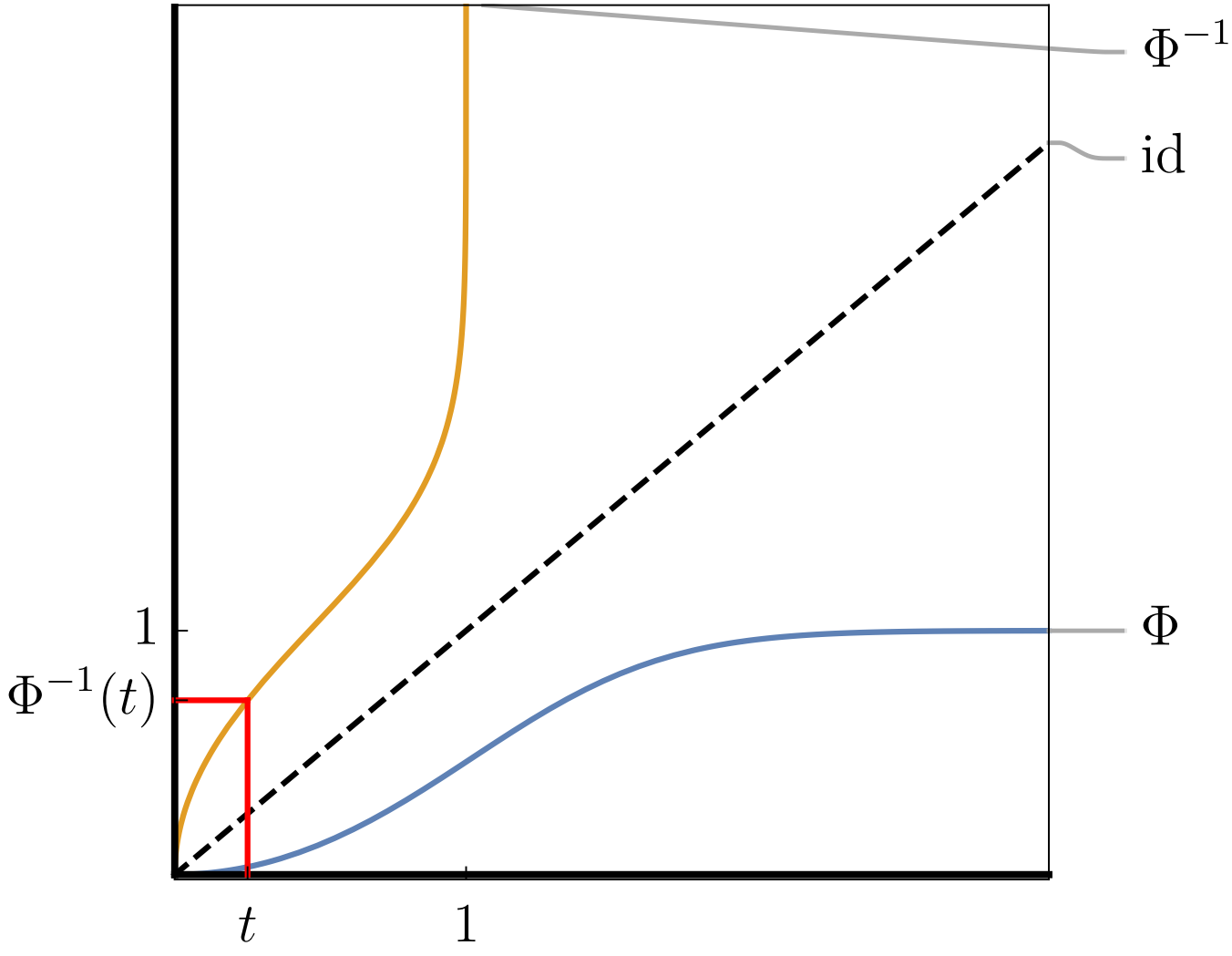
1. Draw $t \sim \mathcal{U}(0,1)$.
2. Calculate $x = \Phi^{-1}(t)$.

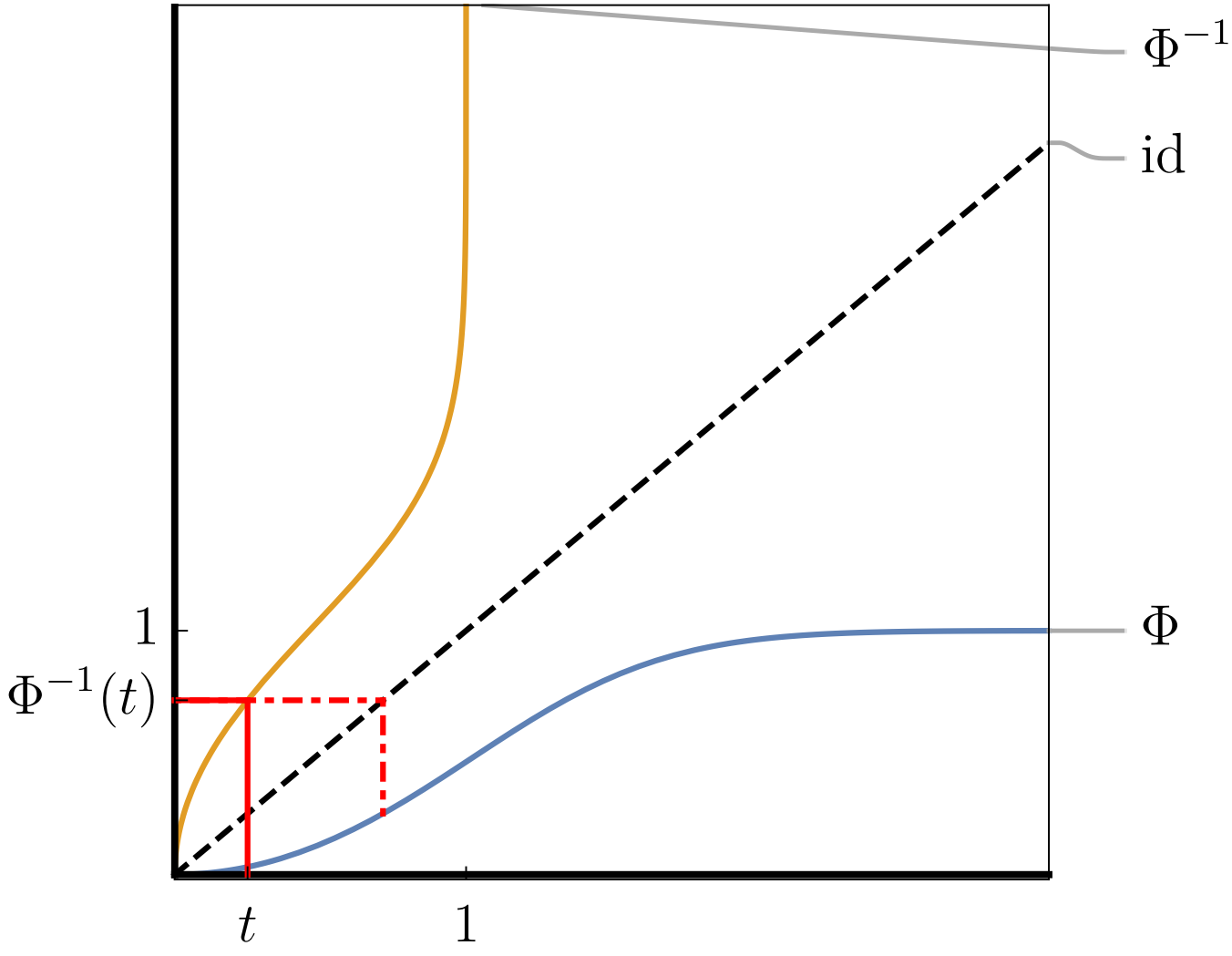
If a closed form expression for the inverse CDF is not available, then a computationally attractive formula for obtaining the value $\Phi^{-1}(t)$ at a point $t \in (0,1)$ is based on the identity

