Bayesian Methods and Uncertainty Quantification for Linear Inverse Problems

John Bardsley, University of Montana

Collaborators:

S. Agapiou, M. Howard, A. Luttman, J. Nagy, A. Stuart

Technical University Denmark, December 2016

Outline

- Characteristics of inverse problems.
- Prior Modeling Using Gaussian Markov random fields.
- Hierarchical Models and MCMC
  - The Gibbs sampler and performance characteristics;
  - The partially collapsed Gibbs sampler;
  - The Marginal-then-Conditional sampler.
  - Gradient Scan Gibbs Sampler.

General Statistical Model

Consider the linear, Gaussian statistical model

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\epsilon},$$

where

- $\mathbf{y} \in \mathbb{R}^m$ is the vector of observations;
- $\mathbf{x} \in \mathbb{R}^n$ is the vector of unknown parameters;
- $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^m$;
- $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \lambda^{-1}\mathbf{I})$, i.e., $\boldsymbol{\epsilon}$ is i.i.d. Gaussian with mean 0 and variance λ^{-1}.

Numerical Discretization of a Physical Model

For us, $\mathbf{y} = [y_1, \ldots, y_m]^T$, with

$$\begin{aligned} y_i &= y(s_i) \\ &= \int_\Omega a(s_i, s')x(s')ds' \quad \left(\stackrel{\text{def}}{=} [\mathcal{A}_m x]_i\right) \end{aligned}$$

Numerical Discretization of a Physical Model

For us, $\mathbf{y} = [y_1, \ldots, y_m]^T$, with

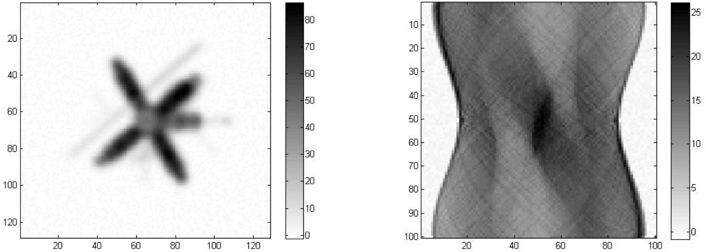
$$\begin{aligned} y_i &= y(s_i) \\ &= \int_\Omega a(s_i, s')x(s')ds' \quad \left(\stackrel{\text{def}}{=} [\mathcal{A}_m x]_i\right) \\ &\approx \frac{1}{\Delta s'} \sum_{j=1}^{n} a(s_i, s'_j)x(s'_j) \quad \text{(numerical quadrature)} \\ &= [\mathbf{A}\mathbf{x}]_i \quad \left([\mathbf{A}]_{ij} = \frac{1}{n} a(s_i, s'_j) \text{ and } \mathbf{x} = [x_1, \ldots, x_n]^T\right), \end{aligned}$$

where $\Omega = [0,1]$ or $[0,1] \times [0,1]$, defines the equation

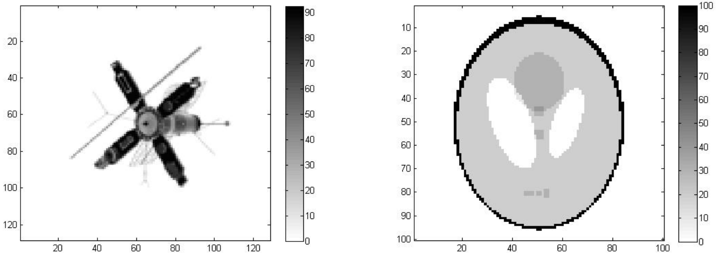
$$\mathbf{y} = \mathbf{A}\mathbf{x}.$$

Synthetic Examples

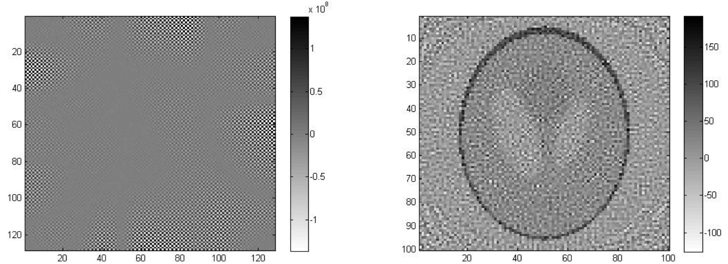
Data $\mathbf{y}$ examples:



