Bayesian Scientific Computing
Day 2

Daniela Calvetti, Erkki Somersalo

Case Western Reserve University
Department of Mathematics, Applied Mathematics and Statistics

Lyngby, December 2019

Bayes' Formula

Let

$$\begin{aligned} X &= \text{random variable representing the unknown of interest,} \\ B &= \text{random variable representing the observed quantity,} \\ b &= \text{measured data, realization of } B. \end{aligned}$$

Bayes' formula,

$$\boxed{\pi_{X|B}(x \mid b) \propto \pi_X(x)\pi_{B|X}(b \mid x).}$$

To solve the inverse problem in the Bayesian framework, we need

- Encode the forward model and uncertainties in the observation process in $\pi_{B|X}(b \mid x)$ (likelihood).
- Find a way to encode prior beliefs in $\pi_X(x)$ (prior),

Likelihood

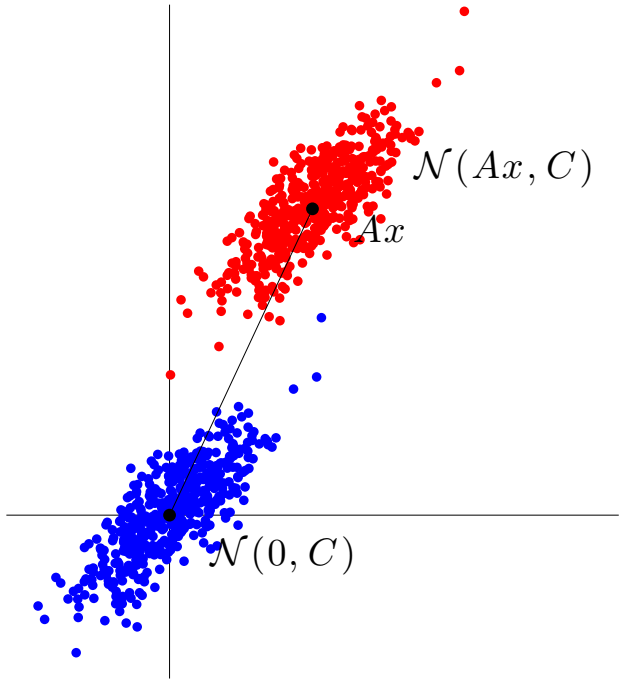
The most common and elementary model assumes that

- the observation noise is additive,
- the noise is independent of the unknown.

$$b = f(x) + \varepsilon, \quad \varepsilon \sim \pi_E,$$

where f and π_E are known.

Likelihood



Example: Linear model, $b = \mathrm{A}x + \varepsilon$, where ε is Gaussian. The "noise cloud" is simply shifted around the presumably known point $\mathrm{A}x$.

Likelihood

Assuming that x, and therefore $f(x)$ is known, b has the same distribution as ε but with a shifted mean,

$$\pi_{B|X}(b \mid x) = \pi_E(b - f(x)).$$

Formally, we solve for ε:

$$\varepsilon = b - f(x) \sim \pi_E, \quad x \text{ fixed.}$$

Likelihood

Most commonly used likelihood model is based on additive Gaussian noise model:

$$b = f(x) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \Sigma),$$

where $\Sigma \in \mathbb{R}^{m \times m}$ is symmetric positive definite matrix.

$$\pi_{B|X}(b \mid x) = \left(\frac{1}{(2\pi)^n |\Sigma|} \right)^{1/2} \exp\left(-\frac{1}{2} (b - f(x))^{\mathsf{T}} \Sigma^{-1} (b - f(x)) \right),$$

where $|\Sigma|$ is the determinant of Σ.
If $\Sigma = \sigma^2 \mathrm{I}_m$, this reduces further to

$$\pi_{B|X}(b \mid x) = \left(\frac{1}{(2\pi)^n \sigma^{2n}} \right)^{1/2} \exp\left(-\frac{1}{2\sigma^2} \|b - f(x)\|^2 \right).$$

Likelihood

More generally, when passing from one random variable to another, we must remember that **probability densities represent measures.**

Start with a one-dimensional change of variables. Assume that we have two real-valued random variables X, Z that are related to each other through a formula

$$X = \phi(Z),$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is a one-to-one mapping. For simplicity, assume that ϕ is strictly increasing.

Question: *If the pdf π_X is given, what is the pdf π_Z?*

First, note that since ϕ is increasing, we have

$$a < Z < b \Leftrightarrow a' = \phi(a) < \phi(Z) = X < \phi(b) = b'.$$

Likelihood

Therefore

$$P\{a' < X < b'\} = P\{a < Z < b\}.$$

The probability density of Z therefore satisfies

$$\int_a^b \pi_Z(z)dz = \int_{a'}^{b'} \pi_X(x)dx.$$

In the latter integral, make the change of variables

$$x = \phi(z), \quad dx = \frac{d\phi}{dz}(z)dz,$$

and so

$$\int_a^b \pi_Z(z)dz = \int_a^b \pi_X(\phi(z))\frac{d\phi}{dz}(z)dz.$$

Likelihood

We conclude that

$$\pi_Z(z) = \pi_X(\phi(z)) \frac{d\phi}{dz}(z).$$

Here, we assumed that Φ was increasing. If it is decreasing, the derivative is negative. The density needs to be positive, so we write, in general,

$$\boxed{\pi_Z(z) = \pi_X(\phi(z)) \left| \frac{d\phi}{dz}(z) \right|.}$$

Likelihood

Change of variables formula in $\mathbb{R}^n$: If

$$X = \phi(Z), \quad X, Z \in \mathbb{R}^m,$$

then

$$\pi_Z(z) = \pi_X\big(\phi(z)\big) \left| \frac{\partial \phi}{\partial z} \right|,$$

where the Jacobian determinant is

$$\frac{\partial \phi}{\partial z} = \begin{vmatrix} \frac{\partial \phi_1}{\partial z_1} & \cdots & \frac{\partial \phi_1}{\partial z_n} \\ \vdots & & \vdots \\ \frac{\partial \phi_n}{\partial z_1} & \cdots & \frac{\partial \phi_n}{\partial z_n} \end{vmatrix}$$

Likelihood

Example: Multiplicative noise:

$$b_j = \varepsilon_j \, f_j(x), \quad f_j : \mathbb{R}^n \to \mathbb{R}, \quad 1 \leq j \leq m.$$

Think of a noisy amplifier: The louder the signal, the more noise. Assume that we know

$$f_j(x) > 0, \quad \varepsilon \sim \pi_E.$$

Solve for the noise vector:

$$\varepsilon_j = \frac{b_j}{f_j(x)} = \phi_j(b), \text{ (recall that } x \text{ is a fixed parameter here)},$$

and think of this as a change of variables, $b \to \varepsilon$.

Likelihood

Compute the Jacobian:

$$\frac{\partial \phi}{\partial b} = \begin{vmatrix} \frac{\partial \phi_1}{\partial b_1} & \cdots & \frac{\partial \phi_1}{\partial b_n} \\ \vdots & & \vdots \\ \frac{\partial \phi_n}{\partial b_1} & \cdots & \frac{\partial \phi_n}{\partial b_n} \end{vmatrix} = \begin{vmatrix} \frac{1}{f_1(x)} & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & \frac{1}{f_n(x)} \end{vmatrix}$$

$$= \frac{1}{f_1(x) f_2(x) \cdots f_n(x)}.$$

Conclusion: The likelihood model is given by

$$\pi_{B|X}(b \mid x) = \frac{1}{f_1(x) f_2(x) \cdots f_n(x)} \pi_E \left(\frac{b_1}{f_1(x)}, \ldots, \frac{b_n}{f_n(x)} \right).$$

Likelihood

Example: Counting data.

- Assume that the data consist of particle counts (photons, electrons).
- Every time we measure, the count is slightly different even if the target and measurement configuration does not change.

The model predicts that

$$E\{b_j\} = f_j(x), \quad 1 \leq j \leq m.$$

Under mild assumptions (stationarity, independence of increments, zero probability of coincidence), we may assume that b_j is Poisson distributed.

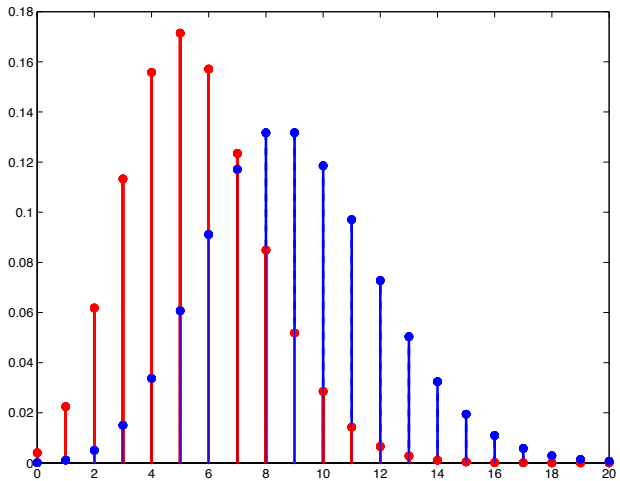
Likelihood

Poisson distribution: An integer valued random variable B is Poisson distributed,

$$B \sim \text{Poisson}(\lambda),$$

if

$$P\{B = n\} = \frac{\lambda^n}{n!} e^{-\lambda}, \quad n = 0, 1, 2, \ldots$$



Likelihood

If the observations b_j are **conditionally independent**, we may therefore write

$$P\{B_j = n \mid X = x\} = \frac{f_j(x)^n}{n!} e^{-f_j(x)},$$

and therefore, the Poisson likelihood is

$$\pi_{B|X}(b \mid x) = \prod_{j=1}^{m} \frac{f_j(x)^{b_j}}{b_j!} e^{-f_j(x)},$$

Likelihood

Sometimes the data are corrupted by several types of noise. For example, we may have additive and multiplicative noise:

$$b = \varepsilon f(x) + w, \quad \varepsilon \sim \pi_E, \quad w \sim \pi_W.$$

Proceed in two phases:

1. Solve the conditional density of B given both X and W.
2. Then, find the conditional distribution of the pair (B, W).
3. Finally, include the randomness of W, and marginalize over W.

$$\pi_{B|X}(b \mid x) = \int \pi_{B,W|X}(b, w \mid x)\pi_W(w)dw.$$

Likelihood

We have

$$B - w = Ef(x), \quad E \sim \pi_E,$$

so, treating x and w as parameters, we have

$$E = \frac{1}{f(x)}(B - w),$$

and therefore,

$$\pi_{B|X,W}(b \mid x, w) = \frac{1}{f(x)} \pi_E \left(\frac{b - w}{f(x)} \right)$$

Likelihood

Now, write

$$\begin{aligned}\pi_{B,W|X}(b, w \mid x) &= \pi_{B|X,W}(b \mid x, w)\pi_W(w) \\ &= \frac{1}{f(x)}\pi_E\left(\frac{b-w}{f(x)}\right)\pi_W(w),\end{aligned}$$

and therefore

$$\pi_{B|X}(b \mid x) = \int \frac{1}{f(x)}\pi_E\left(\frac{b-w}{f(x)}\right)\pi_W(w)dw.$$

Prior models

- The prior expresses what is **believed** to be true about the unknown *prior* to analyzing the data.
- Unless the prior information is certain, the prior should **promote, but not force** desired properties in the solution.
- Bayes' formula

$$\mu_{X|B}(x \mid b) \propto \pi_X(x)\pi_{B|X}(b \mid x)$$

represents an **updating process** of information based on data.