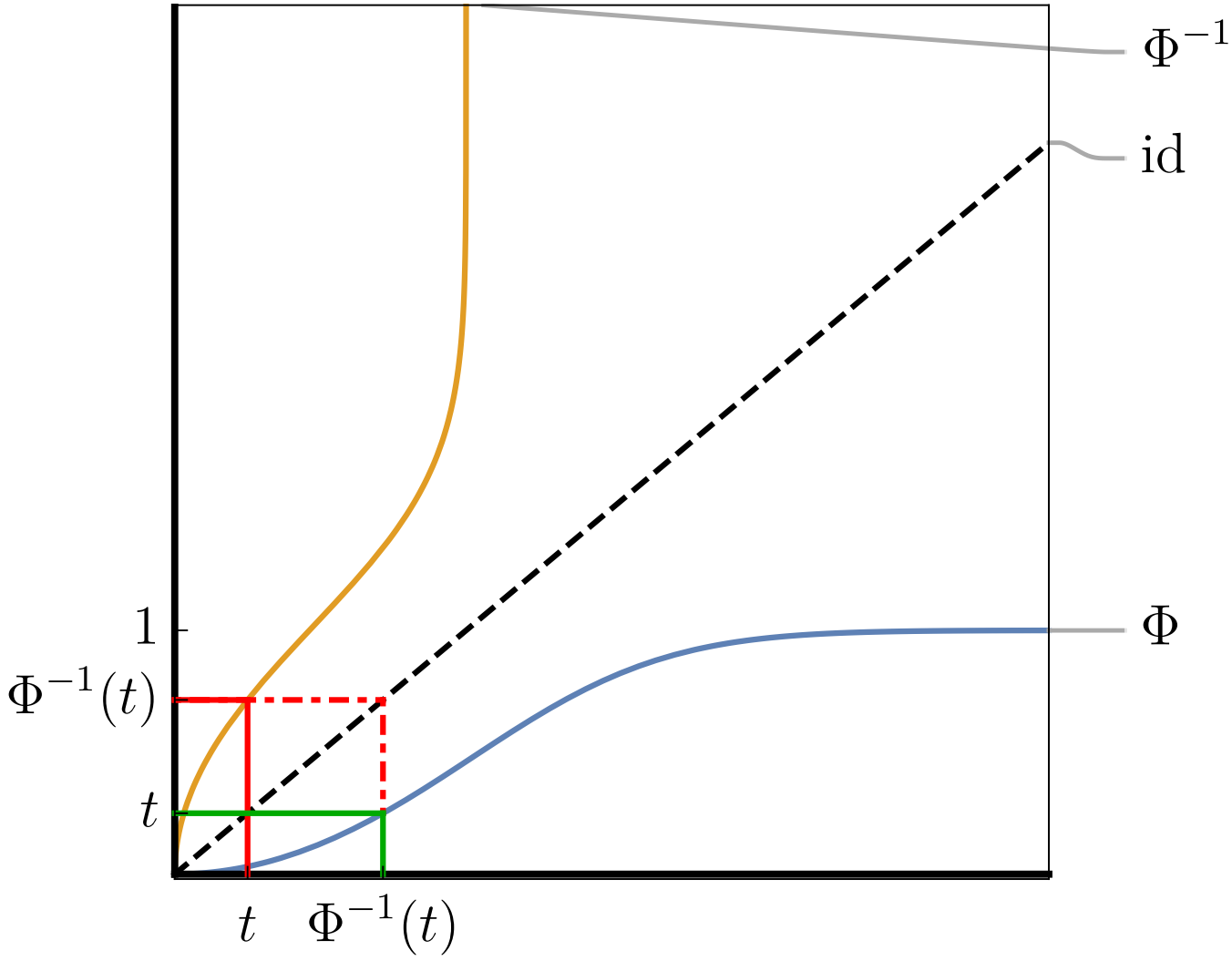
$$\Phi^{-1}(t) = \inf\{x \mid \Phi(x) \geq t\}.$$

