Bayesian Inversion; Examples and Computational Aspects

Ville Kolehmainen

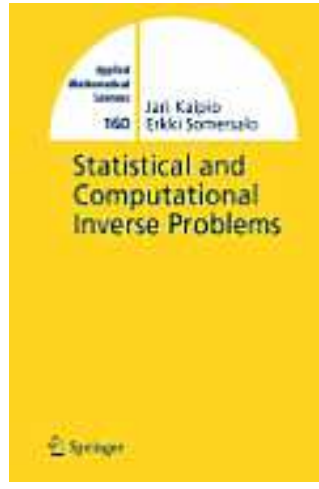
[1] Department of Physics

University of Kuopio

Finland

A few references to Bayesian inversion;

[KS] Jari Kaipio and Erkki Somersalo, *"Statistical and Computational Inverse Problems"*



[FNT] *"Physics 707 (Inverse Problems)"* lecture notes by Colin Fox, Geoff Nicholls and S.M. Tan (Univ. of Auckland) (available at `http://www.math.auckland.ac.nz/ phy707/`)

What are inverse problems?

- Typical “measurement problems”; estimate the quantity of interest $x \in \mathbb{R}^n$ from (noisy) measurement of

$$A(x) \in \mathbb{R}^{n_m},$$

where A is a known mapping.

- Inverse problems are those where the relation from the measured data to the sought quantity x is *ill-posed*;

1. The problem is non-unique (and/or)
2. Solution is extremely sensitive (instable) w.r.t measurement errors.

Examples of inverse problems; 2D deconvolution (image deblurring)

- Given noisy and blurred image

$$m = Ax + e, \quad m \in \mathbb{R}^n$$

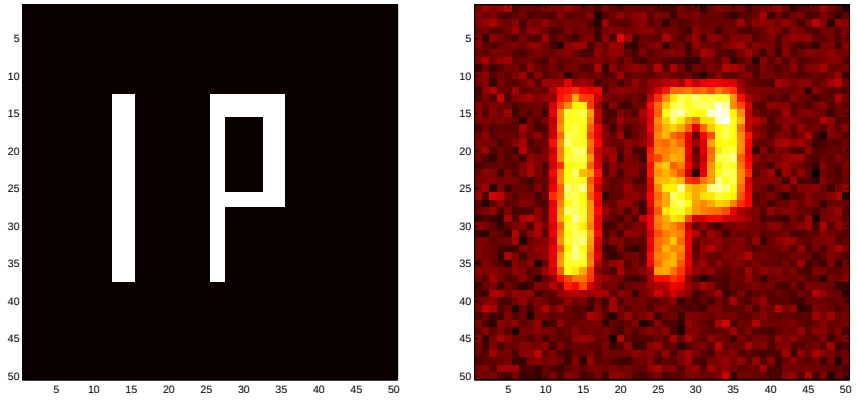
the objective is to reconstruct original image $x \in \mathbb{R}^n$.

- Forward model

$$x \mapsto Ax$$

implements discrete convolution (here the convolution kernel is Gaussian blurring kernel with std of 3 pixels). The forward problem is well-posed.

- Applications; photography, remote sensing, etc ...



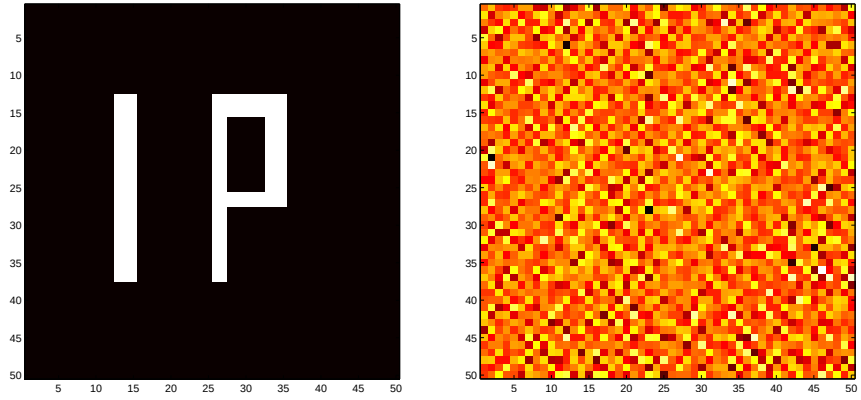
Left; true image x. Right; Noisy, blurred image $m = Ax + e$.

Example (cont.)

- Inverse problem is ill-posed; matrix A has trivial null-space $\mathrm{null}(A) = \{0\} \Rightarrow \exists A^{-1}$ but the solution is sensitive w.r.t measurement noise (singular values of A decay "exponentially").
- Figure shows the LS-solution

$$x_{\mathrm{LS}} = \arg\min_x \|m - Ax\|^2 \quad \Rightarrow x_{\mathrm{LS}} = A^{-1}m$$

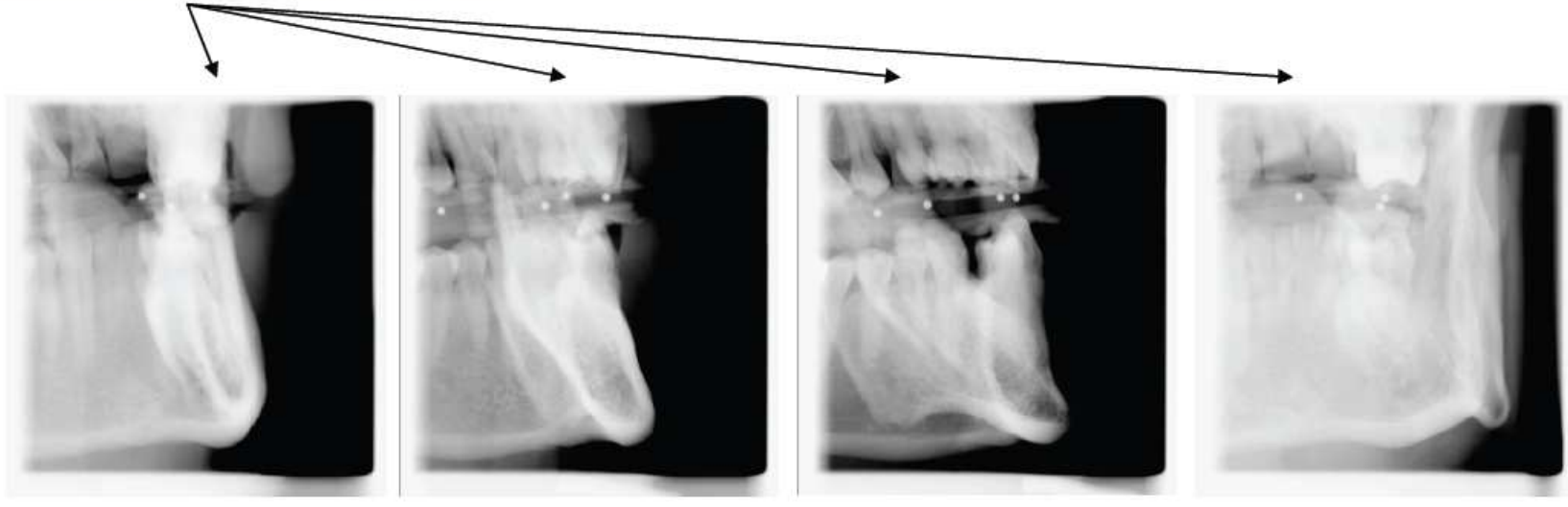
- The least squares solution is totally useless!



Left; true image x. Right; LS solution $x_{\mathrm{LS}} = \arg\min_x \|m - Ax\|^2$

Examples of inverse problems; limited angle x-ray tomography in dental implantology

- Finite number of x-ray projection images of the target are taken from an opening angle less than 180 degrees
- The goal is to reconstruct the 3D x-ray attenuation function from these images.
- Applications; dental radiology, mammography, surgical imaging, etc ...



Limited angle projection data from the mandibular area

Forward model of limited angle x-ray tomography

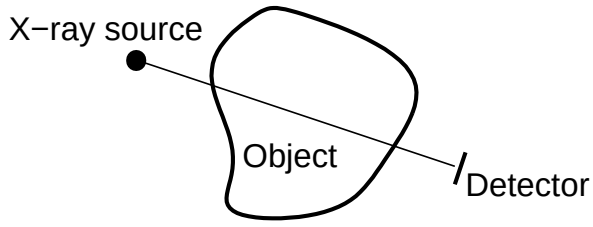
- Each pixel value in the projection model correpond to the total attenuation

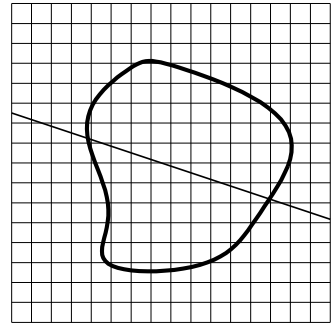
$$\log I_0 - \log I_j = \int_{L_j} x(s)\mathrm{d}s$$

along the line L_j between the source and detector pixel. The whole data set consists of $j = 1, \ldots, n_m$ line integrals ($n_m = \#\,\text{projection directions} \times \#\,\text{detector pixels}$)

- Discretization of attenuation function $x(s)$; $x(s) \approx \sum_{i=1}^{n} x_i \chi_{\Omega_i}(s) \Rightarrow$ Line integrals become

$$\int_{L_j} x(s)\mathrm{d}s = \sum_{i=1}^{n} x_i \, |\Omega_i \cap L_j|$$





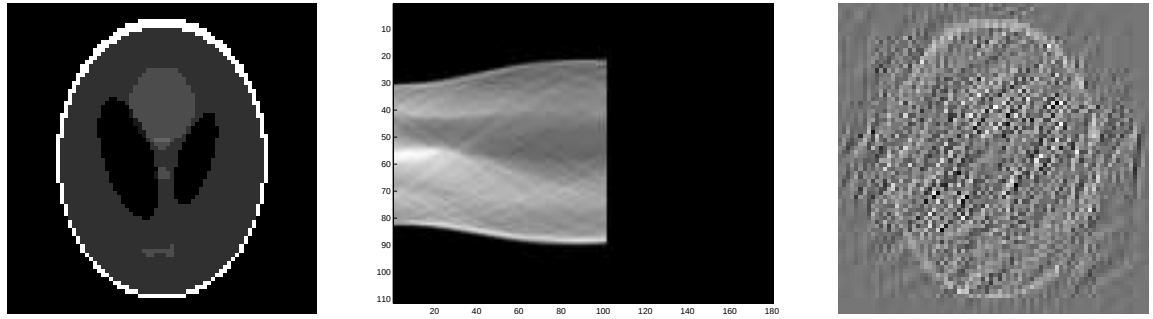
Left; Illustration of the projection model. Right; discretization of the image volume to n disjoint volume elements ($\Omega = \bigcup_{i=1}^{n} \Omega_i$).

- Arranging all the n_m observed pixel values into a single vector, we obtain forward model in matrix form

$$x \mapsto Ax, \quad A : \mathbb{R}^n \mapsto \mathbb{R}^{n_m}$$

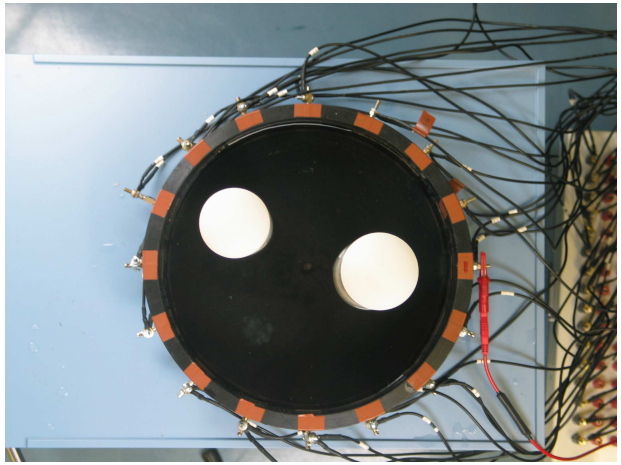
  Forward problem is well-posed.

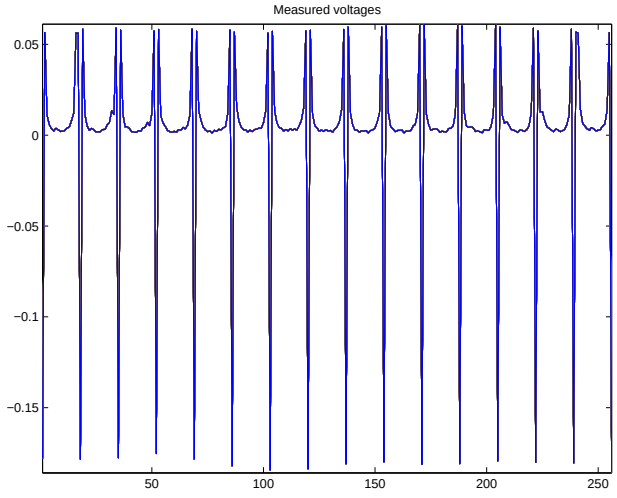
- Inverse problem (i.e, reconstruction of x, given noisy observation of Ax) is ill-posed;
  - Typically underdetermined ($n_m \ll n$).
  - Even if we had enough measurements for full column rank of A, the solution is still sensitive w.r.t measurement noise (see Figure).



Limited angle tomography. Left; true image x. Middle; noisy projection data $m = Ax + e$ (in sinogram form, 100 projections from opening angle of $100°$). Right; LS-solution.

Examples of inverse problems; EIT





- Electrical impedance tomography (EIT);
  - L electrodes are attached on the surface $\partial\Omega$ of the body Ω.
  - Electric currents are injected through the electrodes into the body Ω, resulting voltages are measured using the same electrodes.
  - Goal is to reconstruct the conductivity inside Ω based on the voltage measurements at the boundary.
- Top left; phantom in a laboratory experiment. Bottom left; measured voltages ($V \in \mathbb{R}^{256}$).

- Forward model of EIT;

$$
\begin{aligned}
\nabla \cdot (\sigma \nabla u) &= 0 \text{ in } \Omega \\
u + z_l \sigma \frac{\partial u}{\partial \nu} &= U_l, \quad \text{on } e_l \subset \partial\Omega,\ l = 1, 2, \ldots, L \\
\int_{e_l} \sigma \frac{\partial u}{\partial \nu} \mathrm{d}S &= I_l, \quad \text{on } e_l,\ l = 1, 2, \ldots, L \\
\sigma \frac{\partial u}{\partial \nu} &= 0, \quad \text{on } \partial\Omega \setminus \bigcup_{l=1}^{L} e_l \ ,
\end{aligned}
$$

  with the conditions

$$
\sum_{l=1}^{L} I_l = 0, \quad \sum_{l=1}^{L} U_l = 0 \ .
$$

- The forward problem of EIT; compute the electrode voltages $\{U_\ell\}$, given conductivity σ and the injected currents $\{I_\ell\}$.

- Discretization of the PDE system by the finite element method (FEM). Conductivity approximated in a finite dimensional pixel basis $\sigma = \sum_{i=1}^{n} \sigma_i \chi_{\Omega_i}$.

  Forward solution becomes

$$\sigma \mapsto U(\sigma), \quad U : \mathbb{R}^n \mapsto \mathbb{R}^{n_m}$$

  Forward problem is well-posed.

- Inverse problem; reconstruct σ given noisy observation of $U(\sigma)$. The problem is ill-posed.

Solutions to inverse problems; Regularization techniques

- The ill posed problem is replaced by a well posed approximation. Solution “hopefully” close to the true solution.
- Typically modifications of the associated least squares problem

$$\|m - A(x)\|^2.$$

- Examples of methods; TSVD, Tikhonov regularization, truncated iterations, etc ...
- Example; Consider the problem

$$x_{\mathrm{LS}} = \arg\min_x \{\|m - Ax\|_2^2\} \quad \Rightarrow \underbrace{(A^{\mathrm{T}}A)}_{B} x_{\mathrm{LS}} = A^{\mathrm{T}}m$$

Uniqueness; $\exists B^{-1}$ if $\mathrm{null}(A) = \{0\}$. Stability of the solution?

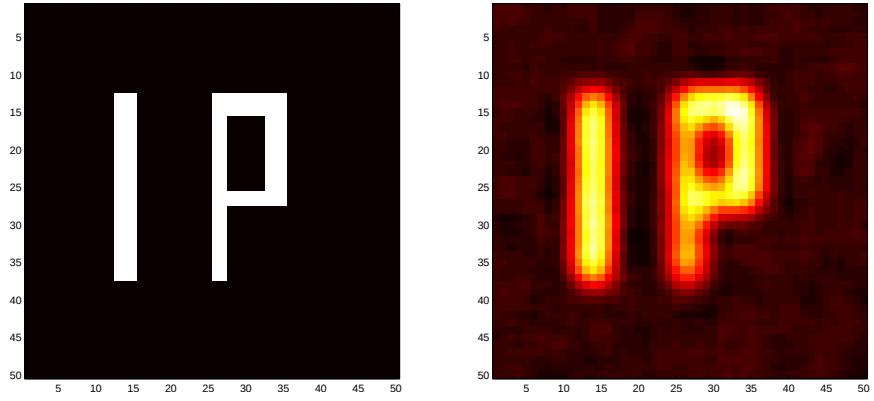
Solutions to inverse problems; Regularization techniques

- Example (cont.); In Tikhonov regularization, the original LS problem is replaced with

$$x_{\mathrm{TIK}} = \arg\min_x \{\|m - Ax\|_2^2 + \alpha\|x\|_2^2\} \quad \Rightarrow \underbrace{(A^{\mathrm{T}}A + \alpha I)}_{\tilde{B}} x_{\mathrm{LS}} = A^{\mathrm{T}}m$$

Uniqueness; $\alpha > 0 \Rightarrow \mathrm{null}(\tilde{B}) = \{0\} \Rightarrow \exists \tilde{B}^{-1}$. Stability of the solution quaranteed by choosing α "large enough".

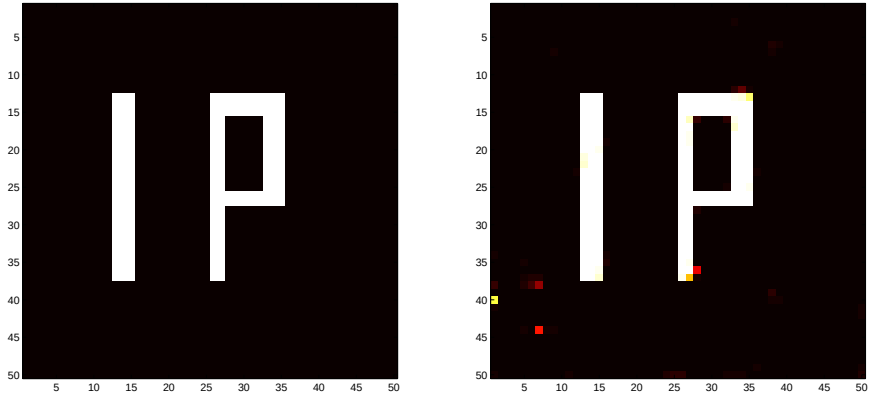
- Regularization poses always (implicit) prior assumptions about the solution!. These assumptions are often well hidden.



Left; true image x. Right; Tikhonov solution x_{TIK}.

Solutions to inverse problems; Bayesian inversion.

- The inverse problem is reformulated as a problem of Bayesian inference.
- The key idea is to extract information and assess uncertainty about the unknowns based on
  - Measured data
  - Model of measurement process
  - Model of *a priori* information
- Ill posedness removed by explicit use of prior models!
- These lectures give introduction to Bayesian inversion. Focus on computational aspects.



Left; true image x. Right; Bayesian point estimate x_{CM} for the 2D deconvolution problem (Model and implementation will be explained later).

Bayesian inversion.

