Computational inverse problems

Nuutti Hyvönen, Jenni Heino,
and Juha-Pekka Puska

nuutti.hyvonen@aalto.fi, juha-pekka.puska@aalto.fi

Seventh lecture, March 22, 2021.

Example: Random variables waiting for the train

Assume that every day, except on Sundays, a train for your destination leaves every S minutes from the station. On Sundays, the interval between trains is $2S$ minutes. You arrive at the station with no information about the timetable of the trains (or of the day!!). What is your expected waiting time?

Define a random variable, $T =$ waiting time, whose distribution on working days is

$$T \sim \pi(t \,|\, \text{working day}) = \frac{1}{S}\chi_S(t), \qquad \chi_S(t) = \begin{cases} 1, & 0 \le t < S, \\ 0, & \text{otherwise.} \end{cases}$$

On Sundays, the distribution of T is

$$T \sim \pi(t \,|\, \text{Sunday}) = \frac{1}{2S}\chi_{2S}(t).$$

On a working day, the expected waiting time is

$$E\{T \,|\, \text{working day}\} = \int t\pi(t \,|\, \text{working day})dt = \frac{1}{S}\int_0^S tdt = \frac{S}{2}.$$

On Sundays, the expected waiting time is two times as long.

If you have no idea which day of the week it is, you can give equal probability to each day. Thus,

$$\pi(\text{working day}) = \frac{6}{7}, \quad \pi(\text{Sunday}) = \frac{1}{7}.$$

To get the expected waiting time regardless of the day of the week, marginalize over the days of the week:

$$\begin{aligned} E\{T\} &= E\{T \,|\, \text{working day}\}\pi(\text{working day}) + E\{T \,|\, \text{Sunday}\}\pi(\text{Sunday}) \\ &= \frac{3S}{7} + \frac{S}{7} = \frac{4S}{7}. \end{aligned}$$

Example: Poisson distribution

A weak light source emits photons that are counted with a CCD (*Charged Coupled Device*). The counting process $N(t)$,

$$N(t) = \text{number of particles observed in } [0, t] \in \mathbb{N}$$

is an integer-valued random variable.

Under some assumptions, it can be shown that N is a *Poisson process*:

$$P\{N(t) = n\} = \frac{(\lambda t)^n}{n!} e^{-\lambda t}, \quad \lambda > 0.$$

We now fix $t = T =$ the recording time, define a random variable $N = N(T)$, and let $\theta = \lambda T$. We write

$$N \sim \text{Poisson}(\theta).$$

We want to calculate the expectation and variance of this Poisson random variable. Since the discrete probability density is

$$\pi(n) = P\{N = n\} = \frac{\theta^n}{n!} e^{-\theta}, \quad \theta > 0,$$

and our random variable takes on discrete values, in the definition of the expectation we have an infinite sum instead of an integral (a countable number of probability masses), that is

$$
\begin{aligned}
\mathrm{E}\{N\} &= \sum_{n=0}^{\infty} n\pi(n) = e^{-\theta} \sum_{n=0}^{\infty} n \frac{\theta^n}{n!} \\
&= e^{-\theta} \sum_{n=1}^{\infty} \frac{\theta^n}{(n-1)!} = e^{-\theta} \sum_{n=0}^{\infty} \frac{\theta^{n+1}}{n!} \\
&= \theta e^{-\theta} \underbrace{\sum_{n=0}^{\infty} \frac{\theta^n}{n!}}_{e^{\theta}} = \theta.
\end{aligned}
$$

We calculate the variance of a Poisson random variable in a similar way, writing first

$$\begin{aligned}\mathrm{var}(N) = \mathrm{E}\{(N-\theta)^2\} &= \mathrm{E}\{N^2\} - 2\theta \underbrace{\mathrm{E}\{N\}}_{=\theta} + \theta^2 \\ &= \mathrm{E}\{N^2\} - \theta^2 \\ &= \sum_{n=0}^{\infty} n^2 \pi(n) - \theta^2.\end{aligned}$$

Substituting the expression of $\pi(n)$, we thus get

$$
\begin{aligned}
\operatorname{var}(N) &= e^{-\theta} \sum_{n=0}^{\infty} n^2 \frac{\theta^n}{n!} - \theta^2 = e^{-\theta} \sum_{n=1}^{\infty} n \frac{\theta^n}{(n-1)!} - \theta^2 \\
&= e^{-\theta} \sum_{n=0}^{\infty} (n+1) \frac{\theta^{n+1}}{n!} - \theta^2 \\
&= \theta e^{-\theta} \sum_{n=0}^{\infty} n \frac{\theta^n}{n!} + \theta e^{-\theta} \sum_{n=0}^{\infty} \frac{\theta^n}{n!} - \theta^2 \\
&= \theta e^{-\theta} \left((\theta + 1) e^{\theta} \right) - \theta^2 \\
&= \theta,
\end{aligned}
$$

that is, the mean and the variance coincide.