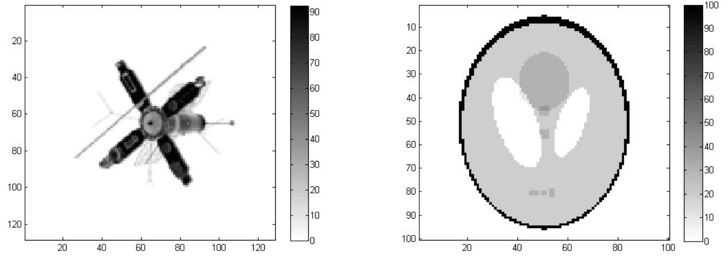
Corresponding true images $\mathbf{x}$:



Naive Solutions: $\mathbf{x}_{\mathrm{LS}} = \mathbf{A}^{\dagger}\mathbf{y}$



Corresponding true images $\mathbf{x}$:



Properties of the model matrix $\mathbf{A}$

It is typical in inverse problems that if the matrix $\mathbf{A}$ has SVD

$$\mathbf{A} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T, \quad \text{where} \quad r = \text{rank}(\mathbf{A}).$$

Characteristics of Inverse Problems:

- the σ_i's decay to 0 as $i \to r$;
- the $\{\mathbf{u}_i, \mathbf{v}_i\}$'s become increasingly oscillatory as $i \to n$.

Properties of the model matrix $\mathbf{A}$

It is typical in inverse problems that if the matrix $\mathbf{A}$ has SVD

$$\mathbf{A} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T, \quad \text{where} \quad r = \text{rank}(\mathbf{A}).$$

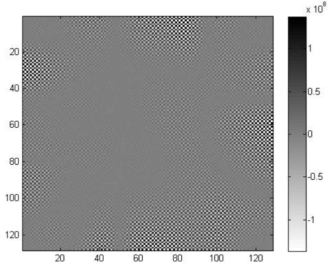
Characteristics of Inverse Problems:

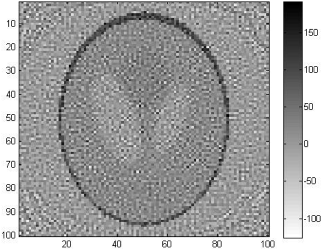
- the σ_i's decay to 0 as $i \to r$;
- the $\{\mathbf{u}_i, \mathbf{v}_i\}$'s become increasingly oscillatory as $i \to n$.

The least squares solution can then be written

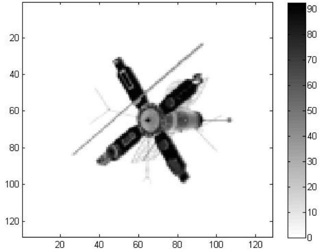
$$\begin{aligned} \mathbf{A}^\dagger \mathbf{y} &= \mathbf{A}^\dagger(\mathbf{A}\mathbf{x} + \boldsymbol{\epsilon}) \\ &= \underbrace{\sum_{i=1}^{r} (\mathbf{v}_i^T \mathbf{x}) \mathbf{v}_i}_{\text{portion due to signal}} + \underbrace{\sum_{i=1}^{r} \left(\frac{\mathbf{u}_i^T \boldsymbol{\epsilon}}{\sigma_i} \right) \mathbf{v}_i}_{\text{portion due to noise}} \end{aligned}$$

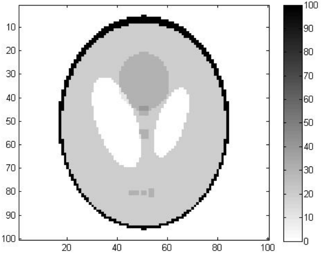
Naive Solutions: $\mathbf{x}_{\mathrm{naive}} = \mathbf{A}^{\dagger}\mathbf{y}$



