Towards a Statistical Problem Setting

Traditional setup:

- We want to estimate a parameter $x \in \mathbb{R}^n$ that we cannot observe directly.
- We may or may not know something about x, e.g., $x \in B$.
- We observe another vector $y \in \mathbb{R}^k$ that depends on x through a mathematical model:
$$y = f(x).$$
- Find an estimate x having the desired properties so that the above equation is *approximately* true. Use, e.g., constrained optimization:
$$\text{minimize } \|y - f(x)\| \text{ subject to constraint } x \in B.$$

BAYESIAN SETTING

We have

- *a priori* beliefs of the qualities of the unknown,
- a reasonable model that explains the observation, with all uncertainties included

We need to

- express x as a parameter that defines the distribution of y; (*construction of the likelihood model*)
- incorporate prior information into the model; (*construction of the prior model*).

Basic Principles and Techniques

Randomness means *lack of information*.

Basic principle: Everything that is not known for sure is a random variable.

Basic techniques are

- *conditioning*: take *one* unknown at a time and pretend that you know the rest:
$$\pi(x, y) = \pi(x \mid y)\pi(y) = \pi(y \mid x)\pi(x),$$
- *marginalization*: if a variable is of no interest, integrate it out:
$$\pi(x, y) = \int \pi(x, y, v)dv.$$

Construction of Likelihood

Likelihood answers to the question: *Assuming that we knew the unknown x, how would the measurement be distributed?*

Randomness of the measurement y, *provided that x is known*, is due to

1. measurement noise
2. any incompleteness in the computational model:
   (a) discretization
   (b) incomplete description of "reality" (to the best of our understanding)
   (c) unknown nuisance parameters

EXAMPLE

Assume a functional dependence,

$$y = f(x),$$

when no errors in the observations.

A frequently used model is the *additive noise model*,

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

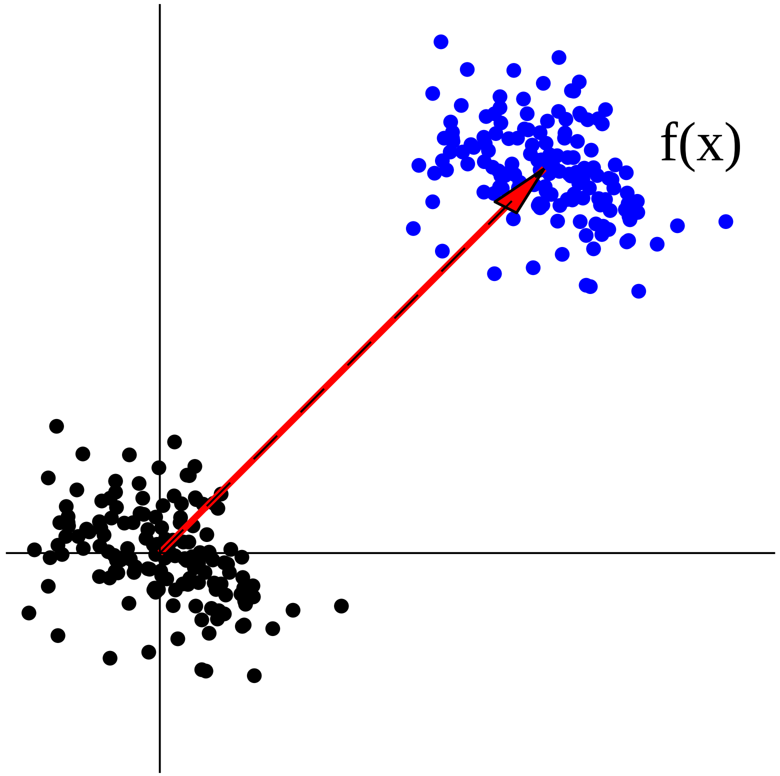
where the distribution of the error is

$$E \sim \pi_{\text{noise}}(e).$$

Assume π_{noise} known.