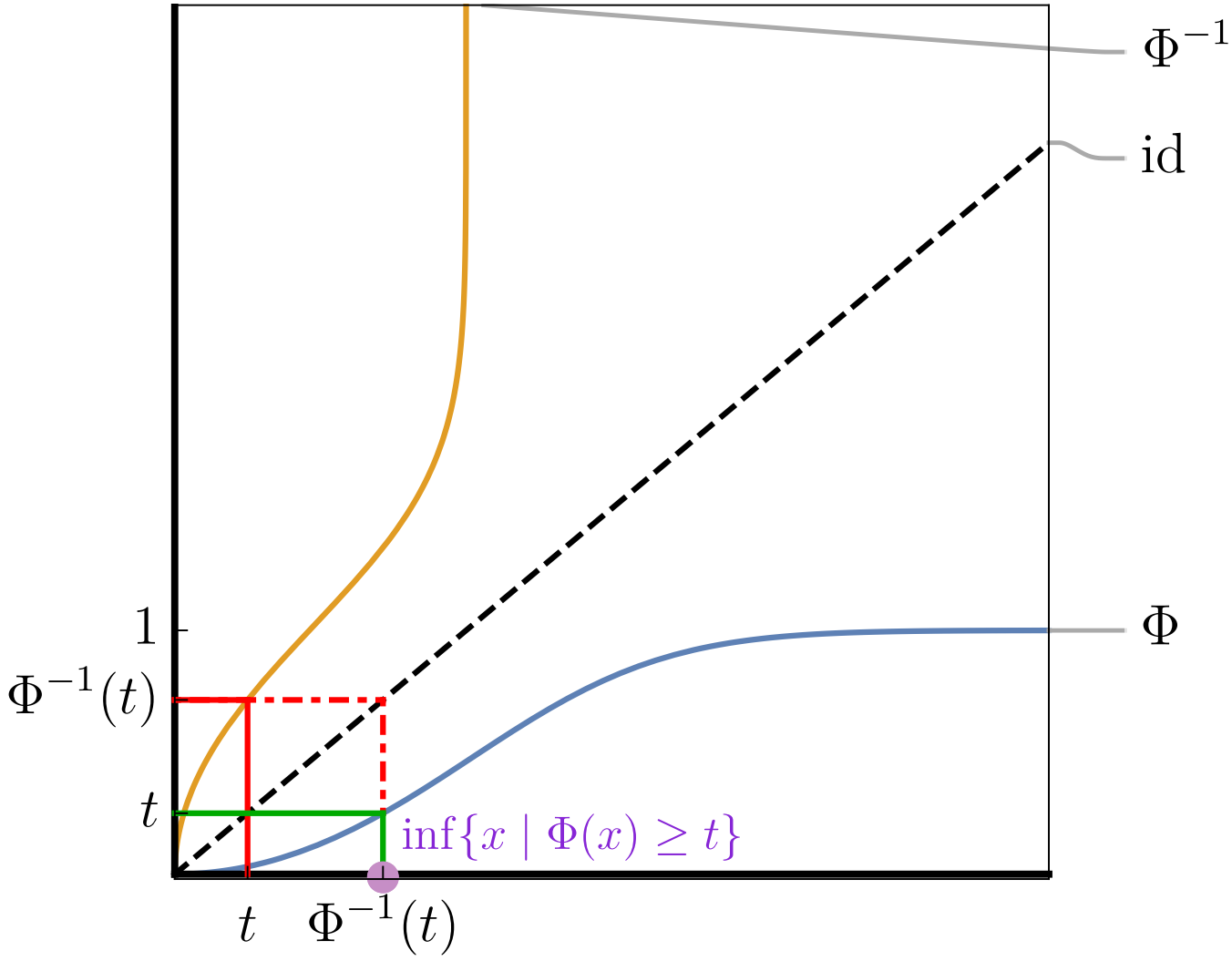
Remark: The above formula is the expression for the *generalized inverse CDF*: the formula with the infimum is valid even in the general case of weakly monotonic and right-continuous CDFs.


Φ^{-1}
id
Φ
1
1


Φ^{-1}
id
Φ
1
$\Phi^{-1}(t)$
t
1


Φ^{-1}
id
Φ
1
$\Phi^{-1}(t)$
t
1


Φ^{-1}
id
Φ
1
$\Phi^{-1}(t)$
t
t
$\Phi^{-1}(t)$



"Draw $t \sim \mathcal{U}(0,1)$ and find the smallest value of x such that $\Phi(x) \geq t$."

Remarks:

- The inverse transform sampling method can be used to sample univariate densities $\pi(u)$. However, if the components of a multivariate density are *mutually independent*, i.e., $\pi(u_1, \ldots, u_n) = \pi(u_1) \cdots \pi(u_n)$ holds a.e., then inverse transform sampling can be used to generate samples componentwise.
- Unfortunately, the components of multivariate posterior distributions are generally *not mutually independent*. In the next two weeks, we will discuss importance sampling and MCMC methods for sampling high-dimensional (posterior) distributions. These methods are applicable even when the components of multivariate distributions are **not** mutually independent.

Example

Suppose that we have the PDF $\pi(x) := (6x - 6x^2)\chi_{(0,1)}(x)$. We can design the following simple scheme based on inverse transform sampling to draw samples from this distribution.

MATLAB implementation:

```
n = 1e5; % sample size
x = linspace(0,1);
p = @(x) 6*x-6*x.^2; % PDF
P = cumsum(p(x)); P = P/P(end); % "empirical" CDF of p
samples = [];
for iter = 1:n
  u = rand; % realization of U(0,1)
  ind = find(u <= P,1,'first'); % inverse CDF rule
  samples = [samples,x(ind)]; % store sample
end
histogram(samples,'Normalization','pdf'); % draw a histogram
hold on, plot(x,p(x),'LineWidth',3), legend('samples','pdf');
hold off;
```

Python implementation:

```
import numpy as np
import matplotlib.pyplot as plt
n = int(1e5) # sample size
x = np.linspace(0,1,1000)
p = lambda x: 6*x-6*x**2 # PDF
P = np.cumsum(p(x)); P = P/P[-1] # "empirical" CDF of p
samples = []
for iter in range(n):
    u = np.random.uniform() # realization of U(0,1)
    ind = np.where(u<=P)[0][0] # inverse CDF rule
    samples.append(x[ind]) # store sample
plt.hist(samples,bins='auto',
         density=True,label='samples') # draw a histogram
plt.plot(x,p(x),linewidth=2,label='pdf')
plt.legend()
plt.show()
# Thanks to Subodh Khanger for the Python implementation!
```

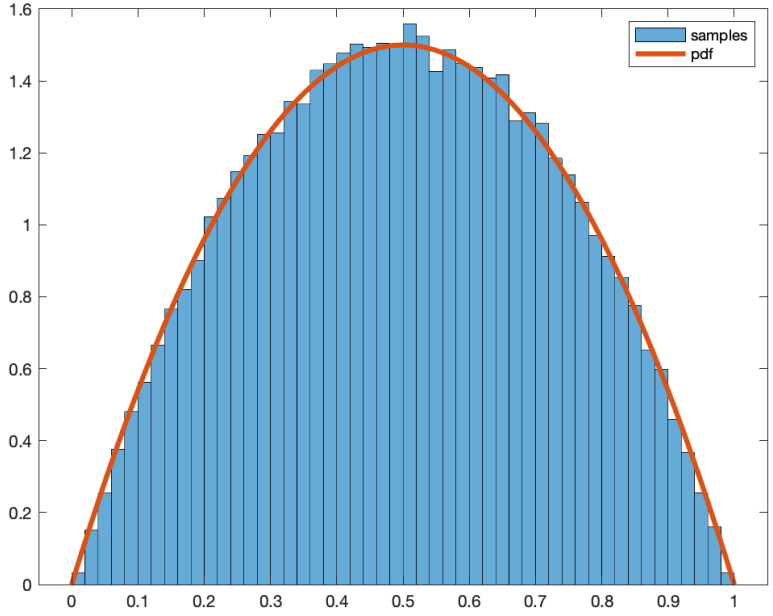


Figure: 10^5 samples drawn from the distribution given on the previous page organized as a histogram.

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

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

are the most used prior distribution in statistical inverse problems. They 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 usually be generated directly, for example in MATLAB via `randn` or `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.