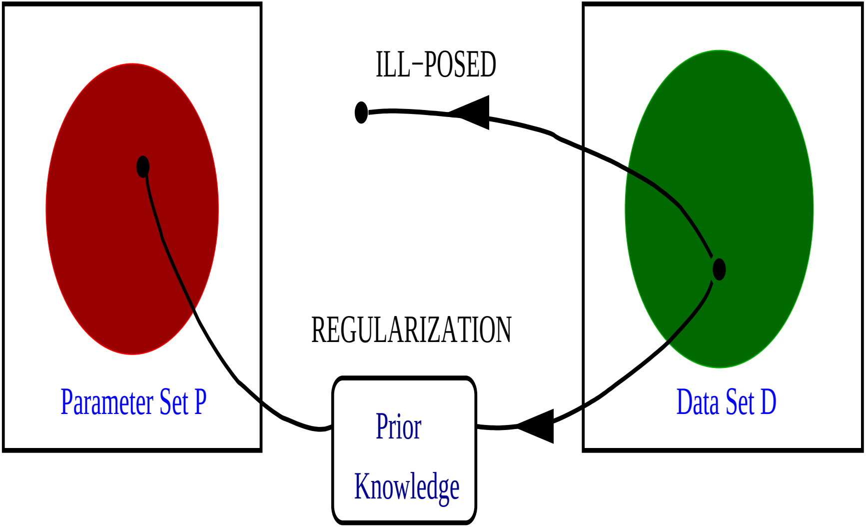


Corresponding true images $\mathbf{x}$:





The Fix: Regularization



Bayes Law:

$$\underbrace{p(\mathbf{x}|\mathbf{y},\lambda,\delta)}_{\text{posterior}} \propto \underbrace{p(\mathbf{y}|\mathbf{x},\lambda)}_{\text{likelihood}}\underbrace{p(\mathbf{x}|\delta)}_{\text{prior}}.$$

For our assumed statistical model,

$$p(\mathbf{y}|\mathbf{x},\lambda) \propto \exp\left(-\frac{\lambda}{2}\|\mathbf{A}\mathbf{x}-\mathbf{y}\|^2\right).$$

Bayes Law:

$$\underbrace{p(\mathbf{x}|\mathbf{y},\lambda,\delta)}_{\text{posterior}} \propto \underbrace{p(\mathbf{y}|\mathbf{x},\lambda)}_{\text{likelihood}}\underbrace{p(\mathbf{x}|\delta)}_{\text{prior}}.$$

For our assumed statistical model,

$$p(\mathbf{y}|\mathbf{x},\lambda) \propto \exp\left(-\frac{\lambda}{2}\|\mathbf{A}\mathbf{x}-\mathbf{y}\|^2\right).$$

In this talk, we will assume a Gaussian prior

$$p(\mathbf{x}|\delta) \propto \exp\left(-\frac{\delta}{2}\mathbf{x}^T\mathbf{L}\mathbf{x}\right),$$

so that

$$p(\mathbf{x}|\mathbf{y},\lambda,\delta) \propto \exp\left(-\frac{\lambda}{2}\|\mathbf{A}\mathbf{x}-\mathbf{y}\|^2 - \frac{\delta}{2}\,\mathbf{x}^T\mathbf{L}\mathbf{x}\right).$$

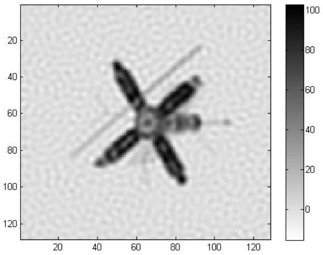
Maximum a Posteriori (MAP) Estimation

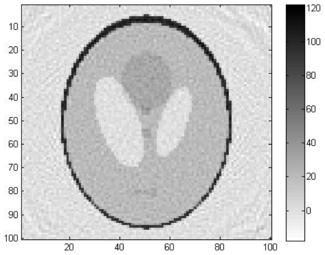
The maximizer of the posterior density is

$$\mathbf{x}_{\mathrm{MAP}} = \arg\min_{\mathbf{x}} \left\{ \frac{\lambda}{2} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2 + \frac{\delta}{2}\, \mathbf{x}^T \mathbf{L} \mathbf{x} \right\}$$

