# Bayesian inversion

Lecture notes, spring 2016 course

**Samuli Siltanen**

March 16, 2016

## Contents

# 1 Real-valued random variables

## 1.1 Basic definitions

Let $\pi : \mathbb{R} \to \mathbb{R}^+$ be a probability density function satisfying

$$\int_{-\infty}^{\infty} \pi(x)dx = 1. \tag{1}$$

We consider a random variable $X$, taking values in $\mathbb{R}$, whose probability distribution is described by the function $\pi$. The interpretation is that the probability of a randomly sampled value of $X$ belonging to the interval $[a, b]$ is given by the integral $\int_a^b \pi(x)dx$. More generally, if $E \subset \mathbb{R}$ is a Lebesgue measurable set, then

$$\Pr(X \in E) = \int_E \pi(x)dx.$$

For more general cases, when the probability distribution of $X$ is described by a measure (and not necessarily by a probability density function), see for example [1].

Define the cumulative distribution function $F : \mathbb{R} \to [0, 1]$ by

$$F(x) := \int_{-\infty}^{x} \pi(x)dx. \tag{2}$$

## 1.2 Sampling a random variable

We are interested in producing algorithmically a random sequence $x^{(1)}, x^{(2)}, x^{(3)}, \ldots, x^{(N)}$ of real numbers in such a way that their distribution follows a given probability density function $\pi : \mathbb{R} \to \mathbb{R}^+$ satisfying condition (1).

The trick is two-fold:

1. Use the Matlab command `rand` to produce a sequence $t^{(1)}, t^{(2)}, t^{(3)}, \ldots, t^{(N)}$ of floating point numbers (pseudo-)randomly picked from the uniform probability distribution on the interval $[0, 1]$.
2. Define $x^{(\ell)} := F^{-1}(t^{(\ell)})$ for $\ell = 1, \ldots, N$.

Of course, there are several things to check for ensuring that the above trick works. Most notably, is it well-defined to apply the inverse function $F^{-1}$? Also, how to make sure that the resulting points are correctly distributed? These are left as exercises.

## 1.3 Computational example

Let us consider the special case

$$\pi(x) = \begin{cases} \frac{3}{4}(1 - x^2), & \text{for } |x| < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

See Figure 1 for plots of the above probability density function and the corresponding cumulative distribution function. Figure 2 shows the cumulative distribution function and its inverse. See Figure 3 for a comparison of sample histograms with the probability density function.