Smoothness Prior Models

Example: A discretized signal

$$x_j = g(t_j), \quad t_j = \frac{j}{n}, \quad 0 \le j \le n,$$

needs to be estimated from given data (not specified yet).
Consider two prior models:

1. We know that $x_0 = 0$, and believe that the absolute value of the slope of g is bounded by some $m_1 > 0$.
2. We know that $x_0 = x_n = 0$ and believe that the curvature of g is bounded by some $m_2 > 0$.

Smoothness Prior Models

1. Slope:

$$g'(t_j) \approx \frac{x_j - x_{j-1}}{h}, \quad h = \frac{1}{n},$$

Prior information: We believe that

$$|x_j - x_{j-1}| \leq h\, m_1 \text{ with some uncertainty.}$$

2. Curvature:

$$g''(t_j) \approx \frac{x_{j-1} - 2x_j + x_{j+1}}{h^2}.$$

Prior information: We believe that

$$|x_{j-1} - 2x_j + x_{j+1}| \leq h^2 m_2 \text{ with some uncertainty.}$$

Smoothness Prior Models

In both cases, we assume that x_j is a realization of a random variable X_j.
Boundary conditions:

1. $X_0 = 0$ with certainty. Probabilistic model for X_j, $1 \leq j \leq n$.
2. $X_0 = X_n = 0$ with certainty. Probabilistic model for X_j, $1 \leq j \leq n-1$.

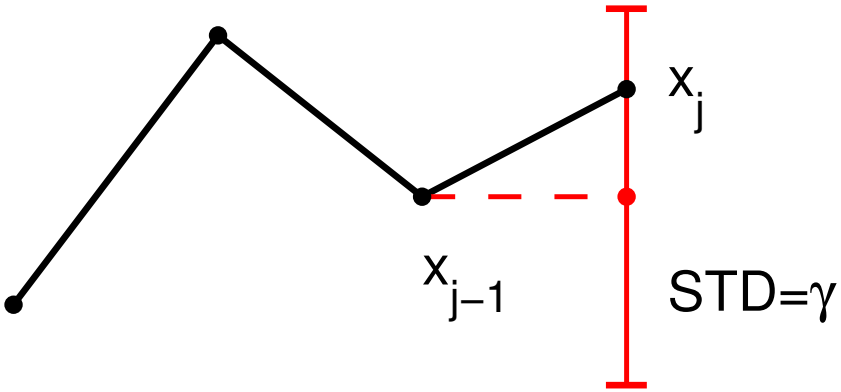
Smoothness Prior Models

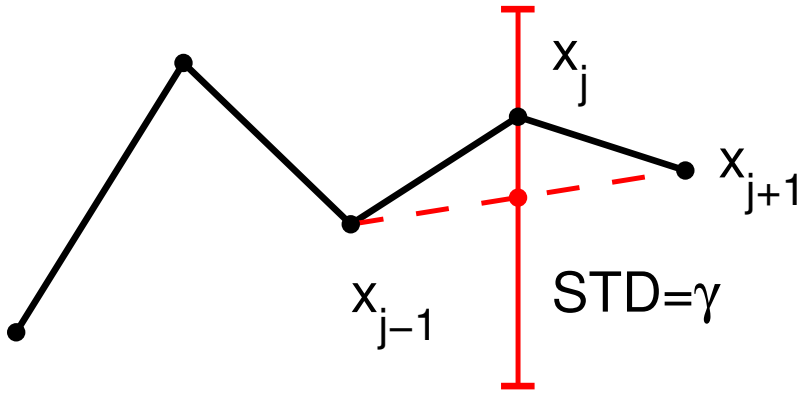
1. First order prior:

$$X_j = X_{j-1} + \gamma W_j, \quad W_j \sim \mathcal{N}(0,1), \quad \gamma = h\, m_1.$$

2. Second order prior:

$$X_j = \frac{1}{2}(X_{j-1} + X_{j+1}) + \gamma W_j, \quad W_j \sim \mathcal{N}(0,1), \quad \gamma = \frac{1}{2} h^2\, m_2.$$





Matrix form: first order model

System of equations:

$$\begin{aligned} X_1 = X_1 - X_0 &= \gamma W_1 \\ X_2 - X_1 &= \gamma W_2 \\ \vdots &\quad \vdots \\ X_n - X_{n-1} &= \gamma W_n \end{aligned}$$

$$\mathsf{L}_1 = \begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix} \in \mathbb{R}^{n \times n}, \quad X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}, \quad W = \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_n \end{bmatrix}.$$

$$\mathsf{L}_1 X = \gamma W, \quad W \sim \mathcal{N}(0, \gamma^2 \mathsf{I}_n),$$

Matrix form: second order model

System of equations:

$$\begin{aligned} X_2 - 2X_1 = X_2 - 2X_1 + X_0 &= \gamma W_1 \\ X_3 - 2X_2 + X_1 &= \gamma W_2 \\ \vdots &\quad \vdots \\ -2X_{n-1} - X_{n-2} = X_n - 2X_{n-1} + X_{n-2} &= \gamma W_{n-1} \end{aligned}$$

$$\mathsf{L}_2 = \begin{bmatrix} -2 & 1 & & \\ 1 & -2 & 1 & \\ & \ddots & \ddots & \\ & & 1 & -2 \end{bmatrix} \in \mathbb{R}^{(n-1)\times(n-1)}, \quad X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_{n-1} \end{bmatrix},$$

$$\mathsf{L}_2 X = \gamma W, \quad W \sim \mathcal{N}(0, \gamma^2 \mathsf{I}_{n-1}),$$

Testing a Prior

Given a formula

$$X = \mathsf{L}^{-1} W, \quad W \sim \mathcal{N}(0, \mathsf{I}),$$

the covariance C of X is

$$\mathsf{C} = \mathsf{E}\{XX^{\mathsf{T}}\} = \mathsf{L}^{-1}\mathsf{E}\{WW^{\mathsf{T}}\}\mathsf{L}^{-\mathsf{T}} = \mathsf{L}^{-1}\mathsf{L}^{-\mathsf{T}},$$

or, in terms of the *precision matrix*,

$$\mathsf{C}^{-1} = \mathsf{L}^{\mathsf{T}}\mathsf{L}.$$

The transformation

$$X \to \mathsf{L}X$$

is the *whitening* transformation (or Mahalanobis transformation).

Testing a Prior

To test how well a prior corresponds to the underlying prior assumptions, we may produce random draws from the prior.

$$X \sim \mathcal{N}(x_0, \mathsf{C}),$$

compute any symmetric factorization of the precision matrix,

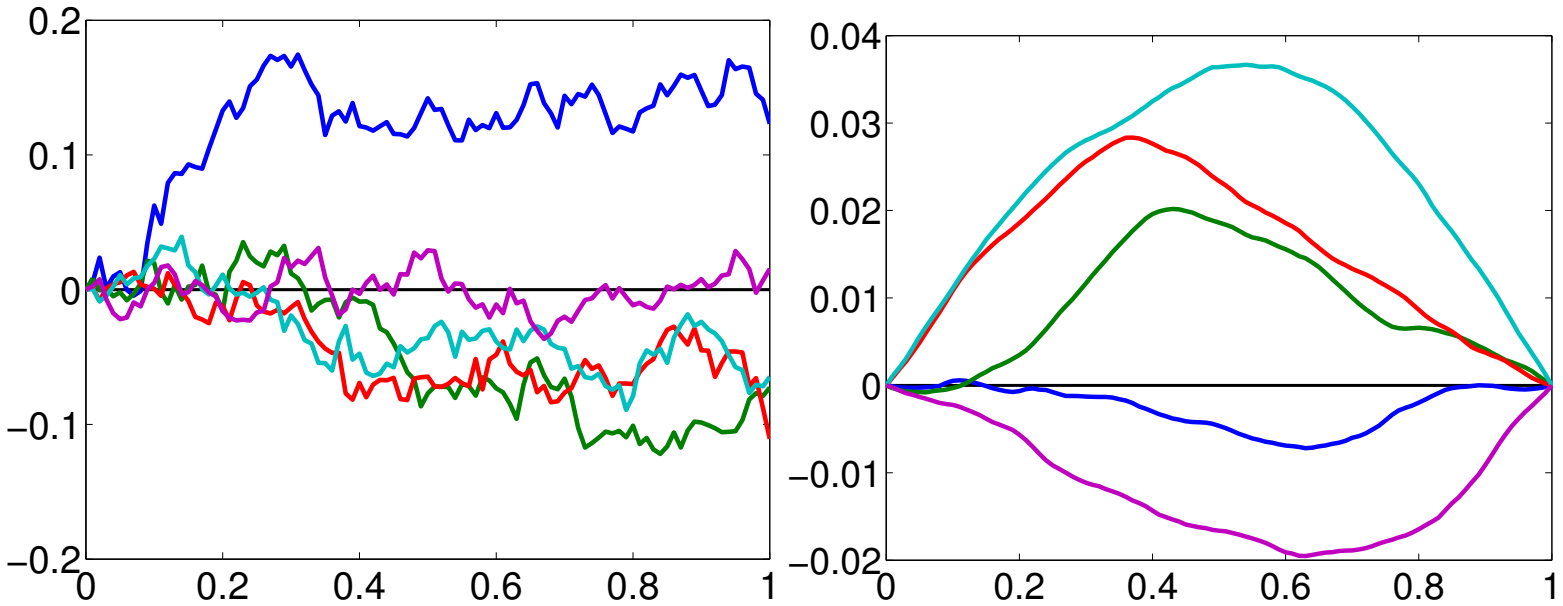
$$\mathsf{C}^{-1} = \mathsf{L}^{\mathsf{T}}\mathsf{L}.$$

Sampling from Gaussian densities

Repeat N times:

1. Draw a realization $w \sim \mathcal{N}(0, \mathsf{I}_n)$
2. Solve $\mathsf{L}(x - x_0) = w$.

Plots of the random draws



Smoothness Priors, Multidimensional

In higher dimensions,

1. Define a differential operator $\mathscr{L}$ over $\Omega \subset \mathbb{R}^d$,
2. Define boundary conditions to have a well-posed BVP
3. Write a discrete approximation
4. Solve the system with d-dimensional random noise (e.g., Gaussian white noise)