$f_{9,0}$	$f_{9,1}$	$f_{9,2}$	$f_{9,3}$	$f_{9,4}$	$f_{9,5}$	$f_{9,6}$	$f_{9,7}$	$f_{9,8}$	$f_{9,9}$
$f_{8,0}$	$f_{8,1}$	$f_{8,2}$	$f_{8,3}$	$f_{8,4}$	$f_{8,5}$	$f_{8,6}$	$f_{8,7}$	$f_{8,8}$	$f_{8,9}$
$f_{7,0}$	$f_{7,1}$	$f_{7,2}$	$f_{7,3}$	$f_{7,4}$	$f_{7,5}$	$f_{7,6}$	$f_{7,7}$	$f_{7,8}$	$f_{7,9}$
$f_{6,0}$	$f_{6,1}$	$f_{6,2}$	$f_{6,3}$	$f_{6,4}$	$f_{6,5}$	$f_{6,6}$	$f_{6,7}$	$f_{6,8}$	$f_{6,9}$
$f_{5,0}$	$f_{5,1}$	$f_{5,2}$	$f_{5,3}$	$f_{5,4}$	$f_{5,5}$	$f_{5,6}$	$f_{5,7}$	$f_{5,8}$	$f_{5,9}$
$f_{4,0}$	$f_{4,1}$	$f_{4,2}$	$f_{4,3}$	$f_{4,4}$	$f_{4,5}$	$f_{4,6}$	$f_{4,7}$	$f_{4,8}$	$f_{4,9}$
$f_{3,0}$	$f_{3,1}$	$f_{3,2}$	$f_{3,3}$	$f_{3,4}$	$f_{3,5}$	$f_{3,6}$	$f_{3,7}$	$f_{3,8}$	$f_{3,9}$
$f_{2,0}$	$f_{2,1}$	$f_{2,2}$	$f_{2,3}$	$f_{2,4}$	$f_{2,5}$	$f_{2,6}$	$f_{2,7}$	$f_{2,8}$	$f_{2,9}$
$f_{1,0}$	$f_{1,1}$	$f_{1,2}$	$f_{1,3}$	$f_{1,4}$	$f_{1,5}$	$f_{1,6}$	$f_{1,7}$	$f_{1,8}$	$f_{1,9}$
$f_{0,0}$	$f_{0,1}$	$f_{0,2}$	$f_{0,3}$	$f_{0,4}$	$f_{0,5}$	$f_{0,6}$	$f_{0,7}$	$f_{0,8}$	$f_{0,9}$

Let us consider an image. We divide this region into $n \times n$ pixels and label the pixels $f_{i,j}$ for $i, j \in \{0, \ldots, n-1\}$.

x_{90}	x_{91}	x_{92}	x_{93}	x_{94}	x_{95}	x_{96}	x_{97}	x_{98}	x_{99}
x_{80}	x_{81}	x_{82}	x_{83}	x_{84}	x_{85}	x_{86}	x_{87}	x_{88}	x_{89}
x_{70}	x_{71}	x_{72}	x_{73}	x_{74}	x_{75}	x_{76}	x_{77}	x_{78}	x_{79}
x_{60}	x_{61}	x_{62}	x_{63}	x_{64}	x_{65}	x_{66}	x_{67}	x_{68}	x_{69}
x_{50}	x_{51}	x_{52}	x_{53}	x_{54}	x_{55}	x_{56}	x_{57}	x_{58}	x_{59}
x_{40}	x_{41}	x_{42}	x_{43}	x_{44}	x_{45}	x_{46}	x_{47}	x_{48}	x_{49}
x_{30}	x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	x_{36}	x_{37}	x_{38}	x_{39}
x_{20}	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}	x_{28}	x_{29}
x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}
x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9

It is convenient to reshape the matrix/image $(f_{i,j})$ into a vector x of length $d = n^2$ so that

$$x_{in+j} = f_{i,j}, \quad i, j \in \{0, \ldots, n-1\}.$$

The image on the left illustrates the new numbering corresponding to the pixels.

Note that `x = f.reshape((n*n,1))` and `f = x.reshape((n,n))`.
(In MATLAB: `x = f(:)` and `f = reshape(x,n,n)`.)

As an example, consider a problem where the unknown is a two-dimensional pixel image, arranged as a vector $x \in \mathbb{R}^d$. The components x_j represent the intensity of the j^{th} pixel. Since we consider images it is natural to add a positivity constraint to our prior. Assuming that x_i and x_j are independent for $i \neq j$, the Gaussian white noise density with positivity constraint is

$$\pi(x) \propto \chi_+(x) \exp\left(-\frac{1}{2\alpha^2}\|x\|^2\right),$$

where $\chi_+(x) = 1$ if $x_j > 0$ for all j and $\chi_+(x) = 0$ otherwise.

Since we assumed that each component is independent of the others, random draws can be performed componentwise.

Impulse priors

We assume again that the unknown is a two-dimensional pixel image.

Assume that our prior information is that the image contains small and well localized objects 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 pixel image, where the component x_j is the intensity of the j^{th} pixel. 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

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

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 pixels and a low-amplitude background is the *Cauchy density*

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

with $\alpha > 0$.

Since we consider images we add a positivity constraint to our prior. For the ℓ^1 prior, we set

$$\pi(x) = \alpha^d \chi_+(x) \exp(-\alpha \|x\|_1),$$

where $\chi_+(x) = 1$ if $x_j > 0$ for all j and $\chi_+(x) = 0$ otherwise. The components x_j are independent and each have the cumulative distribution function

$$\Phi(t) = \alpha \int_0^t \mathrm{e}^{-\alpha s} \, \mathrm{d}s = 1 - \mathrm{e}^{-\alpha t} \quad \text{for all } t \geq 0.$$