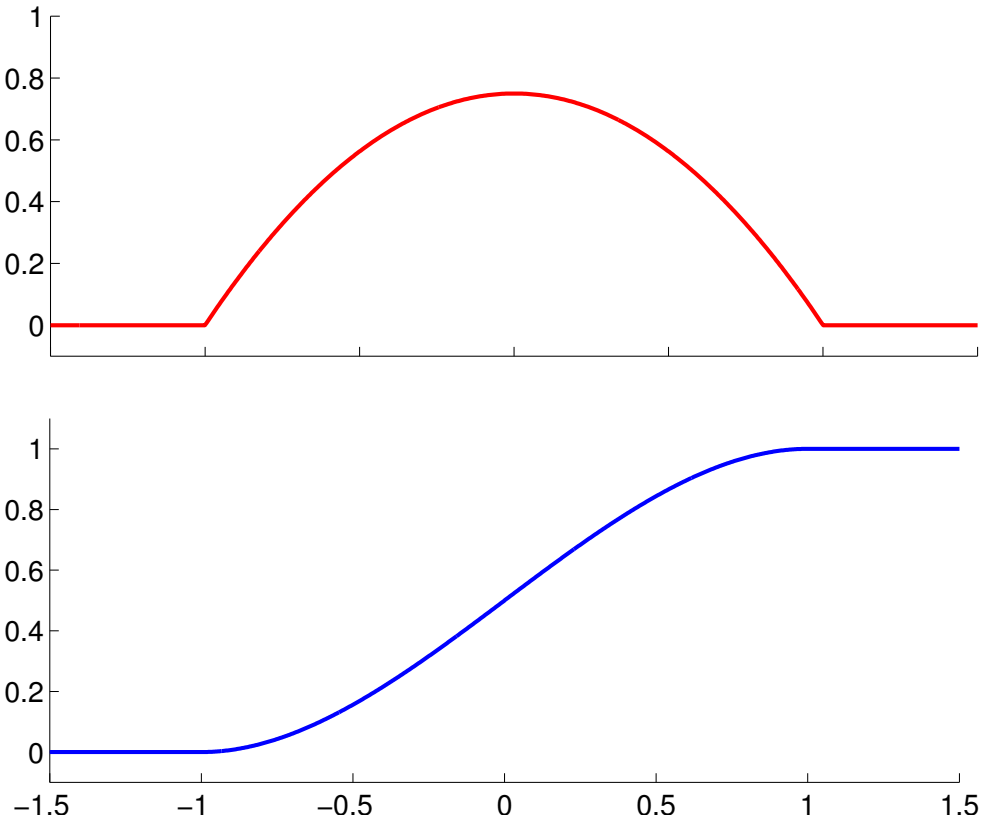


Figure 1: Top: Probability density function $\pi(x)$. Bottom: Cumulative distribution function.

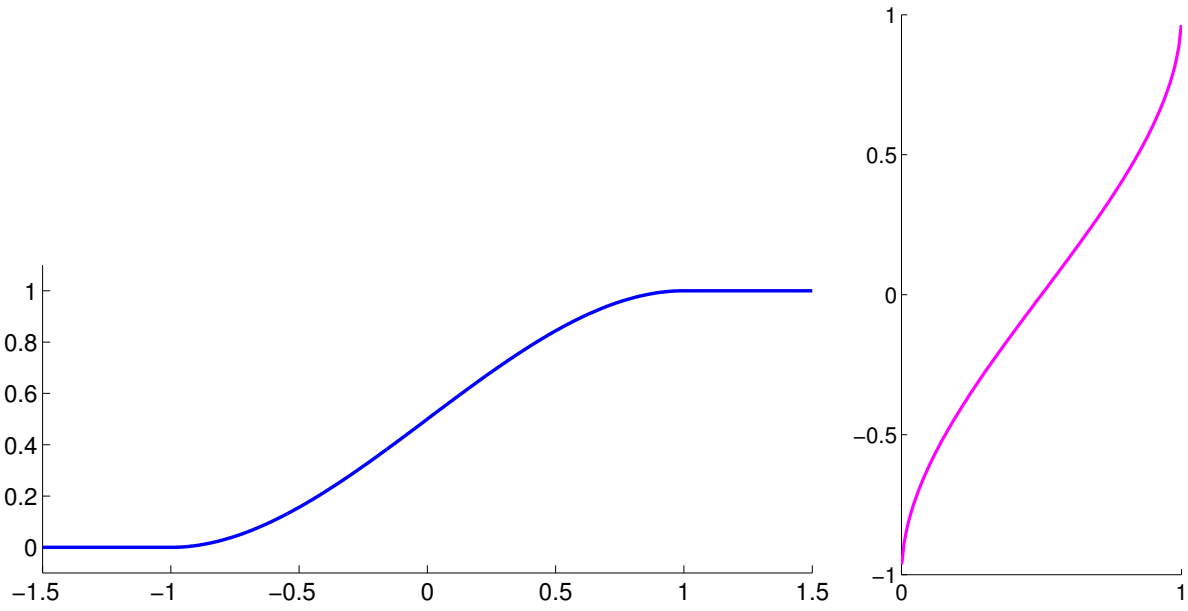


Figure 2: Left: Cumulative distribution function $F : \mathbb{R} \to [0, 1]$. Right: Inverse $F^{-1} : [0, 1] \to \mathbb{R}$ of the cumulative distribution function.