If E and X are mutually independent,

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


f(x)

The noise distribution may depend on unknown parameters θ:

$$\pi_{\rm noise}(e) = \pi_{\rm noise}(e \mid \theta).$$

Likelihood in this case:

$$\pi(y \mid x, \theta) = \pi_{\rm noise}(y - f(x) \mid \theta).$$

Example: E is zero mean Gaussian with unknown variance σ^2,

$$E \sim \mathcal{N}(0, \sigma^2 I),$$

where $I \in \mathbb{R}^{m\times m}$ is the identity matrix. In this case,

$$\pi(y \mid x, \sigma^2) = \frac{1}{(2\pi)^{m/2}\sigma^m} \exp\left(-\frac{1}{2\sigma^2}\|y - f(x)\|^2\right).$$

Example

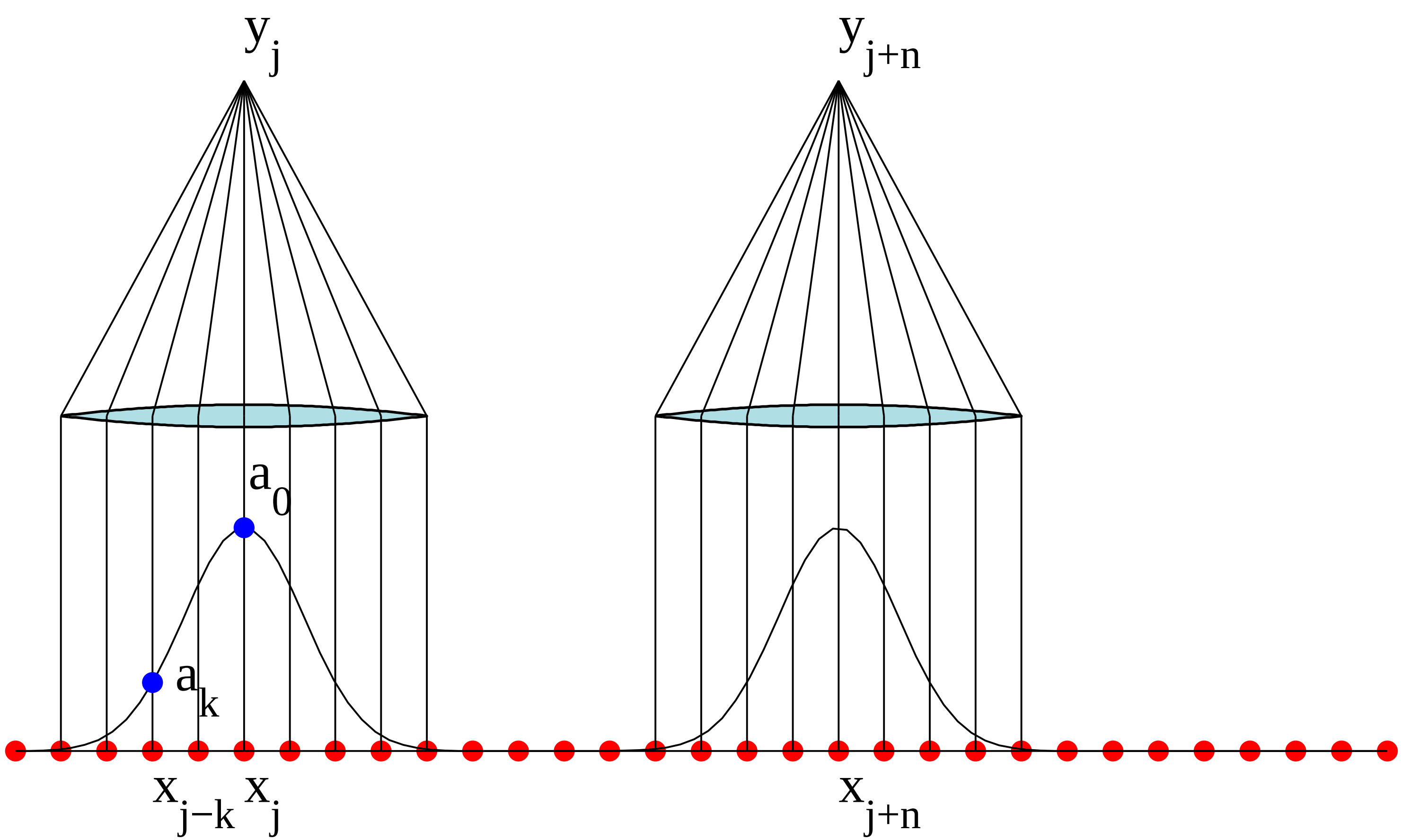
Assume that

- the device consists of a collecting lens and a photon counter,
- the photons come from N emitting sources.