Now, we can draw samples of x_j using

$$x_j = \Phi^{-1}(u_j) = -\frac{1}{\alpha} \ln(1 - u_j),$$

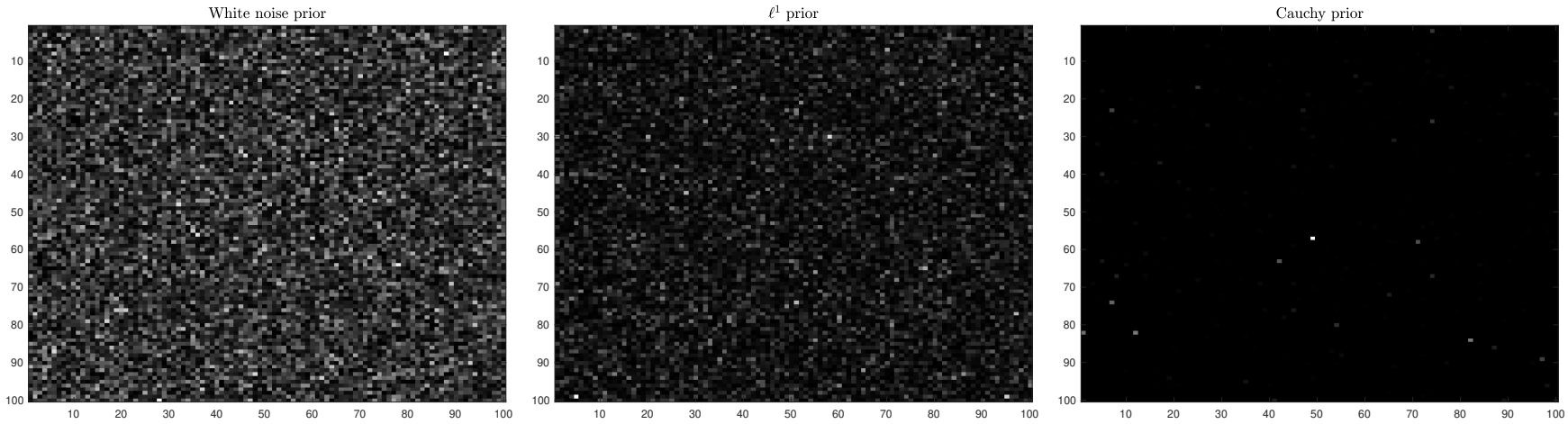
where the u_j are independent random draws from the uniform distribution $\mathcal{U}(0, 1)$.

Similarly, the components x_j of the Cauchy prior with positivity constraint are independent and have the CDF

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

so that the inverse cumulative distribution is $\Phi^{-1}(t) = \frac{1}{\alpha} \tan\left(\frac{\pi t}{2}\right)$.

Random draws from the white noise prior with positivity constraint, the impulse (ℓ^1) prior, and the Cauchy prior:



Note that as along as all components are independent, drawing can be done componentwise using inverse transform sampling. Here, for each pixel x_j, we draw t_j from $\mathcal{U}(0,1)$ and calculate $x_j = \Phi^{-1}(t_j)$.

Discontinuities

Assume that we want to estimate a one-dimensional signal $f\colon [0,1] \to \mathbb{R}$ with $f(0) = 0$ 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 f and assigning an impulsive noise distribution to it. Let us discretize the interval $[0,1]$ by points $t_j = j/d$ and write $x_j = f(t_j)$. Consider the density

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

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

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

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 MATLAB, it is more efficient to implement this as `x = cumsum(u)` (similarly `x = numpy.cumsum(u)` in Python).

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

$$\mathbb{P}(x|\alpha).$$

We model the unknown α with a *hyperprior* $\mathbb{P}(\alpha) = \pi_{\mathsf{h}}(\alpha)$ and write the joint distribution of x and α as

$$\mathbb{P}(x, \alpha) = \mathbb{P}(x|\alpha)\,\mathbb{P}(\alpha).$$

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

$$\mathbb{P}(x, \alpha|y) \propto \mathbb{P}(y|x, \alpha)\, \mathbb{P}(x, \alpha) = \mathbb{P}(y|x, \alpha)\, \mathbb{P}(x|\alpha)\, \mathbb{P}(\alpha).$$

The hyperprior density π_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 assumed to be more robust with respect to fixing a value for the hyperparameter α_0 than fixing a value for α.

The linear Gaussian setting

In this chapter we study the linear Gaussian setting, where the forward map F is linear and both the prior distribution and the distribution of the observational noise η are Gaussian.

For several reasons, it plays a central role in the study of inverse problems.

It arises frequently in applications, either directly or in the form of posterior distributions that are asymptotically Gaussian in the large data limit. It also allows computing explicit solutions which can be used to gain a general understanding. Apart from that, many methods employed in a nonlinear or non-Gaussian setting build on ideas from the linear Gaussian case by performing linearization or Gaussian approximation.

Let us suppose that the unknown $x \in \mathbb{R}^d$ and the data $y \in \mathbb{R}^k$ follow the relation

$$y = Ax + \eta, \tag{1}$$

where

1. The forward model is linear, i.e., $A \in \mathbb{R}^{k \times d}$.
2. The prior distribution is Gaussian: $x \sim \pi = \mathcal{N}(x_0, \Gamma_{\mathrm{pr}})$, where Γ_{pr} is symmetric and positive definite.
3. The noise is Gaussian: $\eta \sim \nu = \mathcal{N}(\eta_0, \Gamma_{\mathrm{n}})$, where Γ_{n} is symmetric and positive definite.
4. x and η are independent.

Theorem

Under assumptions 1–4, the posterior distribution corresponding to (1) is Gaussian with $x|y \sim \mathcal{N}(\mu_{\mathrm{post}}, \Gamma_{\mathrm{post}})$, *where we have the posterior mean*

$$\mu_{\mathrm{post}} = (\Gamma_{\mathrm{pr}}^{-1} + A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}A)^{-1}(A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}(y - \eta_0) + \Gamma_{\mathrm{pr}}^{-1}x_0)$$

and covariance

$$\Gamma_{\mathrm{post}} = (\Gamma_{\mathrm{pr}}^{-1} + A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}A)^{-1}.$$

Proof. Noting that $\Gamma_{\mathrm{post}} = (\Gamma_{\mathrm{pr}}^{-1} + A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}A)^{-1}$ and $\mu_{\mathrm{post}} = \Gamma_{\mathrm{post}}(A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}(y-\eta_0) + \Gamma_{\mathrm{pr}}^{-1}x_0)$, we obtain

$$\pi^y(x) \propto \exp\left(-\frac{1}{2}(y-Ax-\eta_0)^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}(y-Ax-\eta_0)\right)\exp\left(-\frac{1}{2}(x-x_0)^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}(x-x_0)\right)$$

$$= \exp\Bigg(-\frac{1}{2}\big(y^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}y - y^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}Ax - y^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}\eta_0$$
$$- x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}y + x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}Ax + x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}\eta_0$$
$$- \eta_0^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}y + \eta_0^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}Ax + \eta_0^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}\eta_0$$
$$+ x^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}x - 2x^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}x_0 + x_0^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}x_0\big)\Bigg)$$

Proof. Noting that $\Gamma_{\mathrm{post}} = (\Gamma_{\mathrm{pr}}^{-1} + A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}A)^{-1}$ and $\mu_{\mathrm{post}} = \Gamma_{\mathrm{post}}(A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}(y-\eta_0) + \Gamma_{\mathrm{pr}}^{-1}x_0)$, we obtain

$$\pi^y(x) \propto \exp\left(-\frac{1}{2}(y-Ax-\eta_0)^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}(y-Ax-\eta_0)\right)\exp\left(-\frac{1}{2}(x-x_0)^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}(x-x_0)\right)$$

$$\propto \exp\Bigg(-\frac{1}{2}\Big(\quad -x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}y$$

$$-x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}y + x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}Ax + x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}\eta_0$$

$$+x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}\eta_0^{\mathrm{T}}$$

$$+x^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}x - 2x^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}x_0 \qquad \Big)\Bigg)$$