Normal distributions

A random variable $X \in \mathbb{R}$ is normally distributed, or Gaussian, i.e.,

$$X \sim \mathcal{N}(x_0, \sigma^2),$$

if

$$P\{X \leq t\} = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{t} \exp\left(-\frac{1}{2\sigma^2}(x - x_0)^2\right) dx.$$

For $X \sim \mathcal{N}(x_0, \sigma^2)$, it holds that

$$\mathrm{E}\{X\} = x_0, \quad \mathrm{var}(X) = \sigma^2.$$

As a generalization, $X \in \mathbb{R}^n$ is Gaussian if its probability density is

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

where $x_0 \in \mathbb{R}^n$, and $\Gamma \in \mathbb{R}^{n \times n}$ is symmetric and positive definite.

Gaussian random variables are widely used in statistics. They appear naturally when *macroscopic* measurements are averages of individual *microscopic* random effects.

Examples: pressure and temperature.

The Central Limit Theorem sheds light on this:

Central Limit Theorem: Assume that real valued random variables $X_1, X_2, \ldots$ are independent and identically distributed, each with expectation μ and variance σ^2. Then the distribution of

$$Z_n = \frac{1}{\sigma\sqrt{n}}(X_1 + X_2 + \ldots + X_n - n\mu)$$

converges to the distribution of a standard normal random variable

$$\lim_{n\to\infty} P\{Z_n \leq x\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt.$$

Another interpretation of the Central Limit Theorem: If

$$Y_n = \frac{1}{n}\sum_{j=1}^{n} X_j,$$

then for large n a good approximation for the probability distribution of Y is

$$Y \sim \mathcal{N}\Big(\mu, \frac{\sigma^2}{n}\Big).$$

Example: Poisson distribution (revisited)

One implication of the Central Limit Theorem is that the Poisson distribution can be approximated with a Gaussian distribution if the expectation θ is large.

Intuitive reasoning based on the CCD camera: Assume for simplicity that the expectation θ is a positive integer. The total photon count can then be viewed as a sum of sub-counts on $\theta \in \mathbb{N}$ smaller counter units of equal size. These sub-counts can in turn be viewed as mutually independent Poisson distributed random variables with expectation (and variance) 1. Now, it follows from the Central Limit Theorem that as θ increases, the sum of the sub-counts approaches a normally distributed variable with mean and variance θ.