$$\text{Average photon emission/observation time} = x_j,\ 1 \leq j \leq N.$$

The geometry of the lens:

$$\text{Average total count} = \text{weighted sum of the individual contributions.}$$



y_j
y_{j+n}
a_0
a_k
x_{j-k}
x_j
x_{j+n}

Expected output defined by the geometry:

$$\overline{y}_j = \mathrm{E}\{Y_j\} = \sum_{k=-L}^{L} a_k x_{j-k},$$

where

- weights a_j determined by the geometry of the lens
- index L is related to the width of the lens

Here, $x_j = 0$ if $j < 1$ or $j > N$.

Repeating the reasoning over each source point, we arrive at a matrix model

$$\overline{y} = \mathrm{E}\{Y\} = Ax,$$

where $A \in \mathbb{R}^{n\times n}$ is a Toeplitz matrix,

$$A = \begin{bmatrix} a_0 & a_{-1} & \cdots & a_{-L} & & \\ a_1 & a_0 & & & \ddots & \\ \vdots & & \ddots & & & a_{-L} \\ a_L & & & \ddots & & \vdots \\ & \ddots & & & a_0 & a_{-1} \\ & & a_L & \cdots & a_1 & a_0 \end{bmatrix}.$$

The parameter L defines the *bandwidth* of the matrix.

Weak, the observation model described is a photon counting process:

$$Y_j \sim \text{Poisson}\big((Ax)_j\big),$$

that is,

$$\pi(y_j \mid x) = \frac{(Ax)_j^{y_j}}{y_j!} \exp\big(-(Ax)_j\big).$$

Consecutive measurements are independent, $Y \in \mathbb{R}^N$ has the density

$$\pi(y \mid x) = \prod_{j=1}^{N} \pi(y_j \mid x) = \prod_{j=1}^{L} \frac{(Ax)_j^{y_j}}{y_j!} \exp\big(-(Ax)_j\big).$$