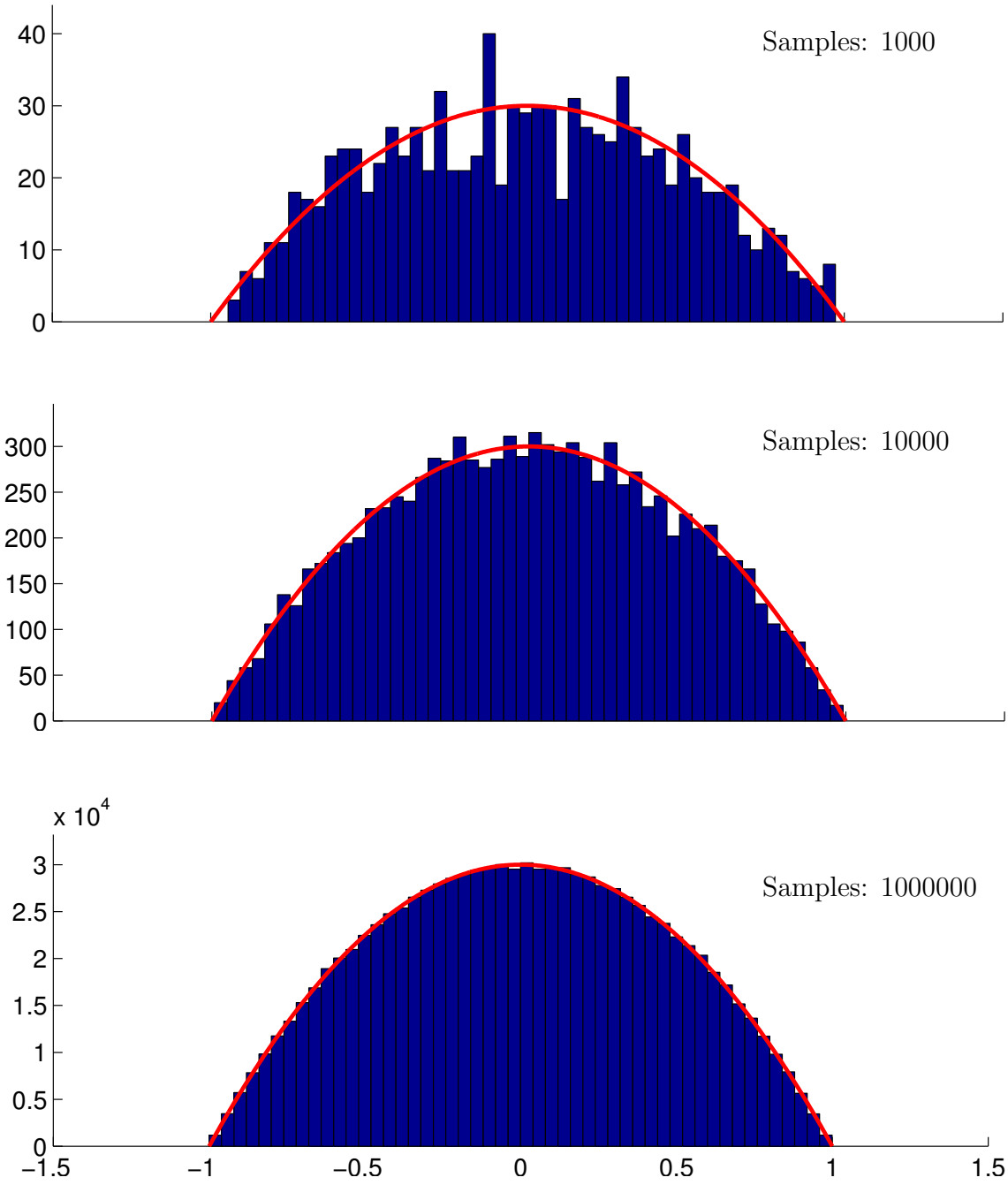


Figure 3: Top: histogram of 1000 random samples. Middle: histogram of 10000 random samples. Bottom: histogram of 1000000 random samples.

### 1.4 The Bayes formula

Let us consider a joint probability density $\pi_{XY} : \mathbb{R}^2 \to \mathbb{R}$ of two $\mathbb{R}$-valued random variables $X$ and $Y$. We must have

$$\begin{aligned} \pi_{XY}(x,y) &\geq 0 \text{ for all } x, y \in \mathbb{R}, && (4) \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \pi_{XY}(x,y)dxdy &= 1. && (5) \end{aligned}$$

Now the probability that a sampled pair $(x^{(1)}, y^{(1)})$ belongs to the rectangle $[a,b] \times [c,d]$ is given by the integral

$$\Pr(a \leq x^{(1)} \leq b \text{ and } c \leq y^{(1)} \leq d) = \int_a^b \int_c^d \pi_{XY}(x,y)dxdy.$$