Proof. Noting that $\Gamma_{\mathrm{post}} = (\Gamma_{\mathrm{pr}}^{-1} + A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}A)^{-1}$ and $\mu_{\mathrm{post}} = \Gamma_{\mathrm{post}}(A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}(y-\eta_0) + \Gamma_{\mathrm{pr}}^{-1}x_0)$, we obtain

$$\pi^y(x) \propto \exp\left(-\frac{1}{2}(y-Ax-\eta_0)^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}(y-Ax-\eta_0)\right)\exp\left(-\frac{1}{2}(x-x_0)^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}(x-x_0)\right)$$

$$\propto \exp\Bigg(-\frac{1}{2}\Big(\quad -x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}y$$

$$-x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}y + x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}Ax + x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}\eta_0$$

$$+x^{\mathrm{T}}A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}\eta_0^{\mathrm{T}}$$

$$+x^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}x - 2x^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}x_0 \quad \Big)\Bigg)$$

$$= \exp\left(-\frac{1}{2}(x^{\mathrm{T}}\underbrace{(\Gamma_{\mathrm{pr}}^{-1} + A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}A)}_{=\Gamma_{\mathrm{post}}^{-1}}x - 2x^{\mathrm{T}}\underbrace{(A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}(y-\eta_0) + \Gamma_{\mathrm{pr}}^{-1}x_0)}_{=\Gamma_{\mathrm{post}}^{-1}\mu_{\mathrm{post}}})\right).$$

On the previous slide, we arrived at

$$\pi^y(x) \propto \exp\left(-\frac{1}{2}\left(x^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}x - 2x^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}\mu_{\mathrm{post}}\right)\right).$$

To finish the proof, we "complete the square" by multiplying and dividing by $\exp(-\frac{1}{2}\mu_{\mathrm{post}}^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}\mu_{\mathrm{post}})$. Since this term does not depend on x, we can absorb the denominator into the implied coefficient to obtain

$$\begin{aligned}
\pi^y(x) &\propto \exp\left(-\frac{1}{2}\left(x^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}x - 2x^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}\mu_{\mathrm{post}}\right)\right) \exp\left(-\frac{1}{2}\mu_{\mathrm{post}}^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}\mu_{\mathrm{post}}\right) \\
&= \exp\left(-\frac{1}{2}\left(x^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}x - 2x^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}\mu_{\mathrm{post}} + \mu_{\mathrm{post}}^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}\mu_{\mathrm{post}}\right)\right) \\
&= \exp\left(-\frac{1}{2}\left((x-\mu_{\mathrm{post}})^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}(x-\mu_{\mathrm{post}}) + 2x^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}\mu_{\mathrm{post}} - 2x^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}\mu_{\mathrm{post}}\right)\right) \\
&= \exp\left(-\frac{1}{2}\left((x-\mu_{\mathrm{post}})^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}(x-\mu_{\mathrm{post}})\right)\right),
\end{aligned}$$

as desired. □

Remark: The previous proof shows that if $x \sim \mathcal{N}(x_0, \Gamma_{\mathrm{pr}})$ and $\eta \sim \mathcal{N}(\eta_0, \Gamma_{\mathrm{n}})$, then

$$x|y \sim \mathcal{N}(\mu_{\mathrm{post}}, \Gamma_{\mathrm{post}}),$$

where

$$\Gamma_{\mathrm{post}} = (\Gamma_{\mathrm{pr}}^{-1} + A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}A)^{-1} \tag{2}$$

and

$$\mu_{\mathrm{post}} = \Gamma_{\mathrm{post}}(A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}(y - \eta_0) + \Gamma_{\mathrm{pr}}^{-1}x_0). \tag{3}$$

One also has the following alternative representations for the posterior mean

$$\mu_{\mathrm{post}} = x_0 + \Gamma_{\mathrm{pr}}A^{\mathrm{T}}(A\Gamma_{\mathrm{pr}}A^{\mathrm{T}} + \Gamma_{\mathrm{n}})^{-1}(y - Ax_0 - \eta_0) \tag{4}$$

and the posterior covariance

$$\Gamma_{\mathrm{post}} = \Gamma_{\mathrm{pr}} - \Gamma_{\mathrm{pr}}A^{\mathrm{T}}(A\Gamma_{\mathrm{pr}}A^{\mathrm{T}} + \Gamma_{\mathrm{n}})^{-1}A\Gamma_{\mathrm{pr}}. \tag{5}$$

Formula (5) can be proved, e.g., by using the Sherman–Morrison–Woodbury formula on (2). Formula (4) can be proved by plugging the formula (5) into (3) and simplifying the expression (homework).

As the posterior distribution is Gaussian, its mean and its mode coincide. This means that the conditional mean estimator and the MAP estimator coincide in the linear Gaussian setting.

Corollary

The conditional mean estimator and the maximum a posteriori estimator coincide in the linear Gaussian setting and are given by
$\hat{x}_{\mathrm{CM}} = \hat{x}_{\mathrm{MAP}} = \mu_{\mathrm{post}}$.

Example

Let $\Gamma_{\mathrm{n}} = \gamma^2 I$, $\eta_0 = 0$, $\Gamma_{\mathrm{pr}} = \sigma^2 I$, $x_0 = 0$, and set $\lambda = \frac{\gamma^2}{\sigma^2}$. Then μ_{post} minimizes

$$J_\lambda(x) := \|y - Ax\|^2 + \lambda \|x\|^2.$$

and therefore satisfies

$$(A^{\mathrm{T}} A + \lambda I)\mu_{\mathrm{post}} = A^{\mathrm{T}} y. \tag{6}$$

This example provides a connection between Bayesian inference and variational regularization: J_λ can be interpreted as the objective functional in a linear regression model with a regularization term $\lambda\|x\|^2$. Equation (6) for μ_{post} is then exactly the normal equation. In the general case, equation $\mu_{\mathrm{post}} = (\Gamma_{\mathrm{pr}}^{-1} + A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}A)^{-1}(A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}(y - \eta_0) + \Gamma_{\mathrm{pr}}^{-1}x_0)$ can thus be viewed as a generalized normal equation. This point of view helps to understand the structure of Bayesian regularization by linking it to well-understood optimization approaches for inverse problems.