We express this relation simply as

$$Y \sim \text{Poisson}(Ax).$$

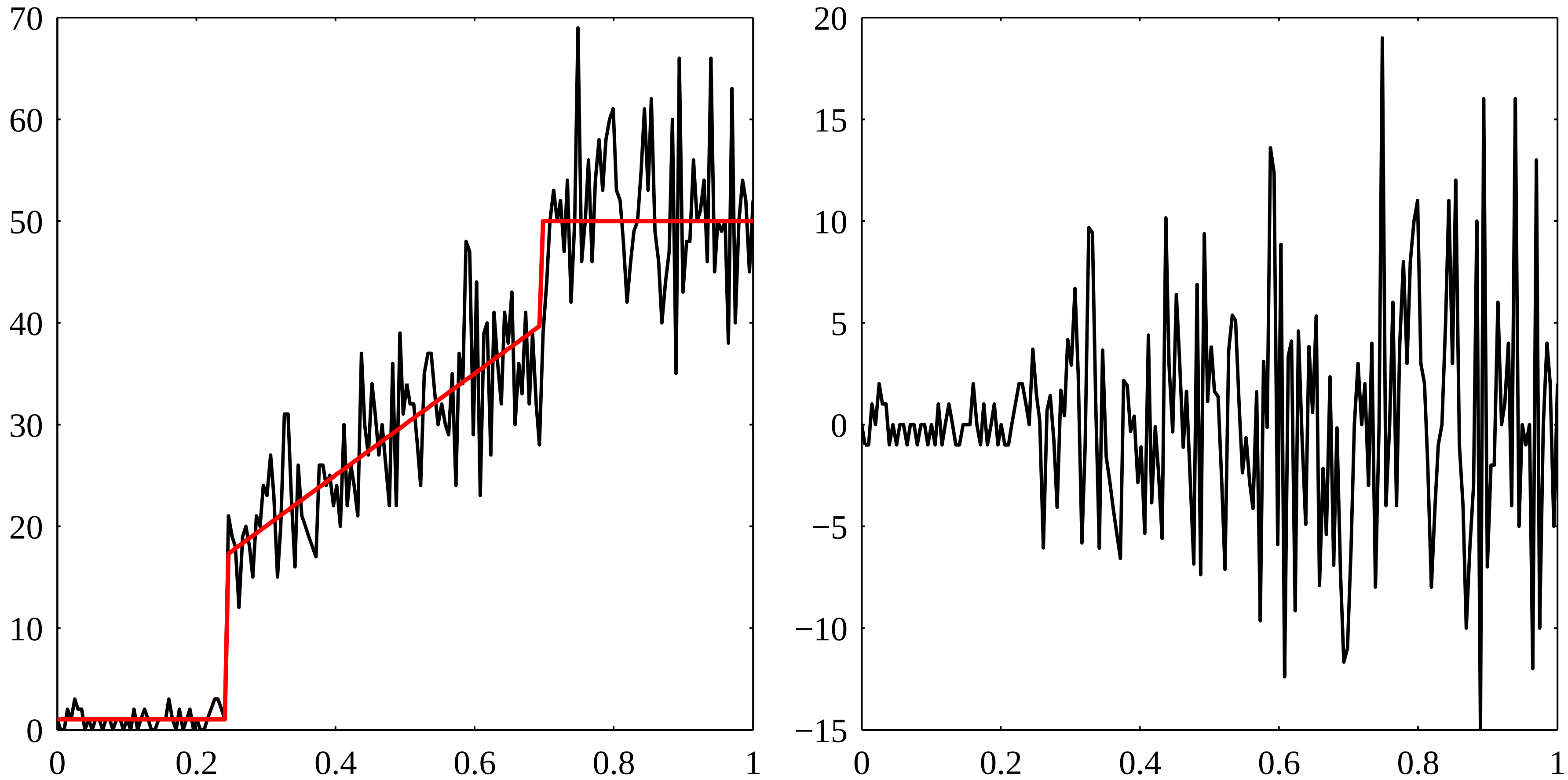
Gaussian approximation

Assuming that the count is high, we may write

$$
\begin{aligned}
\pi(y \mid x) &\approx \prod_{\ell=1}^{L} \left(\frac{1}{2\pi (Ax)_\ell}\right)^{1/2} \exp\left(-\frac{1}{2(Ax)_\ell}\big(y_\ell - (Ax)_\ell\big)^2\right) \\
&= \left(\frac{1}{(2\pi)^L \det(\Gamma)}\right)^{1/2} \exp\left(-\frac{1}{2}(y - Ax)^{\mathrm{T}}\Gamma^{-1}(y - Ax)\right),
\end{aligned}
$$

$$
\Gamma = \Gamma(x) = \mathrm{diag}\big(Ax\big).
$$

The higher the signal, the higher the noise.


70
60
50
40
30
20
10
0
0
0.2
0.4
0.6
0.8
1
20
15
10
5
0
−5
−10
−15
0
0.2
0.4
0.6
0.8
1

Change of variables

Random variables X and Y in $\mathbb{R}^n$,

$$Y = f(X),$$

where f is a differentiable function, and the probability distribution of Y is known:

$$\pi(y) = p(y).$$

Probability density of X?

$$\pi(y)dy = p(y)dy = p(f(x))|\mathrm{det}(Df(x))|dx,$$

Identify

$$\pi(x) = p(f(x))|\mathrm{det}(Df(x))|.$$

EXAMPLE

Noisy amplifier: input $f(t)$ amplified by a factor $\alpha > 1$.

Ideal model for the output signal:

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

Noise: α fluctuates.