Now we can define the *marginal distributions* of $X$ and $Y$ by

$$\pi_X(x) = \int_{-\infty}^{\infty} \pi_{XY}(x,y)dy, \qquad \pi_Y(y) = \int_{-\infty}^{\infty} \pi_{XY}(x,y)dx,$$

respectively. Furthermore, the conditional probability of $X$ given a fixed value of $Y$ is defined by

$$\pi_{X|Y}(x|y) = \frac{\pi_{XY}(x,y)}{\pi_Y(y)}. \tag{6}$$

It is easy to check that

$$\int_{-\infty}^{\infty} \pi_{X|Y}(x|y)dx = 1.$$

Similarly we define the conditional probability of $M$ given a fixed value of $X$ by

$$\pi_{Y|X}(y|x) = \frac{\pi_{XY}(x,y)}{\pi_X(x)}. \tag{7}$$

A combination of (6) and (7) yields the Bayes formula

$$\pi_{X|Y}(x|y) = \frac{\pi_X(x)\,\pi_{Y|X}(y|x)}{\pi_Y(y)}. \tag{8}$$

# 2 Principle of Bayesian inversion

Inverse problems arise in situations where noisy data is measured from an object. The direct problem is *given object, what is the data?* The inverse problem is *given noisy data, recover information about the object.*

For more concrete explanation we need to specify a mathematical model of the measurement.

## 2.1 Measurement model

Consider the equation

$$M = AF + \mathcal{E} \tag{9}$$

that models indirect linear measurement contaminated with additive noise. (We do not discuss nonlinear measuremernts or other than additive noise in this course.) In equation (9)

- The measurement $M$ is a random vector taking values in $\mathbb{R}^k$,
- The noise $\mathcal{E}$ is a random vector taking values in $\mathbb{R}^k$, with probability density function $\pi_{\mathcal{E}}$,
- The object $F$ is a random vector taking values in $\mathbb{R}^n$,
- The deterministic (in other words, not random) $k \times n$ matrix $A$ models the measurement process.

The randomness of $\mathcal{E}$, and therefore that of $M$, arises from the nature of inevitable random errors present in practical measurement devices. The randomness of the object $F$ models our lack of information about it. This is not seen here as a philosophical issue but rather as a pragmatic modelling choice.

## 2.2 The inverse problem

Assume that your measurement device produces a vector $m \in \mathbb{R}^k$. Further, assuming that equation (9) is a reasonably accurate model of the measurement process, we can formulate the inverse problem as follows:

*Given a realization $m$ of the random variable $M$, estimate the object $F$.*

Often the estimate takes the form of a suitable $n$-dimensional vector, possibly augmented with uncertainty quantification such as credibility intervals. Most important examples of useful estimates include the *maximum a posteriori (MAP)* estimate and the *conditional mean (CM)* estimate, which will be defined below.

## 2.3 Posterior density as the solution of inverse problem

Recall that in ill-posed inverse problems the measurement information is typically not sufficient for stable recovery of information about the object. Therefore, robust inversion is always based on complementing measurement data with *a priori* information. This can be done using the Bayes formula.

Given the measurement model (9), we can write the Bayes formula connecting the object and measurement:

$$\pi_{\text{post}}(f|m) = \frac{\pi_F(f)\,\pi_{\text{like}}(m|f)}{\pi_M(m)}. \tag{10}$$

In equation (10), each function on the right-hand side has a special meaning:

- The *prior model* $\pi_F(f)$ describes *a priori information*. The function $\pi_F(f)$ should assign high probability to objects $f$ that are typical in light of *a priori information*, and low probability to unexpected $f$. It is a central challenge in Bayesian inversion to construct a function $\pi_F$ that describes *a priori information* accurately and is quick to evaluate computationally.

- The *likelihood model* $\pi_{\text{like}}(m|f)$ processes measurement information. It gives low probability to objects that produce simulated data which is very different from the measured data.

- The number $\pi_M(m)$ can be seen as a normalization constant.

In ill-posed inverse problems there may be infinitely many objects giving roughly the maximum likelihood probability, making the use of a prior model necessary.

## 2.4 A simple example: measuring temperature

Consider checking the temperature in the morning before leaving for the university. Assume that your thermometer showed 3°C. (This is Helsinki, Finland, so it's not very warm. I recommend the Java island for friends of warmth.)