- All variables in the model are considered as random variables. The randomness reflects our uncertainty of their actual values.
- The degree of uncertainty is coded in the probability distribution models of these variables.
- The complete solution of the inverse problem is the posterior distribution

$$\pi(x \mid m) = \frac{\pi_{\mathrm{pr}}(x)\pi(m \mid x)}{\pi(m)}, \quad m = m_{\mathrm{observed}} \tag{1}$$

where

- $\pi(m \mid x)$ is the likelihood density (based on models of the measurement and noise processes)
- $\pi_{\mathrm{pr}}(x)$ is the prior density (model for the prior information)
- $\pi(m)$ is normalization constant

- The posterior density is a function on n dimensional space;

$$\pi(x \mid m) : \mathbb{R}^n \mapsto \mathbb{R}_+$$

- Dimension n usually large for inverse problems. Direct visual interpretation of the posterior not possible.
- Even if n is small, the whole posterior may not be convenient for analysis & decisions.

$\Rightarrow$ The posterior need to be summarized by different estimates. Examples of typical questions;

- "What is the most probable value of x, given the data and prior? "
- "In what interval are the values of x with 90% probability, given the data and prior? "

Typical summary estimates

- "Representative solutions";
  - *Maximum a posteriori* (MAP) estimate:

$$\pi(x_{\mathrm{MAP}} \mid m) = \arg\max_x \pi(x \mid m).$$

  - Conditional mean (CM) estimate:

$$x_{\mathrm{CM}} = \int_{\mathbb{R}^n} x\pi(x \mid m)\mathrm{d}x.$$

- Uncertainty & interval estimates;
  - Covariance:

$$\Gamma_{x|m} = \int_{\mathbb{R}^n} (x - x_{\mathrm{CM}})(x - x_{\mathrm{CM}})^{\mathrm{T}} \pi(x \mid m)\,\mathrm{d}x$$

- Uncertainty & interval estimates (cont.);
  - Confidence intervals; Given $0 < \tau < 100$, compute a_k and b_k s.t.

$$\int_{-\infty}^{a_k} \pi_k(x_k)\,\mathrm{d}x_k = \int_{b_k}^{\infty} \pi_k(x_k)\,\mathrm{d}x_k = \frac{100-\tau}{200}$$

where $\pi_k(x_k)$ is the marginal posterior density

$$\pi_k(x_k) = \int_{\mathbb{R}^{n-1}} \pi(x \mid m)\,\mathrm{d}x_1 \cdots \mathrm{d}x_{k-1}\mathrm{d}x_{k+1} \cdots \mathrm{d}x_n.$$

The interval $I_k(\tau) = [a_k\ b_k]$ contains τ % of the mass of the marginal density.

Summary of Bayesian Inversion

- In summary, Bayesian solution of an inverse problem consist of the following steps:
  1. Construct the likelihood model $\pi(m|x)$. This step includes the development of the model $x \mapsto A(x)$ and modeling the measurement noise.
  2. Based on the available *a priori* information, construct the prior model $\pi_{\mathrm{pri}}(x)$.
  3. Develop computation methods for the summary statistics.

Summary of Bayesian Inversion (cont)

- The posterior $\pi(x \mid m)$ is our model for all the available information and uncertainty we have on the unknowns x. The soundness of this model is dictated by the likelihood and prior models.
  - The construction of the likelihood is often the most "straightforward" task; model of measurement process (i.e., $A(x)$) is usually well known and noise properties can be measured.
  - The construction of the prior model is typically the most challenging task; The prior knowledge is typically in qualitative form. The problem is then how to translate this information into a (computationally feasible) mathematical model.
- The computation of the summary statistics require large scale optimization & numerical integration. Practicality of the solution becomes a problem of computational efficiency and resources available (computation time, computer resources).

Examples of likelihood models:

- Additive noise model:

$$m = A(x) + e, \quad e \sim \pi_e(e).$$

- If x and e mutually independent, we get

$$\pi(m \mid x) \propto \pi_e(m - A(x))$$

- Most commonly used noise model; Gaussian noise $e \sim \mathcal{N}(e_*, \Gamma_e) \Rightarrow$

$$\pi(m \,|\, x) \propto \exp\left(-\frac{1}{2}(m - A(x) - e_*)^{\mathrm{T}} \Gamma_e^{-1} (m - A(x) - e_*)\right).$$

Let $L_e^{\mathrm{T}} L_e = \Gamma_e^{-1}$, then

$$\pi(m \,|\, x) \propto \exp\left(-\frac{1}{2}\|L_e(m - A(x) - e_*)\|_2^2\right)$$

Examples of likelihood models (cont.):

- Counting process data; Poisson distibution

$$\pi(m_j \mid x) = \frac{(A(x))_j^{m_j}}{m_j!} \exp(-(A(x))_j), \quad (A(x))_j \geq 0$$

If the components m_j have mutually independent fluctuation, we get

$$\begin{aligned} \pi(m \mid x) &= \prod_{j=1}^{n_m} \frac{(A(x))_j^{m_j}}{m_j!} \exp(-(A(x))_j) \\ &\propto \exp(m^{\mathrm{T}} \log(A(x)) - \mathbf{1}^{\mathrm{T}}(A(x))) \end{aligned}$$

- Examples; SPECT, PET, Particle counters (e.g., measuring athmospheric aerosol concentrations), etc ...

Examples of prior models: Positivity prior

- Non-negativity prior:

$$\pi_+(x) = \prod_{k=1}^{n} \theta(x_k), \quad \theta(t) = \begin{cases} 1, & t \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

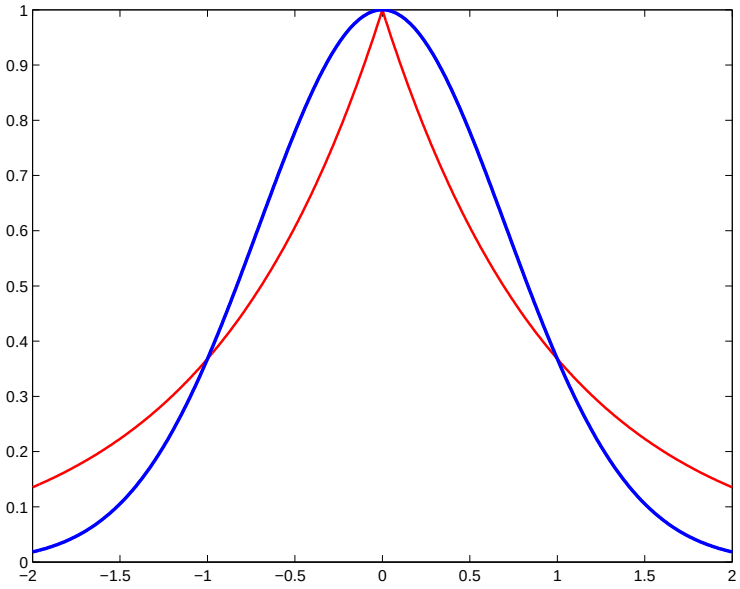
Simple, has often significant impact and applies for most physical parameters!

- Obviously, $\theta(t)$ may be replaced by

$$\theta_k(t) = \begin{cases} 1, & t \in [a_k,\ b_k] \\ 0, & \text{otherwise} \end{cases}$$

if range of possible values for x is known.

Examples of prior models: Generic (qualitative) models.



L^1 (red), L^2 (blue).

- White noise prior (L^2 prior):

$$\begin{aligned}\pi_{\text{pri}}(x) &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k - x_{*,k})^2\right)\\ &= \exp\left(-\frac{1}{2\sigma^2}\|x - x_*\|_2^2\right)\end{aligned}$$

where x_* is the prior mean.

- Outlier prior (L^1-prior):

$$\begin{aligned}\pi_{\text{pri}}(x) &\propto \exp\left(-\alpha\sum_{k=1}^{n}|x_k - x_{*,k}|\right)\\ &= \exp\left(-\alpha\|x - x_*\|_1\right)\end{aligned}$$

Other impulse priors: maximum entropy, Cauchy distribution prior.

- Neighborhood priors (MRF priors). General form;

$$\pi_{\text{pri}}(x) \propto \exp\left(-\alpha \sum_{k=1}^{n} \sum_{j \in \mathcal{N}_k} h(x_k, x_j)\right)$$

where $\mathcal{N}_k \subset \{1, 2, \ldots, n\}$ is the neighborhood of x_k. For example, in a $N \times N = n$ image, the usual 4-point neighborhood is $\mathcal{N}_k = \{k-1, k+1, k-N, k+N\}$.

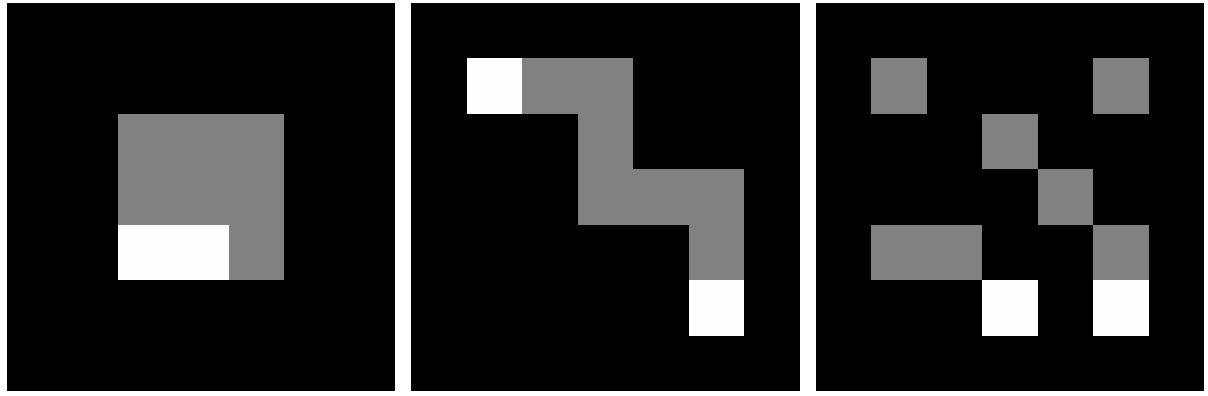
- Choosing $h(x_k, x_j) = 0.5(x_k - x_j)^2$, we get the smoothness prior

$$\begin{aligned} \pi_{\text{pri}}(x) &\propto \exp\left(-\frac{\alpha}{2} \sum_{k=1}^{n} \sum_{j \in \mathcal{N}_k} (x_k - x_j)^2\right) \\ &= \exp\left(-\frac{1}{2}\|Lx\|_2^2\right) \end{aligned}$$

- Total variation (TV) prior;

$$\pi_{\mathrm{pri}}(x) \propto \exp\left(-\alpha \mathrm{TV}(x)\right), \quad \mathrm{TV}(x) = \sum_{k=1}^{n} \sum_{j \in \mathcal{N}_k} |x_k - x_j|$$

- TV-prior model is concentrated around piecewise regular ("blocky") images, i.e., the pixels are clustered in blocks with almost equal value and short boundary. See Figure.



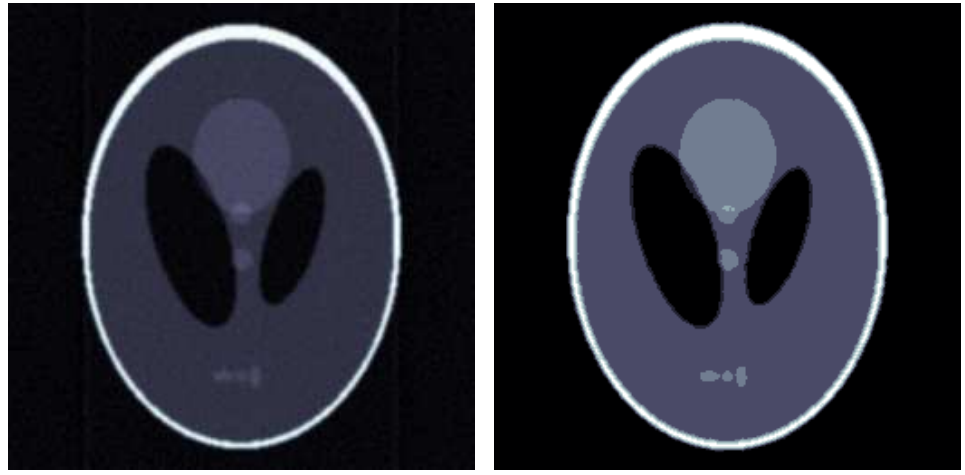
Images having total variations (from left to right) 18, 28 and 40.

- Of the previous models, smoothness and TV priors are improper.
- Proper (integrable) distributions can be obtained by conditioning

$$\pi_{\text{pri}}(x) = \pi_{\text{pri}}(x_{\mathcal{I}} \mid x_{-\mathcal{I}})\pi_{\text{pri}}(x_{-\mathcal{I}})$$

  where
  - $x_{\mathcal{I}} \in \mathbb{R}^{n-p}$ is vector of "non-specified" elements.
  - $x_{-\mathcal{I}} \in \mathbb{R}^{p}$ vector of "specified" elements (e.g., boundary elements of an image).
- The model $\pi_{\text{pri}}(x_{-\mathcal{I}})$ can be used to include information about the quantitative values of $x_{-\mathcal{I}}$ into the prior model (e.g., the boundary values of an image may be approximately known).

- MRF priors are useful for designing *structural priors*.
- Consider an example of multimodality CT/SPECT imaging. Segmentation of the target body to different tissue types (organs) can be obtained from the CT reconstruction.
- Reasonable to expect jumps in the SPECT image on the same locations as in the CT image (radionuclide concentration has different magnitude in different organs).
- Let variable $\tau_k \in \{1, 2, \ldots, N_{\mathrm{org}}\}$ denote the tissue type in image pixel k.



Left: CT reconstruction. Right: Segmentation of the CT image.

- Based on the segmentation of the CT-image, we construct an anisotropic non-homogeneous smoothness prior

$$\begin{aligned} \pi_{\mathrm{pri}}(x) &\propto \exp\left(-\frac{\alpha}{2}\sum_{k=1}^{n}\sum_{j\in\mathcal{N}_k}\lambda_{k,j}(x_k-x_j)^2\right) \\ &= \exp\left(-\frac{1}{2}\|\Lambda L x\|_2^2\right) \end{aligned}$$

where

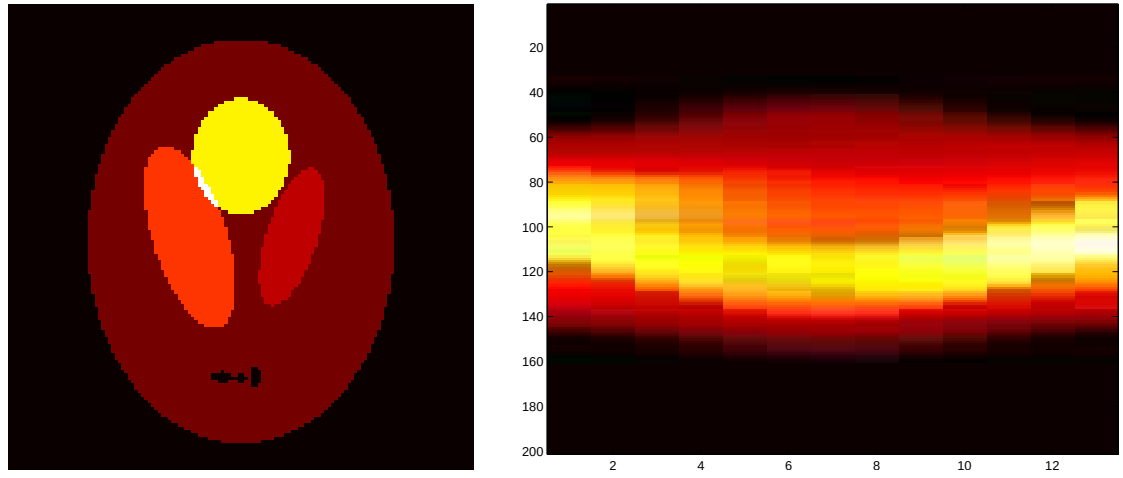
$$\lambda_{k,j} = \begin{cases} 1, & \tau_k = \tau_j \\ \beta, & \tau_k \neq \tau_j \end{cases}$$

and $\beta \ll 1$ (in the following example $\beta = 10^{-3}$).

- SPECT data acquired concurrently with the CT data. SPECT data consist of 12 projections from total view-angle of $165°$.
- Gaussian likelihood model

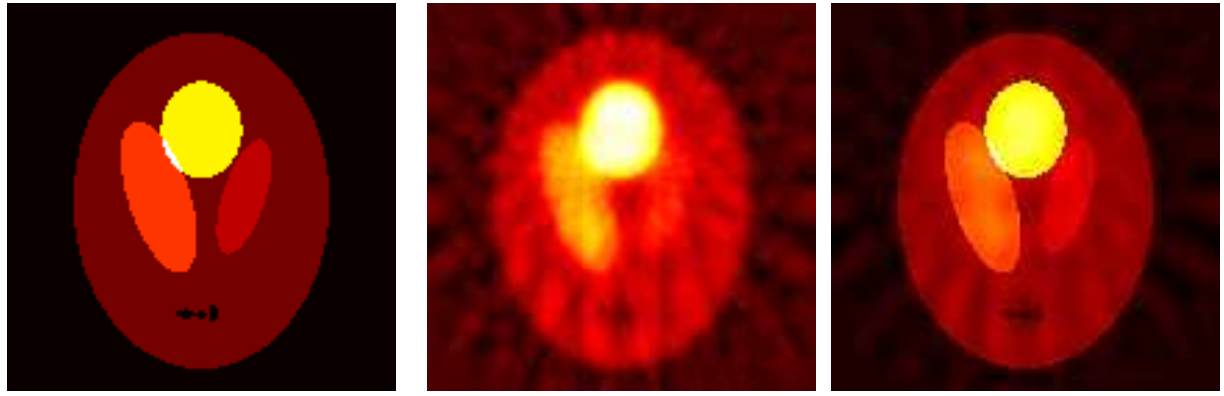
$$\pi(m \mid x) \propto \exp\left(-\frac{1}{2}\|L_e(m - A(x))\|_2^2\right)$$

with noise variance $\sigma_i = \sqrt{N_i}$.