Example: Whittle-Matérn prior in a domain $\Omega \subset \mathbb{R}^d$,

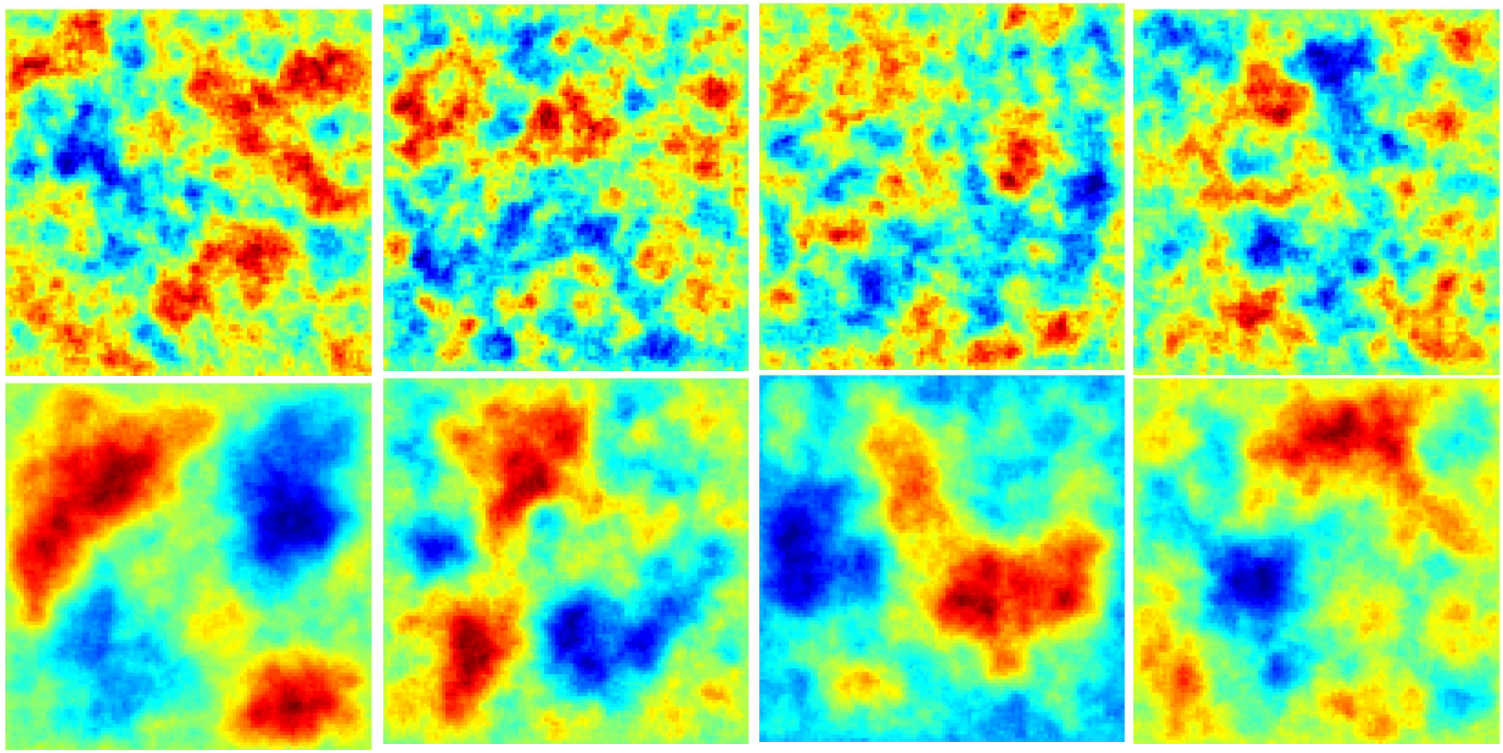
$$\mathscr{L} = -\Delta_D \quad \text{(Dirichlet Laplacian)}.$$

Discretize the Laplacian, e.g., using FEM or FD approximations ($\Delta_D \to \mathsf{L}_D$), and solve

$$(-\mathsf{L}_D + \lambda^{-2}\mathsf{I})^{\beta} X = \gamma W,$$

where $\lambda > 0$ is the *correlation length*, β is the smoothness order.

Smoothness Priors, Multidimensional



Random draws with two different correlation lengths ($\lambda = 0.05$, $\lambda = 0.5$), smoothness parameter $\beta = 1$.

Adding structure

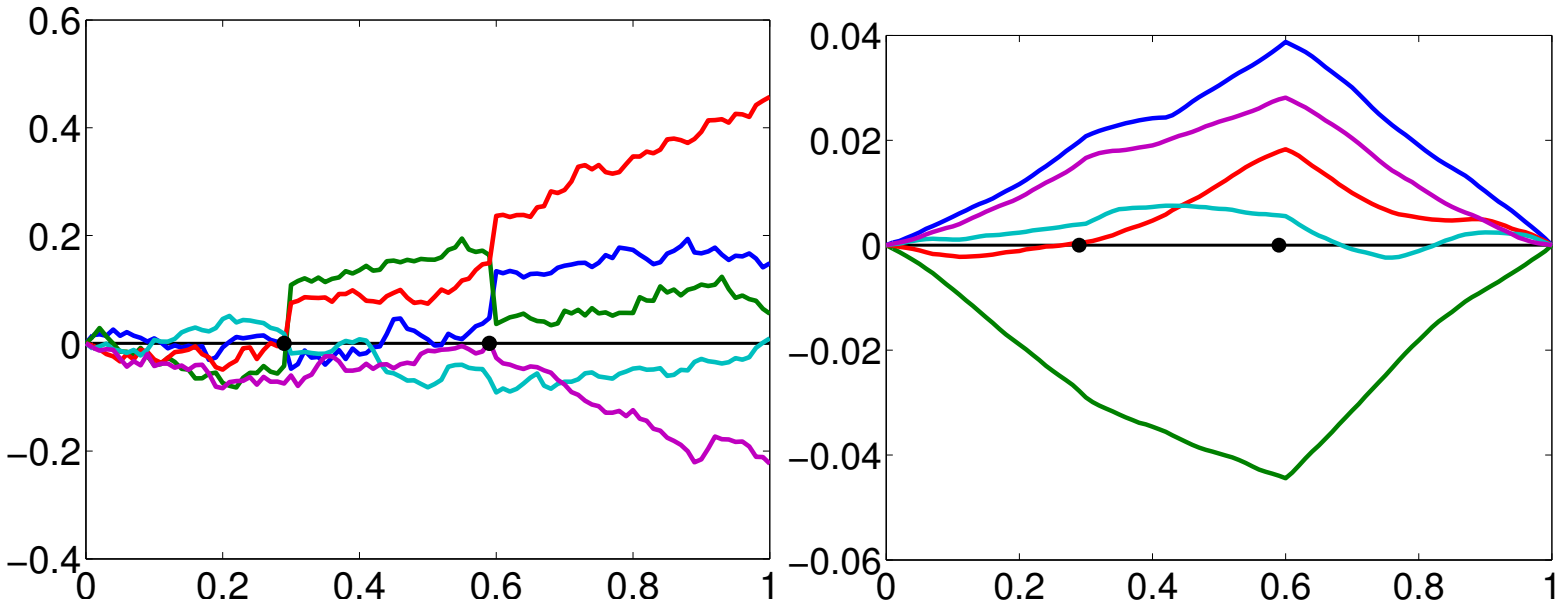
Assume that there is a reason to believe that the slope (or the curvature) may be 10 times higher at isolated points.
If t_k is such point, we replace the condition

$$X_k - X_{k-1} = \gamma W_k$$

by the modified condition

$$X_k - X_{k-1} = 10\gamma W_k.$$

Plots of the random draws



The jumps (or kinks) are allowed, but not forced.

Hypermodels

If the bound on the slope (curvature) is not known

- we can model it also as a random variable θ.
- The prior takes on the form

$$X_k - X_{k-1} = \theta_k^{1/2} W_k$$

- The parameter θ_k is the variance of the Gaussian innovation
- θ_k quantifies the uncertainty in going from X_{k-1} to X_k

Matrix form of hypermodels

In matrix-vector terms

$$\mathsf{L}X = \mathsf{D}_\theta^{1/2} W$$

where

$$\mathsf{D}_\theta = diag\{\theta_1, \theta_2, \ldots, \theta_n\}.$$

Since W is an n-variate standard normal, we can write the probability density of X as

$$\pi_X(x) \propto \exp\left(-\frac{1}{2}\|\mathsf{D}_\theta^{-1/2}\mathsf{L}x\|^2\right).$$

Quantitative prior

If we have information about the

- location
- number
- expected amplitude

of the jumps, it should be encoded in the first order Markov model by setting the corresponding θs.

Qualitative prior

If we only know that *jumps may occur* but no information about how many, where and how big is available, then

- The variance of the innovation is unknown.
- The variance is modeled as a random variable
- The estimation of the variance of the Markov process is part of the inverse problem
- The prior for the problem is the joint prior for X and Θ

$$\pi_{X,\Theta}(x,\theta) = \pi_{X|\Theta}(x \mid \theta)\pi_{\Theta}(\theta)$$

Conditional smoothness prior

If we had the variance information, the original smoothness prior for X would be determined. Since the variance vector is unknown, we cannot ignore the normalizing factor:

$$\pi_{X|\Theta}(x \mid \theta) = \left(\frac{det(\mathsf{L}^T \mathsf{D}_\theta^{-1} \mathsf{L})}{(2\pi)^n} \right)^{1/2} \exp\left(-\frac{1}{2} \|\mathsf{D}_\theta^{-1/2} \mathsf{L} x\|^2 \right)$$

If L is invertible, there is an analytic expression for the determinant,

$$det(\mathsf{L}^T \mathsf{D}_\theta^{-1} \mathsf{L}) \propto \frac{1}{\prod_{j=1}^n \theta_j},$$

$$\pi_{X|\Theta}(x \mid \theta) \propto \exp\left(-\frac{1}{2} \|\mathsf{D}_\theta^{-1/2} \mathsf{L} x\|^2 - \frac{1}{2} \sum_{j=1}^n \log \theta_j \right).$$

Conditional smoothness prior

- Choosing the hyperprior π_Θ so that it promotes sparse solution is a topic that will be discussed later.
- The resulting model is referred to as a hierarchical model
- Care must be taken if L is not an invertible matrix.

Data driven prior

Given a sample of typical solutions,

$$\mathscr{S} = \{x^{(1)}, x^{(2)}, \ldots, x^{(p)}\}, \quad x^{(j)} \in \mathbb{R}^n,$$

find a Gaussian prior π_X such that

- π_X is concentrated in the affine subspace $\mathscr{H}$ spanned by $\{x^{(1)}, \ldots, x^{(p)}\}$,
- $\pi_X\big|_{\mathscr{H}}$ is distributed according to the sample.

Note: The prior may be **degenerate**, that is, supported in a proper affine subspace.