Numerical example: one-dimensional deconvolution

Let us revisit the deconvolution example from last week: we are interested in estimating a signal $f : [0, 1] \to \mathbb{R}$ from noisy, blurred observations modeled as

$$y_i = y(s_i) = \int_0^1 K(s_i, t) f(t) \, \mathrm{d}t + \eta_i, \quad i \in \{1, \ldots, k\},$$

where the blurring kernel is

$$K(s, t) = \exp\left(-\frac{1}{2\omega^2}(s - t)^2 \right), \quad \omega = 0.5,$$

and we have Gaussian measurement noise $\eta \sim \mathcal{N}(\eta_0, \Gamma_{\text{noise}})$ with a symmetric, positive definite covariance matrix Γ_{noise}.

If $s_i = \frac{i}{k} - \frac{1}{2k}$ for $i \in \{1, \ldots, k\}$ and we discretize the integral using the midpoint rule with $t_j = \frac{j}{d} - \frac{1}{2d}$ and $x_j = f(t_j)$ for $j \in \{1, \ldots, d\}$, then we have the discrete linear model

$$y = Ax + \eta, \quad \text{where } A_{i,j} = \frac{1}{d} K(s_i, t_j).$$

Linear Gaussian setting

Suppose that we set a Gaussian prior for the unknown $x \sim \mathcal{N}(x_0, \Gamma_{\mathrm{pr}})$, where Γ_{pr} is a symmetric, positive definite covariance matrix.

Now the posterior probability density of x given the measurement y is

$$\pi^y(x) \propto \exp\left(-\frac{1}{2}(x-\overline{x})^{\mathrm{T}}\Gamma_{\mathrm{post}}^{-1}(x-\overline{x})\right),$$

where we have the posterior mean

$$\overline{x} = x_0 + \Gamma_{\mathrm{pr}}A^{\mathrm{T}}(A\Gamma_{\mathrm{pr}}A^{\mathrm{T}} + \Gamma_{\mathrm{noise}})^{-1}(y - Ax_0 - \eta_0)$$

and posterior covariance

$$\Gamma_{\mathrm{post}} = \Gamma_{\mathrm{pr}} - \Gamma_{\mathrm{pr}}A^{\mathrm{T}}(A\Gamma_{\mathrm{pr}}A^{\mathrm{T}} + \Gamma_{\mathrm{noise}})^{-1}A\Gamma_{\mathrm{pr}}.$$

With additive noise $\eta \sim \nu(\eta) = \mathcal{N}(\eta_0, \sigma^2 I)$, we have the likelihood

$$\mathbb{P}(y|x) = \nu(y - Ax) \propto \exp\left(-\frac{1}{2\sigma^2}\|y - Ax - \eta_0\|^2 \right).$$

Let $L = \mathrm{tridiag}(-1, 2, -1)$ and consider the following priors

$$\pi_{\mathrm{pr},1}(x) \propto \exp\left(-\frac{1}{2\gamma^2}\|x - x_0\|^2 \right) \quad \text{with covariance } \Gamma_{\mathrm{pr},1} = \gamma^2 I;$$

$$\pi_{\mathrm{pr},2}(F) \propto \exp\left(-\frac{1}{2\gamma^2}\|L(x - x_0)\|^2 \right)$$

$$= \exp\left(-\frac{1}{2\gamma^2}(x - x_0)^{\mathrm{T}}(L^{\mathrm{T}}L)(x - x_0) \right) \quad \text{with covariance } \Gamma_{\mathrm{pr},2} = \gamma^2(L^{\mathrm{T}}L)^{-1},$$

where $x_0 \in \mathbb{R}^d$ is the prior mean (assumed to be the same in both cases). Hence (from the previous page)

$$\overline{x_j} = x_0 + \Gamma_{\mathrm{pr},j} A^{\mathrm{T}} G_j^{-1}(y - Ax_0 - \eta_0),$$

$$\Gamma_{\mathrm{post},j} = \Gamma_{\mathrm{pr},j} - \Gamma_{\mathrm{pr},j} A^{\mathrm{T}} G_j^{-1} A \Gamma_{\mathrm{pr},j},$$

where $G_j = A\Gamma_{\mathrm{pr},j}A^{\mathrm{T}} + \Gamma_{\mathrm{noise}}$ and $\Gamma_{\mathrm{noise}} = \sigma^2 I$.

For the numerical experiment, we simulate measurements using the (smooth) ground truth signal

$$f(t) = 8t^3 - 16t^2 + 8t,$$

which satisfies $f(0) = f(1) = 0$. The measurements are contaminated with zero-mean 10% *relative* noise ($\sigma \approx 0.0618$) and we set $d = k = 120$.

Remark: When we simulate the measurement data, it is important to avoid the *inverse crime*. One way to do this is to generate the measurement data using a denser grid and then interpolate the forward solution onto a coarser computational grid, which is actually used to compute the reconstruction.

Since both the prior and the posterior are now Gaussian, we can use the coloring transformation to draw samples from the prior and posterior.

See the scripts `deconv.m` / `deconv.py` on the course webpage!