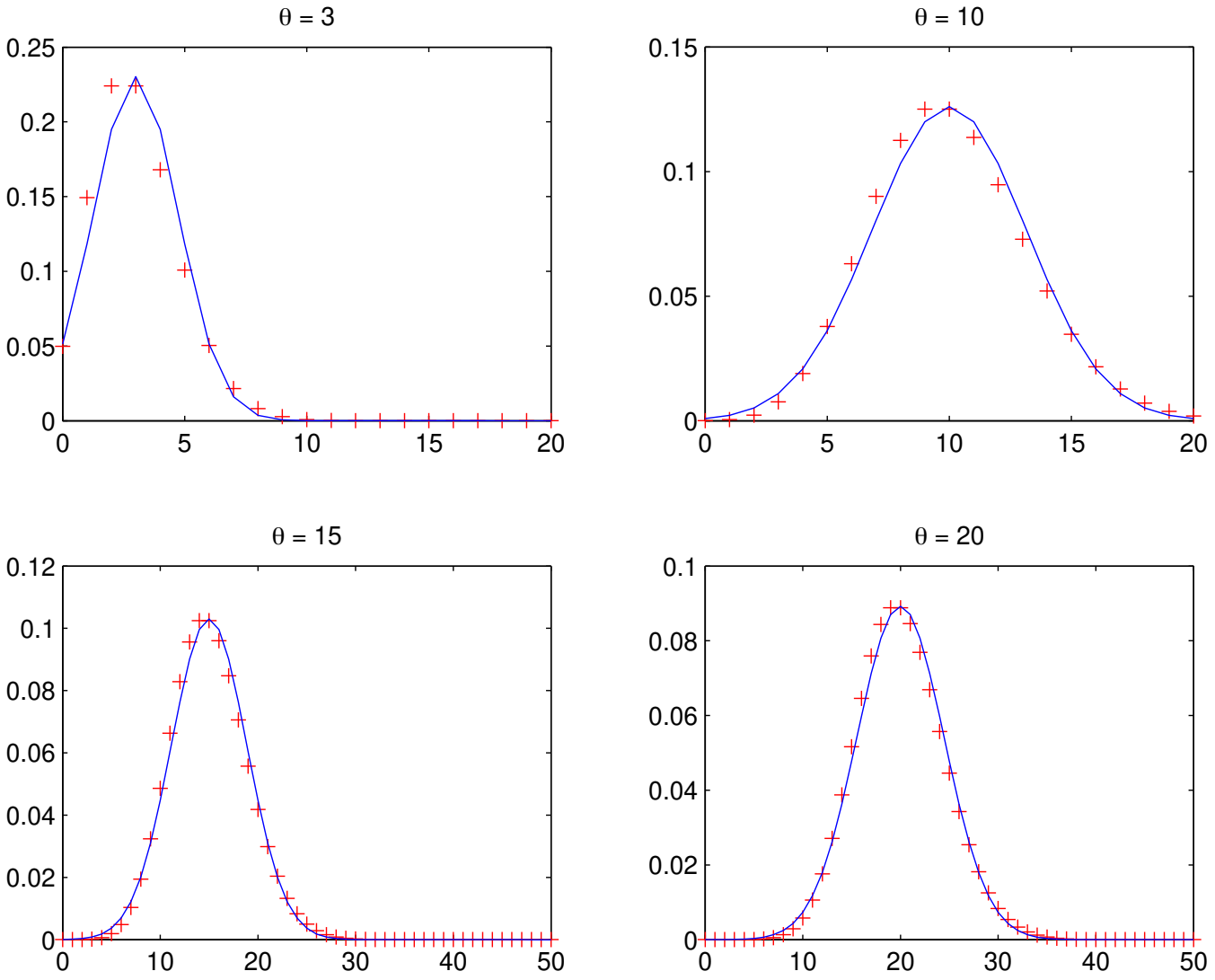
Let us test this hypothesis numerically. We plot the Poisson probability distribution

$$\pi_{\mathrm{Poisson}}(n \,|\, \theta) = \frac{\theta^n}{n!} e^{-\theta}$$

as a function of $n \in \mathbb{N}$, and compare it to the Gaussian approximation

$$\pi_{\mathrm{Gaussian}}(x \,|\, \theta, \theta) = \frac{1}{\sqrt{2\pi\theta}} \exp\left(-\frac{1}{2\theta}(x-\theta)^2\right)$$

as a function of $x \in \mathbb{R}_+$, for increasing values of $\theta > 0$.


θ = 3
0.25
0.2
0.15
0.1
0.05
0
0
5
10
15
20
θ = 10
0.15
0.1
0.05
0
0
5
10
15
20
θ = 15
0.12
0.1
0.08
0.06
0.04
0.02
0
0
10
20
30
40
50
θ = 20
0.1
0.08
0.06
0.04
0.02
0
0
10
20
30
40
50

Inverse problems and Bayes' formula

Classical setup for inverse problems:

$$y = f(x, e),$$

where

- $y \in \mathbb{R}^m$ is the measured quantity,
- $x \in \mathbb{R}^n$ is the quantity we seek to get information about,
- $e \in \mathbb{R}^k$ contains the poorly known parameters and noise, and
- $f : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^m$ is the model.

In the statistical setup, all parameters are viewed as random variables, and the classical model is replaced by

$$Y = f(X, E).$$

Notice that the probability distributions of the three random variables X, Y and E depend on each other.

Nomenclature:
Y is called the *measurement*, and its realization y_{obs} the *data*.
X is the unobservable variable of primary interest and called the *unknown*.
The other variables E that are neither observable nor of primary interest are called *parameters* or *noise*.

Prior density

Even before performing the measurement, we typically have some knowledge about the variable X. This information is coded in a probability density $x \mapsto \pi_{\mathrm{pr}}(x)$ called the *prior density*.

Likelihood function

The conditional probability density of Y in case we know the value of the unknown, i.e., $X = x$, is called the *likelihood function*:

$$\pi(y \,|\, x) = \frac{\pi(x, y)}{\pi_{\mathrm{pr}}(x)}, \qquad \text{if } \pi_{\mathrm{pr}}(x) \neq 0.$$

Posterior density

Given the measurement data $Y = y_{\text{obs}}$, the conditional probability density

$$\pi(x \,|\, y_{\text{obs}}) = \frac{\pi(x, y_{\text{obs}})}{\pi(y_{\text{obs}})}, \qquad \text{if } \pi(y_{\text{obs}}) = \int_{\mathbb{R}^n} \pi(x, y_{\text{obs}}) dx \neq 0,$$

is called the *posterior density* of X.