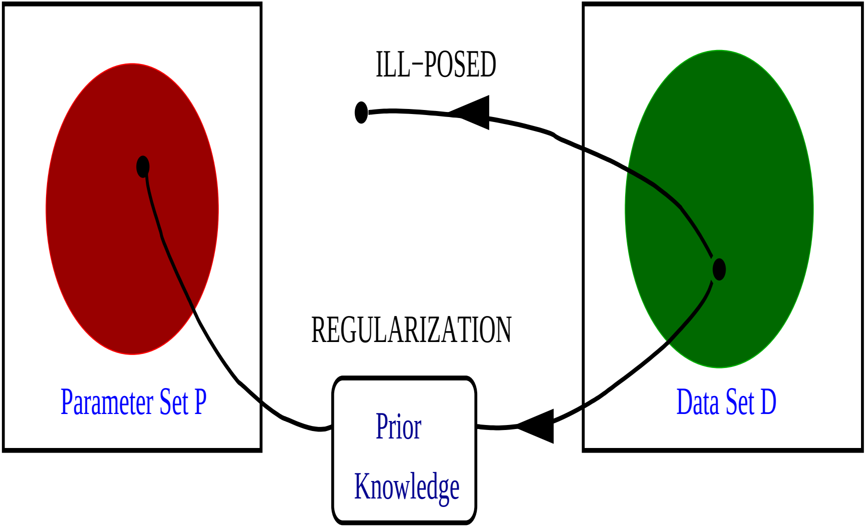
which is the regularized solution $\mathbf{x}_\alpha$ with $\alpha = \delta/\lambda$.

$$\alpha = 2.5 \times 10^{-4} \qquad \alpha = 1.05 \times 10^{-4}.$$





Modeling the Prior $p(\mathbf{x}|\delta)$



Gaussian Markov Random field (GMRF) priors

The neighbor values for x_{ij} are below (in black)

$$\begin{aligned} \mathbf{x}_{\partial_{ij}} &= \{x_{i-1,j}, x_{i,j-1}, x_{i+1,j}, x_{i,j+1}\} \\ &= \begin{bmatrix} & x_{i,j+1} & \\ x_{i-1,j} & x_{ij} & x_{i+1,j} \\ & x_{i,j-1} & \end{bmatrix}. \end{aligned}$$

Gaussian Markov Random field (GMRF) priors

The neighbor values for x_{ij} are below (in black)

$$\begin{aligned}
\mathbf{x}_{\partial_{ij}} &= \{x_{i-1,j}, x_{i,j-1}, x_{i+1,j}, x_{i,j+1}\} \\
&= \begin{bmatrix} & x_{i,j+1} & \\ x_{i-1,j} & x_{ij} & x_{i+1,j} \\ & x_{i,j-1} & \end{bmatrix}.
\end{aligned}$$

Then we assume

$$x_{ij}|\mathbf{x}_{\partial_{ij}} \sim \mathcal{N}\left(\bar{x}_{\partial_{ij}}, (\delta n_{ij})^{-1}\right),$$

where $\bar{x}_{\partial_{ij}} = \frac{1}{n_{ij}} \sum_{(r,s)\in\partial_{ij}} x_{rs}$ and $n_{ij} = |\partial_{ij}|$.

Gaussian Markov Random field (GMRF) priors

This leads to the prior

$$p(\mathbf{x}|\delta) \propto \delta^n \exp\left(-\frac{\delta}{2}\mathbf{x}^T\mathbf{L}\mathbf{x}\right),$$

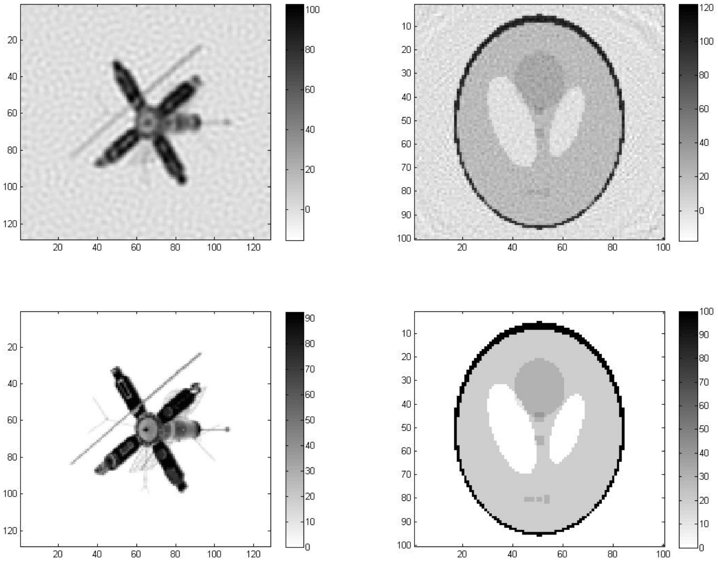
where if $r = (i, j)$ after column-stacking 2D arrays