Data driven prior

Arrange the data in a matrix

$$\mathsf{X} = \left[\begin{array}{cccc} x^{(1)} & x^{(2)} & \cdots & x^{(p)} \end{array} \right] \in \mathbb{R}^{n \times p}.$$

Compute the mean value

$$\overline{x} = \frac{1}{p} \sum_{j=1}^{p} x^{(j)}.$$

Center the data:

$$\mathsf{X}_c = \left[\begin{array}{cccc} x_c^{(1)} & x_c^{(2)} & \cdots & x_c^{(p)} \end{array} \right], \quad x_c^{(j)} = x^{(j)} - \overline{x}.$$

Compute the empirical covariance

$$\Gamma = \frac{1}{p} \sum_{j=1}^{p} x_c^{(j)} \left(x_c^{(j)} \right)^{\mathsf{T}} = \frac{1}{p} \mathsf{X}_c \mathsf{X}_c^{\mathsf{T}}.$$

Data driven prior

To draw from the density:

- Compute

$$X_c = U\Sigma V^T.$$

- Given

$$\xi \sim \mathcal{N}(0, I_p),$$

define a random variable

$$X = \overline{x} + \frac{1}{\sqrt{p}} U\Sigma\xi.$$

Data driven prior

This X is a Gaussian random variable with

- Mean

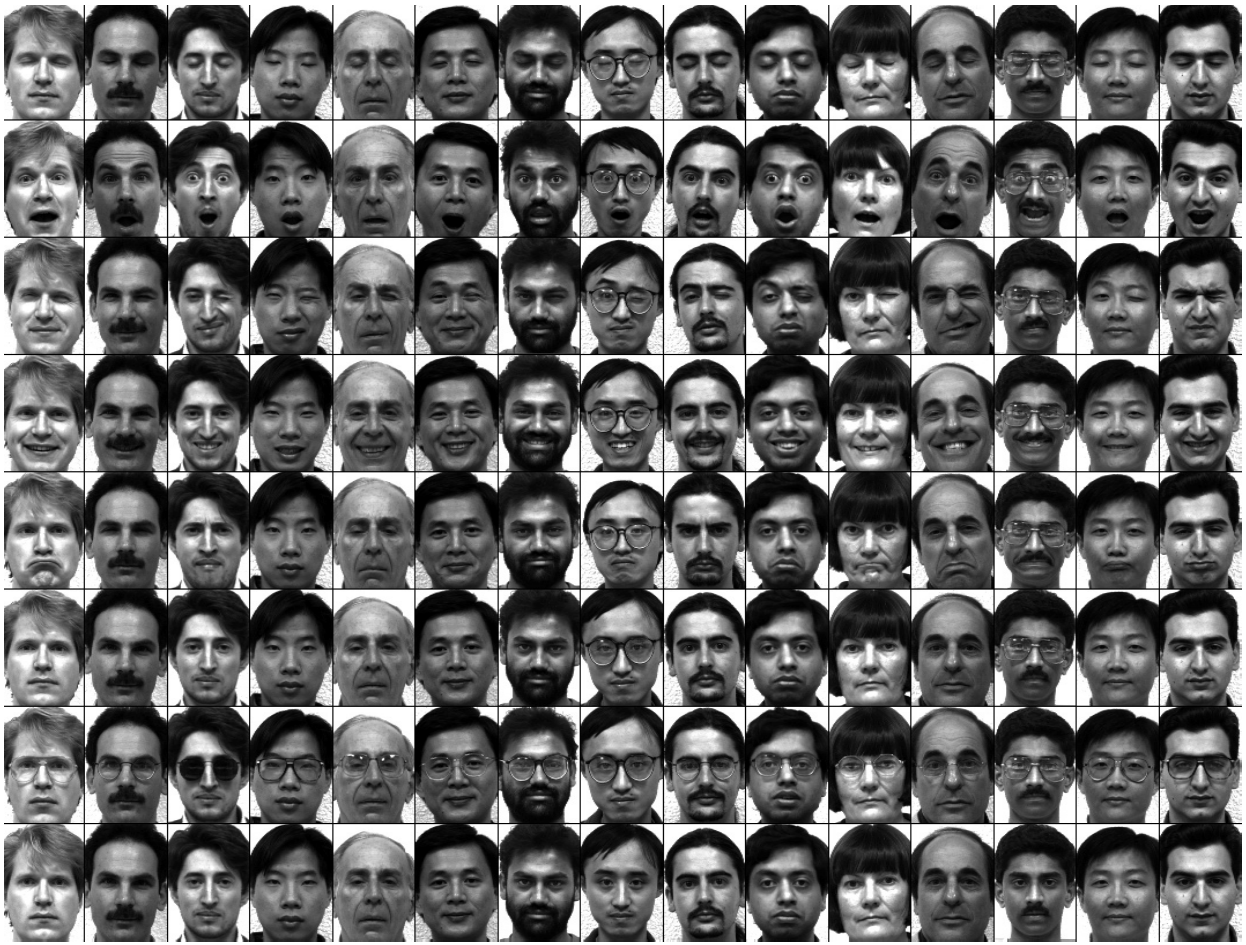
$$\mathrm{E}\{X\} = \overline{x},$$

- Covariance

$$\begin{aligned} \mathrm{Cov}(X) &= \mathrm{E}\{(X-\overline{x})(X-\overline{x})^{\mathsf{T}}\} = \frac{1}{p}\mathsf{U}\Sigma \underbrace{\mathrm{E}\{\xi\xi^{\mathsf{T}}\}}_{\mathsf{I}_p=\mathsf{V}^{\mathsf{T}}\mathsf{V}} \Sigma^{\mathsf{T}}\mathsf{U}^{\mathsf{T}} \\ &= \frac{1}{p}\mathsf{U}\Sigma\mathsf{V}^{\mathsf{T}}\mathsf{V}\Sigma^{\mathsf{T}}\mathsf{U}^{\mathsf{T}} = \frac{1}{p}\mathsf{X}_c\mathsf{X}_c^{\mathsf{T}} \\ &= \mathsf{D}. \end{aligned}$$

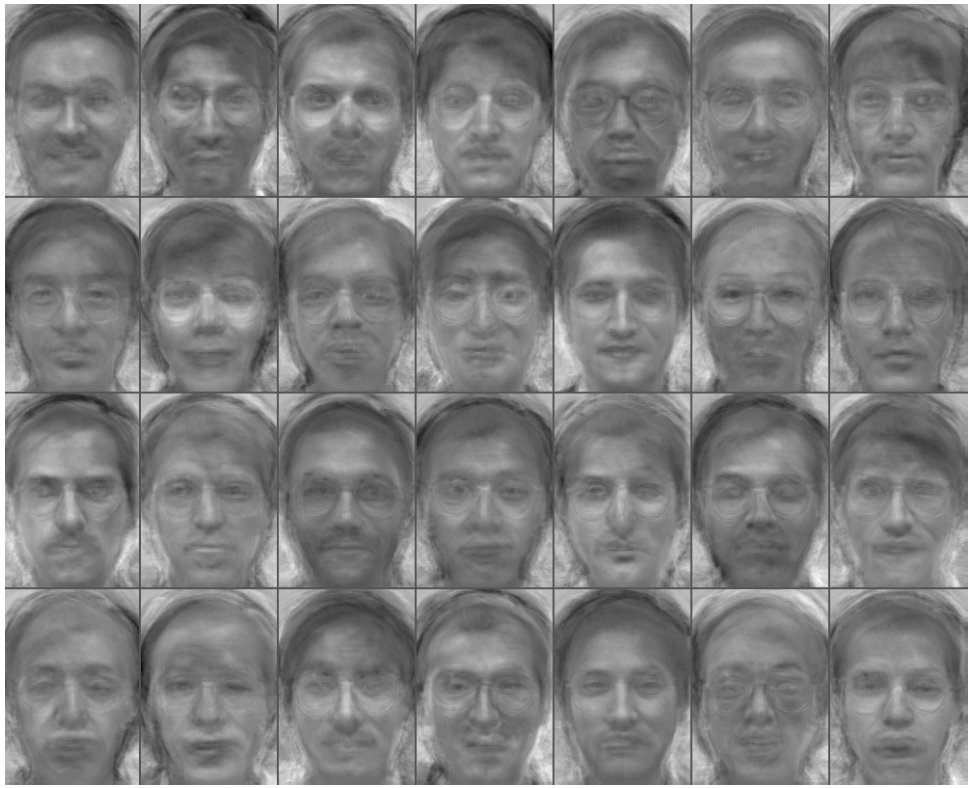
Observe: D may be symmetric positive **semidefinite**, since the sample vectors span only a subspace.

Data driven prior: An Example



From Yale Face Database
(http://cvc.cs.yale.edu/cvc/projects/yalefaces/yalefaces.html)

Sample-based prior: An Example



Mean vector (left) and 28 random draws from the density (right)

Interlude: Drawing from distributions

So far, we have dealt with Gaussian distributions:

$$X \sim \mathcal{N}(\mu, \mathsf{C}).$$

To draw from a Gaussian, take **any** symmetric factorization of C,

$$\mathsf{C} = \mathsf{R}^{\mathsf{T}}\mathsf{R} \quad \text{(e.g., Cholesky factorization)}.$$

Write

$$X = \mathsf{R}^{\mathsf{T}} W, \quad W \sim \mathcal{N}(0, \mathsf{I}_n).$$

Check:

$$\mathrm{cov}(X) = E\{XX^{\mathsf{T}}\} = \mathsf{R}^{\mathsf{T}} \underbrace{E\{WW^{\mathsf{T}}\}}_{=\mathsf{I}} \mathsf{R} = \mathsf{R}^{\mathsf{T}}\mathsf{R} = \mathsf{C}.$$

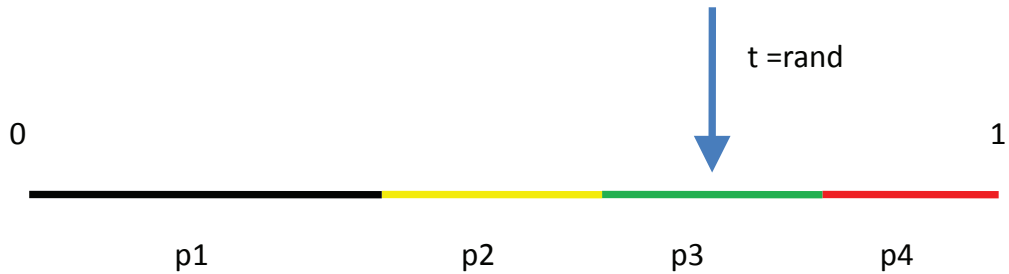
Interlude: Drawing from distributions

Consider a finite state space: X takes on values $\{e_1, e_2, \ldots, e_n\}$, and

$$p_j = P\{X = e_j\} = \text{probability of the event } X = e_j,$$

where

$$p_j \geq 0, \quad \sum_{j=1}^{n} p_j = 1.$$



Interlude: Drawing from distributions

Matlab code:

```
Phi = 0;
ell = 0;
xi = rand;
while Phi < xi
    ell = ell+1;
    Phi = Phi + p(ell);
end
```

Set $x = e_\ell$.

Interlude: Drawing from distributions

Poisson distribution: Infinite discrete state space.
We have

$$p_{n+1} = \frac{\lambda^{n+1}}{(n+1)!} e^{\lambda} = \frac{\lambda}{n+1} p_n, \quad p_0 = e^{-\lambda}.$$

Matlab code:

```
Phi = 0;
ell = 0;
xi = rand;
p = exp(-lambda);
while Phi < xi
    ell = ell+1;
    p = lambda/ell*p;
    Phi = Phi + p;
end
```

Set $x = \ell$.