We model the reading of the thermometer like this:

$$M = F + \mathcal{E}, \tag{11}$$

where $M, F$ and $\mathcal{E}$ are random variables all taking values in $\mathbb{R}$. Furthermore, we assume that the additive noise $\mathcal{E}$ is normally distributed with mean $\mu = 0$ and standard deviation $\sigma > 0$:

$$\pi_{\mathcal{E}}(\varepsilon) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{1}{2\sigma^2}(\varepsilon - \mu)^2). \tag{12}$$

Note that formula (12) describes the classical Gaussian bell curve.

Then the likelihood model takes the form

$$\pi_{\text{like}}(m|f) = \pi_{\mathcal{E}}(m - f) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{1}{2\sigma^2}(m - f)^2), \tag{13}$$

where we used equations (11) and (12).

In this very simple example we use the "flat prior" $\pi_F(f) \equiv 1$, although it does not integrate to 1. We could do something more sensible but just don't. Nevertheless, the posterior distribution will behave well.

Then the posterior distribution takes the following form, up to a constant $C$:

$$\pi_{\text{post}}(f|m) = \frac{\pi_F(f)\,\pi_{\text{like}}(m|f)}{\pi_M(m)} = C \exp(-\frac{1}{2\sigma^2}(m - f)^2). \tag{14}$$

See Figure Now assume that $\sigma = 0.4$ and recall that $m = 3$. The posterior is

$$\pi_{\text{post}}(f|3) = Ce^{-3.125(f-3)^2}. \tag{15}$$

See Figure 5 for a plot.

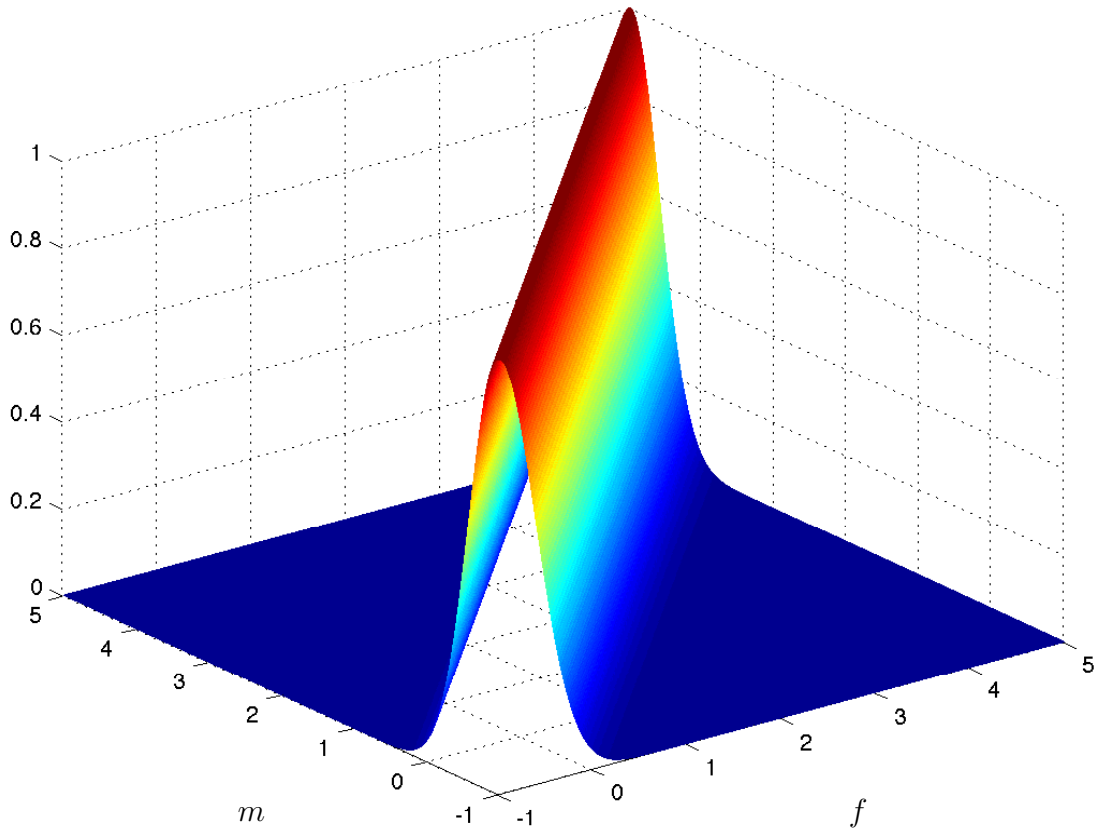


Figure 4: Likelihood function $\pi_{\text{post}}(f|m)$ of equation (13).