$$[\mathbf{L}]_{rs} = \begin{cases} 4 & s = r, \\ -1 & s \in \partial_r, \\ 0 & \text{otherwise.} \end{cases}$$

NOTE: $\mathbf{L}$ = 2D discrete unscaled neg-Laplacian. Recall the MAP estimator

$$\mathbf{x}_\alpha = \arg\min_{\mathbf{x}} \left\{ \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2 + \frac{\alpha}{2}\,\mathbf{x}^T\mathbf{L}\mathbf{x} \right\}$$

$\alpha = 2.5 \times 10^{-4}$ $\alpha = 1.05 \times 10^{-4}$



2D GMRF Increment Models

For a 2D signal, suppose

$$
\begin{aligned}
x_{i+1,j} - x_{ij} &\sim \mathcal{N}(0, (w_{ij}^h \delta)^{-1}), \quad i, j = 1, \ldots, \sqrt{n} \\
x_{i,j+1} - x_{ij} &\sim \mathcal{N}(0, (w_{ij}^v \delta)^{-1}), \quad i, j = 1, \ldots, \sqrt{n}.
\end{aligned}
$$

2D GMRF Increment Models

For a 2D signal, suppose

$$\begin{aligned} x_{i+1,j} - x_{ij} &\sim \mathcal{N}(0, (w_{ij}^h \delta)^{-1}), \quad i,j = 1, \ldots, \sqrt{n} \\ x_{i,j+1} - x_{ij} &\sim \mathcal{N}(0, (w_{ij}^v \delta)^{-1}), \quad i,j = 1, \ldots, \sqrt{n}. \end{aligned}$$

Assuming independence the density function for $\mathbf{x}$ has the form

$$\begin{aligned} p(\mathbf{x}|\delta) &\propto \delta^{n/2} \exp\left(-\frac{\delta}{2} \sum_{i,j=1}^{\sqrt{n}} w_{ij}^h (x_{i+1,j} - x_{ij})^2 \right) \times \\ &\qquad \exp\left(-\frac{\delta}{2} \sum_{i,j=1}^{\sqrt{n}} w_{ij}^v (x_{i,j+1} - x_{ij})^2 \right) \\ &= \delta^{n/2} \exp\left(-\frac{\delta}{2} \mathbf{x}^T (\mathbf{D}_h^T \boldsymbol{\Lambda}_h \mathbf{D}_h + \mathbf{D}_v^T \boldsymbol{\Lambda}_v \mathbf{D}_v) \mathbf{x} \right), \end{aligned}$$

2D GMRF Increment Models

For a 2D signal, suppose

$$\begin{aligned} x_{i+1,j} - x_{ij} &\sim \mathcal{N}(0, (w_{ij}^h \delta)^{-1}), \quad i,j = 1, \ldots, \sqrt{n} \\ x_{i,j+1} - x_{ij} &\sim \mathcal{N}(0, (w_{ij}^v \delta)^{-1}), \quad i,j = 1, \ldots, \sqrt{n}. \end{aligned}$$