The posterior density expresses what we know about X after realizing the observation $Y = y_{\text{obs}}$.

Inverse problem in the Bayesian framework

Given the data $Y = y_{\mathrm{obs}}$*, find the conditional probability density* $\pi(x \,|\, y_{\mathrm{obs}})$ *of the variable* X*.*

Bayes theorem of inverse problems

Assume that the random variable $X \in \mathbb{R}^n$ has a known prior probability density $\pi_{\mathrm{pr}}(x)$ and the data consist of the observed value y_{obs} of an observable random variable $Y \in \mathbb{R}^m$ such that $\pi(y_{\mathrm{obs}}) > 0$. Then, the posterior probability density of X, given the data y_{obs}, is

$$\pi_{\mathrm{post}}(x) = \pi(x \,|\, y_{\mathrm{obs}}) = \frac{\pi_{\mathrm{pr}}(x)\pi(y_{\mathrm{obs}} \,|\, x)}{\pi(y_{\mathrm{obs}})}.$$

In practice, the marginal density $\pi(y_{\mathrm{obs}})$ plays a role of a norming constant and is often not important.

Solving an inverse problem in the Bayesian framework

1. Based on all available prior information on the unknown X, find a prior probability density π_{pr} that reflects this information as well as possible.
2. Find the likelihood function $\pi(y \mid x)$ that describes the interrelation between the observation and the unknown.
3. Develop methods to explore the posterior probability density.

Estimators

Maximum a posteriori estimate (MAP)

$$x_{\mathrm{MAP}} = \arg\max_{x \in \mathbb{R}^n} \pi(x \,|\, y)$$

Existence or uniqueness is not guaranteed.

Finding the MAP estimate requires solution of an optimization problem, using, e.g, iterative gradient-based methods.

Conditional mean (CM) estimate is defined as

$$x_{\mathrm{CM}} = E\{x \mid y\} = \int_{\mathbb{R}^n} x\pi(x \,|\, y)dx$$

provided that the integral converges.

Requires solving an integration problem. In high-dimensional spaces, this may require special techniques (sampling).

Maximum likelihood (ML) estimate

$$x_{\mathrm{ML}} = \arg \max_{x \in \mathbb{R}^n} \pi(y \,|\, x)$$

Answers the question: *Which value of the unknown is most likely to produce the measured data?*

The ML estimate is a non-Bayesian estimate, and in the case of ill-posed inverse problems, often not useful. Loosely speaking, it corresponds to solving a classical inverse problem without regularization.

Conditional covariance is a 'spread estimator':

$$\operatorname{cov}(x \,|\, y) = \int_{\mathbb{R}^n} (x - x_{\mathrm{CM}})(x - x_{\mathrm{CM}})^{\mathrm{T}} \pi(x \,|\, y) dx \in \mathbb{R}^{n \times n}$$

Requires solving an integration problem.

Bayesian credibility set

Given $p,\ 0 < p < 100$, the credibility set D_p of $p\%$ is defined through the conditions

$$\int_{D_p} \pi(x \,|\, y) dx = \frac{p}{100}, \qquad \pi(x \,|\, y)\big|_{x \in \partial D_p} = \text{constant},$$

and $\pi(x \,|\, y) \geq \pi(z \,|\, y)$ for all $x \in D_p$ and $z \notin D_p$. The boundary of D_p is an equiprobability hypersurface enclosing $p\%$ of the mass of the posterior distribution. (Notice that D_p is not necessarily well defined.)

For a single component, one can look at the symmetric interval of a given credibility: The conditional marginal density of the kth component X_k of X is obtained as

$$\pi(x_k \,|\, y) = \int_{\mathbb{R}^{n-1}} \pi(x_1, \ldots, x_n \,|\, y) dx_1 \cdots dx_{k-1} dx_{k+1} \cdots dx_n.$$

The end points a and b, $a < b$, of the credibility interval $I_k(p) \subset \mathbb{R}$ with a given p, $0 < p < 100$, are determined from the conditions

$$\int_{-\infty}^{a} \pi(x_k \,|\, y) dx_k = \int_{b}^{\infty} \pi(x_k \,|\, y) dx_k = \frac{1}{2} - \frac{p}{200}.$$

(Unfortunately, these conditions do not always define $I_k(p)$ uniquely.)