Discrete signal:

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

Amplification at $t = t_j$ is a_j:

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

Stochastic extension:

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

or in the vector notation as

$$Y = A.X, \tag{1}$$

Assume: A has the probability density

$$A \sim \pi_{\rm noise}(a),$$

Likelihood density for Y, conditioned on $X = x$, is

$$\pi(y \mid x) \propto \pi_{\rm noise}\left(\frac{y.}{x}\right),$$

Normalizing:

$$\pi(y \mid x) = \frac{1}{x_1 x_2 \cdots x_n} \pi_{\rm noise}\left(\frac{y.}{x}\right), \tag{2}$$

Formally:

$$y = a.x, \quad \text{or} \quad a = \frac{y.}{x}, \quad x \text{ fixed},$$

or

$$a_j = \frac{y_j}{x_j}, \quad da_j = \frac{dy_j}{x_j}.$$

$$\begin{aligned} p(a)da &= p(a)da_1 \cdots da_n = p\left(\frac{y.}{x}\right)\frac{dy_1}{x_1}\cdots\frac{dy_n}{x_n} \\ &= \underbrace{\left(\frac{1}{x_1 x_2 \cdots x_n} p\left(\frac{y.}{x}\right)\right)}_{=\pi(y)} dy_1 \cdots dy_n. \end{aligned}$$

Example: all the variables are positive, and A is *log-normally distributed*:

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

components mutually independent.

Note: *the probability distributions transform as densities, not as functions!*

$$\mathrm{P}\big\{W_i = \log A_i < t\big\} = \mathrm{P}\big\{A_i < e^t\big\}. \tag{3}$$

L.h.s. as an integral:

$$\mathrm{P}\big\{W_i < t\big\} = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{t} \exp\left(-\frac{1}{2\sigma^2}(w_i - w_0)^2\right) dw_i.$$

Change of variables:

$$w_i = \log a_i, \quad dw_i = \frac{1}{a_i} da_i,$$

and substitute $w_0 = \log \alpha_0$:

$$\begin{aligned} \mathrm{P}\{W_i < t\} &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_0^{e^t} \frac{1}{a_i} \exp\left(-\frac{1}{2\sigma^2}(\log a_i - \log \alpha_0)^2\right) da_i \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_0^{e^t} \frac{1}{a_i} \exp\left(-\frac{1}{2\sigma^2}\left(\log \frac{a_i}{\alpha_0}\right)^2\right) da_i. \end{aligned}$$

Compare to the r.h.s. to identify

$$\pi(a_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{a_i} \exp\left(-\frac{1}{2\sigma^2}\left(\log \frac{a_i}{\alpha_0}\right)^2\right),$$

which is the one-dimensional log-normal density.

Independent components:

$$\begin{aligned}\pi(y \mid x) &= \pi(y_1 \mid x)\cdots\pi(y_n \mid x) \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2}\frac{1}{y_1 y_2\cdots y_n}\exp\left(\frac{1}{2\sigma^2}\sum_{j=1}^{n}\left(\log\frac{y_j}{\alpha_0 x_j}\right)^2\right).\end{aligned}$$

Remark: Alternative approach:

$$\log Y = \log X + \log A = \log X + W,$$

and we may write the conditional density for $\log Y$, as

$$\pi(\log y \mid x) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(\log y_j - \log x_j - \log\alpha_0)^2\right).$$

EXAMPLE

Poisson noise and additive Gaussian noise:

$$Y = Z + E, \quad Z \sim \text{Poisson}(Ax), \quad E \sim \mathcal{N}(0, \sigma^2 I).$$

First step: assume that $X = x$ and $Z = z$ are known, giving

$$\pi(y_j \mid z_j, x) \propto \exp\left(-\frac{1}{2\sigma^2}(y_j - z_j)^2\right).$$

Conditioning:

$$\pi(y_j, z_j \mid x) = \pi(y_j \mid z_j, x)\pi(z_j \mid x).$$

The value of z_j (integer) is not of interest here, so

$$\begin{aligned} \pi(y_j \mid x) &= \sum_{z_j=0}^{\infty} \pi(y_j, z_j \mid x) \\ &\propto \sum_{z_j=0}^{\infty} \pi(z_j \mid x) \exp\left(-\frac{1}{2\sigma^2}(y_j - z_j)^2\right). \end{aligned}$$

CONSTRUCTION OF PRIORS

EXAMPLE: Assume that we try to determine the hemoglobin level x in blood by near-infrared (NIR) measurement at the patients finger.