Interlude: Drawing from distributions

Drawing from probability density over $\mathbb{R}$: Define the cumulative distribution function (CDF),

$$\Phi_X(x) = \int_{-\infty}^{x} \pi_X(t)dt, \quad \Phi'_X(x) = \pi_X(x).$$

Observe: Φ_X is non-decreasing, and

$$0 = \lim_{x\to-\infty} \Phi_X(x) \leq \Phi_X(x) \leq \lim_{x\to\infty} \Phi_X(x) = 1.$$

Define a new random variable

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

Interlude: Drawing from distributions

For simplicity, assume that Φ_X is strictly increasing, so we may write

$$X = \Phi_X^{-1}(T).$$

For any α, $0 < \alpha < 1$, we calculate

$$P\{T < x\} = P\{X < \Phi_X^{-1}(\alpha)\} = \int_{-\infty}^{\Phi_X^{-1}(\alpha)} \pi_X(x)dx.$$

Change of variables:

$$t = \Phi_X(x), \quad dt = \Phi_X'(x)dx = \pi_X(x)dx,$$

so we obtain

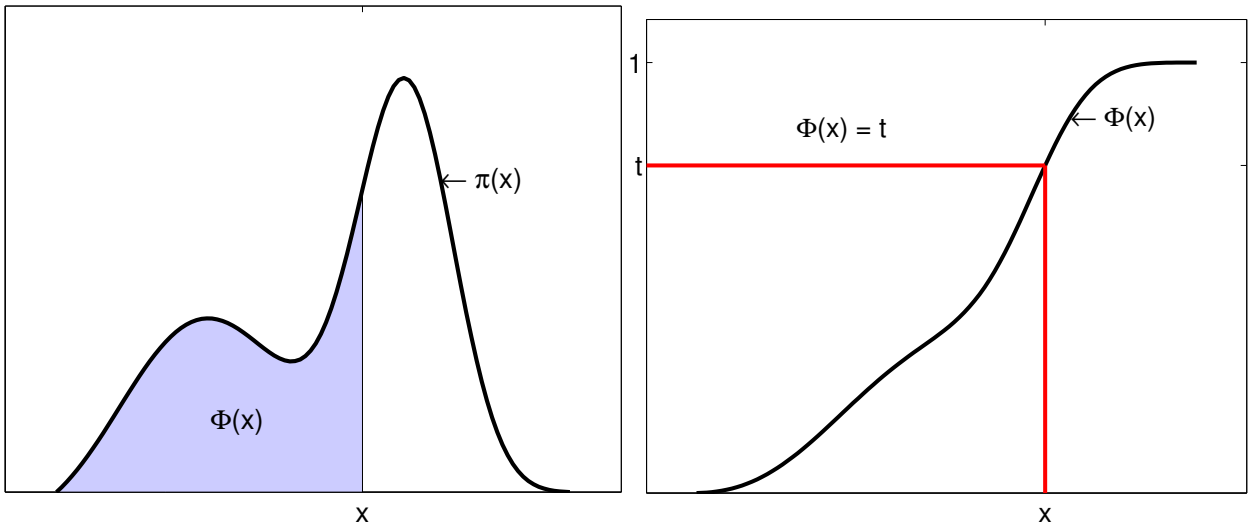
$$P\{T < x\} = \int_0^{\alpha} dt = \alpha.$$

Interlude: Drawing from distributions

Conclusion: T is uniformly distributed over $[0, 1]$.

Algorithm for drawing from the density π_X:

1. Draw $T \sim \mathrm{Uniform}([0, 1])$, $(\texttt{t = rand}) \rightarrow t \in [0, 1]$;
2. Solve x from $\Phi_X(x) = t$.



Posterior density, Gaussian models

Consider the linear model

$$b = \mathsf{A}x + e, \quad e \sim \mathcal{N}(0, \Sigma).$$

Likelihood:

$$\pi_{B|X}(b \mid x) \propto \exp\left(-\frac{1}{2}(b - \mathsf{A}x)^{\mathsf{T}}\Sigma^{-1}(b - \mathsf{A}x)\right).$$

Assume a Gaussian prior:

$$X \sim \mathcal{N}(0, \mathsf{D}),$$

or, in terms of densities,

$$\pi_X(x) \propto \exp\left(-\frac{1}{2}x^{\mathsf{T}}\mathsf{D}^{-1}x\right).$$

Posterior density, Gaussian models

Bayes' formula:

$$\begin{aligned} p_{X|B}(x \mid b) &\propto p_X(x) p_{B|X}(b \mid x) \\ &= \exp\left(-\frac{1}{2} x^{\mathsf{T}} \mathsf{D}^{-1} x - \frac{1}{2} (b - \mathsf{A}x)^{\mathsf{T}} \Sigma^{-1} (b - \mathsf{A}x) \right). \end{aligned}$$

- The covariance matrices D and Σ are symmetric positive definite (SPD).
- This implies that the precision matrices D^{-1} and Σ^{-1} are SPD.
- Therefore, they allow symmetric factorizations (e.g. Cholesky):

$$\mathsf{D}^{-1} = \mathsf{L}^{\mathsf{T}} \mathsf{L}, \quad \Sigma^{-1} = \mathsf{S}^{\mathsf{T}} \mathsf{S}.$$

Posterior density, Gaussian models

We have

$$x^{\mathsf{T}} \mathsf{D}^{-1} x = x^{\mathsf{T}} \mathsf{L}^{\mathsf{T}} \mathsf{L} x = \|\mathsf{L}x\|^2,$$

and

$$\begin{aligned} (b - \mathsf{A}x)^{\mathsf{T}} \Sigma^{-1} (b - \mathsf{A}x) &= (b - \mathsf{A}x)^{\mathsf{T}} \mathsf{S}^{\mathsf{T}} \mathsf{S} (b - \mathsf{A}x) \\ &= \|\mathsf{S}(b - \mathsf{A}x)\|^2. \end{aligned}$$

Hence, the posterior density is

$$\pi_{X|B}(x \mid b) \propto \exp\left(-\frac{1}{2}\|\mathsf{L}x\|^2 - \frac{1}{2}\|\mathsf{S}(b - \mathsf{A}x)\|^2 \right).$$

Posterior density, Gaussian models

Connection to classical regularization theory: Assume that

$$\Sigma = \sigma^2 \mathsf{I} \Rightarrow \mathsf{S} = \frac{1}{\sigma}\mathsf{I}.$$

Then

$$\pi_{X|B}(x \mid b) \propto \exp\left(-\frac{1}{2}\|\mathsf{L}x\|^2 - \frac{1}{2\sigma^2}\|b - \mathsf{A}x\|^2\right).$$

The maximizer of the posterior density is called the **Maximum A Posteriori** (MAP) estimate, and we see that

$$\begin{aligned} x_{\mathrm{MAP}} &= \operatorname{argmax}\{\pi_{X|B}(x \mid b)\} \\ &= \operatorname{argmin}\{\|b - \mathsf{A}x\|^2 + \sigma^2\|\mathsf{L}x\|^2\}, \end{aligned}$$

which is the Tikhonov regularized solution with regularization parameter σ.

What about sparsity?

A **sparse vector** in $\mathbb{R}^n$ is a vector x with most components equal to zero.

- The **support of** x is the index set $I \subset \{1, 2, \ldots, n\}$ corresponding to the non-zero components. We write

$$I = \mathrm{supp}(x),$$

- Notation:

$$\|x\|_0 = \text{cardinality of } \mathrm{supp}(x), \text{ or number of non-zero entries.}$$

Sparsity-promoting priors favor solutions x such that

$$\|x\|_0 \ll n.$$

Sparsity Considerations

Sparsity means a signal with a *sparse representation*

- The sparse vector in that case contains the coefficients of a suitable representation, for example
- Wavelet basis
- Fourier basis
- First order differencing matrix for piecewise constant signals in terms of their increments

ℓ_p-priors

Write the ℓ_p-norm of a vector as

$$\|x\|_p = \left(\sum_{j=1}^{n} |x_j|^p \right)^{1/p}.$$

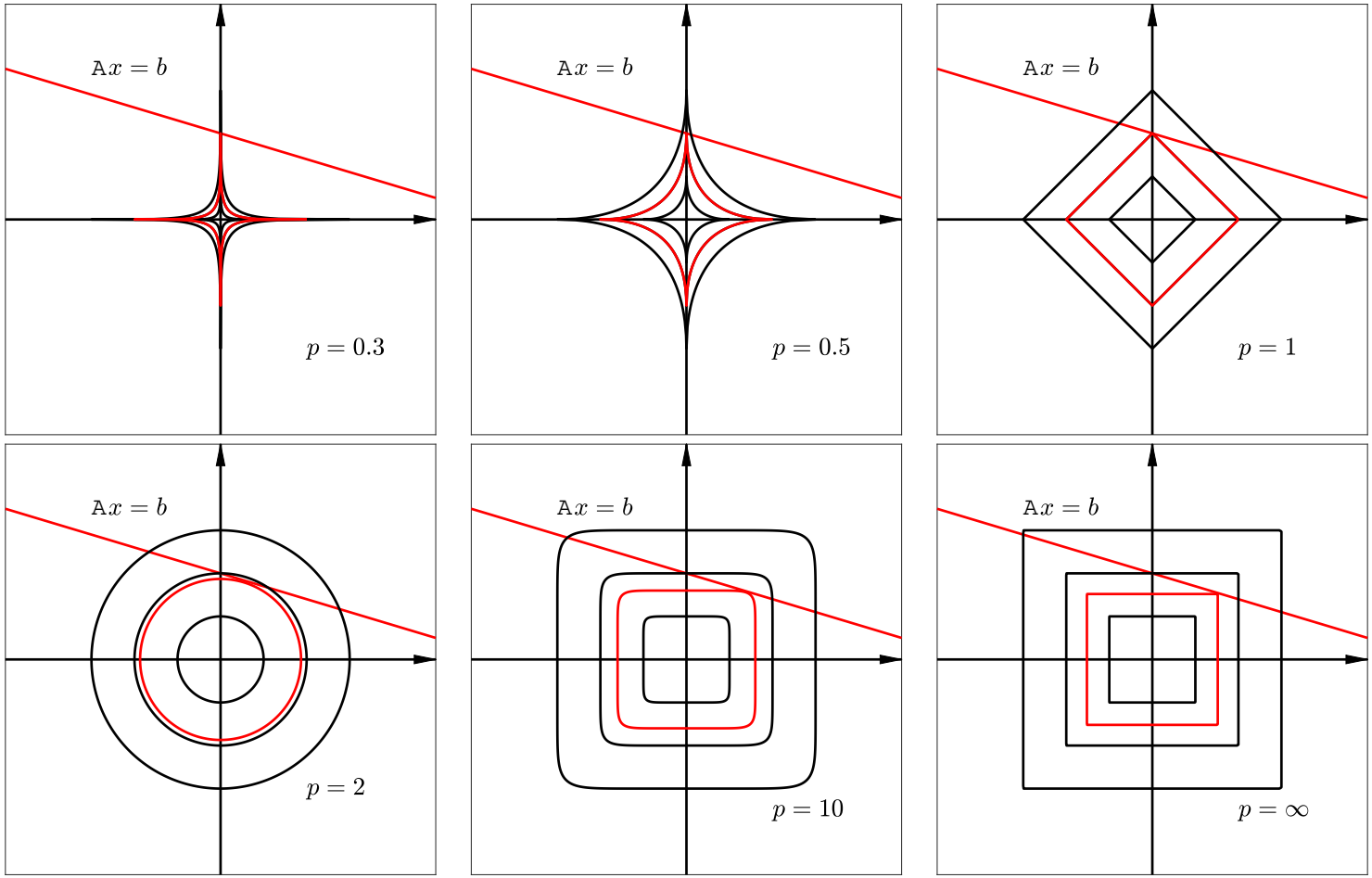
Writing a non-Gaussian prior density

$$\pi_X(x) \propto \exp\left(-\alpha \|x\|_p\right),$$

combined with a Gaussian likelihood leads to a posterior density

$$\pi_{X|B}(x \mid b) \propto \exp\left(-\frac{1}{2}\|\mathsf{S}(b - \mathsf{A}x)\|^2 - \alpha \|x\|_p \right).$$