Assuming independence the density function for $\mathbf{x}$ has the form

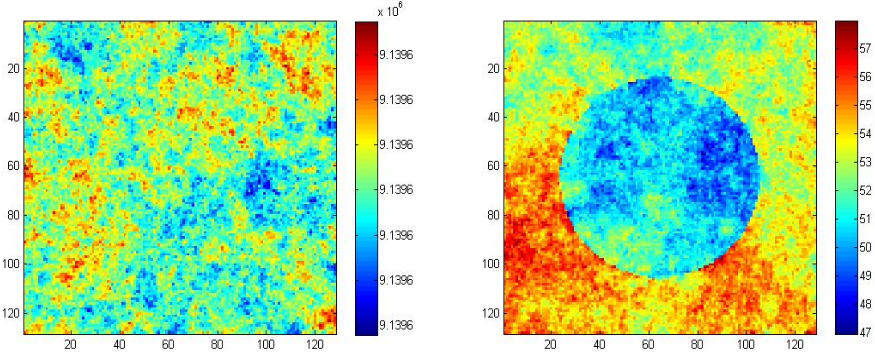
$$\begin{aligned} p(\mathbf{x}|\delta) &\propto \delta^{n/2} \exp\left(-\frac{\delta}{2} \sum_{i,j=1}^{\sqrt{n}} w_{ij}^h (x_{i+1,j} - x_{ij})^2\right) \times \\ &\qquad \exp\left(-\frac{\delta}{2} \sum_{i,j=1}^{\sqrt{n}} w_{ij}^v (x_{i,j+1} - x_{ij})^2\right) \\ &= \delta^{n/2} \exp\left(-\frac{\delta}{2} \mathbf{x}^T (\mathbf{D}_h^T \mathbf{\Lambda}_h \mathbf{D}_h + \mathbf{D}_v^T \mathbf{\Lambda}_v \mathbf{D}_v) \mathbf{x}\right), \end{aligned}$$

- $\mathbf{D}_h = \mathbf{I} \otimes \mathbf{D}$, $\mathbf{D}_v = \mathbf{D} \otimes \mathbf{I}$, where $\mathbf{D}$ = 1D difference matrix;
- $\mathbf{\Lambda}_h = \text{diag}(\text{vec}(\{w_{ij}^h\}_{ij=1}^{\sqrt{n}}))$, $\mathbf{\Lambda}_v = \text{diag}(\text{vec}(\{w_{ij}^v\}_{ij=1}^{\sqrt{n}}))$.

2D GMRF Increment Models

The matrix $\frac{1}{\Delta s^2}\mathbf{D}_h^T\mathbf{\Lambda}_h\mathbf{D}_h + \frac{1}{\Delta t^2}\mathbf{D}_v^T\mathbf{\Lambda}_v\mathbf{D}_v$ is a discretization of

$$-\frac{\partial}{\partial s}\left(w_h(s,t)\frac{\partial}{\partial s}\right) - \frac{\partial}{\partial t}\left(w_v(s,t)\frac{\partial}{\partial t}\right)$$



Left: $w^h_{ij} = w^v_{ij} = 1$ for all ij.

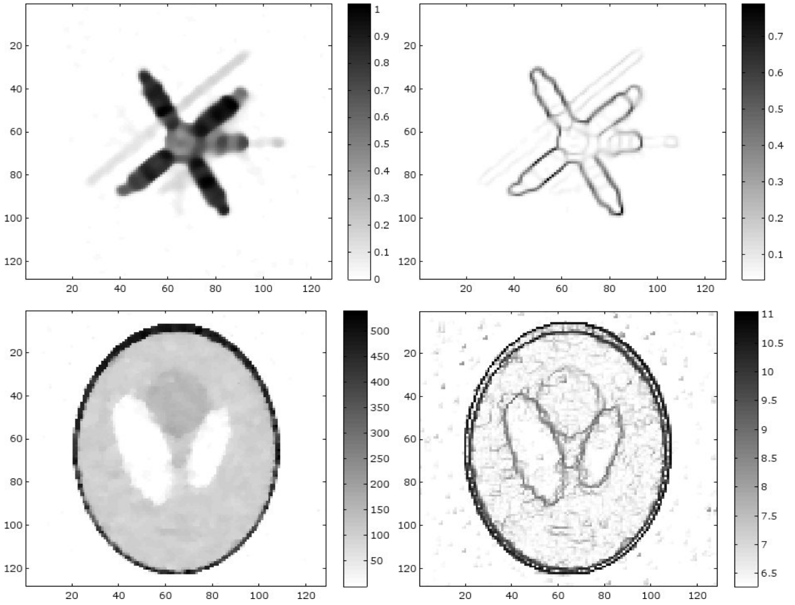
Right: $w^h_{ij} = w^v_{ij} = 0.01$ for ij on the circle boundary.