Previous measurements directly from the patient's blood,

$$S = \{x_1, \ldots, x_N\}.$$

Think as *realizations* of a random variable with an unknown distribution.

- *Non-parametric* approach: Look at a histogram based on S.
- *Parametric* approach: Justify a parametric model, find the ML estimate of the model parameters.

Let us assume that

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

From previous analysis, the ML estimate for x_0 is

$$x_{0,\mathrm{ML}} = \frac{1}{N}\sum_{j=1}^{N} x_j,$$

and for σ^2,

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{j=1}^{N}(x_j - x_{0,\mathrm{ML}})^2.$$

Any future value x will be another realization from the same distribution.

Postulate:

- The unknown X is a random variable, whose probability distribution is denoted as $\pi_{\rm pr}(x)$ and called the *prior distribution*,
- By prior experience, and assuming that the Gaussian approximation of the prior is justifiable, we use the parametric model

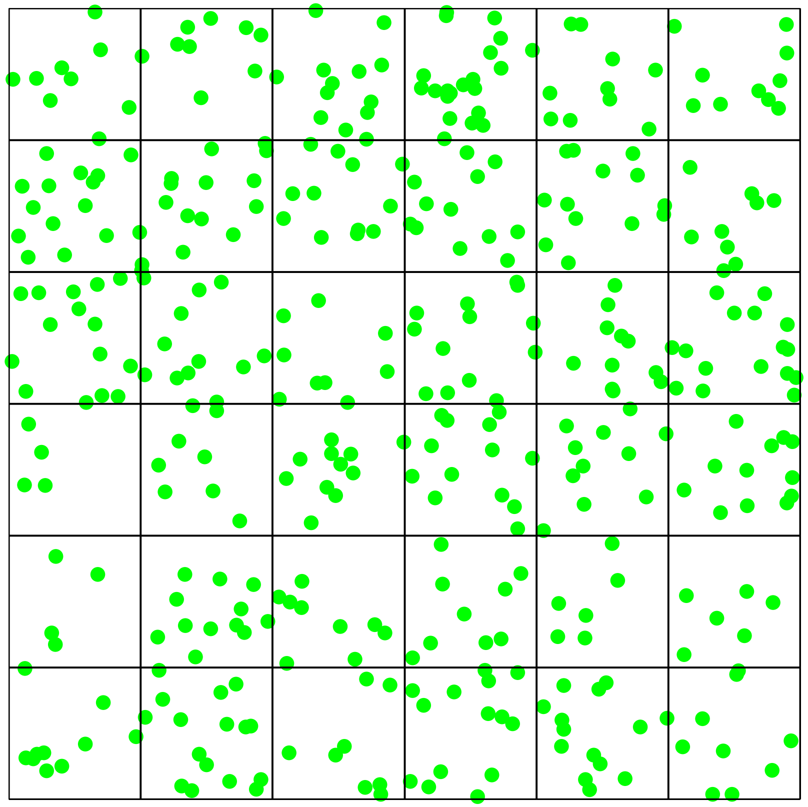
$$\pi_{\rm pr}(x) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{1}{2\sigma^2}(x - x_0)^2\right),$$

  where x_0 and σ^2 are determined experimentally from S by the formulas above.

The above approach, where the prior is defined through previous experience, is called *empirical Bayes* approach.

EXAMPLE

Rectangular array of squares. Each square contains a number of bacteria.