Promote sparsity via norm penalization



Alternative: Conditionally Gaussian priors

Recall:

1. A conditionally Gaussian prior

$$X \sim \mathcal{N}(0, \mathrm{D}_\theta), \quad \mathrm{D}_\theta = \mathrm{diag}(\theta_1, \ldots, \theta_n),$$

2. where the unknown variances $\theta_j > 0$ are mutually independent random variables following a gamma distribution,

$$\Theta_j \sim \mathrm{Gamma}(\beta, \theta_j^*) \propto \theta_j^{\beta-1} \exp\left(-\frac{\theta_j}{\theta_j^*}\right), \quad 1 \leq j \leq n.$$

Alternative: Conditionally Gaussian priors

Why gamma distribution?

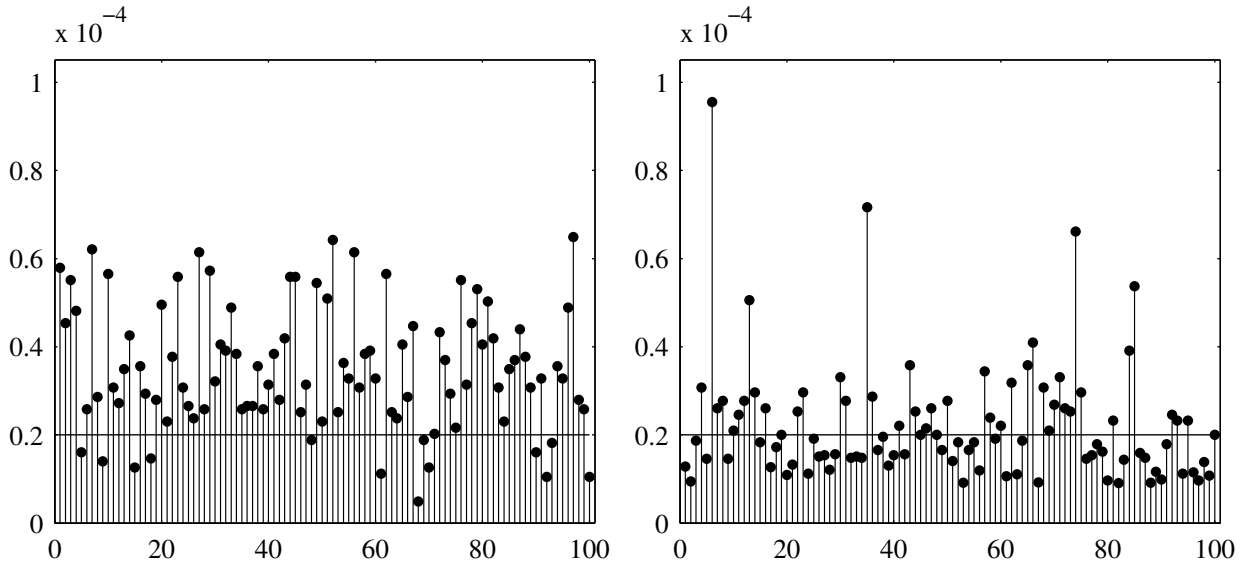
- Gamma distribution has a "fat tail": It goes to zero relatively slowly (compared to Gaussians).
- Draws from a fat-tailed (or "leptokurtic") distributions typically produce outliers.
- The possibility of outliers allows to have occasionally a large variance, and therefore a large component x_j.

Note: Gamma distribution is not the only choice, fat-tailed distributions are common. Gamma distribution facilitates computations.
One possible alternative: Inverse gamma:

$$\Theta_j \sim \text{InvGamma}(\beta, \theta_j^*) \propto \theta_j^{-\beta-1} \exp\left(-\frac{\theta_j^*}{\theta_j}\right).$$

Alternative: Conditionally Gaussian priors

Random draws from gamma and inverse gamma distributions:



Alternative: Conditionally Gaussian priors

Hierarchical model: Treat the pair $(X\Theta)$ as the unknown, and write a hierarchical prior model

$$\pi_{X,\Theta}(x,\theta) = \pi_{X|\Theta}(x \mid \theta)\pi_{\Theta}(\theta),$$

and then the posterior density as

$$\begin{aligned}
\pi_{X,\Theta|B}(x,\theta) &\propto \pi_{X,\Theta}(x,\theta)\pi_{B|X}(b \mid x) \\
&\propto \exp\left(-\frac{1}{2}\|\mathsf{S}(b-\mathsf{A}x)\|^2 - \frac{1}{2}\sum_{j=1}^{n}\frac{x_j^2}{\theta_j} + \eta\sum_{j=1}^{n}\log\theta_j - \sum_{j=1}^{n}\frac{\theta_j}{\theta_j^*}\right)
\end{aligned}$$

where $\eta = \beta - 3/2 > 0$.

Iterared Alternating Sequential (IAS) algorithm

To compute x_{MAP} we minimize the Gibbs energy

$$\mathscr{E}(x;\theta) = \overbrace{\frac{1}{2}\|\mathsf{S}(b-\mathsf{A}x)\|^2 + \sum_{j=1}^{n}\frac{x_j^2}{2\theta_j}}^{(a)} \underbrace{- \sum_{j=1}^{n}\left(\eta\log\theta_j - \frac{\theta_j}{\theta_j^*}\right)}_{(b)} \tag{1}$$

Given the initial value $\theta^0 = \theta^*$, $x^0 = 0$, and $k = 0$, iterate until convergence:

(a) Update $x^k \to x^{k+1}$ by minimizing $\mathscr{E}(x;\theta^k)$;

(b) Update $\theta^k \to \theta^{k+1}$ by minimizing $\mathscr{E}(x^{k+1};\theta)$;

(c) Increase $k \to k+1$.

IAS algorithm

Initialize: $k = 0, \ \theta_0 = \theta^*$;

While $\|\theta_k - \theta_{k-1}\| > \text{tol}$

1. Update x; Set $\theta = \theta_k$, and $x_{k+1} = \text{argmin} \left\{ \|\mathsf{S}(b - \mathsf{A}x)\|^2 + \|\mathsf{D}_\theta^{-1/2} x\|^2 \right\}$ by solving

$$\begin{bmatrix} \mathsf{SA} \\ \mathsf{D}_\theta^{-1/2} \end{bmatrix} x = \begin{bmatrix} \mathsf{S}b \\ 0 \end{bmatrix}$$

   in the least squares sense.
2. Update θ; Set $x = x_{k+1}$, update the components of θ_{k+1} according to the formula

$$\theta_j = \theta_j^* \left(\frac{\eta}{2} + \sqrt{\frac{\eta^2}{4} + \frac{x_j^2}{2\theta_j^*}} \right)$$

Iterated Sequential Alternative (IAS) minimization algorithm.

IAS algorithm

One can prove that

1. The minimization problem of the Gibbs energy has a unique minimizer,
2. The IAS algorithm converges to that minimizer,
3. If a sparse solution exists, the convergence outside the support is quadratic, and in general, at least linear.

Moreover, as $\eta \to 0+$, the solution x_η of the IAS algorithm converges to x_0, the minimizer of the functional

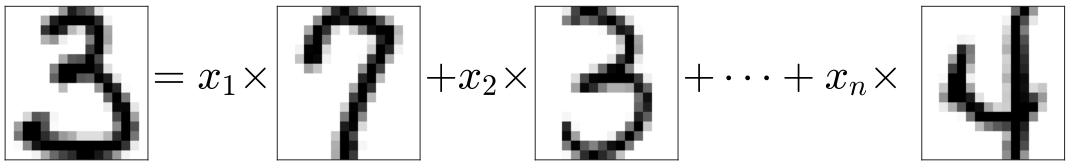
$$\|\mathsf{S}(b - \mathsf{A}x)\|^2 + \sqrt{2}\sum_{j=1}^{n} \frac{|x_j|}{\sqrt{\theta_j^*}},$$

and the values θ_j^* are related to the sensitivity of the data to different components x_j.

MNIST Data: Handwritten digits

Dictionary learning:

- Dictionary consists of N 16 gray scale images of handwritten digits.
- Annotation: $c_j \in \{0, 1, 2, \ldots, 9\}$.
- b is a handwritten digits not in the dictionary set.
- The annotation of b can be based on the coefficients x_j.