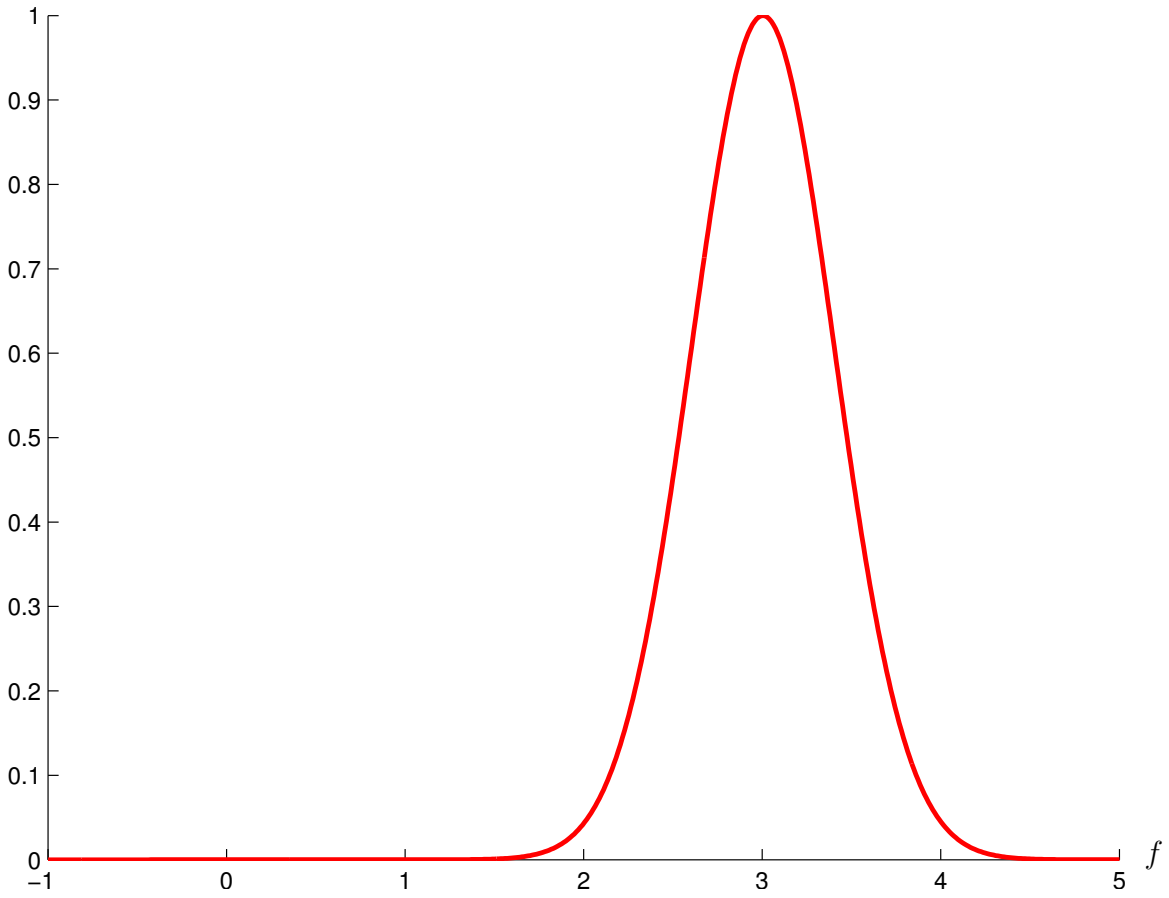


Figure 5: Posterior distribution $\pi_{\text{post}}(f|3)$ of equation (15).

## 2.5 Drawing estimates from the posterior

It is often impractical to offer the full posterior distribution as the solution of the inverse problem. In practice it is desirable to calculate some estimates, and perhaps some quantitative information about their reliability, from the posterior.