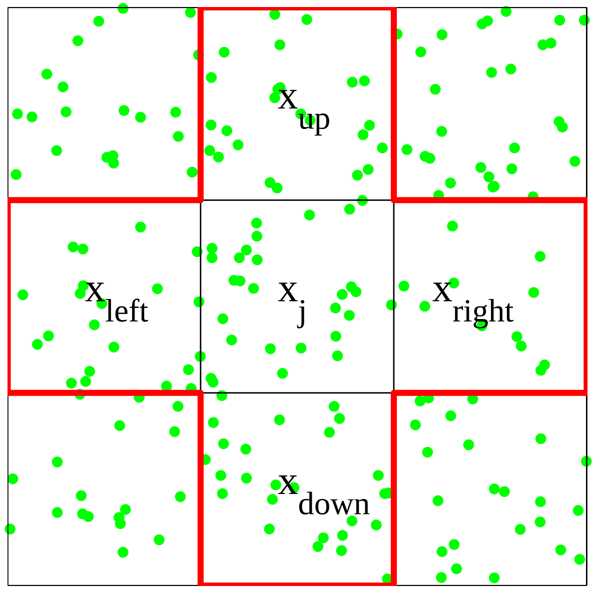
The inverse problem: estimate the density of the bacteria from some indirect measurements.

Set up a model based on your belief how bacteria grow:

Number of bacteria in a box ≈ average of neighbours,

or

$$x_j \approx \frac{1}{4}(x_{\mathrm{left},j} + x_{\mathrm{right},j} + x_{\mathrm{up},j} + x_{\mathrm{down},j}).$$



Modification at boundary pixels: Define $x_j = 0$ for pixels outside the square.

Matrix $A \in \mathbb{R}^{N\times N}$, $N =$ number of pixels,

$$\begin{array}{l} \qquad\qquad\qquad\quad \text{(up)} \quad \text{(down)} \quad \text{(left)} \quad \text{(right)} \\ A(j,\,:\,) = \big[0\cdots 1/4 \cdots \ 1/4 \ \cdots \ 1/4\cdots \quad 1/4\cdots 0\big], \end{array}$$

Absolute certainty of your model, ($\approx \longrightarrow =$):

$$x = Ax. \tag{4}$$

But this does not work: write (4) as

$$(I - A)x = 0 \Rightarrow x = 0,$$

since

$$\det(I - A) \neq 0.$$

Solution: relax the model and write

$$x = Ax + r, \quad r = \text{uncertainty of the model.} \tag{5}$$

Since r is not known, model it as a random variable.

Postulate a distribution to it,

$$r \sim \pi_{\rm mod.error}(r).$$

From $x - Ax = r$ follows a natural prior model,

$$\pi_{\rm prior}(x) = \pi_{\rm mod.error}(x - Ax).$$

The model (5) is referred to as *autoregressive Markov model*, and r is an *innovation process*.

In particular, if r is a Gaussian variable with mutually independent and equally distributed components,

$$r \sim \mathcal{N}(0, \sigma^2 I),$$

we obtain the prior model

$$\begin{aligned}
\pi_{\rm prior}(x \mid \sigma^2) &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\|x - Ax\|^2\right) \\
&= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\|Lx\|^2\right),
\end{aligned}$$

where

$$L = I - A.$$

Note: if σ^2 is not known (as it usually isn't), it is part of the estimation problem. *Hierarchical models* discussed later.

Observe that L is a second order finite difference matrix with the mask

$$\begin{bmatrix} & -1/4 & \\ -1/4 & 1 & -1/4 \\ & -1/4 & \end{bmatrix}.$$

The model leads to what is often referred to as the *second order smoothness prior*.

Another derivation: Assume that

$$x_j = f(p_j), \quad p_j = \text{point in the } j\text{th pixel.}$$

Finite difference approximation,

$$\Delta f(p_j) \approx \frac{1}{h^2}\big(Ax\big)_j,$$

where h = discretization size.

Sparse matrices in Matlab

```
 n = 50;  % Number of pixels per directions

% Creating an index matrix to enumerate the pixels

 I = reshape([1:n^2],n,n);

% Right neighbors of each pixel

 Icurr = I(:,1:n-1);
 Ineigh = I(:,2:n);
 rows = Icurr(:);
 cols = Ineigh(:);
 vals = ones(n*(n-1),1);

% Left neighbors of each pixel
```

```
 Icurr = I(:,2:n);
 Ineigh = I(:,1:n-1);
 rows = [rows;Icurr(:)];
 cols = [cols;Ineigh(:)];
 vals = [vals;ones(n*(n-1),1)];

% Upper neighbors of each pixel

 Icurr = I(2:n-1,:);
 Ineigh = I(1:n-1,:);
 rows = [rows;Icurr(:)];
 cols = [cols;Ineigh(:)];
 vals = [vals;ones(n*(n-1),1)];

% Lower neighbors of each pixel

 Icurr = I(1:n-1,:);
 Ineigh = I(2:n,:);
```

```
rows = [rows;Icurr(:)];
cols = [cols;Ineigh(:)];
vals = [vals;ones(n*(n-1),1)];

A = 1/4*sparse(rows,cols,vals);
L = speye(n^2) - A;
```