Example: Dictionary Learning

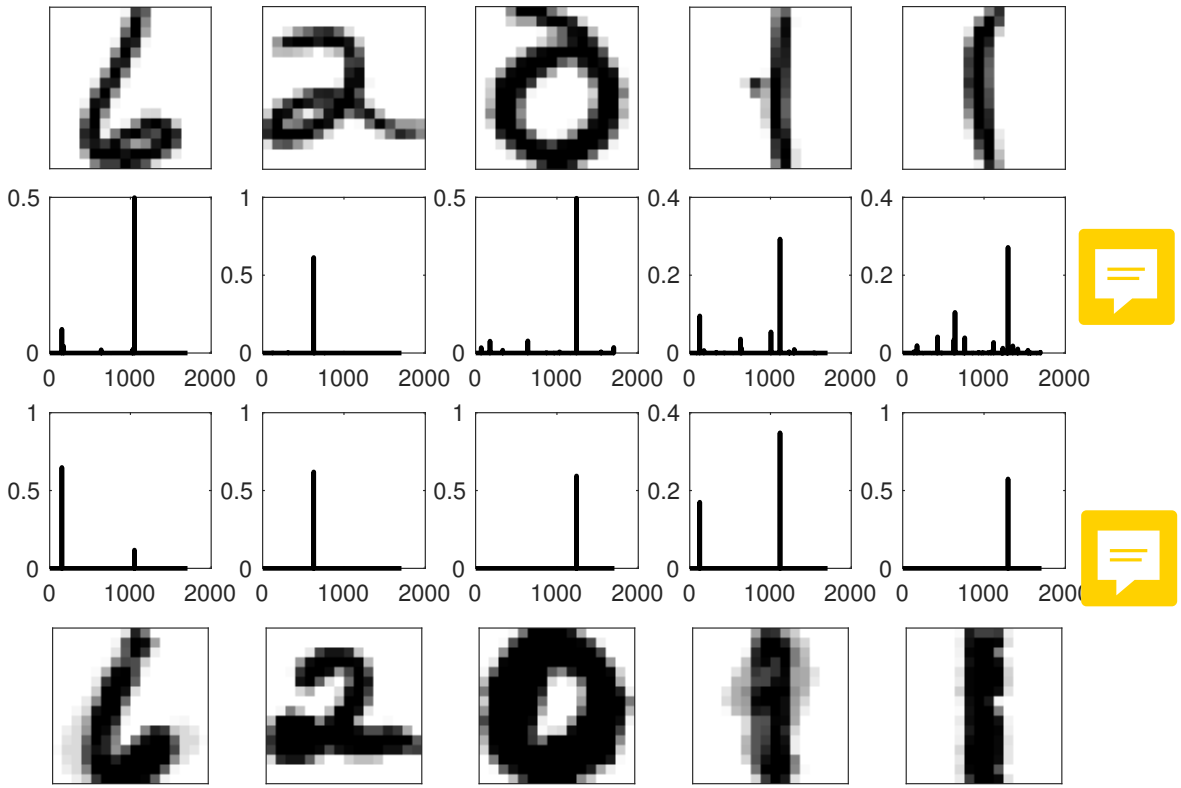
- MNIST data: Training set $n = 1707$ annotated hand-written digits $\mathsf{V}^{(j)} \in \mathbb{R}^{16\times 16}$.
- $\mathsf{B} \in \mathbb{R}^{16\times 16}$ drawn from an independent set of digits.
- Set

$$b = \mathrm{vec}(\mathsf{B}) \in \mathbb{R}^{256}, \quad v^{(j)} \in \mathbb{R}^{256}.$$

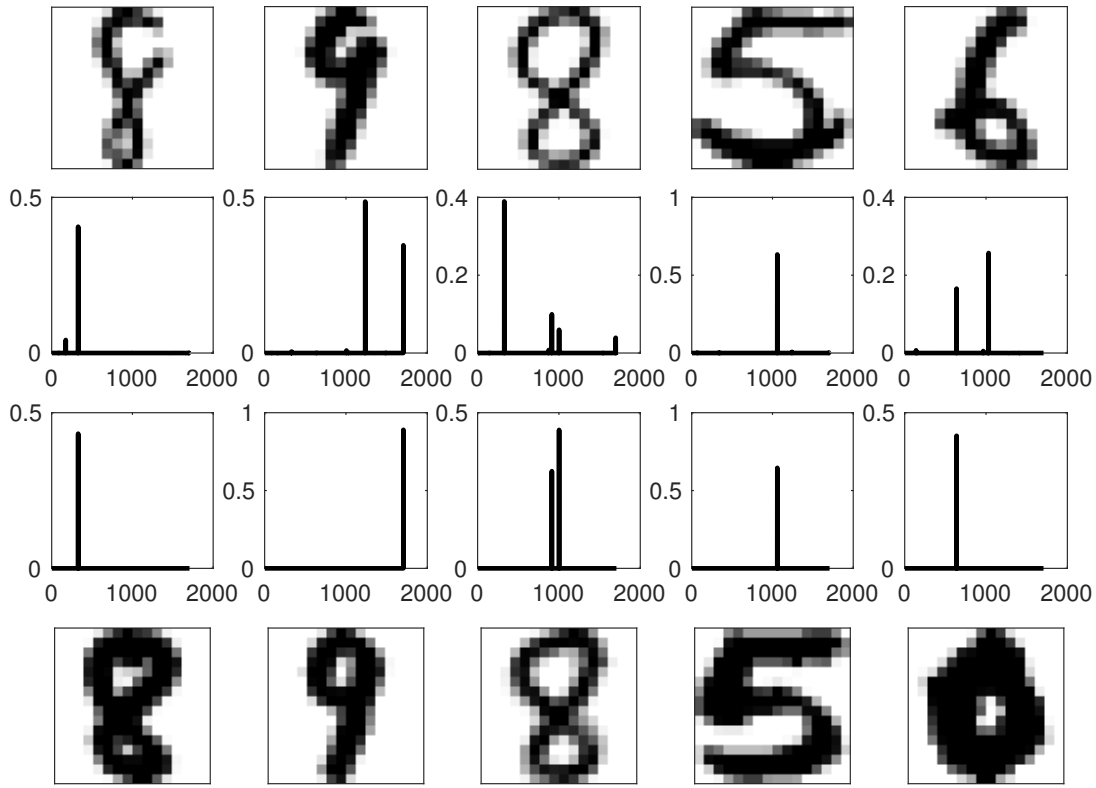
- Set the noise distribution

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathsf{I}), \quad \sigma = 0.1 \quad \text{(educated guess)}.$$

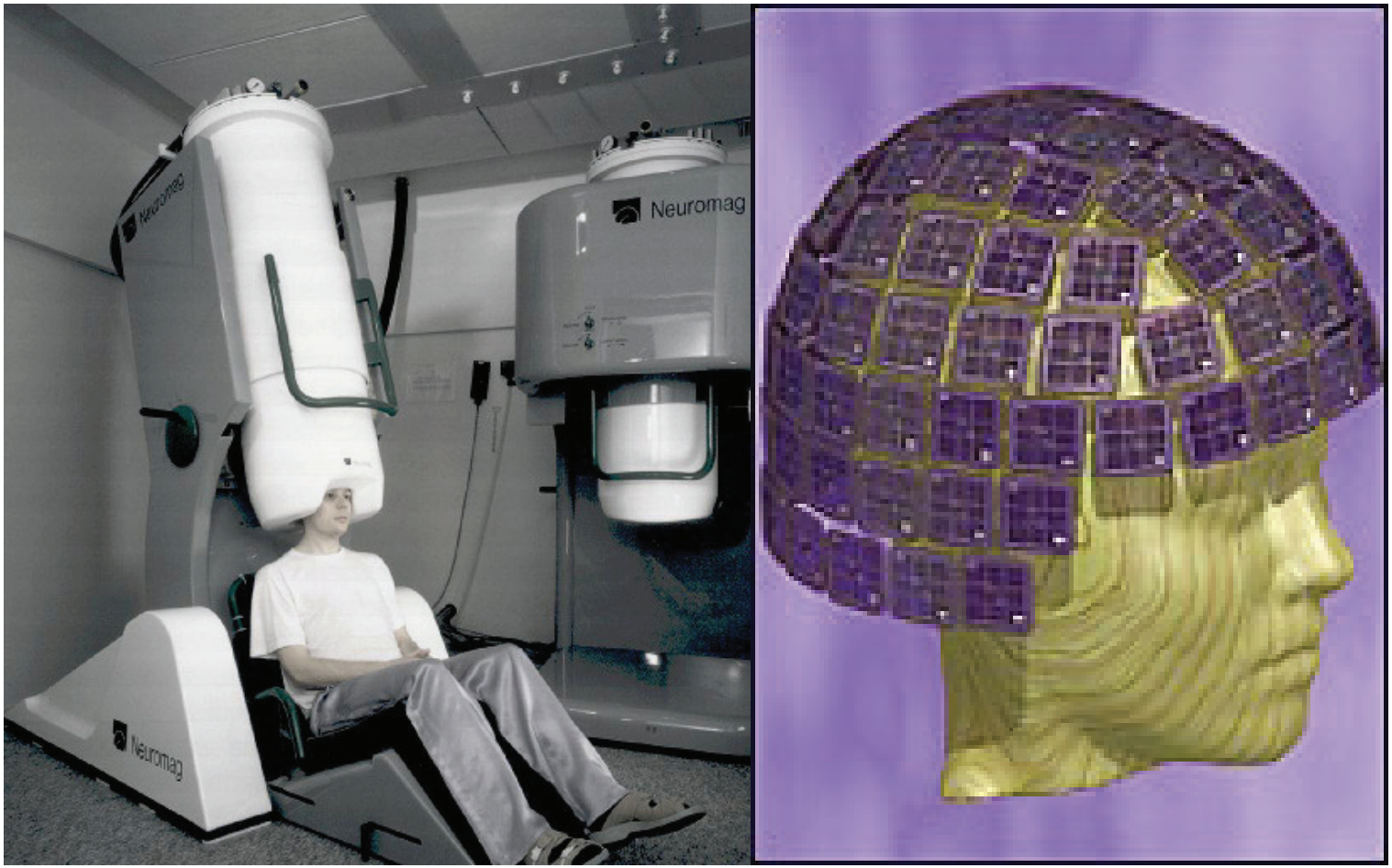
Dictionary Learning



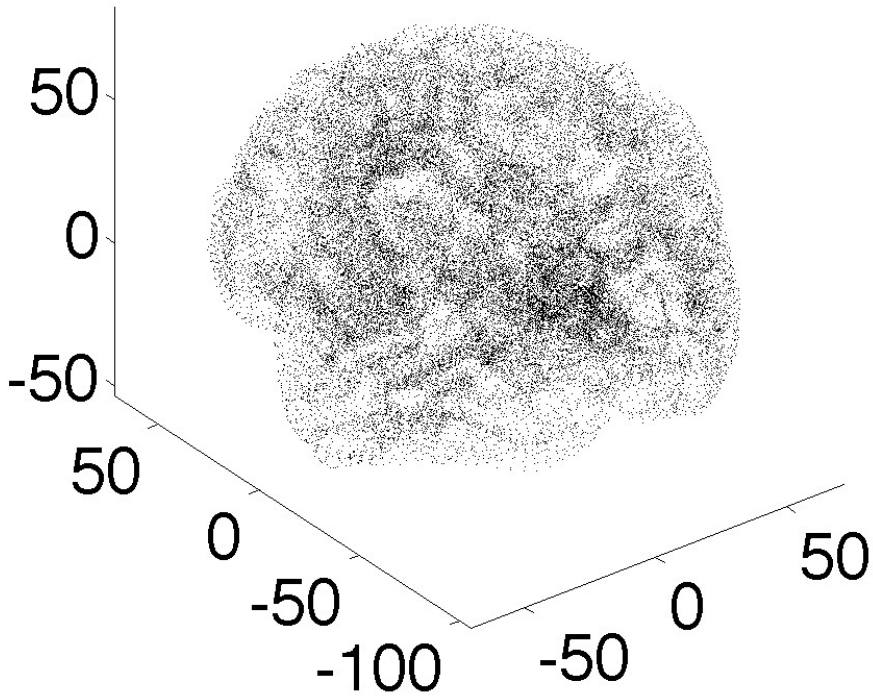
Dictionary Learning



Magnetoencephalography (MEG) Revisited



Discretization: Dipole model



Brain model based on segmented MRI image. The grid points $\vec{r}_j$ represent the gray matter, and are the possible dipole locations.