An Example: x_{MAP} and x_{CM} estimates

In this example, we compare the x_{MAP} and x_{CM} estimates in a simple one-dimensional case. Let $X \in \mathbb{R}$ and assume that the posterior density $\pi_{\mathrm{post}}(x)$ of X is given by

$$\pi_{\mathsf{post}}(x) = \frac{\alpha}{\sigma_0}\phi\left(\frac{x}{\sigma_0}\right) + \frac{1-\alpha}{\sigma_1}\phi\left(\frac{x-1}{\sigma_1}\right),$$

where $0 < \alpha < 1$, $\sigma_0, \sigma_1 > 0$, and ϕ is the standard Gaussian density,

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

In this case, we have

$$x_{\mathrm{CM}} = 1 - \alpha,$$

and for small σ_0 and σ_1 it is a good estimate that

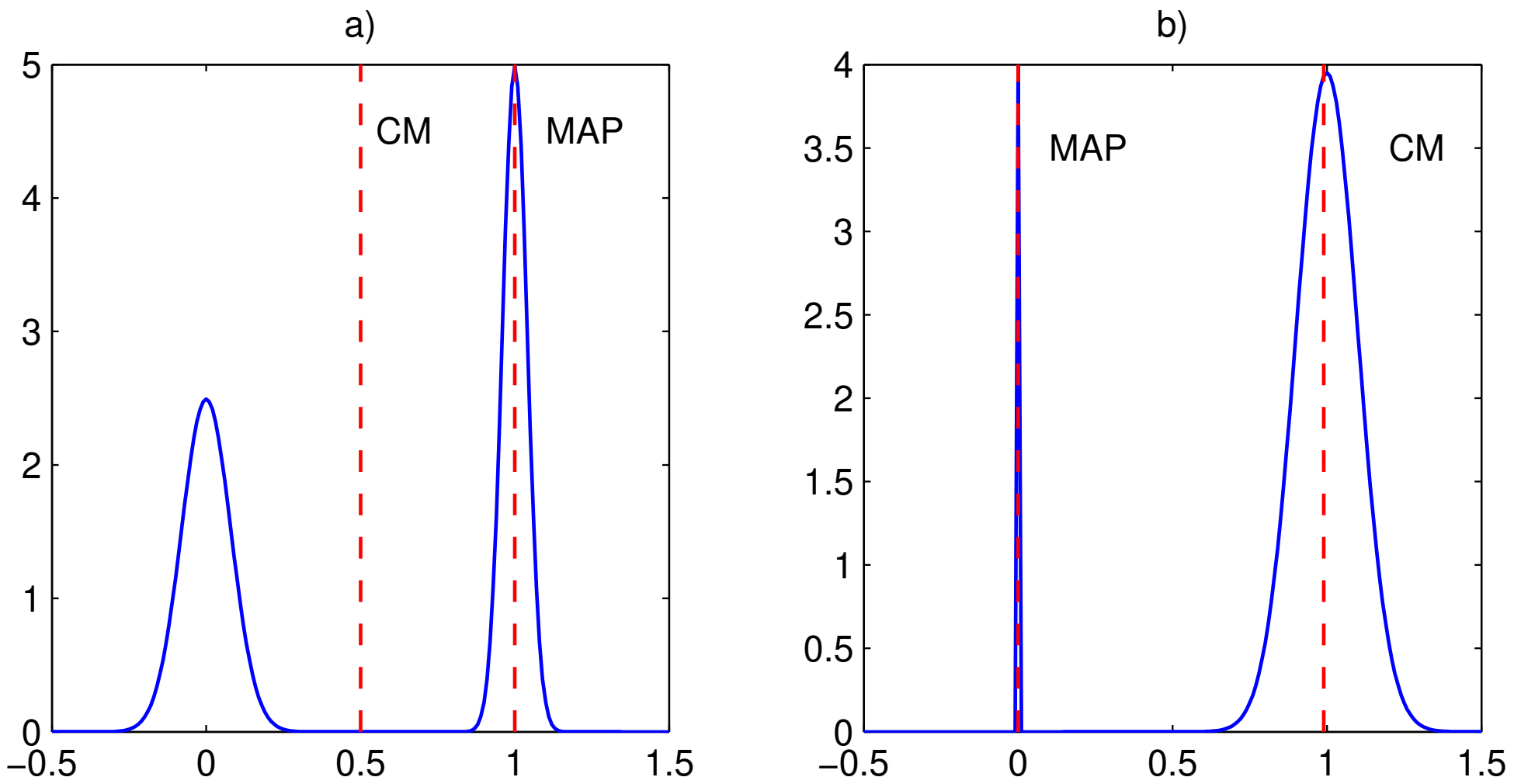
$$x_{\mathrm{MAP}} \approx \begin{cases} 0 & \text{if } \alpha/\sigma_0 \gtrsim (1-\alpha)/\sigma_1, \\ 1 & \text{if } \alpha/\sigma_0 \lesssim (1-\alpha)/\sigma_1. \end{cases}$$