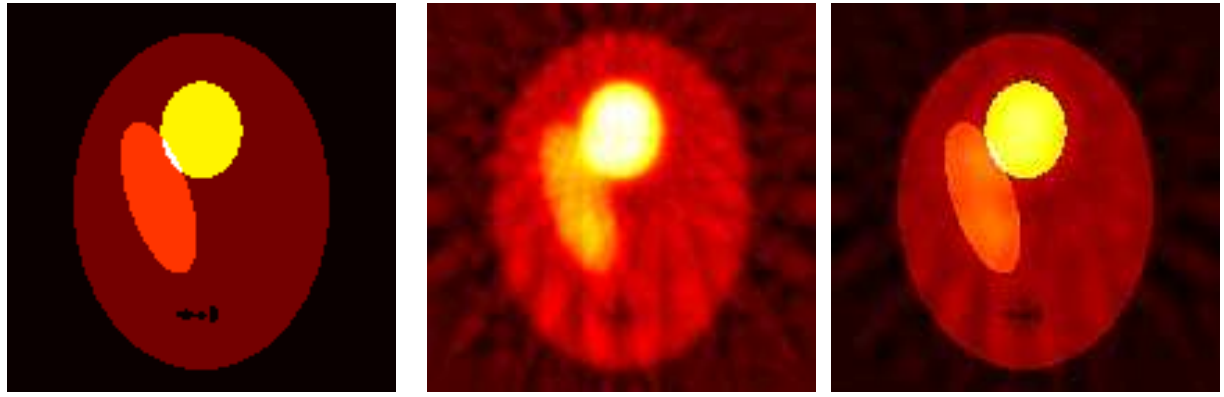


Left: Example of activity distribution. Right: Noisy SPECT data (sinogram)

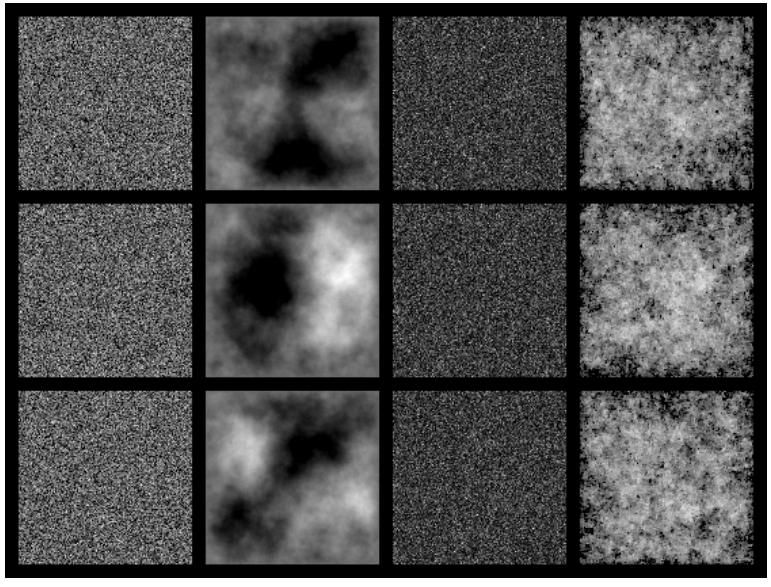
- Simulation 1: “Healthy case”.
- $x_{\mathrm{MAP}} = \arg\max_x \pi(m \mid x)\pi_{\mathrm{pri}}(x)$
- Left: True activity distribution.
- Middle: conventional (homogeneous isotropic) smoothness prior
  $\pi_{\mathrm{pri}}(x) = \exp\left(-\frac{1}{2}\|Lx\|_2^2\right)$
- Right: structural (non-homogeneous anisotropic) smoothness prior
  $\pi_{\mathrm{pri}}(x) = \exp\left(-\frac{1}{2}\|\Lambda Lx\|_2^2\right)$



- Simulation 2: “Diseased case” (right lung not ventilated).
- $x_{\mathrm{MAP}} = \arg\max_x \pi(m \mid x)\pi_{\mathrm{pri}}(x)$
- Left: True activity distribution.
- Middle: conventional (homogeneous isotropic) smoothness prior
  $\pi_{\mathrm{pri}}(x) = \exp\left(-\frac{1}{2}\|Lx\|_2^2\right)$
- Right: structural (non-homogeneous anisotropic) smoothness prior
  $\pi_{\mathrm{pri}}(x) = \exp\left(-\frac{1}{2}\|\Lambda Lx\|_2^2\right)$ (same as above, i.e., based on segmented CT that shows the right lung!)



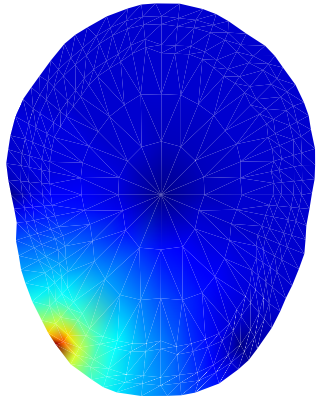
Visualization of priors.

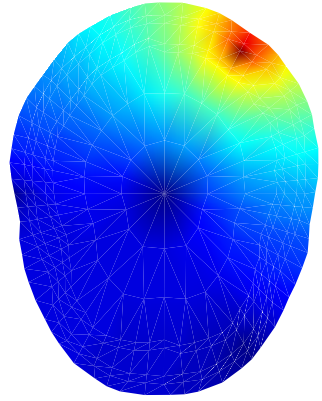


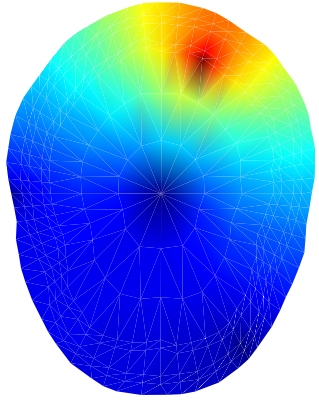
- Sampling and visualization can be used to verify the qualitative nature of the prior model $\pi_{\mathrm{pri}}(x)$.
- Left: random samples from prior model $\pi_{\mathrm{pri}}(x)$. Columns from left to right;
  1. white noise prior
  2. smoothness prior
  3. impulse (L^1) prior
  4. TV-prior.
- The samples of smoothness and TV prior were drawn from

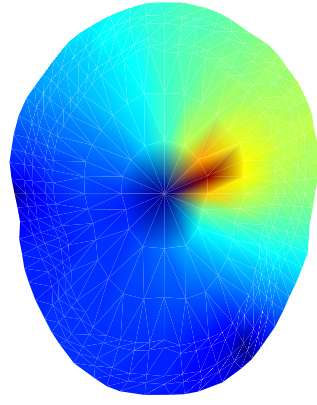
$$\pi_{\mathrm{pri}}(x_{\mathcal{I}} \mid x_{-\mathcal{I}})$$

with boundary pixels $x_{-\mathcal{I}} = 0$.









- Visualization of prior covariances

$$(\mathrm{cov}(x_i, x_1), \mathrm{cov}(x_i, x_2), \ldots, \mathrm{cov}(x_i, x_n))^{\mathrm{T}} \in \mathbb{R}^n$$

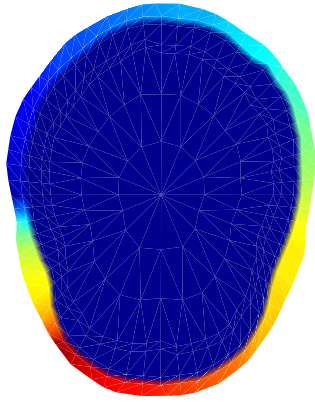
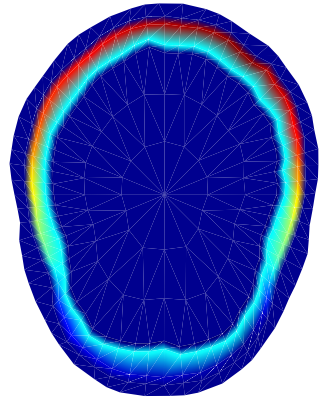
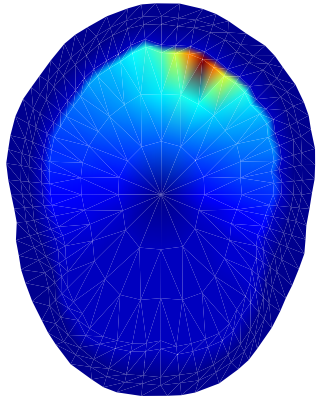
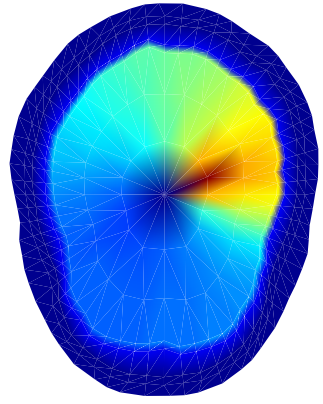
reveal the prior covariance structures.

- Left; examples of (conditional) covariances

$$\mathrm{cov}(x_{\mathcal{I}} \mid x_{-\mathcal{I}})$$

corresponding to the smoothness prior in an EIT application.

- The covariance values of the conditioned pixels $x_{-\mathcal{I}}$ are set to zero (dark blue) and the highest correlation (autocorrelation) is printed with red.

- Examples of (conditional) covariances

$$\mathrm{cov}(x_{\mathcal{I}} \mid x_{-\mathcal{I}})$$

corresponding to a nonhomogeneous anisotropic smoothness prior in an EIT application.

- The locations of the scalp-skull and skull-brain interfaces were assumed known from MRI scan in the constuction of the prior.
- The covariance values of the conditioned pixels $x_{-\mathcal{I}}$ are set to zero (dark blue) and the highest correlation (autocorrelation) is printed with red.

Computation of the estimates

- Gaussian distributions; General form

$$\begin{aligned}\pi(x \mid m) &\propto \pi(m \mid x)\pi_{\text{pri}}(x) \\ &= \exp\left(-\frac{1}{2}(m - Ax - e_*)^{\mathrm{T}}\Gamma_e^{-1}(m - Ax - e_*) - \frac{1}{2}(x - x_*)^{\mathrm{T}}\Gamma_x^{-1}(x - x_*)\right)\end{aligned}$$

- MAP and CM same point for Gaussian posterior.
- Posterior $\pi(x \mid m)$ fully determined by mean and covariance:

$$\begin{aligned}x_{\text{CM}} &= \Gamma_{x|m}(A^{\mathrm{T}}\Gamma_e^{-1}(m - e_*) + \Gamma_x^{-1}x_*) \\ \Gamma_{x|m} &= (A^{\mathrm{T}}\Gamma_e^{-1}A + \Gamma_x^{-1})^{-1}\end{aligned}$$

- Let $L_e^{\mathrm{T}}L_e = \Gamma_e^{-1}$ and $L_x^{\mathrm{T}}L_x = \Gamma_x^{-1}$, then we can write $\pi(x \mid m) \propto \exp\{-\frac{1}{2}F(x)\}$, where

$$F(x) = \|L_e(m - Ax - e_*)\|_2^2 + \|L_x(x - x_*)\|_2^2$$

and

$$x_{\text{MAP}} = \arg\min_x F(x)$$

$\Rightarrow$ Connection to Tikhonov regularization!

Example

- Consider the original form of Tikhonov regularization;

$$x_{\mathrm{TIK}} = \arg\min_x \{\|m - Ax\|_2^2 + \alpha\|x\|_2^2\} \quad \Rightarrow \quad x_{\mathrm{TIK}} = (A^{\mathrm{T}}A + \alpha I)^{-1} A^{\mathrm{T}} m$$

- From the Bayesian viewpoint, x_{TIK} correspond to the posterior mode with the following assumptions
  - Measurement model $m = Ax + e$, x and e mutually independent with $e \sim \mathcal{N}(0, I)$.
  - x is assumed *a priori* mutually independent zero mean white noise with variance $1/\alpha$.
- The original idea of the Tikhonov method was to approximate $A^{\mathrm{T}}A$ with a matrix $A^{\mathrm{T}}A + \alpha I$ that is invertible and produces stable solution.

About solving the Gaussian problem ...

- When dimension n is large ($\sim 10^3$ and more), the inversion of $\Gamma_{x|m}^{-1}$ may not fit to the computer memory. In these cases, the solution can be computed neatly with standard iterative solvers (CG,MINRES,GMRES) applied to linear system

$$\underbrace{\Gamma_{x|m}^{-1}}_{n\times n \text{ matrix}} \quad x_{\mathrm{CM}} = \underbrace{(A^{\mathrm{T}}\Gamma_e^{-1}(m-e_*)+\Gamma_x^{-1}x_*)}_{n\times 1 \text{ vector}}$$

- Columns of $\Gamma_{x|m}$ respectively; to obtain column j, solve

$$\Gamma_{x|m}^{-1}c = b, \quad b = (\underbrace{0,\ldots,0}_{1:j-1},1,\underbrace{0,\ldots,0}_{j+1:n})^{\mathrm{T}}\mathbb{R}^n$$

- Can be implemented matrix free; needs only function that returns product

$$\Gamma_{x|m}^{-1}x$$

- Examples; use of GPUs in x-ray tomography, FFTs in image deconvolution, etc..

Computation of the estimates: MAP

- Generally, solving MAP estimate

$$x_{\mathrm{MAP}} = \arg\max_x \pi(x \mid m)$$

  is an optimization problem.

- Example;

$$\pi(x \mid m) \propto \pi_+(x) \exp\left(-\frac{1}{2}\|L_e(m - A(x))\|_2^2 - W(x)\right)$$

  Then

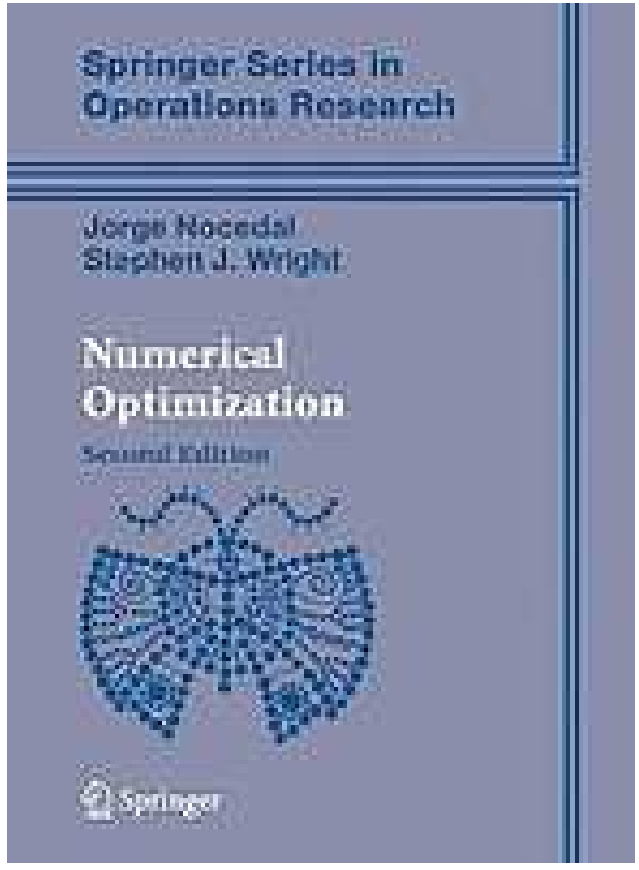
$$x_{\mathrm{MAP}} = \arg\min_{x \geq 0}\{\frac{1}{2}\|L_e(m - A(x))\|_2^2 + W(x)\}$$

  $\Rightarrow$ Similarity to (Tikhonov) regularized LS-estimation.

- Large variety of applicable solution methods.

Computation of the estimates: MAP

- The literature on (numerical) optimization methods is vast. As general reference see e.g.,



Computation of the estimates; Integration based estimates

- Most of the summary estimates are of the form

$$\bar{f}(x) = \int_{\mathbb{R}^n} f(x)\pi(x \mid m)\mathrm{d}x$$

- Examples:
  - $f(x) = x \quad \rightsquigarrow x_{\mathrm{CM}}$
  - $f(x) = (x - x_{\mathrm{CM}})(x - x_{\mathrm{CM}})^{\mathrm{T}} \quad \rightsquigarrow \Gamma_{x|m}$
  - etc ...
- Analytical evaluation in most cases impossible
- Traditional numerical quadratures not applicable when n is large (number of points needed unreasonably large, support of $\pi(x \mid m)$ may not be well known)
- $\Rightarrow$ Monte Carlo integration.

Computation of the estimates; Monte Carlo integration

- Monte Carlo integration
  1. Draw an ensemble $\{x^{(k)}, k = 1, 2, \ldots, N\}$ of i.i.d samples from $\pi(x)$.
  2. Estimate

$$\int_{\mathbb{R}^n} f(x)\pi(x)\mathrm{d}x \approx \frac{1}{N}\sum_{k=1}^{N} f(x^{(k)})$$

- Convergence (law of large numbers)

$$\lim_{N\to\infty} \frac{1}{N}\sum_{k=1}^{N} f(x^{(k)}) \to \int_{\mathbb{R}^n} f(x)\pi(x)\mathrm{d}x \quad (\mathrm{a.c})$$

- Variance of the estimator $\bar{f} = \frac{1}{N}\sum_{k=1}^{N} f(x^{(k)})$ reduces

$$\propto \frac{1}{N}$$

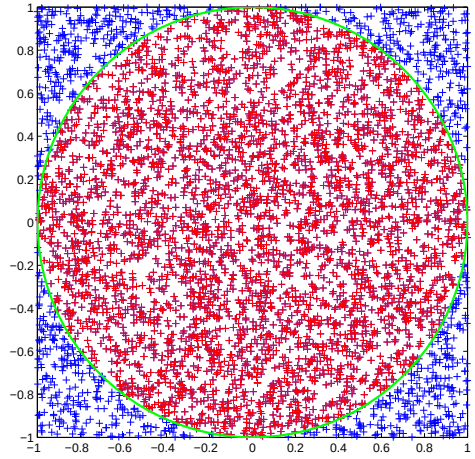
Simple example of Monte Carlo integration

- Let $x \in \mathbb{R}^2$ and

$$\pi(x) = \begin{cases} \frac{1}{4}, & x \in G, \quad G = [-1,\ 1] \times [-1,\ 1] \\ 0, & x \notin G. \end{cases}$$

The task is to compute integral $g(x) = \int_{\mathbb{R}^2} \chi_D \pi(x) \mathrm{d}x$ where D is the unit disk on $\mathbb{R}^2$.

- Using $N = 5000$ samples, we get estimate $g(x) = 0.7814$ (true value $\pi/4 = 0.7854$)



Example of Monte Carlo integration. Red marks denote samples inside D

Computation of the estimates; MCMC

- Direct sampling of the posterior usually not possible $\Rightarrow$ Markov chain Monte Carlo (MCMC).
- MCMC is Monte Carlo integration using Markov chains (dependent samples);
  1. Draw $\{x^{(k)}, k = 1, 2, \ldots, N\} \sim \pi(x)$ by simulating a Markov chain (with equilibrium distribution $\pi(x)$).
  2. Estimate

$$\int_{\mathbb{R}^n} f(x)\pi(x)\mathrm{d}x \approx \frac{1}{N}\sum_{k=1}^{N} f(x^{(k)})$$

- For the conditions that the law of large numbers holds, see e.g. ([KS], Chapter 3.6, Proposition 3.11) or Kaipio et al, "Statistical inversion and Monte Carlo sampling methods in EIT" *Inverse Problems* 16, 1487-1522,2000, Proposition 3.1. (later referred as [Kaipio 2000])
- Algorithms for MCMC;
  1. Metropolis Hastings algorithm
  2. Gibbs sampler

- Variance of the estimator $\bar{f} = \frac{1}{N}\sum_{k=1}^{N} f(x^{(k)})$ reduces as

$$\propto \frac{\tau}{N}$$

where $\tau \geq 1$ is the integrated autocorrelation time of the Markov chain, defined by

$$\tau = 1 + 2\sum_{j=1}^{\infty} \rho(s), \quad \rho(s) = \frac{C(s)}{C(0)}, \quad C(s) = \mathrm{cov}(x^{(\ell)}, x^{(\ell+s)})$$

- Interpretation; τ correspond to the number of samples with same variance reduction power as one independent sample.