A note on marginal Gaussian distributions

Let

$$\pi(x) \propto \exp(-\tfrac{1}{2}(x-\mu)^{\mathrm{T}}\Gamma^{-1}(x-\mu))$$

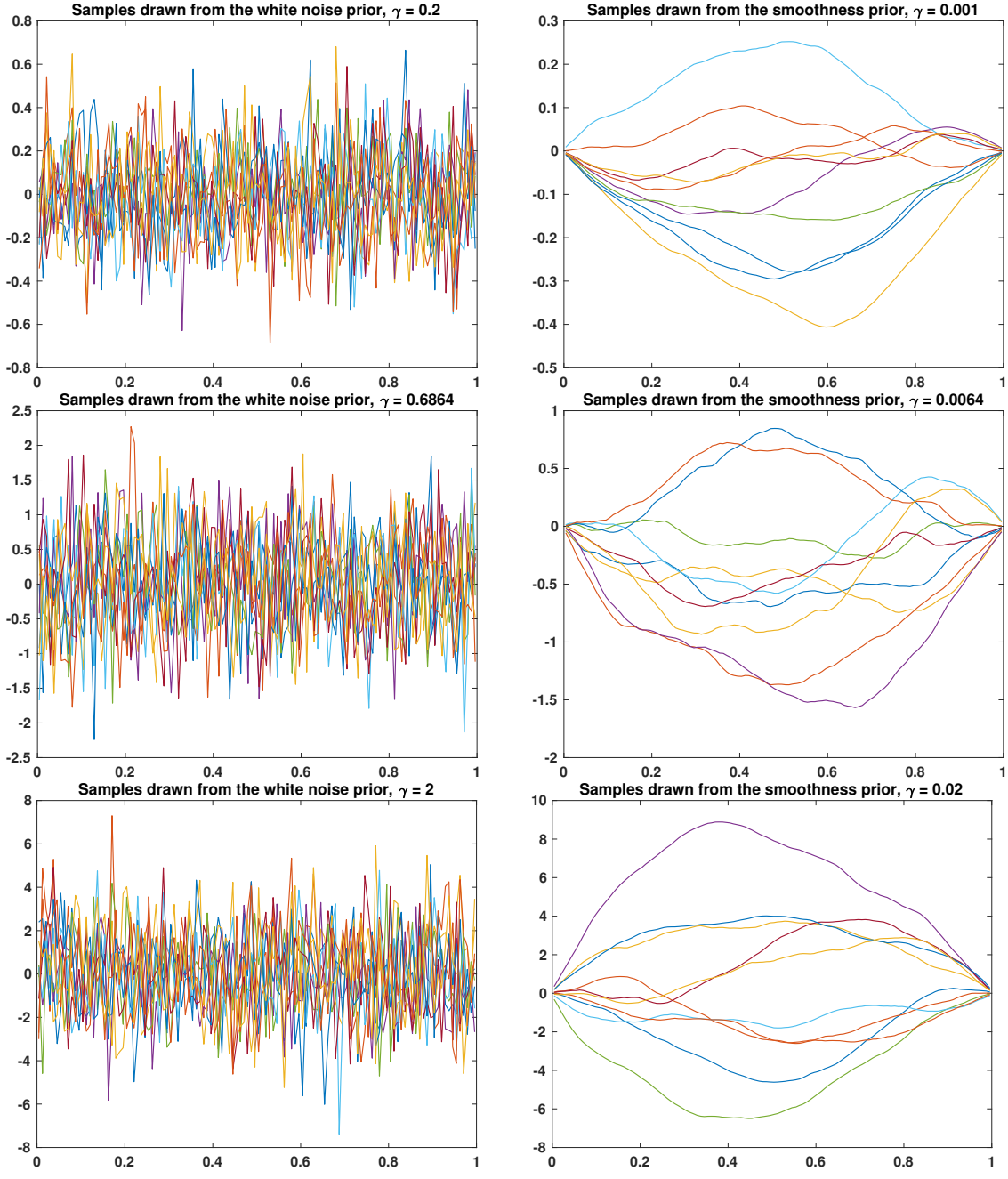
be a multivariate Gaussian PDF with mean μ and positive definite and symmetric covariance matrix Γ.

Q: What is Γ_{ii}?
A: $\sigma_i^2 := \Gamma_{ii}$ is the variance of the marginal distribution with PDF

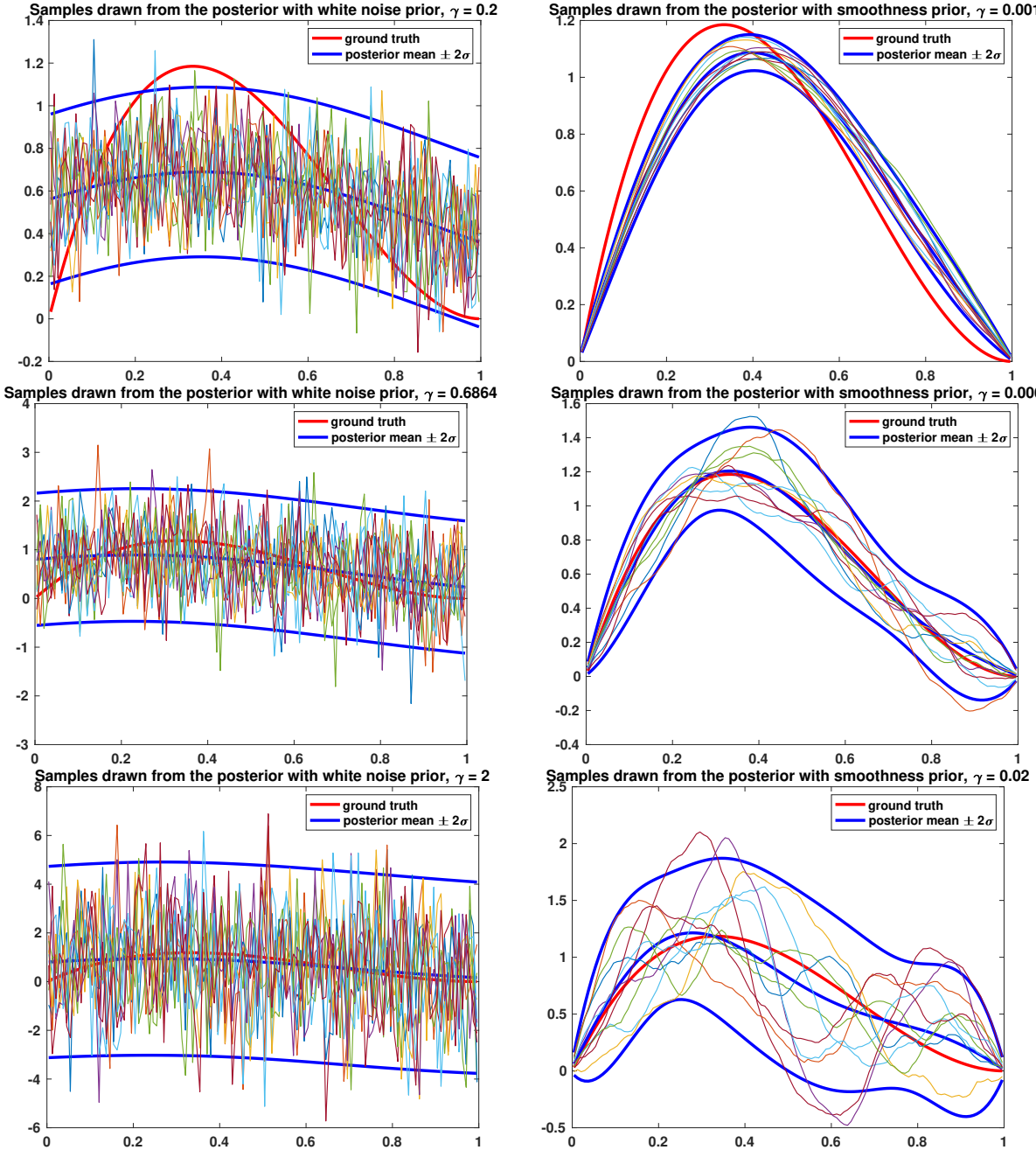
$$\pi(x_i) = \int_{\mathbb{R}^{n-1}} \pi(x_1, \ldots, x_i, \ldots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_{i-1}\, \mathrm{d}x_{i+1} \cdots \mathrm{d}x_n,$$

which is itself a (univariate) Gaussian PDF with mean μ_i.

This is why we can obtain the credibility envelopes by taking the square roots of the diagonal values of $\Gamma_{\mathrm{post},j}$.



Samples drawn from the white noise prior and the smoothness prior for several values of γ.



Samples drawn from the posterior corresponding to both the white noise prior and the smoothness prior for several values of γ. We also plot the ground truth solution and the posterior mean.