We investigate two different choices of the parameters α, σ_0, σ_1, namely

a) $\alpha = 0.5$, $\sigma_0 = 0.08$ and $\sigma_1 = 0.04$,

b) $\alpha = 0.01$, $\sigma_0 = 0.001$ and $\sigma_1 = 0.1$.

Note that in case b), $\alpha = \sigma_0/\sigma_1$, which means that $\alpha/\sigma_0 > (1-\alpha)/\sigma_1$, and thus $x_{\mathrm{MAP}} \approx 0$ should be the valid case. (You can easily verify this fact numerically.)


a)
CM
MAP
5
4
3
2
1
0
–0.5
0
0.5
1
1.5
b)
MAP
CM
4
3.5
3
2.5
2
1.5
1
0.5
0
–0.5
0
0.5
1
1.5

Let us also consider the posterior variance

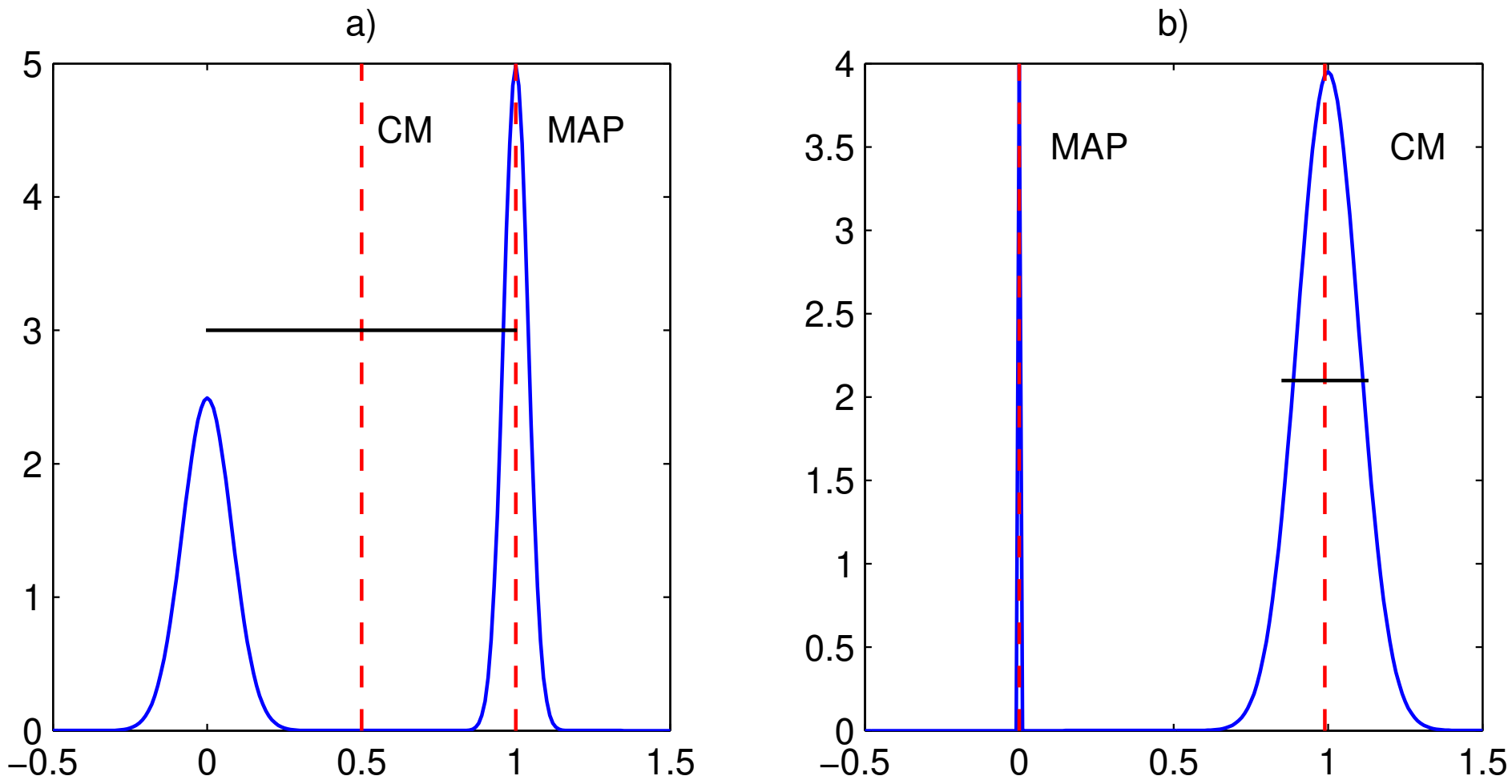
$$\sigma^2 = \int_{-\infty}^{\infty} (x - x_{\mathrm{CM}})^2 \pi_{\mathrm{post}}(x) dx = \int_{-\infty}^{\infty} x^2 \pi_{\mathrm{post}}(x) dx - x_{\mathrm{CM}}^2,$$