Metropolis-Hastings algorihtm (MHMCMC)

- Generation of the ensemble $\{x^{(k)}, k = 1, \ldots, N\} \sim \pi(x)$ using Metropolis Hastings algorithm;

  1. Pick an initial value $x^{(1)} \in \mathbb{R}^n$ and set $\ell = 1$
  2. Set $x = x^{(\ell)}$.
  3. Draw a candidate sample x' from proposal density

     $$x' \sim q(x, x')$$

     and compute the acceptance factor

     $$\alpha(x, x') = \min\left(1, \frac{\pi(x')q(x', x)}{\pi(x)q(x, x')}\right).$$

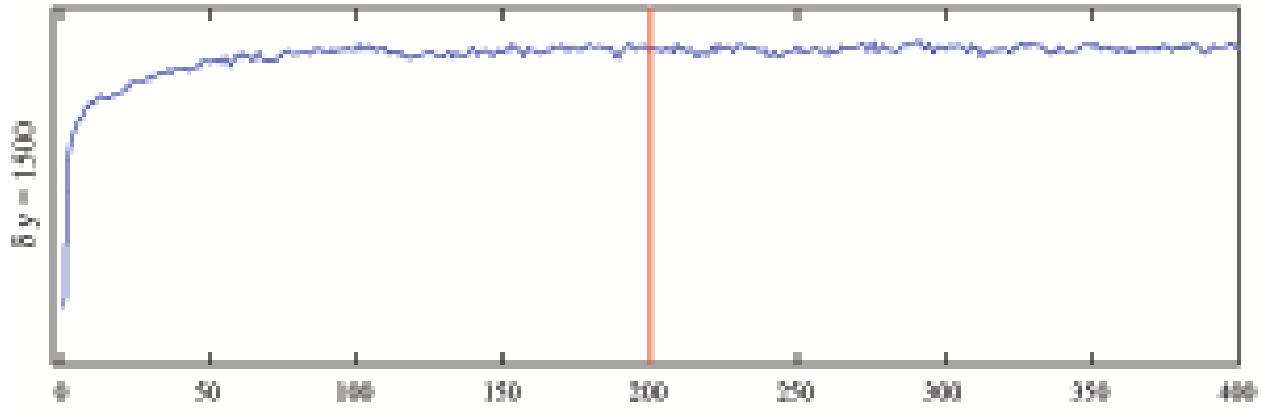
  4. Draw $t \in [0, 1]$ from uniform probability density ($t \sim \mathrm{uni}(0,\ 1)$).
  5. If $\alpha(x, x') \geq t$, set $x^{(\ell+1)} = x'$, else $x^{(\ell+1)} = x$. Increment $\ell \to \ell + 1$.
  6. When $\ell = N$ stop, else repeat from step 2.

- Normalization constant of π do not need to be known.
- Great flexibility in choosing the proposal density $q(x, x')$; almost any density would do the job (eventually). For conditions needed ensuring convergence, see e.g. [KS, Proposition 3.12], [Kaipio 2000, Proposition 3.2]).
- However, the choice of $q(x, x')$ is a crucial part of successful MHMCMC; it determines the efficiency (autocorrelation time τ) of the algorithm $\Rightarrow q(x, x')$ should be s.t. τ as small as possible
- No systematic methods for choosing $q(x, x')$. More based on "intuition & art". The goal to choose $q(x, x')$ s.t.
  1. $q(x, x')$ would respect the known properties of $\pi(x)$ as well as possible (obviously with $q(x, x') = \pi(x)$ we would have $\tau = 1$, i.e, i.i.d samples)
  2. $q(x, x')$ computationally efficient to sample

- The updating may proceed
  - All unknowns at a time
  - Single component x_j at a time
  - A block of unknowns at a time
- The order of updating (in single component and blockwise updating) may be
  - Update elements chosen in random (random scan)
  - Systematic order
- The proposal can be mixture of several densities

$$q(x, x') = \sum_{i=1}^{t} p_i q_i(x, x'), \quad \sum_{i=1}^{t} p_i = 1$$

- Proposal parameteres are usually calibrated by pilot runs; aim at finding parameters giving best efficiency (minimal τ).
- "Burn-in"; The chain takes some time to converge to the desired distribution. The actual collection of the samples begins after the burn-in period.
- Determining the burn-in; Trial runs (e.g, running several chains & checking that they are close by). Often determined from the plot of "posterior trace" $\{\pi(x^{(k)}), k = 1, \ldots, N\}$ (see Figure).
- The acceptance ratio (number of proposals x' accepted per total number of samples) is useful indicator for the efficiency of the chain.



Determination of burn-in period. The burn-in was chosen at the location of the vertical line

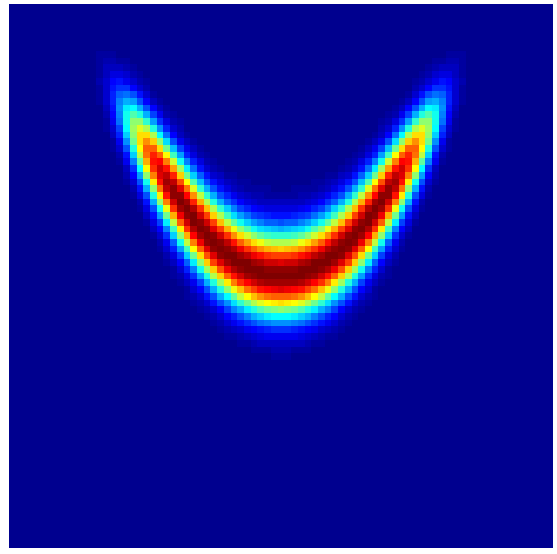
Example 1;

- Let $x \in \mathbb{R}^2$ and posterior

$$\pi \propto \pi_D(x) \exp\left\{-10(x_1^2 - x_2)^2 - (x_2 - \frac{1}{4})^4\right\},$$

where π_D is

$$\pi_D(x) = \begin{cases} 1, & x \in D \\ 0, & \text{otherwise} \end{cases} \qquad D = [-2\ 2] \times [-2\ 2] \subset \mathbb{R}^2.$$



Intensity plot of $\pi(x)$

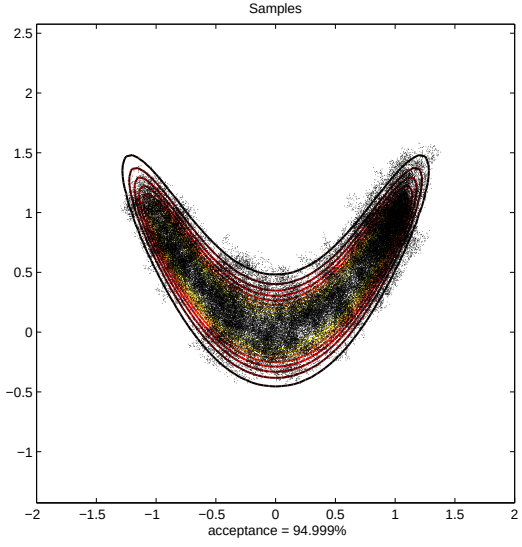
Example 1; Metropolis Hastings

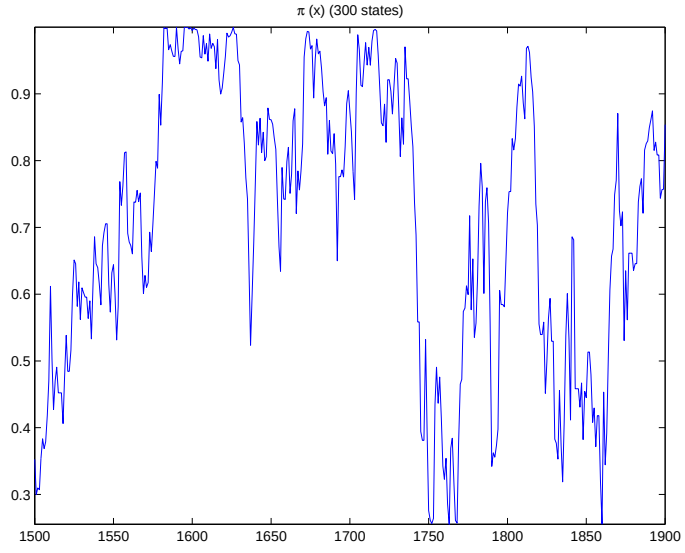
- Sample distribution $\pi(x)$ using the Metropolis Hasting algorithm. Choose proposal as (random walk Metropolis-Hastings)

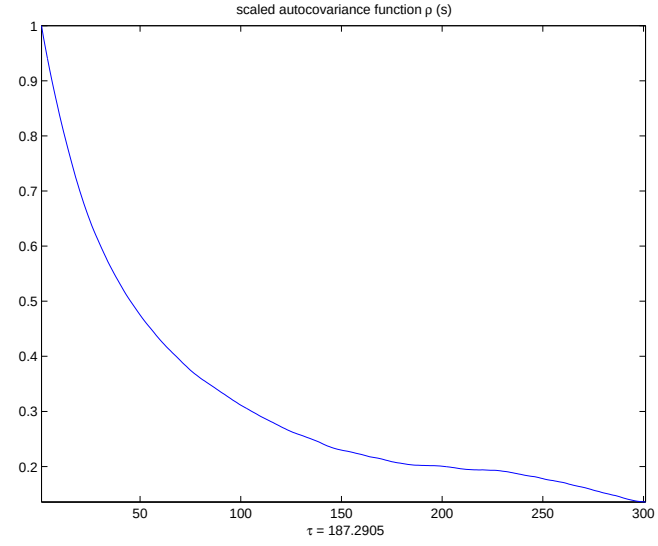
$$q(x, x') = \prod_{i=1}^{2} \frac{1}{\sqrt{2}\gamma} \exp\left\{ -\frac{1}{2\gamma^2}(x_i' - x_i)^2 \right\}.$$

- Consider the effect of the proposal parameter γ on the efficiency of the algorithm.

- Sampling with $\gamma = 0.02$
- $\tau = 187.3$, acceptance ratio 95%.
- Chain moves too slowly (very small steps $\longrightarrow$ long correlation), inefficient.

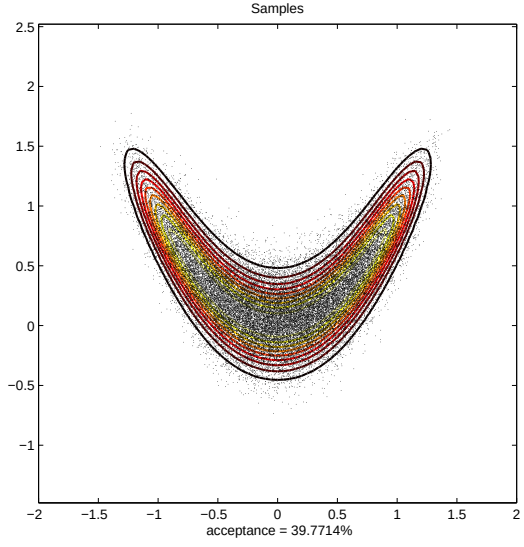


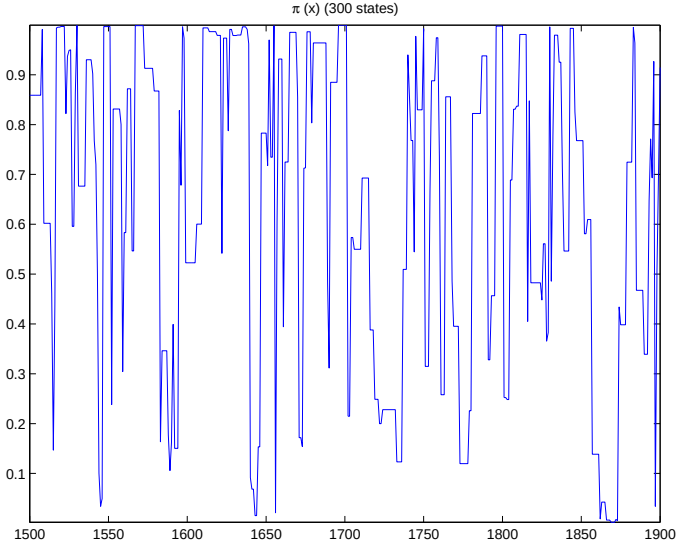


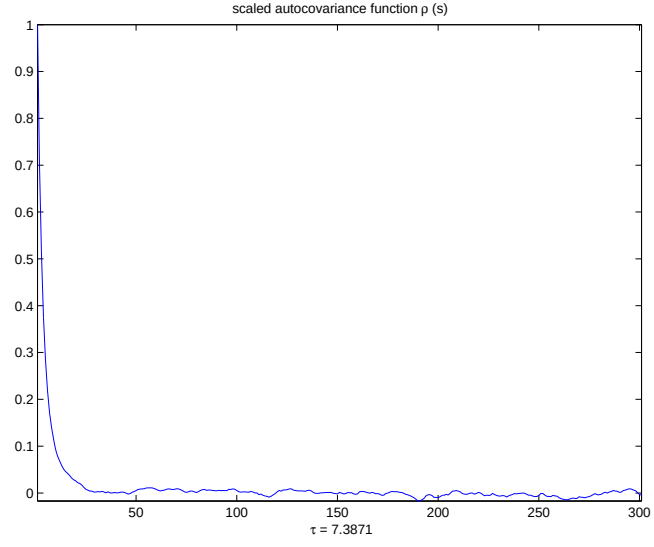


Left; samples. Middle; Posterior trace $\{\pi(x^{(k)})\}$ (400 states). Right; scaled autocovariance function.

- Sampling with $\gamma = 0.4$
- $\tau = 7.4$, acceptance ratio 40%.
- Chain is moving optimally.

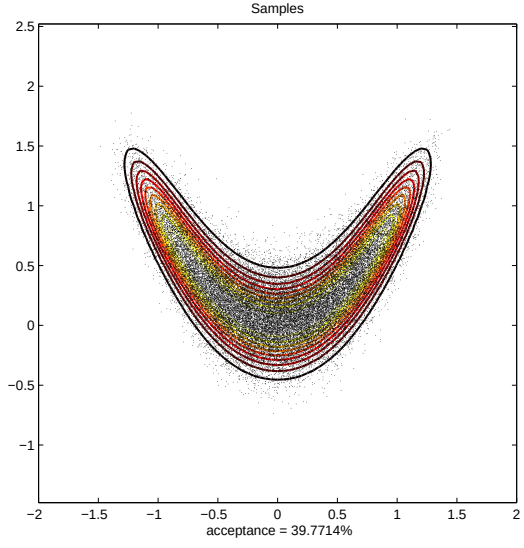


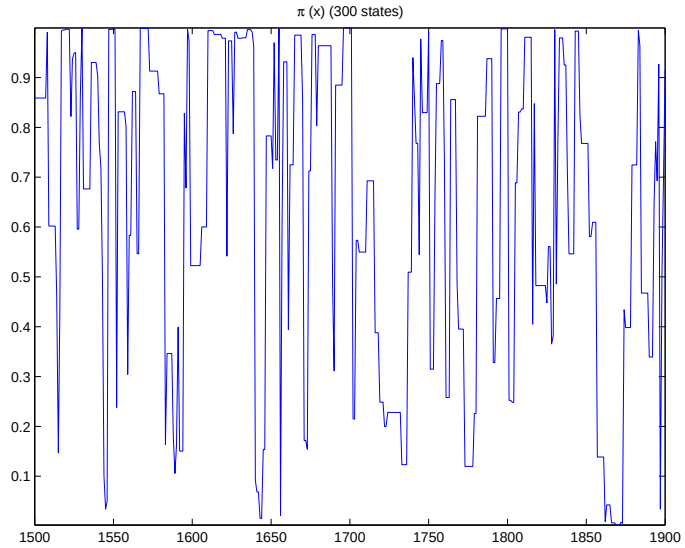


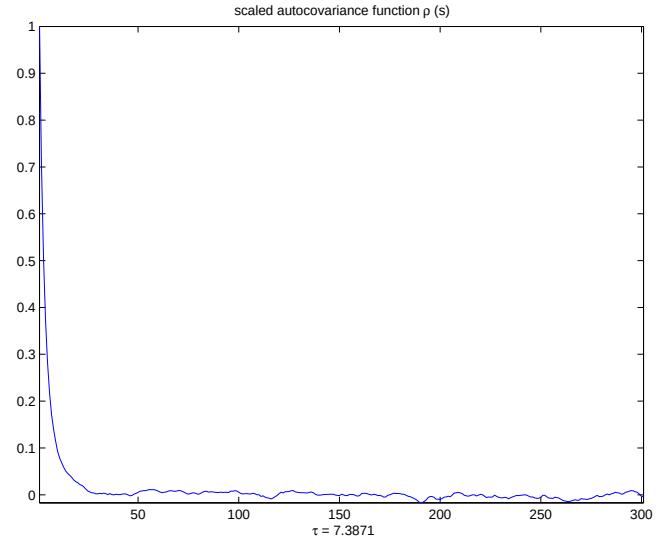


Left; samples. Middle; Posterior trace $\{\pi(x^{(k)})\}$ (400 states). Right; scaled autocovariance function.

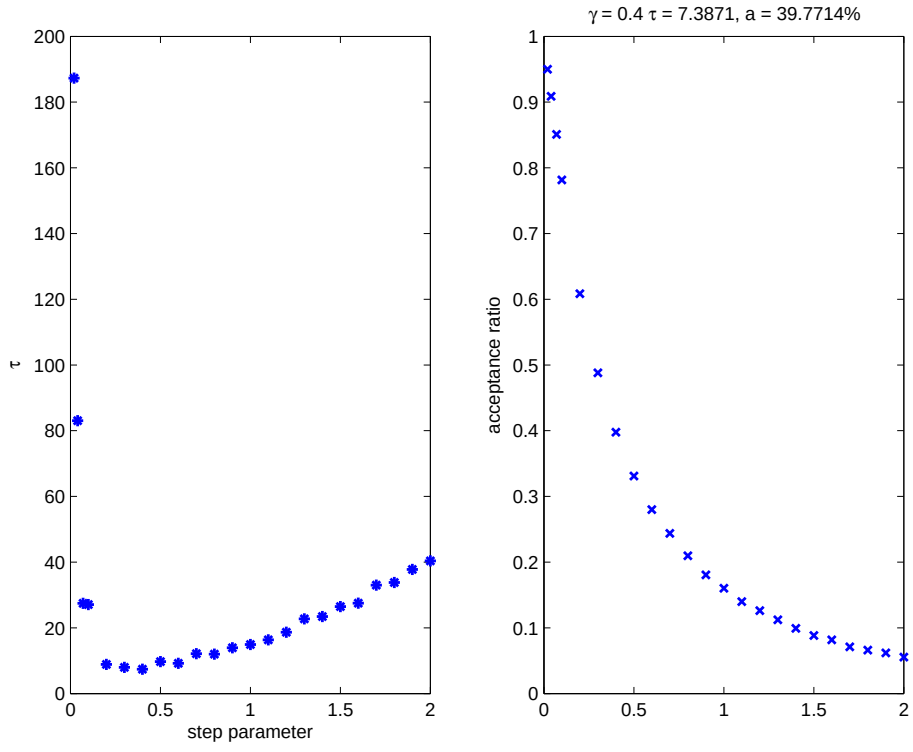
- Sampling with $\gamma = 1.8$
- $\tau = 33.8$, acceptance ratio 7%.
- Proposal too wide; most proposals x' rejected, chain gets "stuck" to single state for long periods ($\rightarrow$ long correlation).







Left; samples. Middle; Posterior trace $\{\pi(x^{(k)})\}$ (400 states). Right; scaled autocovariance function.



- Summary of the example; Left image shows τ vs γ and right image shows acceptance ratio vs. γ.
- Rule of thumb (when $q(x, x')$ Gaussian); Aim at acceptance ratio $20 - 50\%$.

Example 1 (cont.)