The two most popular estimates are *conditional mean* estimate $f^{\mathrm{CM}}$ and the *maximum a posteriori* estimate $f^{\mathrm{MAP}}$, defined as follows:

$$f^{\mathrm{CM}} := \int_{\mathbb{R}^n} f\pi_{\mathrm{post}}(f|m)\,df, \tag{16}$$

$$f^{\mathrm{MAP}} := \arg\max_{f\in\mathbb{R}^n} \pi_{\mathrm{post}}(f|m). \tag{17}$$

## 2.6 Determining the MAP estimate in the Gaussian case

Assume that the likelihood is

$$\pi_{\mathrm{like}}(m|f) = \left(\frac{1}{2\pi\sigma^2}\right)^{k/2} \exp(-\frac{1}{2\sigma^2}\|Af-m\|^2)$$

for some $\sigma > 0$. Further, take the prior to be

$$\pi_F(f) = \left(\frac{1}{2\pi\mathrm{det}(\Gamma)}\right)^{n/2} \exp(-\frac{1}{2}f^T\Gamma^{-1}f)$$

for some symmetric positive-definite covariance matrix $\Gamma = (L^TL)^{-1}$. Now (17) takes the form

$$f^{\mathrm{MAP}} = \arg\max_{f\in\mathbb{R}^n} \left(\exp\Big(-\frac{1}{2}(\|Lf\|^2 + \frac{1}{\sigma^2}\|Af-m\|^2)\Big)\right). \tag{18}$$

Clearly, (18) is equivalent to the minimization problem

$$f^{\mathrm{MAP}} = \arg\min_{f\in\mathbb{R}^n} \left(\frac{1}{\sigma^2}\|Af-m\|^2 + \|Lf\|^2\right). \tag{19}$$

Assume that the quadratic functional $\frac{1}{\sigma^2}\|Af-m\|^2+\|Lf\|^2$ has a unique minimum. (It turns out that the invertibility of the matrix $\frac{1}{\sigma^2}A^TA+L^TL$ is a sufficient condition for that.) Then the minimizer $f^{\mathrm{MAP}}$ satisfies

$$0 = \frac{d}{dt}\left\{\frac{1}{\sigma^2}\|A(f^{\mathrm{MAP}}+tw)-m\|^2 + \|L(f^{\mathrm{MAP}}+tw)\|^2\right\}\Big|_{t=0}$$

for any $w\in\mathbb{R}^n$. Here $t\in\mathbb{R}$.

Compute

$$\begin{aligned}
&\frac{d}{dt}\|A(f^{\text{MAP}}+tw)-m\|^2\Big|_{t=0}\\
=&\frac{d}{dt}\langle Af^{\text{MAP}}+tAw-m, Af^{\text{MAP}}+tAw-m\rangle\Big|_{t=0}\\
=&\frac{d}{dt}\Big\{\|Af^{\text{MAP}}\|^2+2t\langle Af^{\text{MAP}}, Aw\rangle+t^2\|Aw\|^2\\
&\qquad -2t\langle m, Aw\rangle-2\langle Af^{\text{MAP}}, m\rangle+\|m\|^2\Big\}\Big|_{t=0}\\
=&2\langle Af^{\text{MAP}}, Aw\rangle-2\langle m, Aw\rangle,
\end{aligned}$$

and

$$\begin{aligned}
&\frac{d}{dt}\langle Lf^{\text{MAP}}+tLw, Lf^{\text{MAP}}+tLw\rangle\Big|_{t=0}\\
=&\frac{d}{dt}\left\{\|Lf^{\text{MAP}}\|^2+2t\langle Lf^{\text{MAP}}, Lw\rangle+t^2\|Lw\|^2\right\}\Big|_{t=0}\\
=&2\langle Lf^{\text{MAP}}, Lw\rangle.
\end{aligned}$$

Thus, we have $\frac{1}{\sigma^2}\langle Af^{\text{MAP}}-m, Aw\rangle+\langle Lf^{\text{MAP}}, Lw\rangle=0$, and by taking transposes,

$$\frac{1}{\sigma^2}\langle A^TAf^{\text{MAP}}-A^Tm, w\rangle+\langle L^TLf^{\text{MAP}}, w\rangle=0.$$

This results in the variational form

$$\langle(\frac{1}{\sigma^2}A^TA+L^TL)f^{\text{MAP}}-\frac{1}{\sigma^2}A^Tm, w\rangle=0. \tag{20}$$

Since (20) holds for any nonzero $w\in\mathbb{R}^n$, we have $(\frac{1}{\sigma^2}A^TA+L^TL)f^{\text{MAP}}=\frac{1}{\sigma^2}A^Tm$. So the MAP estimate satisfies

$$f^{\text{MAP}}=(\frac{1}{\sigma^2}A^TA+L^TL)^{-1}\frac{1}{\sigma^2}A^Tm, \tag{21}$$

and actually (21) can be used for computing $f^{\text{MAP}}$ in case the inverse $(\frac{1}{\sigma^2}A^TA+L^TL)^{-1}$ exists.