MEG forward model

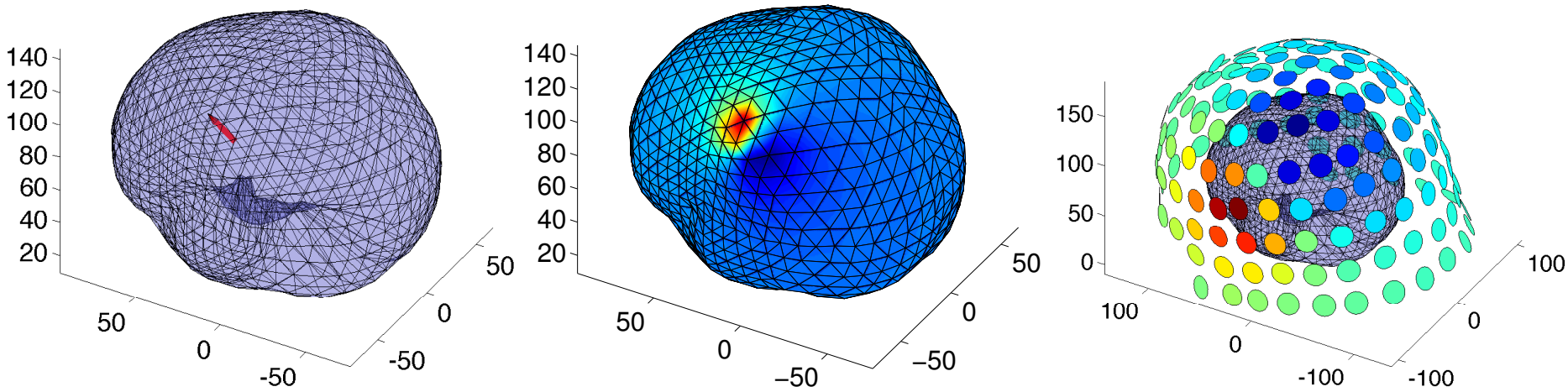
- Electric source currents (neuronal activity) create Ohmic volume currents in the brain tissue,
- All currents induce a magnetic field outside the head,
- The magnetometers measure the weak induced field.

By linearity of Maxwell's equations, it is possible to write

$$\beta_k = \int_\Omega \vec{\mathscr{M}}_k(\vec{r}) \cdot \vec{J}(\vec{r}) d\vec{r},$$

where $\vec{\mathscr{M}}_k$ depends on the geometry and conductivity of the head.
assuming constant conductivity, $\vec{\mathscr{M}}_k$ can be approximated numerically by Boundary Element Method (BEM).

Geometry



Current dipole (left), the surface electric potential computed by BEM (center), and the magnetic field at the magnetometers (right).

Back to Linear Algebra

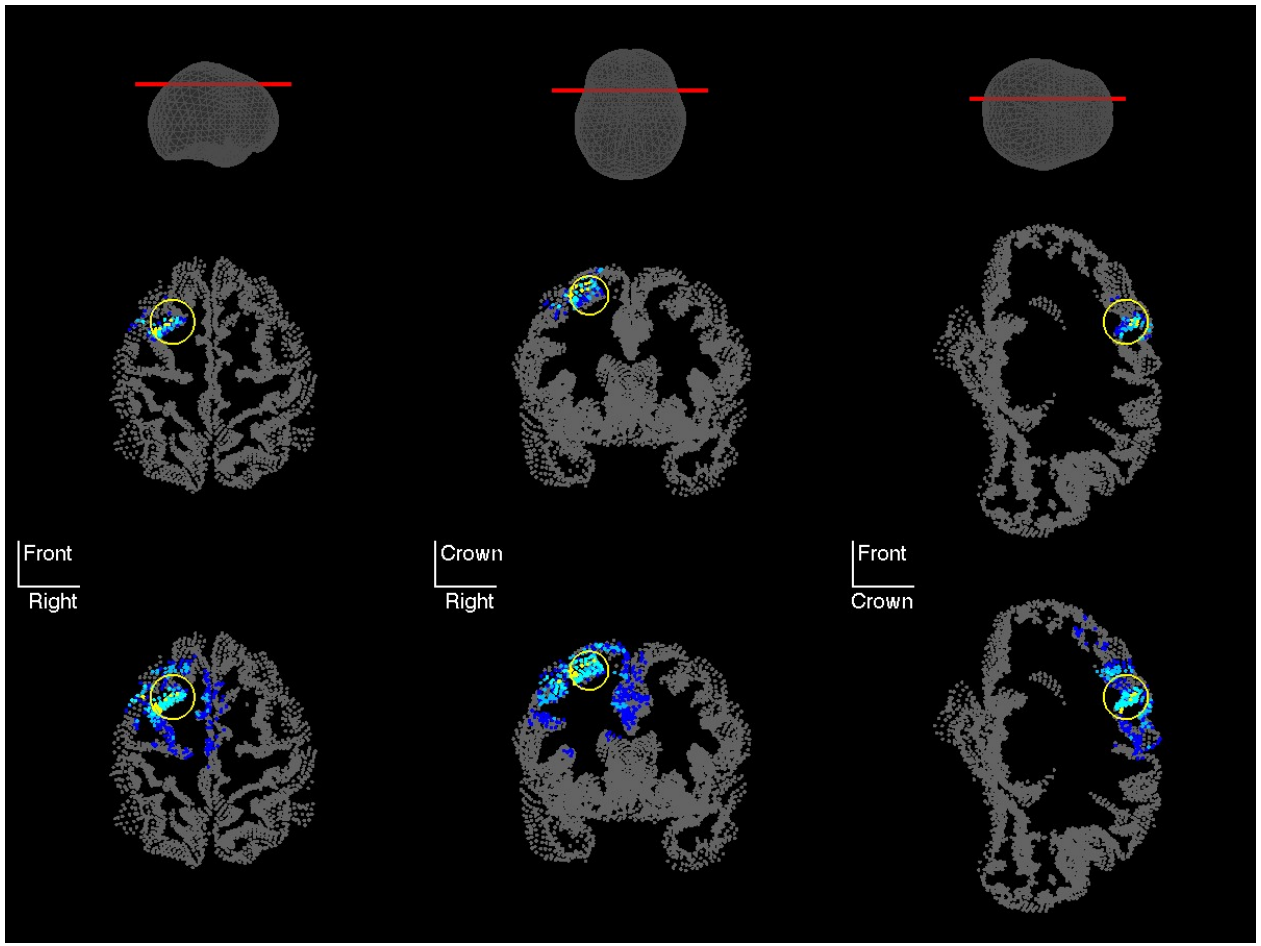
The physical model, when discretized, leads to a linear problem

$$b = Ax + \varepsilon,$$

where $A \in \mathbb{R}^{153 \times 122\,886.}$.
Prior model:

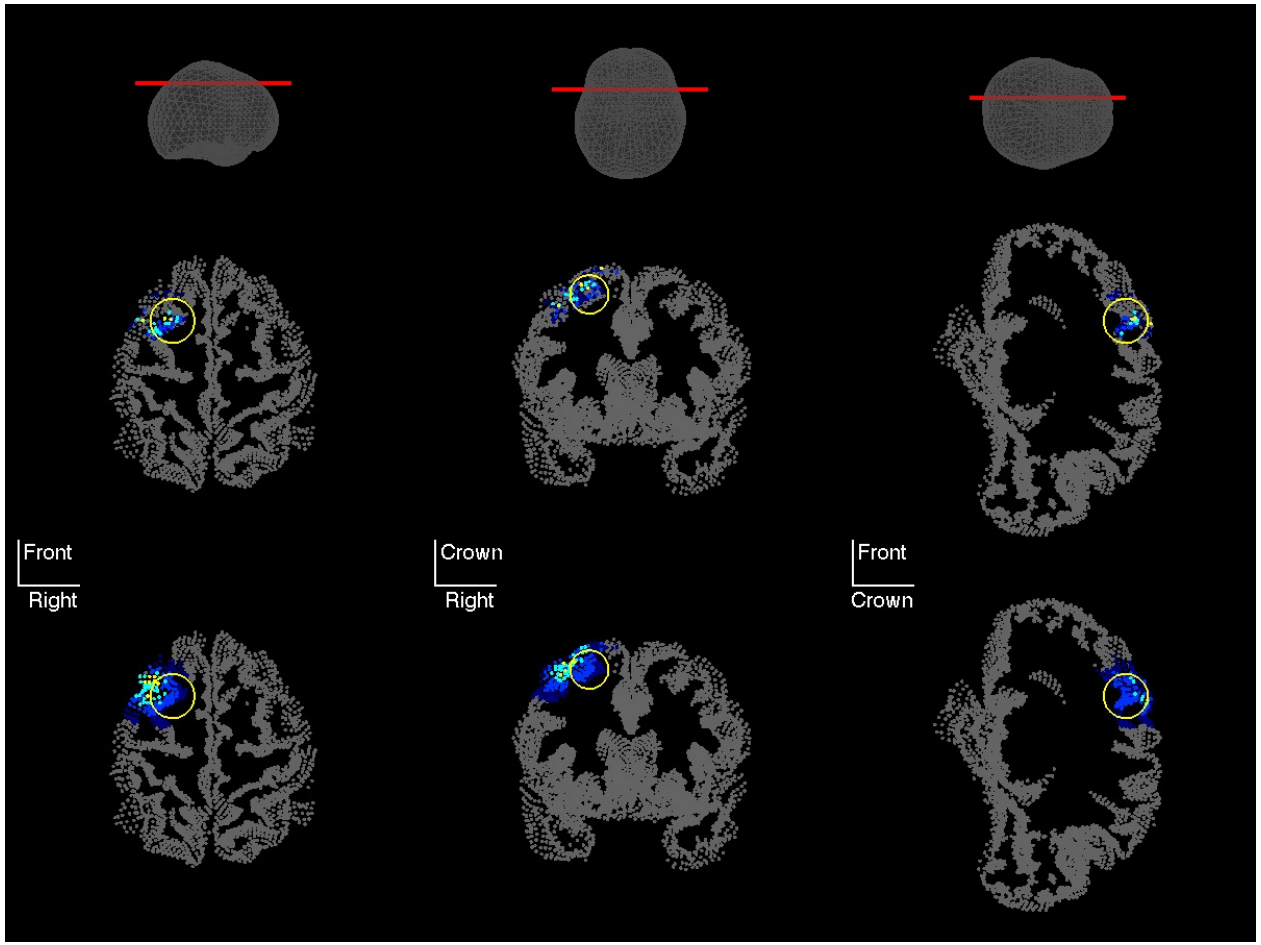
- Sparse solution (focal cerebral activity).
- Include anatomical information (orientation of neurons in cortex).

Computed examples: Effect of hyperparameter



Top row: Prior favoring strongly sparsity.
Bottom row: Prior favoring less strongly the sparsity.

Computed examples: Effect of anatomical prior



Top row: The anatomical information of the brain encoded.
Bottom row: No anatomical information used.