- Improve the proposal. Let us draw samples as

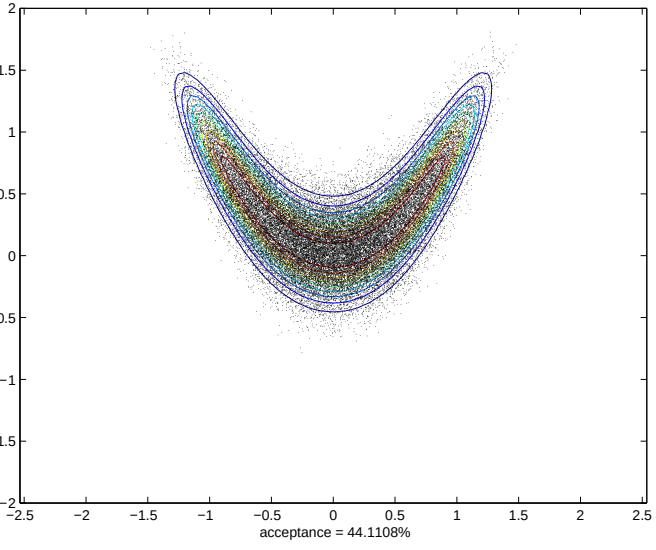
$$x_i' = x_i + \delta_i(x) + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, \gamma^2),\ i = 1, 2,$$

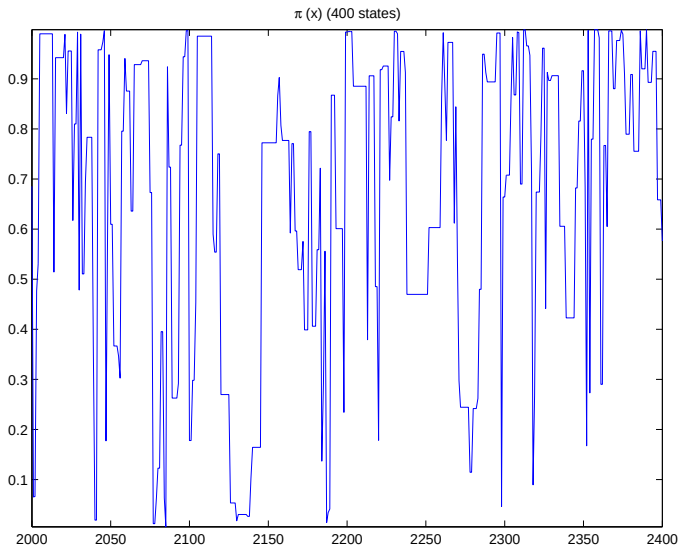
where $\delta_i : \mathbb{R}^2 \mapsto \mathbb{R}$ is deterministic mapping. Now

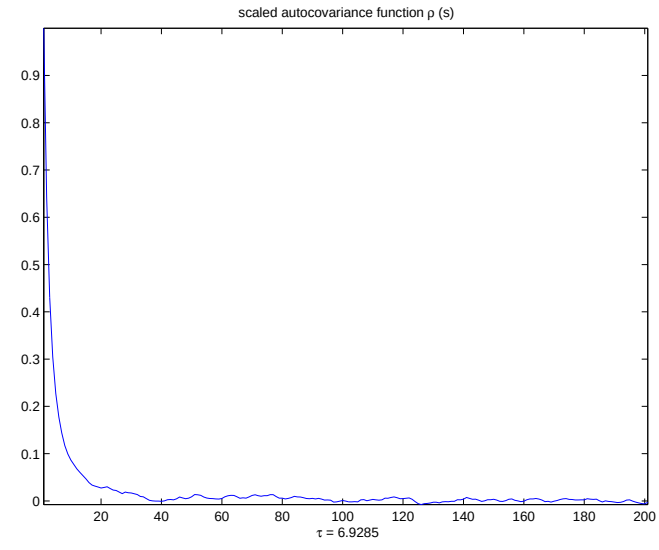
$$q(x, x') = \prod_{i=1}^{2} \frac{1}{\sqrt{2}\gamma} \exp\left\{ -\frac{1}{2\gamma^2}(x_i' - (x_i + \delta_i(x))^2 \right\}.$$

- Choose $\delta_i(x) = (h\nabla\pi(x))_i$, where h is constant.

- Sampling with $\gamma = 0.4$, $h = 0.02$
- $\tau = 6.9$, acceptance ratio 44%.







Left; samples. Middle; Posterior trace $\{\pi(x^{(k)})\}$ (400 states). Right; scaled autocovariance function.

Gibbs sampler algorithm

- Generation of the ensemble $\{x^{(k)}, k = 1, \ldots, N\} \sim \pi(x)$ using Gibbs sampler;

  1. Pick initial value $x^{(0)}$ and set $j = 1$.
  2. Generate $x^{(j)}$ a single variable at a time:

$$
\begin{aligned}
&\text{draw} \quad x_1^{(j)} \quad \text{from the density } t \mapsto \pi(t, x_2^{(j-1)}, \ldots, x_n^{(j-1)}),\\
&\text{draw} \quad x_2^{(j)} \quad \text{from the density } t \mapsto \pi(x_1^{(j)}, t, x_3^{(j-1)}, \ldots, x_n^{(j-1)}),\\
&\qquad\qquad \vdots\\
&\text{draw} \quad x_n^{(j)} \quad \text{from the density } t \mapsto \pi(x_1^{(j)}, \ldots, x_{n-1}^{(j)}, t).
\end{aligned}
$$

  3. If $j = N$, stop, else set $j \leftarrow j + 1$ and go to 2.

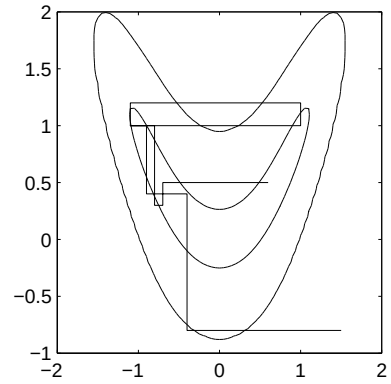


Illustration of the componentwise sampling with the Gibbs sampler.

- Conditionals can be sampled as follows;

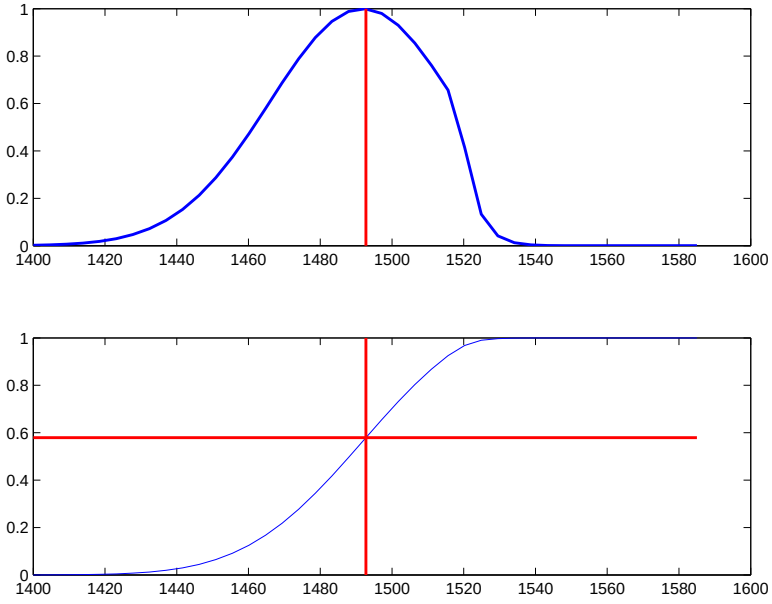
1. Determine the cumulative function

$$\Phi_j(t) = \int_{-\infty}^{t} \pi(x \mid x_{-j}) \mathrm{d}x_j, \quad x_{-j} = (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n) \in \mathbb{R}^{n-1}$$

2. Draw random sample $\xi \sim \mathrm{uni}([0\ 1])$. Sample y_j is obtained as

$$y_j = \Phi_j^{-1}(\xi)$$

Illustration of sampling from the 1D conditionals.



- Top; $\pi(x \mid x_{-j})$
- Bottom; cumulative function $\Phi_j(t) = \int_{-\infty}^{t} \pi(x \mid x_{-j}) \mathrm{d}x_j$.
- $\xi \sim \mathrm{uni}([0\ 1])$ (location of horizontal line)
- Sample $y_j = \Phi_j^{-1}(\xi)$ (location of the vertical line)

- Posterior typically in non-parametric form, normalization constant unknown $\Rightarrow$ Numerical approximation (e.g. trapezoidal quadrature)

$$\Phi_j(t_m) = C \int_a^{t_m} \pi(x_j \mid x_{-j}) \mathrm{d}x \approx C \sum_{k=1}^{r} w_k \pi(t_k \mid x_{-j})$$

  where w_k are the quadrature weights.

- Support $[a,\ b]$ of the conditional $\pi(x_j \mid x_{-j})$ has to be "forked" carefully. C chosen s.t. $\Phi(b) = 1$.

- Sample can be obtained by interpolation. Let $\Phi_j(t_p) < \xi < \Phi_j(t_{p+1})$, then linear interpolation gives

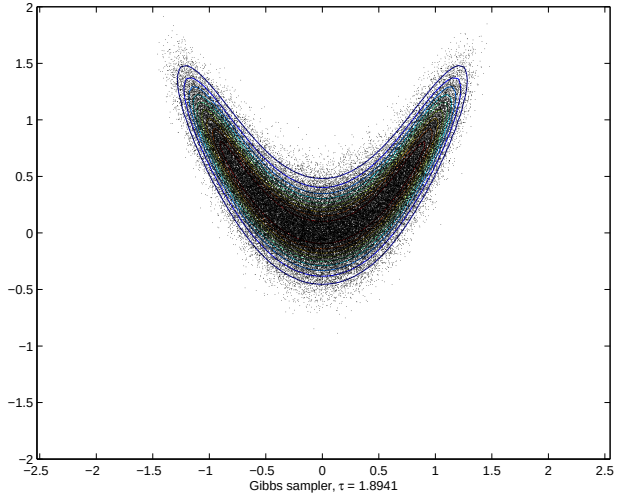
$$y_j = t_p + \frac{\xi - \Phi_j(t_p)}{\Phi_j(t_{p+1}) - \Phi_j(t_p)} (t_{p+1} - t_p)$$

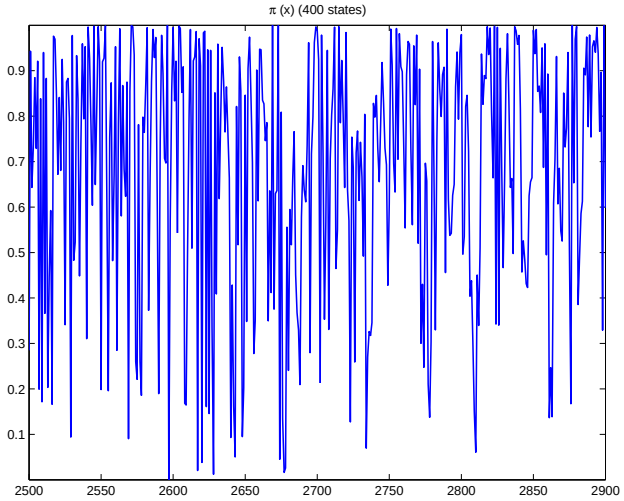
Some features of Gibbs sampler;

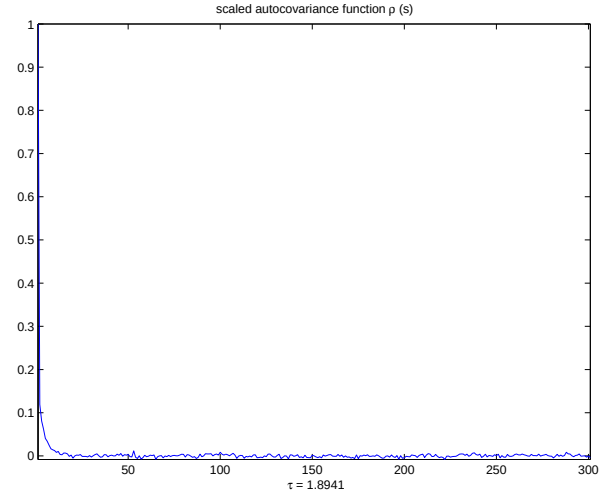
- Acceptance probability always 1. No need to tuning the proposal.
- Determination of burn-in similarly as for Metropolis Hastings.
- If $\pi(x)$ has high correlation, can get inefficient (τ increases)
- Numerical evaluation of the conditionals can get computationally infeasible (e.g. high dimensional cases with PDE based forward models).

Example 1 (cont.); Gibbs sampling

- $\tau = 1.9$.
- Computation time longer than for Metropolis Hastings.



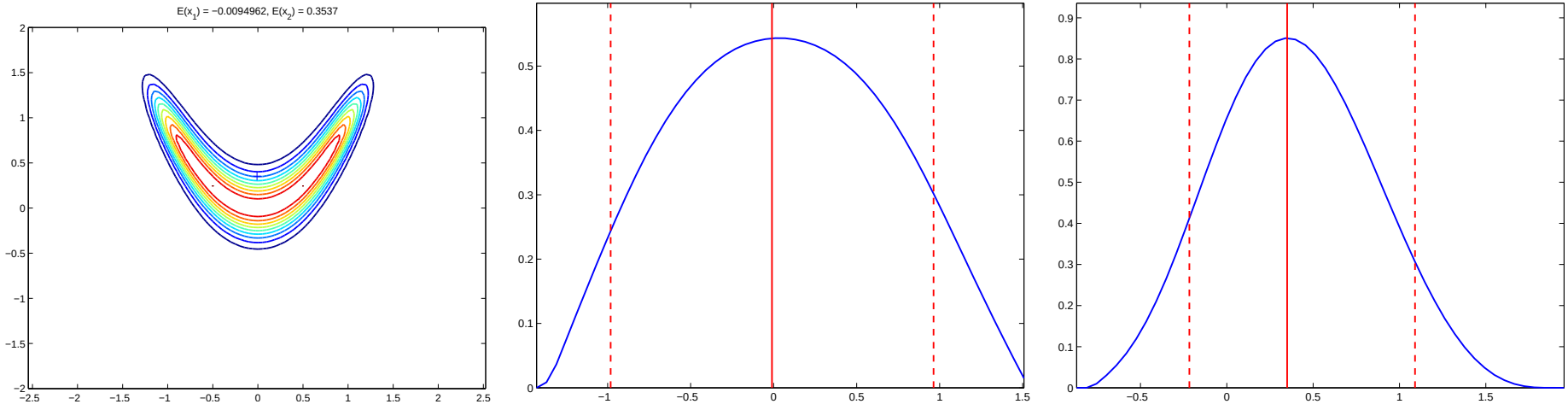




Left; samples. Middle; Posterior trace $\{\pi(x^{(k)})\}$ (400 states). Right; scaled autocovariance function.

Example 1; Estimates (Gibbs sampling)

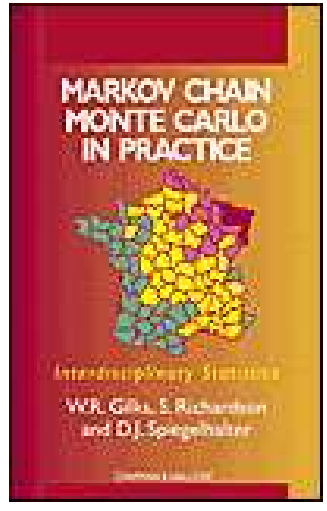
- $x_{\mathrm{CM}} \approx (0.00, 0.35)^{\mathrm{T}}$
- Marginal densities $\pi(x_1) = \int \pi(x)\mathrm{d}x_2$ and $\pi(x_2) = \int \pi(x)\mathrm{d}x_1$
- Solid vertical lines show the mean, dotted the 90% confidence limits.



Left; Contours of $\pi(x)$. x_{CM} is shown with $+$. Middle; marginal density $\pi(x_1)$. Right; marginal density $\pi(x_1)$.

A few references to MCMC;

- The collection by Gilks et al. General overview to MCMC.



- "Physics 707 (Inverse Problems)" lecture notes by Colin Fox, Geoff Nicholls and S.M. Tan (Univ. of Auckland) (available at `http://www.math.auckland.ac.nz/ phy707/`). Presentation of the theory (disrete state space) and practical computations.

Example 2; 1D deconvolution problem

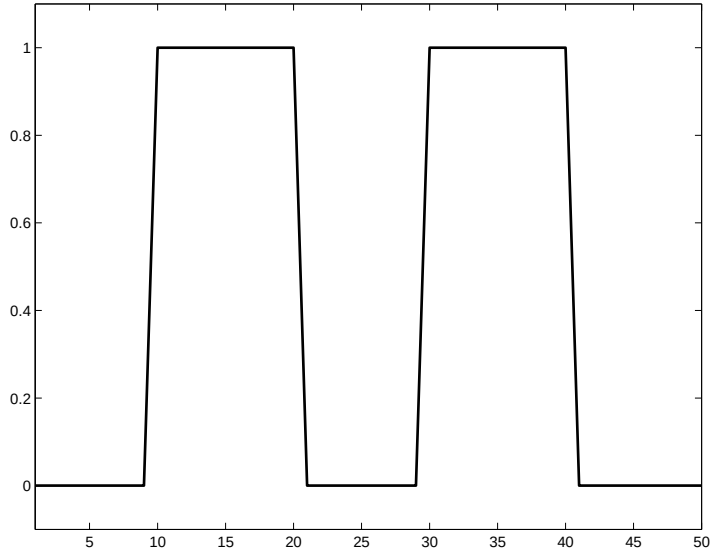
- $x \in \mathbb{R}^{50}, \quad m \in \mathbb{R}^{50}$. Measurement model

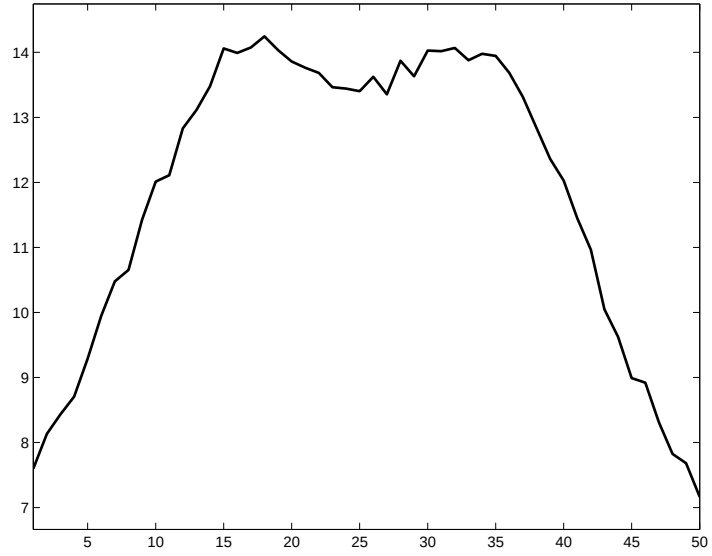
$$m = Ax + e, \quad e \sim \mathcal{N}(0, \Gamma_e)$$

where elements of matrix A are

$$A_{i,j} = \exp\{-0.05|t_i - t_j|\}.$$

- Left figure shows x_{true} and right figure shows noisy, blurred observation m.
- Level of noise e is 1%.

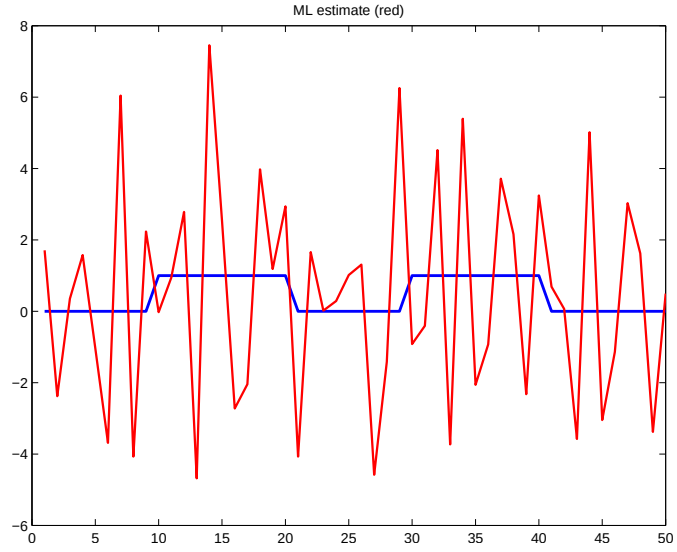




Example 2; 1D deconvolution problem (cont.)