which can be calculated analytically in our simple setting:

$$\sigma^2 = \alpha\sigma_0^2 + (1-\alpha)(\sigma_1^2 + 1) - (1-\alpha)^2.$$

In the following images, we have visualized the intervals of length 2σ, i.e., of length two times the standard deviation, centered at x_{CM} for both sets of parameters.

Notice that when the conditional mean gives a poor estimate, this is reflected as a larger variance.


a)
CM
MAP
5
4
3
2
1
0
−0.5
0
0.5
1
1.5
b)
MAP
CM
4
3.5
3
2.5
2
1.5
1
0.5
0
−0.5
0
0.5
1
1.5

Construction of the likelihood function

The likelihood function answers the question: *If we knew the unknown x, how would the measurements be distributed?*

What makes the data deviate from the predicted value given by our observation model?

Some common sources:

1. measurement noise in the data,
2. incompleteness of the observation model (e.g., discretization errors, the reduced nature of the model as compared to the "reality").

Commonly used techniques in construction of the likelihood function (and priors) include conditioning (inspect one variable at the time) and marginalization (eliminate variables of secondary interest).

Additive noise

Very often, the noise is modelled as additive and independent of X. This means that the stochastic model is

$$Y = f(X) + E.$$

Let us assume that the probability distribution of the noise is known:

$$P\{E \in B\} = \int_B \pi_{\text{noise}}(e)de, \qquad B \subset \mathbb{R}^m.$$

Because X and E are mutually independent, fixing $X = x$ does not alter the probability distribution of E. Hence, Y conditioned on $X = x$ is distributed as E shifted by the constant $f(x)$:

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

If the prior probability density of X is π_{pr}, we thus obtain from the Bayes formula that

$$\pi(x \,|\, y) \propto \pi_{\mathrm{pr}}(x)\pi_{\mathrm{noise}}(y - f(x)).$$

If the unknown X and the noise E are *not* mutually independent, we need to know the conditional density of the noise

$$P\{E \in B \,|\, X = x\} = \int_B \pi_{\mathrm{noise}}(e \,|\, x)de.$$

Then, we may write

$$\pi(y \,|\, x) = \int_{\mathbb{R}^m} \pi(y, e \,|\, x)de = \int_{\mathbb{R}^m} \pi(y \,|\, x, e)\pi_{\mathrm{noise}}(e \,|\, x)de.$$

If both $X = x$ and $E = e$ are fixed, $Y = f(x) + e$, and hence

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

Substituting $\pi(y \,|\, x, e)$ into the last formula of the preceding slide thus yields

$$\pi(y \,|\, x) = \pi_{\mathrm{noise}}(y - f(x) \,|\, x),$$

and once again from the Bayes formula we get that

$$\pi(x \,|\, y) \propto \pi_{\mathrm{pr}}(x)\pi_{\mathrm{noise}}(y - f(x) \,|\, x).$$

Example: Additive independent noise

A simple low-dimensional example: a linear model

$$Y = AX + E,$$

where $X \in \mathbb{R}^2$ and $Y, E \in \mathbb{R}^3$ are random variables, and

$$A = \begin{bmatrix} 1 & -1 \\ 1 & -2 \\ 2 & 1 \end{bmatrix}$$

is deterministic. Assume that E has mutually independent normally distributed components with zero mean and variance $\sigma^2 = 0.09$, i.e.,