Posterior Densities

Fundamental identity:

$$\pi(x, y) = \pi_{\text{prior}}(x)\pi(y \mid x) = \pi(y)\pi(x \mid y),$$

Bayes' formula

$$\pi(x \mid y) = \frac{\pi_{\text{prior}}(x)\pi(y \mid x)}{\pi(y)}, \quad y = y_{\text{observed}}. \tag{6}$$

Here $\pi(x \mid y)$ is the *posterior density*

The posterior density is the solution Bayesian of the inverse problem.

Example

Linear inverse problem, additive noise:

$$y = Ax + e, \quad x \in \mathbb{R}^n, \; y, e \in \mathbb{R}^m, \; A \in \mathbb{R}^{m \times n},$$

Stochastic extension

$$Y = AX + E.$$

Assume that X and E are independent and Gaussian,

$$X \sim \mathcal{N}(0, \gamma^2 \Gamma), \quad E \sim \mathcal{N}(0, \sigma^2 I).$$

The prior density is

$$\pi_{\text{prior}}(x \mid \gamma) \propto \frac{1}{\gamma^n} \exp\left(-\frac{1}{2\gamma^2} x^{\mathrm{T}} \Gamma^{-1} x\right).$$

Observe:

$$\det\left(\gamma^2 \Gamma\right) = \gamma^{2n} \det\left(\Gamma\right).$$

Likelihood:

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

From Bayes' formula:

$$
\begin{aligned}
\pi(x \mid y, \gamma) &\propto \pi_{\rm prior}(x \mid \gamma)\pi(y \mid x) \\
&\propto \frac{1}{\gamma^n}\exp\left(-\frac{1}{2\gamma^2}x^{\rm T}\Gamma^{-1}x - \frac{1}{2\sigma^2}\|y - Ax\|^2\right) \\
&= \frac{1}{\gamma^n}\exp\left(-V(x \mid y, \gamma)\right).
\end{aligned}
$$

The matrix Γ is symmetric positive definite. Cholesky factorization:

$$\Gamma^{-1} = R^{\mathrm{T}} R.$$

where R is upper triangular matrix.

From

$$x^{\mathrm{T}} \Gamma^{-1} x = x^{\mathrm{T}} R^{\mathrm{T}} R x = \|Rx\|^2$$

it follows that

$$T(x) = 2\sigma^2 V(x \mid y, \gamma) = \|y - Ax\|^2 + \delta^2 \|Rx\|^2, \quad \delta = \frac{\sigma}{\gamma}. \tag{7}$$

The functional T is called the *Tikhonov functional*

Maximum A Posteriori (MAP) Estimator

Bayesian analogue of Maximum Likelihood estimator:

$$x_{\rm MAP} = \arg\max\ \pi(x \mid y),$$

or, equivalently,

$$x_{\rm MAP} = \arg\min\ V(x \mid y), \quad V(x \mid y) = -\log\pi(x \mid y).$$

Here,

$$x_{\rm MAP} = \arg\min\left(\|y - Ax\|^2 + \delta^2\|Rx\|^2\right) \tag{8}$$

Maximum Likelihood estimator is the least squares solution of the problem

$$Ax = y, \tag{9}$$

Equivalent characterization of the MAP estimator:

$$\|y - Ax\|^2 + \delta^2 \|Rx\|^2 = \left\| \begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} A \\ \delta R \end{bmatrix} x \right\|^2,$$

so the MAP estimate is the least squares solution of

$$\begin{bmatrix} A \\ \delta R \end{bmatrix} x = \begin{bmatrix} y \\ 0 \end{bmatrix}.$$