- Inverse problem; unique ($\mathrm{null}(A) = \{0\} \Rightarrow \exists A^{-1}$) but sensitive to the measurement errors.
- Figure shows the solution

$$x = \arg\min_x \{\|m - Ax\|_2^2\} \quad (\Rightarrow x = A^{-1}m)$$



x_{true} (blue), $x = \arg\min_x \{\|m - Ax\|_2^2\}$ (red).

Example 2; 1D deconvolution problem (cont.)

- Likelihood model

$$\pi(m|x) \propto \exp\left(-\frac{1}{2}\|L_e(m - Ax)\|_2^2\right),$$

  where $L_e^{\mathrm{T}} L_e = \Gamma_e^{-1}$.

- We assume *a priori* that the target is
  - Non-negative
  - Piecewise constant
- We model this information by prior model

$$\pi_{\mathrm{pri}}(x) \propto \pi_+(x) \exp\left(-\alpha \sum_{j=1}^{n-1} |x_{j+1} - x_j|\right)$$

Example 2; 1D deconvolution problem (cont.)

- Posterior density;

$$\pi(x|m) \propto \pi_+(x) \exp\left(-\frac{1}{2}\|L_e(m - Ax)\|_2^2 - \alpha \sum_{j=1}^{n-1} |x_{j+1} - x_j| \right)$$

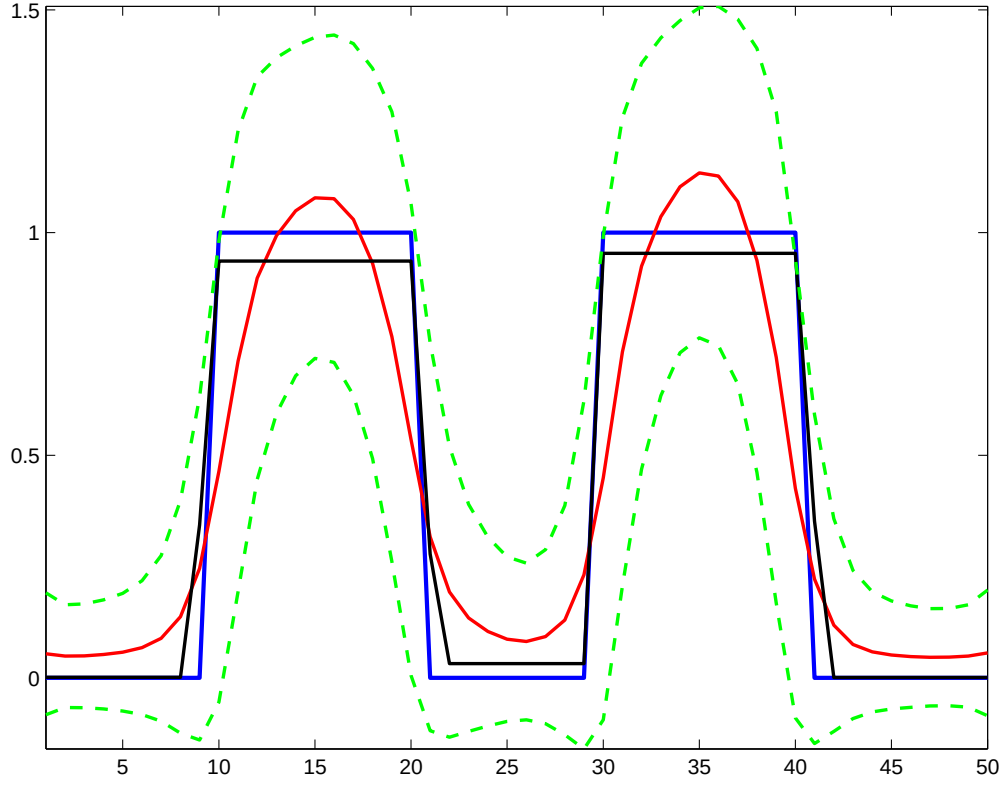
- We compute
  - MAP estimate

$$x_{\mathrm{MAP}} = \arg\min_{x \geq 0} \{ \frac{1}{2}\|L_e(m - Ax)\|_2^2 + \alpha \sum_{j=1}^{n-1} |x_{j+1} - x_j| \}$$

    by quadratic programming.
  - CM estimate, covariances and marginal confidence limits; MCMC (Gibbs sampler).

Example 2; 1D deconvolution problem (cont.)

- CM estimate (red). MAP estimate (black). CM with 3 std confidence limits (green).
- The limits correspond approximately to 90% confidence limits (Chebychev inequality)



x_{true} (blue), x_{MAP} (black), x_{CM} (red), $x_{\mathrm{CM}} \pm 3\sigma_{x|m}$ (green).

Example 2; 1D deconvolution problem (cont.)

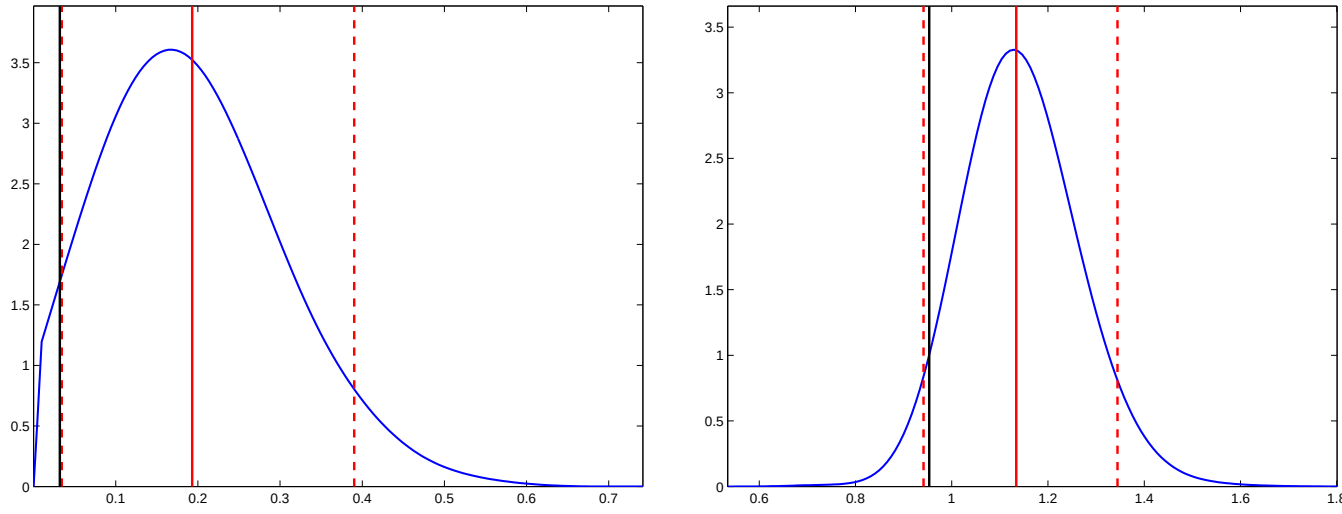
- Marginal posterior densities

$$\pi_k(x_k) = \int_{\mathbb{R}^{n-1}} \pi(x \mid m) \mathrm{d}x_{-k}$$

for elements x_{22} and x_{35}. Value of x_{MAP} (black) and x_{CM} (red) vertical lines.

- Dashed lines show the 90% condidence intervals

$$\int_{-\infty}^{a_k} \pi_k(x_k)\, \mathrm{d}x_k = \int_{b_k}^{\infty} \pi_k(x_k)\, \mathrm{d}x_k = 0.05$$



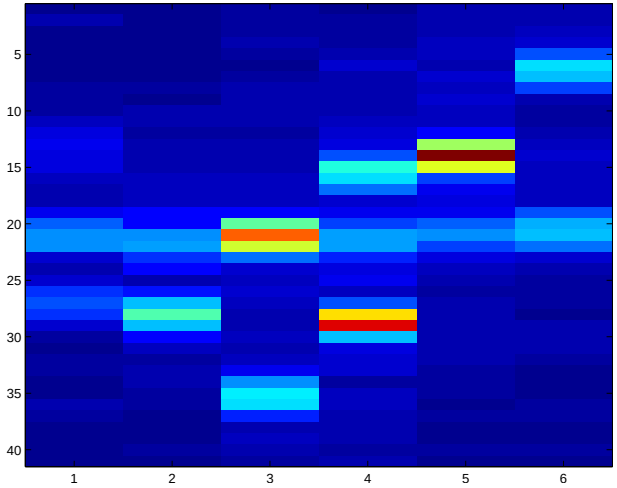
Marginal posterior densities for x_{22} (left) and x_{35} (right).

Example 3; Emission tomography in brachytherapy

Brachytherapy source pellets.

- Brachytherapy (sealed source radiotherapy); radioactive source pellets are placed inside the tumor for radiation treatment.
- Used in treatment of cancers of prostata, breast, head and neck area, etc...
- Thin hollow tubes are placed inside the cancerous tissues. The radioactive source pellets are inside the tubes.
- The tubes are connected to a (pneumatic) loader machine that can be used to control the location of the pellets inside the tubes.

Example 3 (cont.)



TLD emission data (6 projections). Data from Kuopio University hospital.

- Objective of the experiment to test the feasibility of (limited data) emission tomography for verification of the correct placement of the source pellets inside tissues.
- Phantom experiment. 6 projections with projection interval of 30°
- Data collected with a termoluminescent dosimeter (TLD) with a parallel beam collimator geometry. 41 readings in each projection (i.e., $m \in \mathbb{R}^{246}$)
- The model for the expectation of the observed photon count

$$\bar{m}_j = \int_{L_j} x(s)\mathrm{d}s$$

where L_j is the "line of sight" detected from the j:th detector position.

- The model neglects scattering phenomena and attenuation correction.

Example 3 (cont.)

- Discretization; the domain Ω of interest is divided to regular 41×41 pixel grid (i.e., $f \in \mathbb{R}^{1681}$).
- The forward model becomes

$$x \mapsto Ax, \quad A : \mathbb{R}^{1681} \mapsto \mathbb{R}^{246}$$

- The inverse problem is to reconstruct the activity distribution x, given the vector of observed photon counts m.

Example 3 (cont.)

- Likelihood model; the observations $\{m_j, j = 1, \ldots, n_m\}$ are Poisson distributed random variables. The fluctuations are assumed mutually independent $\Rightarrow$ we use likelihood model

$$\begin{aligned} \pi(m \mid x) &= \prod_{j=1}^{n_m} \frac{(Ax)_j^{m_j}}{m_j!} \exp(-(Ax)_j) \\ &\propto \exp(m^{\mathrm{T}} \log(Ax) - \mathbf{1}^{\mathrm{T}}(Ax)) \end{aligned}$$

  The model neglects electronic noise of the data acquisition system.

- We know *a priori* that the activity distribution is;
  - Non-negative
  - The source pellets are small and inside the hollow tubes $\Rightarrow$ all the activity is in small pixel clusters of approximately constant activity.
- We model this knowledge by prior model

$$\pi_{\mathrm{pri}}(x) \propto \pi_+(x) \exp\left(-\alpha \mathrm{TV}(x)\right), \quad \mathrm{TV}(x) = \sum_{k=1}^{n} \sum_{j \in \mathcal{N}_k} |x_k - x_j|$$

Example 3 (cont.)

- Posterior density;

$$\pi(x|m) \propto \pi_+(x) \exp(m^{\mathrm{T}} \log(Ax) - \mathbf{1}^{\mathrm{T}}(Ax) - \alpha \sum_{k=1}^{n} \sum_{j \in \mathcal{N}_k} |x_k - x_j|)$$

- We compute CM estimate by Metropolis-Hastings MCMC.
- Proposals x' are drawn with the following single component random scan;

  Step 1: choose update element $x_i \in \mathbb{R}$ by drawing index i with uniform probability

$$\frac{1}{n}$$

  Step 2: Generate x' by updating x_i s.t.:

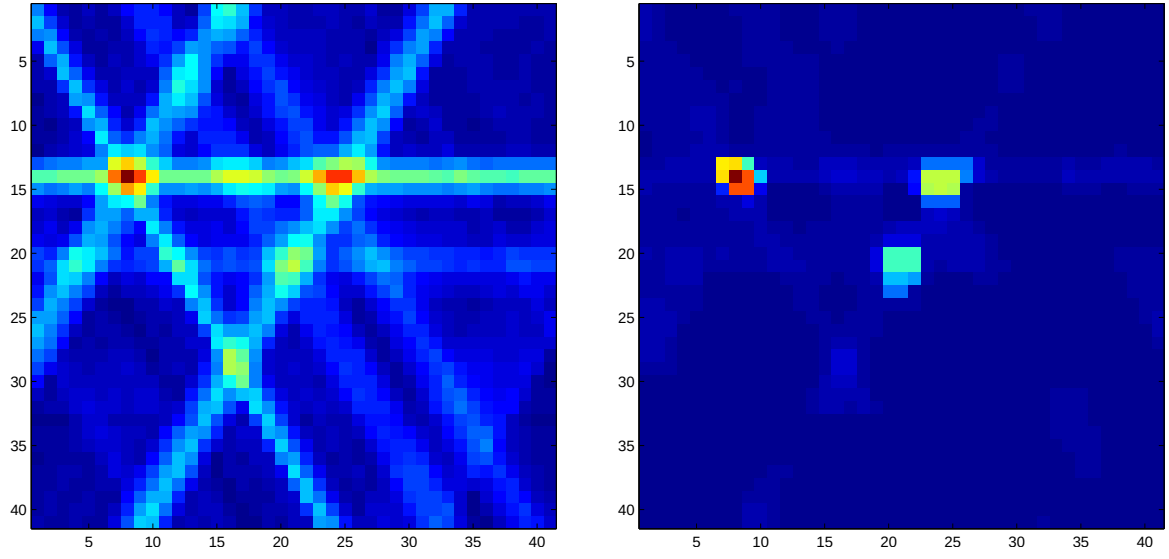
$$x_i' = |x_i + \xi|, \quad \xi \sim \mathcal{N}(0, \epsilon^2)$$

Example 3 (cont.)

- Left image; Tikhonov regularized solution

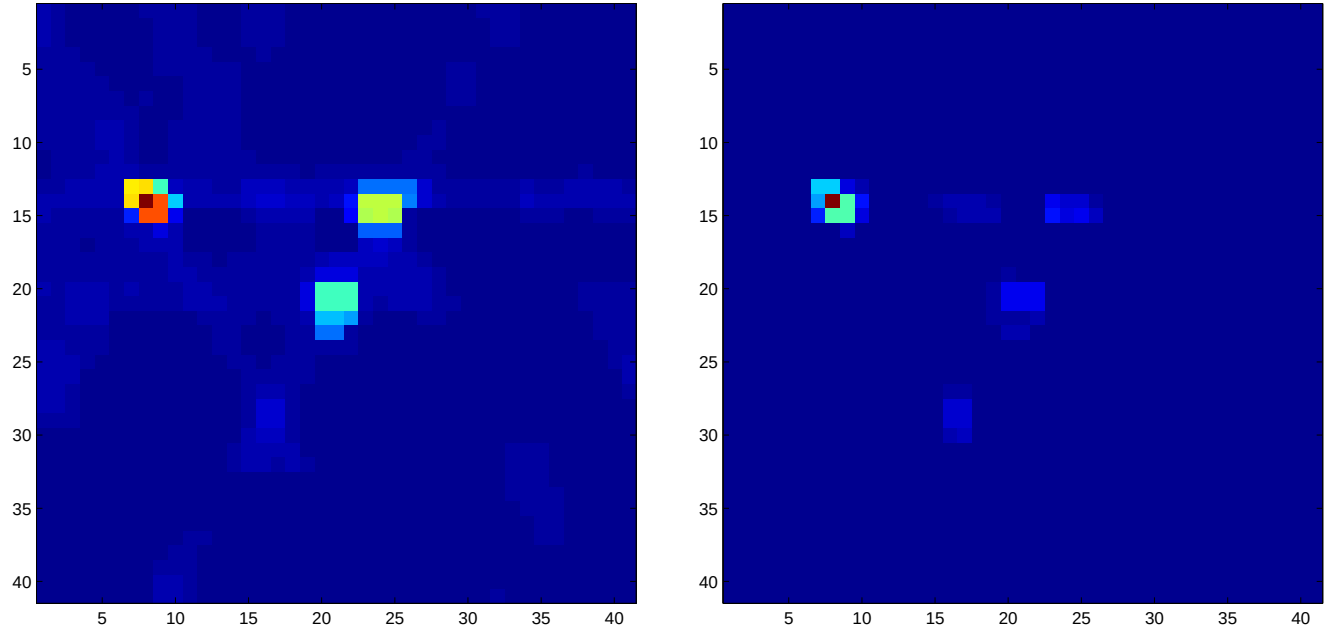
$$x_{\mathrm{TIK}} = \arg\min_x \{\|m - Ax\|_2^2 + \alpha\|x\|_2^2\}$$

- Right image; CM estimate $x_{\mathrm{CM}} = \int_{\mathbb{R}^n} x\pi(x|m)\mathrm{d}x$.
- The source pellets (3 pcs.) are localized correctly in the CM estimate.



Example 3 (cont.)

- Left image; CM estimate x_{CM}.
- Right image; posterior variances $\mathrm{diag}\Gamma_{x|m}$



Example 4; 2D deconvolution (image deblurring)

- We are given a blurred and noisy image

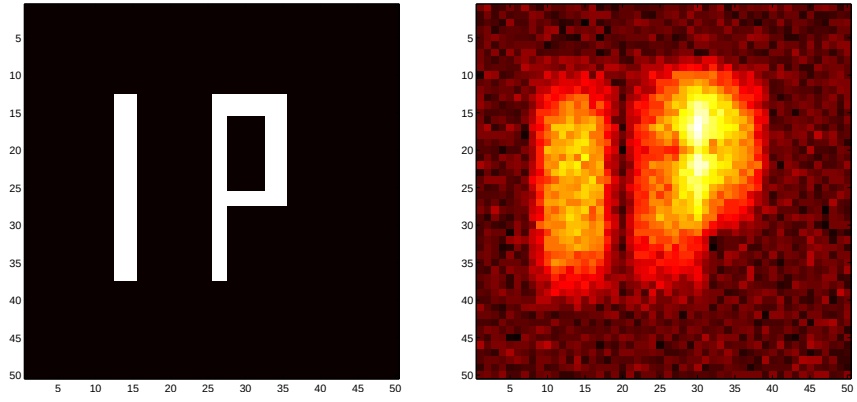
$$m = Ax + e, \quad m \in \mathbb{R}^n$$

- Forward model

$$x \mapsto Ax$$

implements discrete convolution (the convolution kernel is Gaussian blurring kernel with std of 6 pixels).

- Assume that we know *a priori* that the true solution is binary image representing some text.



Left; true image x. Right; Noisy, blurred image $m = Ax + e$.

Example 4 (cont.)

- The prior knowledge is considered consistent with;
  - x can obtain the values $x_j \in \{0, 1\}$
  - Case $x_j = 0$ (black, background), case $x_j = 1$ (white, text).
  - The pixels with value $x_j = 1$ are known to be clustered in alphabets which have clear and short boundaries.
- We model the prior knowledge with

$$\pi_{\text{pri}}(x) \propto \exp\left(\alpha \sum_{i=1}^{n} \sum_{j \in \mathcal{N}_i} \delta(x_i, x_j)\right)$$

where δ is the Kronecker delta

$$\delta(x_i, x_j) = \begin{cases} 1, & x_i = x_j \\ 0, & x_i \neq x_j \end{cases}$$

Example 4 (cont.)

- Noise $e \sim \mathcal{N}(0, \Gamma_e)$, e and x mutually independent. Likelihood model becomes

$$\pi(m|x) \propto \exp\left(-\frac{1}{2}\|L_e(m - Ax)\|_2^2\right),$$

  where $L_e^{\mathrm{T}} L_e = \Gamma_e^{-1}$.