$$\pi_{\text{noise}}(e) \propto \exp\left(-\frac{1}{2\sigma^2}\|e\|^2\right).$$

Our only prior information is that

$$P\{|X_j| > 2\} = 0, \qquad j = 1, 2,$$

which we write in the form of a prior density via

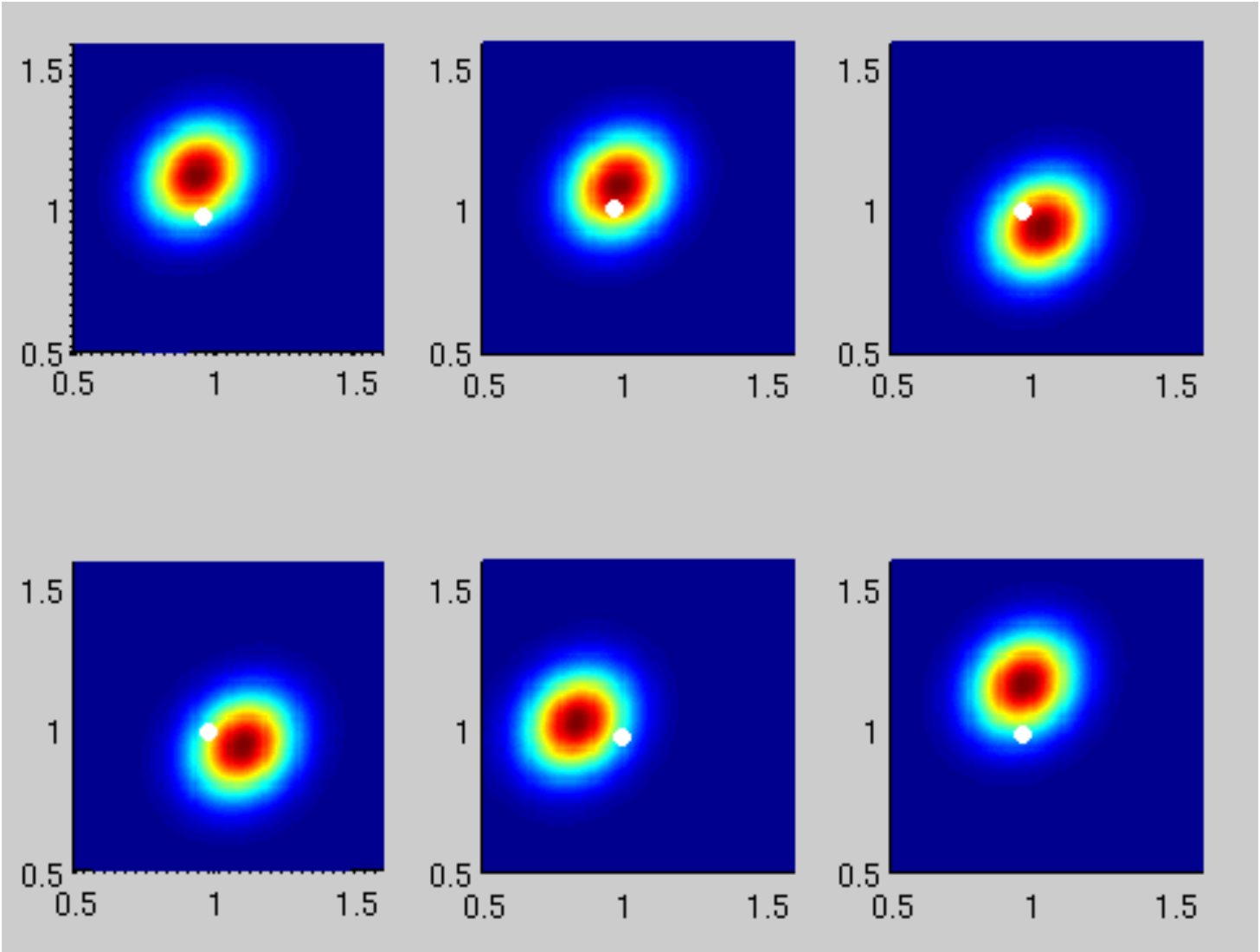
$$\pi_{\mathrm{pr}}(x) = \frac{\chi_Q(x)}{16},$$

where χ_Q is the characteristic function of the square $[-2, 2] \times [-2, 2]$. The posterior density is then

$$\pi(x \,|\, y) \propto \chi_Q(x) \exp\left(-\frac{1}{2\sigma^2} \|y - Ax\|^2\right).$$

Suppose that the true value of X is $x_0 = [1, 1]^{\mathrm{T}}$. We simulate the data through $y = Ax_0 + e$, where e is drawn from π_{noise}.

The following figure illustrates the posterior density with six different realizations of E. Note that in this case the prior hardly plays any role.


1.5
1
0.5
0.5
1
1.5