**Exercise.** Add a nonzero mean to the prior:

$$\pi_F(f)=\left(\frac{1}{2\pi\det(\Gamma)}\right)^{n/2}\exp(-\frac{1}{2}(f-f_0)^T\Gamma^{-1}(f-f_0)).$$

Show that equation (21) takes the form

$$f^{\text{MAP}}=(\frac{1}{\sigma^2}A^TA+L^TL)^{-1}(\Gamma^{-1}f_0+\frac{1}{\sigma^2}A^Tm). \tag{22}$$

## 2.7 MAP estimate in the case of sparsity-promoting priors

We want to compute the vector $f^{\text{MAP}} \in \mathbb{R}^n$ defined by

$$f^{\text{MAP}} = \arg\min_{f \in \mathbb{R}^n} \left( \frac{1}{\sigma^2} \|Af - m\|^2 + \|Lf\|_1 \right). \tag{23}$$

We write the vector $Lf \in \mathbb{R}^n$ in the form

$$\mathbf{v}_+ - \mathbf{v}_- = Lf,$$

where $\mathbf{v}_\pm$ are nonnegative vectors: $\mathbf{v}_\pm \in \mathbb{R}^n_+$, or $(\mathbf{v}_\pm)_j \geq 0$ for all $j = 1, \ldots, n$. Now $f^{\text{MAP}}$ can be seen to be a the minimizer of

$$\frac{1}{\sigma^2} \|Af\|_2^2 - \frac{2}{\sigma^2} m^T Af + \mathbf{1}^T \mathbf{v}_+ + \mathbf{1}^T \mathbf{v}_-,$$

where $\mathbf{1}$ is the vector with all elements equal to one: $\mathbf{1} = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}^T \in \mathbb{R}^n$, and the minimization is taken over $y \in \mathbb{R}^{3n}$ defined by

$$\mathbf{y} = \begin{bmatrix} f \\ \mathbf{v}_+ \\ \mathbf{v}_- \end{bmatrix}, \quad \text{where} \quad \begin{array}{l} f \in \mathbb{R}^n \\ \mathbf{v}_+ \in \mathbb{R}^n_+ \\ \mathbf{v}_- \in \mathbb{R}^n_+ \end{array}.$$

Note the identity $\|Af\|_2^2 = f^T A^T Af$ and write

$$H = \begin{bmatrix} \frac{2}{\sigma^2} A^T A & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad h = \begin{bmatrix} -\frac{2}{\sigma^2} A^T m \\ \alpha \mathbf{1} \\ \alpha \mathbf{1} \end{bmatrix}.$$

We then have the quadratic optimization problem in standard form

$$\arg\min_{\mathbf{y}} \left\{ \frac{1}{2} \mathbf{y}^T H \mathbf{y} + h^T \mathbf{y} \right\} \tag{24}$$

with the constraints

$$L \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} y_{n+1} \\ \vdots \\ y_{2n} \end{bmatrix} - \begin{bmatrix} y_{2n+1} \\ \vdots \\ y_{3n} \end{bmatrix} \tag{25}$$

and

$$y_j \geq 0 \text{ for } j = n+1, \ldots, 3n. \tag{26}$$

Several software packages (such as `quadprog.m` routine in MATLAB's Optimization Toolbox) exist that can deal with a problem of the form (24) with constraints of type (25).

One downside of the above approach is that the optimization problem (24) has $3n$ degrees of freedom, whereas the original problem (23) has only $n$. Numerical optimization becomes harder in higher dimensions. However, the advantage is that (24) is in a well-understood standard form.

# References

[1] Shiryaev, A. N.: *Probability*, Graduate Texts in Mathematics **95**, Second Edition, Springer-Verlag, New York, 1996.