- Posterior model for the deblurring problem;

$$\pi(x|m) \propto \exp\left(-\frac{1}{2}\|L_e(m - Ax)\|_2^2 + \alpha \sum_{i=1}^{n} \sum_{j \in \mathcal{N}_i} \delta(x_i, x_j)\right)$$

- Posterior explored with the Metropolis Hastings algorithm.

Example 4 (cont.)

- The proposal is a mixture $q(x, x') = \sum_{i=1}^{2} \xi_i q_i(x, x')$ of two move types.

Move 1: choose update element $x_i \in \mathbb{R}$ by drawing index i with uniform probability $\frac{1}{n}$ and change the value of x_i.

Move 2: Let $N^*(x)$ denote the set of active edges in image x (edge l_{ij} connects pixels x_i and x_j in the lattice; it is active if $x_i \neq x_j$). Pick an active update edge w.p.

$$\frac{1}{|N^*(x)|}$$

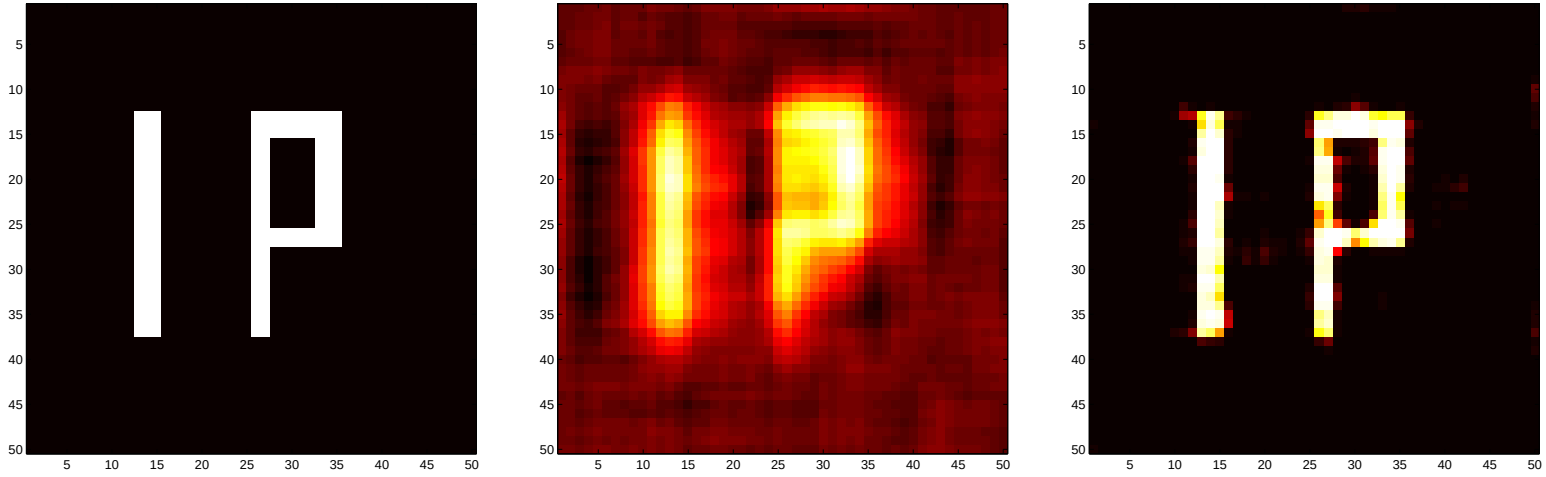
Pick one of the two pixels related to the chosen edge w.p $\frac{1}{2}$ and change the value of the chosen pixel.

Example 4 (cont)

- Left image; True image x
- Middle; Tikhonov regularized solution

$$x_{\mathrm{TIK}} = \arg\min_x \{\|m - Ax\|_2^2 + \alpha\|x\|_2^2\}$$

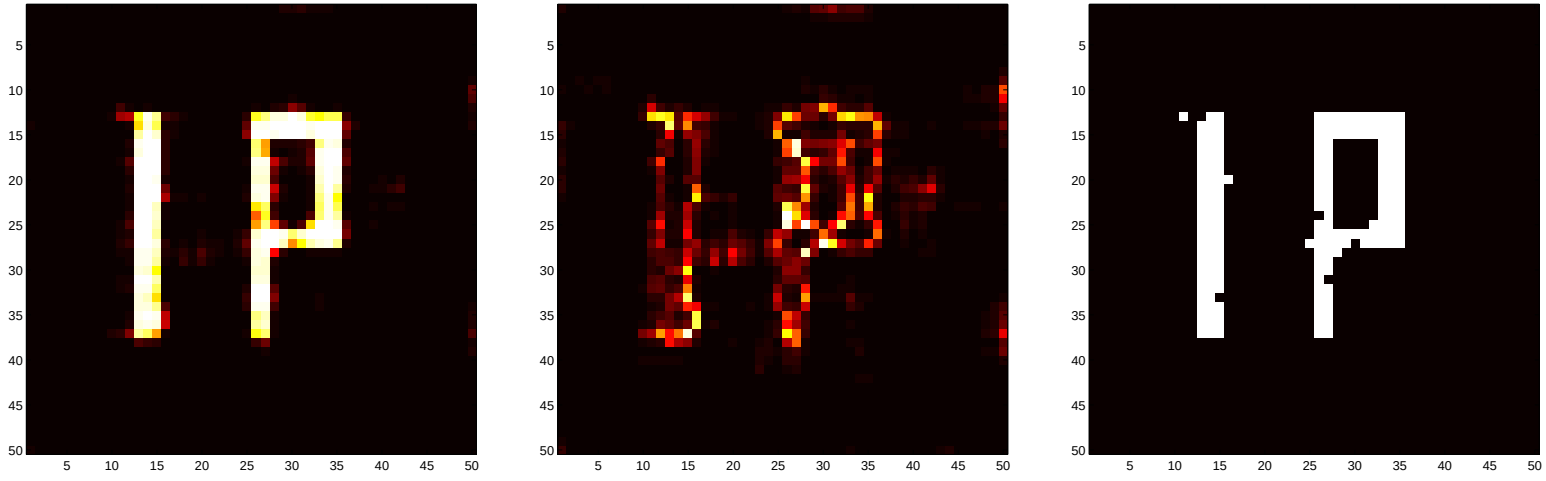
- Right image; CM estimate $x_{\mathrm{CM}} = \int_{\mathbb{R}^n} x\pi(x|m)\mathrm{d}x$.



Left; true image x, Middle; x_{TIK}. Right; x_{CM}.

Example 4 (cont)

- Left image; x_{CM}
- Posterior variances $\mathrm{diag}(\Gamma_{x|m})$
- Right; A sample from $\pi(x \mid m)$

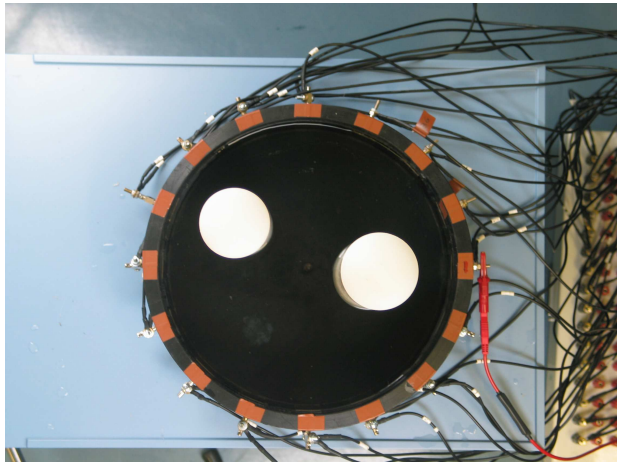


Left; x_{CM}, Middle; posterior variances $\mathrm{diag}(\Gamma_{x|m})$. Right; A sample x from $\pi(x \mid m)$.

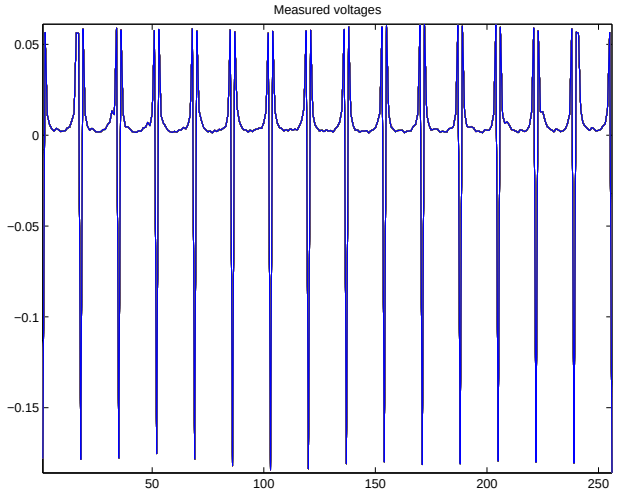
Example 5; Bayesian Inversion of EIT data

Joint work with Colin Fox and Geoff Nicholls

Bayesian inversion of EIT data



Target conductivity



Measured voltages $V \in \mathbb{R}^{256}$

- Measurement model

$$V = U(\sigma) + e, \quad e \sim \mathcal{N}(0, \Gamma_e)$$

where V denotes the measured voltages and $U(\sigma)$ FEM-based forward map.

- Γ_e estimated by repeated measurements.
- Posterior model

$$\pi(\sigma \mid V) \sim \exp\left\{ -\frac{1}{2}(V - U(\sigma))^{\mathrm{T}} \Gamma_e^{-1} (V - U(\sigma)) \right\} \pi_{\mathrm{pr}}(\sigma)$$

- We compute results with four different prior models $\pi_{\mathrm{pri}}(x)$.
- Sampling carried out with the Metropolis-Hasting sampling algorithm.

Prior models

- **Smoothness prior**

$$\pi_{\mathrm{pr}}(\sigma) \propto \pi_{+}(\sigma) \exp\left(-\alpha \sum_{i=1}^{n} H_i(\sigma)\right), \quad H_i(\sigma) = \sum_{j \in \mathcal{N}_i} |\sigma_i - \sigma_j|^2.$$

- **Material type prior (Fox & Nicholls,1998)**
  - The possible material types $\{1, 2, \ldots, C\}$ inside the body are known but the segmentation to these materials is unknown. The material distibution is represented by an (auxliary) image $\tau \in \mathbb{R}^n$ (pixel values $\tau_j \in \{1, 2, \ldots, C\}$).
  - The possible values of conductivity $\sigma(\tau)$ for different materials are known only approximately.
  - The different material types are assumed to be clustered in “blocky structures” (e.g., organs, etc)

- We model this information by

$$\pi_{\mathrm{pr}}(\sigma|\tau) \propto \pi_+(\sigma) \exp\left(-\alpha \sum_{i=1}^{n} G_i(\sigma)\right) \prod_{i=1}^{n} \exp\left(-\frac{1}{2\xi(\tau_i)^2}(\sigma_i - \eta(\tau_i))^2\right),$$

where G_i implements the structural ("segmented") MRF prior:

$$G_i(\sigma) = \sum_i \sum_{j \in \{k | k \in \mathcal{N}_i \,\mathrm{and}\, \tau_k = \tau_i\}} (\sigma_i - \sigma_j)^2$$

- For the material type image τ we use an Ising prior:

$$\pi_{\mathrm{pr}}(\tau) \propto \exp\left(\beta \sum_{i=1}^{n} T_i(\tau)\right), \quad T_i(\tau) = \sum_{j \in \mathcal{N}_i} \delta(\tau_i, \tau_j), \tag{1}$$

where $\delta(\tau_i, \tau_j)$ is the Kronecker delta function.

- **Circle prior**
  - Domain Ω with known (constant) background conductivity σ_{bg} is assumed to contain *unknown number* of circular inclusions
  - The inlcusions have known contrast
  - Size of inclusions unknown
- For the number N of the circular inclusions we write the point process prior:

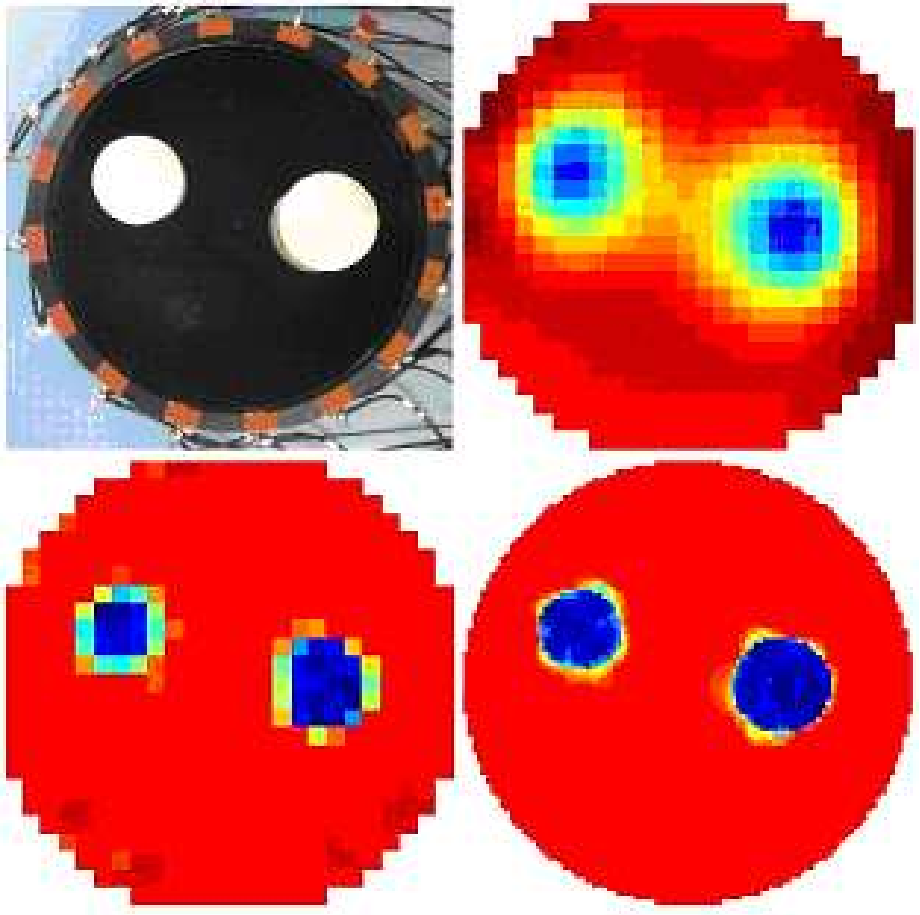
$$\pi_{\mathrm{pr}}(\sigma) = \beta^N \exp(-\beta A)\delta(\sigma \in \mathcal{A}),$$

where

  - $\beta = 1/A$ (A is the area of the domain Ω)
  - $\mathcal{A}$ is set of feasible circle configurations (the circle inclusions are disjoint)
  - δ is the indicator function

$$\delta(\sigma \in \mathcal{A}) = \begin{cases} 1, & \sigma \in \mathcal{A} \\ 0, & \sigma \notin \mathcal{A} \end{cases}$$

CM estimates with the different prior models;

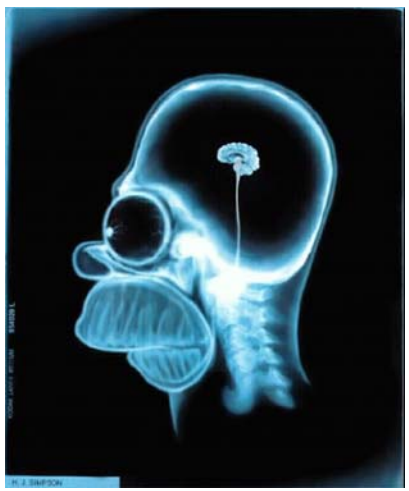


- Top left; target
- Top right; σ_{CM} (smoothness prior)
- Bottom left; σ_{CM} (material type prior)
- Bottom right; σ_{CM} (circle prior)

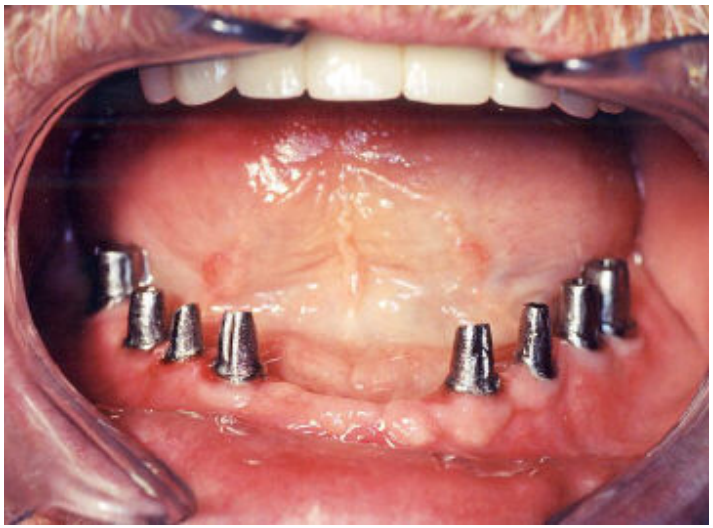
Example 6; Bayesian inversion in 3D dental x-ray imaging

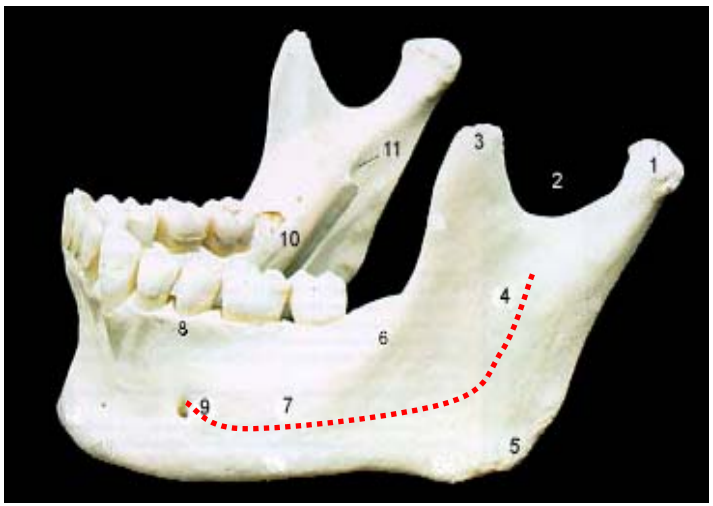
joint work with

Samuli Siltanen, Matti Lassas, Jari Kaipio, Antti Vanne,
Seppo Järvenpää and Martti Kalke

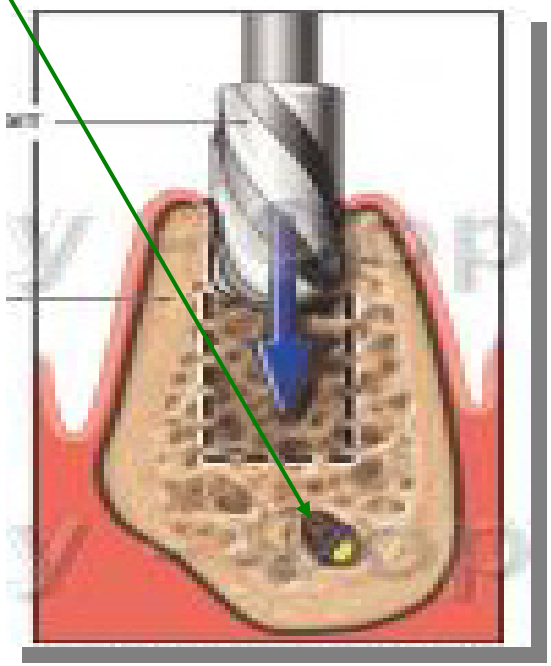


Background; Implant planning in dental implantology

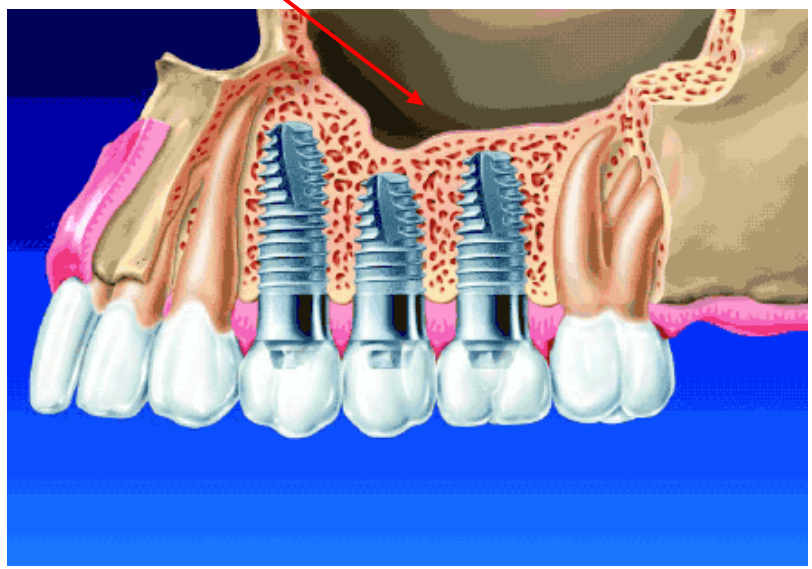




- Accurate measurements are needed for thickness & orientation of solid bone for optimal attachment of the implants
- Mandibular area; the location of the nerve canal need to be known (avoiding damage to the facial nerve)

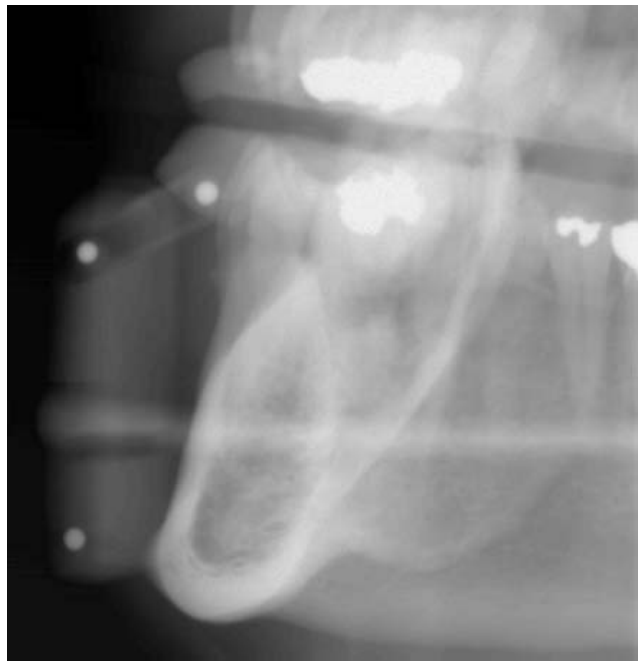


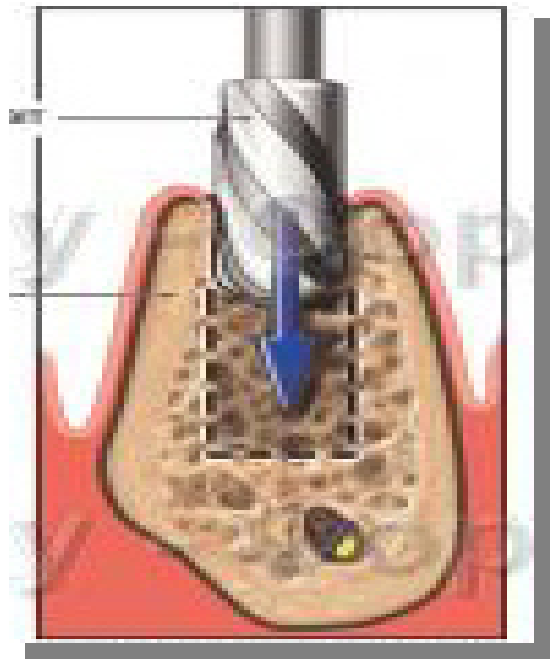
- Maxilla area; distance to sinus cavities need to be known



• These measurements that are needed for safe and successful implantation cannot be obtained from a single projection radiograph.

• Our objective is to retrieve the needed 3D information using regular (digital) dental panoramic x-ray device

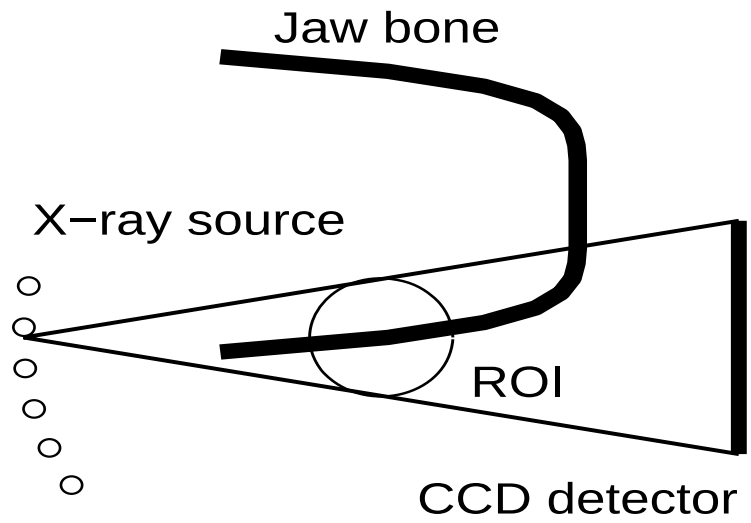




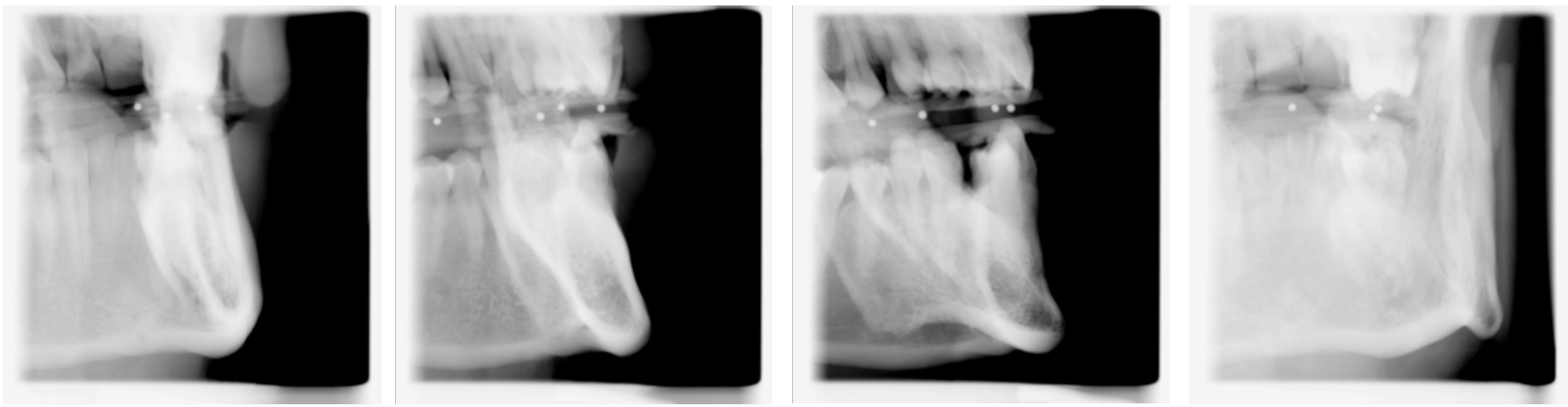
Left: x-ray imaging with digital panoramic device. **Middle:** Projection radiograph of the mandibular area. **Right:** drilling of implant hole.

We consider the following limited angle experiment with the panoramic device:

- 11 projection images of the mandibular area
- 1k x 1k pixels per image



Projection geometry (maximal opening angle available ~40 deg)



Projection images of the mandibular area

- Quantitative and qualitative prior knowledge in dental radiology:

- ✓ Different tissue types (bone, gum, pulp chamber) and possible artificial materials (amalgam, previous implants) are approximately homogeneous
- ✓ Attenuation (density) of tissues is non-negative (X-radiation does not intensify inside tissue)
- ✓ There are crisp boundaries between the different tissues.

- We model the prior knowledge by the following prior models:

Positivity prior:

$$\pi_+(x) = \prod_{k=1}^{n} \theta(x_k), \quad \theta(t) = \begin{cases} 1, & t \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

Approximate total variation (aTV) prior

$$\pi_{\text{pri}}(x) \propto \exp\left(-\alpha \sum_{i=1}^{n} \sum_{j \in \mathcal{N}_i} |x_i - x_j|_\beta\right)$$

where

$$|t|_\beta = \sqrt{t^2 + \beta}, \quad \beta > 0$$

Likelihood model for x-ray imaging

- Measurement model

$$m = A(x) + e, \quad e \sim \mathcal{N}(0, \Gamma_e), \quad L_e^{\mathrm{T}} L_e = \Gamma_e^{-1}$$

where the noise is assumed independent of the unknown

→ likelihood model:

$$\pi(m|x) \propto \exp\left(-\frac{1}{2}\|L_e(m - Ax)\|_2^2\right)$$

For justification of the Gaussian noise model, see Siltanen et al
"Statistical inversion for medical x-ray tomography with few radiograph:
1. General theory, Phys. Med. Biol. 48: 1437-1463 (2003)

Posterior model;

$$\pi(x|m) \propto \pi_+(x) \exp\left(-\frac{1}{2}\|L_e(m - Ax)\|_2^2 - \alpha \sum_{i=1}^{n} \sum_{j \in \mathcal{N}_i} |x_i - x_j|_\beta \right)$$

Computation of the MAP estimate;

$$x_{\text{MAP}} = \arg\min_{x \geq 0} \left\{ \frac{1}{2}\|L_e(m - Ax)\|_2^2 + \alpha \sum_{i=1}^{n} \sum_{j \in \mathcal{N}_i} |x_i - x_j|_\beta \right\}$$

- Number of data ~ 1 million, number of unknowns ~ 7 millions.
- To obtain solution in clinically acceptable time, we
  - solve the problem with gradient based optimization methods. Positivity constraint implemented via an exterior point method
  - Efficient implementation via parallelization / GPU computing

In exterior point method, the constrained optimization problem is replaced with a sequence of unconstrained problems that use asymptotic penalty ("barrier") for the positivity;

$$x_{\mathrm{MAP}}^{(t)} = \arg\min_{x} \underbrace{\left\{ \frac{1}{2}\|L_e(m - Ax)\|_2^2 + \alpha \sum_{i=1}^{n} \sum_{j \in \mathcal{N}_i} |x_i - x_j|_\beta + \Psi^{(t)}(x) \right\}}_{F^{(t)}}$$

where

$$\Psi^{(t)}(x) = \sum_{k=1}^{n} \phi^{(t)}(x_j), \qquad \phi^{(t)}(x_j) = \begin{cases} c_t x_j^2 & , x_j < 0 \\ 0 & , x_j \geq 0 \end{cases}$$

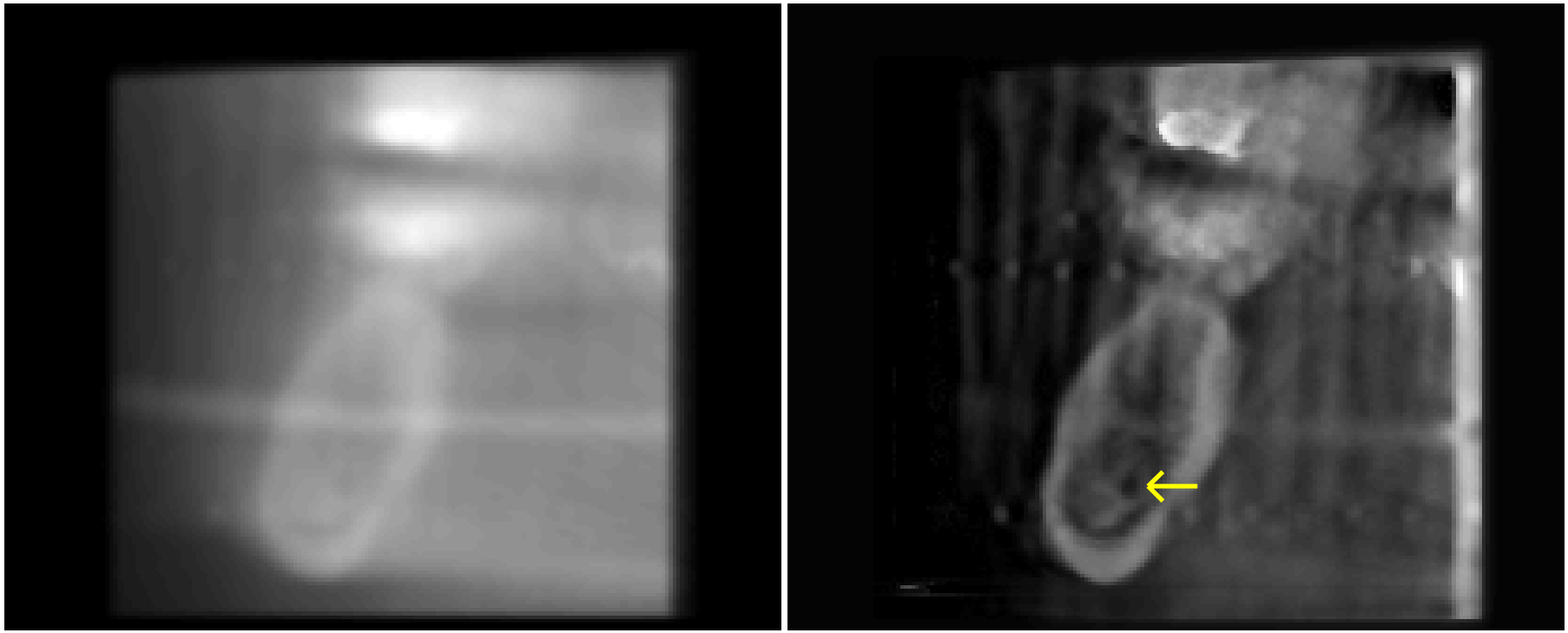
is the "barrier" function and $\{c_t, t = 1, 2, \ldots, n_s\}$ is sequence of increasing parameters.

Solution by the Barzilai-Borwein optimization method;

$$x^{(\ell+1)} = x^{(\ell)} - a_\ell^{-1} \nabla F^{(t)}(x^{(\ell)})$$

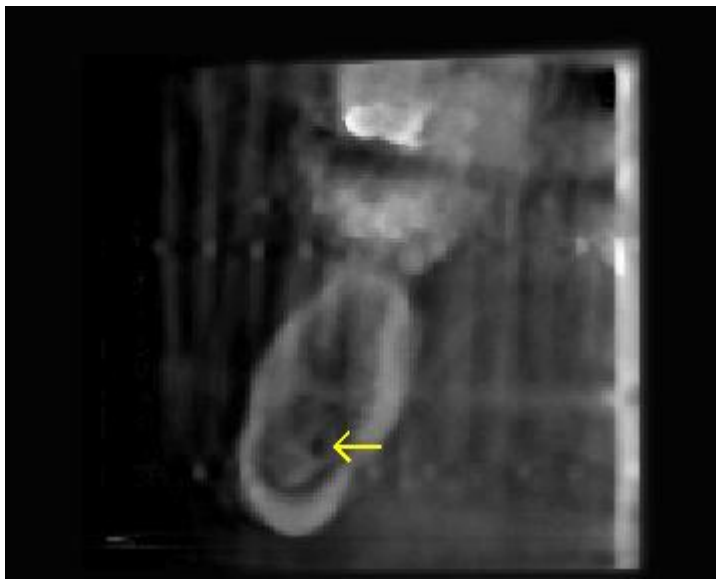
$$a_\ell = \frac{(x^{(\ell)} - x^{(\ell-1)})^{\mathrm{T}} (\nabla F^{(t)}(x^{(\ell)}) - \nabla F^{(t)}(x^{(\ell-1)}))}{(x^{(\ell)} - x^{(\ell-1)})^{\mathrm{T}} (x^{(\ell)} - x^{(\ell-1)})}$$

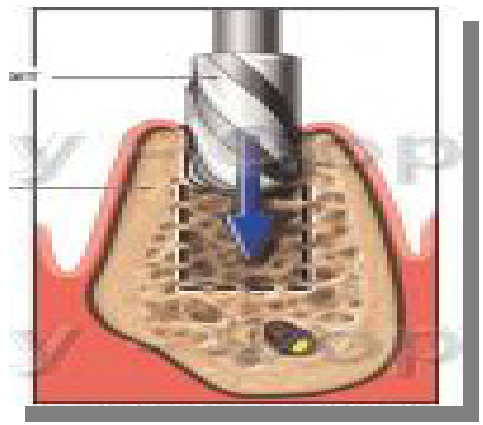
- Reconstruction from the limited angle data of the mandibular area (10 projections from 40 degrees opening angle).



Reconstruction from the limited angle data of the mandibular area. **Left:** Tomosynthetic reconstruction (current standard in dental imaging). **Right:** MAP estimate. Yellow arrow indicates the location of the mandibular nerve canal. The measurements for implantation are reliably available in the MAP estimate.

The 3D limited angle modality for implantology has been implemented as an add-on “volumetric tomography (VT)” product to the panoramic x-ray equipment by PaloDEX group (http://www.instrumentariumdental.com/).





The system has:

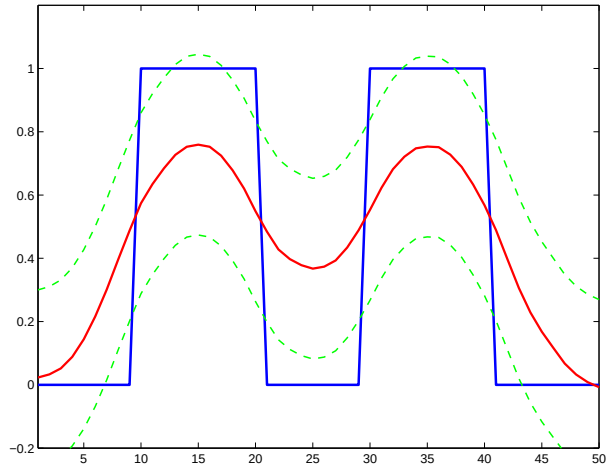
- ✓ Low cost
- ✓ Low dose
- ✓ Well suited to the dental workflow

About interpretation of the results ...

- Consider the following “counter example”; Let m,A, true solution x and e as in the 1D deconvolution example (example 2).
- Assume that our prior model would be Gaussian white noise prior, leading to posterior

$$\pi(x \mid m) \propto \exp\left(-\frac{1}{2}(m - Ax)^{\mathrm{T}}\Gamma_e^{-1}(m - Ax) - \frac{1}{2}x^{\mathrm{T}}\Gamma_x^{-1}x\right)$$

where $\Gamma_e = I$ and $\Gamma_x = \alpha^{-1}I$. Figure shows the CM estimate with 99% posterior confidence limits.



$x_{\mathrm{CM}} \pm 3\sigma_{x|m}$ corresponding to $\sim 99\%$ posterior confidence intervals

About interpretation of the results ...

- Here the true value is outside the 99% posterior confidence limits for most elements of x!
- Keep in mind that the model $\pi(x \mid m)$ reflects our information and uncertainty on x based on measured data and
  - Model of measurement process $\pi(m|x)$
  - Model of *a priori* information $\pi_{\mathrm{pri}}(x)$

  The reliability of the estimates and confidence intervals depend on both models; the confidence intervals can be misleading if the models are not carefully constructed respecting the measurement process and *a priori* information.