GMRF Edge-Preserving Reconstruction

0. Set $\boldsymbol{\Lambda} = \mathbf{I}$.
1. Define $\mathbf{L} = \mathbf{D}_h^T \boldsymbol{\Lambda} \mathbf{D}_h + \mathbf{D}_v^T \boldsymbol{\Lambda} \mathbf{D}_v$, where
2. Compute
$$\mathbf{x}_\alpha = (\mathbf{A}^T \mathbf{A} + \alpha \mathbf{L})^{-1} \mathbf{A}^T \mathbf{y}$$
using PCG with <u>α obtained via L-curve, GCV, etc.</u>
3. Set
$$\boldsymbol{\Lambda}(\mathbf{x}_\alpha) = \text{diag}\left(\frac{\mathbf{1}}{\sqrt{(\mathbf{D}_h \mathbf{x}_\alpha)^2 + (\mathbf{D}_v \mathbf{x}_\alpha)^2 + \beta \mathbf{1}}} \right),$$
$0 < \beta \ll 1$, and return to Step 1.

NOTE: This is just the lagged-diffusivity iteration.

Numerical Results



Infinite Dimnensional Limit

Question: When is

$$p(x|\mathbf{y},\lambda,\delta) \stackrel{\text{def}}{=} \lim_{n\to\infty} p(\mathbf{x}|\mathbf{y},\lambda,\delta)$$

well defined?

Infinite Dimnensional Limit

Question: When is

$$p(x|\mathbf{y},\lambda,\delta) \stackrel{\text{def}}{=} \lim_{n\to\infty} p(\mathbf{x}|\mathbf{y},\lambda,\delta)$$

well defined? First

$$\lim_{n\to\infty} \|\mathbf{A}\mathbf{x}-\mathbf{y}\|^2 = \|\mathcal{A}_m x - \mathbf{y}\|^2,$$

where

$$[\mathcal{A}_m x]_i = \int_\Omega a(s_i, s')x(s')ds', \quad i = 1, \dots, m.$$

Note: $\mathcal{A}_m : C^\infty(\Omega) \to \mathbb{R}^m$, where $\Omega = [0,1]$ or $[0,1]\times[0,1]$, and $C^\infty(\Omega)$ is the space of smooth functions on Ω.

The Infinite Dimnensional Limit

Next,

$$\lim_{n\to\infty} c(n)\mathbf{x}^T\mathbf{L}\mathbf{x} = \langle x, \mathcal{L}x\rangle \stackrel{\text{def}}{=} \int_\Omega x(s')\mathcal{L}x(s')\, ds',$$

The Infinite Dimnensional Limit

Next,

$$\lim_{n\to\infty} c(n)\mathbf{x}^T\mathbf{L}\mathbf{x} = \langle x, \mathcal{L}x\rangle \stackrel{\text{def}}{=} \int_\Omega x(s')\mathcal{L}x(s')\,ds',$$

where $c(n) = n$ in one-dimension and

$$\mathcal{L} = -\frac{d}{ds}\left(w(s)\frac{d}{ds}\right), \quad 0 < s < 1;$$

whereas $c(n) = 1$ in two-dimensions and

$$\mathcal{L} = -\frac{\partial}{\partial s}\left(w_s(s,t)\frac{\partial}{\partial s}\right) - \frac{\partial}{\partial t}\left(w_t(s,t)\frac{\partial}{\partial t}\right), \quad 0 < s, t < 1.$$

The Infinite Dimnensional Limit

$$\lim_{n\to\infty}\left\{\frac{\lambda}{2}\|\mathbf{A}\mathbf{x}-\mathbf{y}\|^2+\frac{\delta c(n)}{2}\mathbf{x}^T\mathbf{L}\mathbf{x}\right\}=\frac{\lambda}{2}\|\mathcal{A}_m x-\mathbf{y}\|^2+\frac{\delta}{2}\langle x,\mathcal{L}x\rangle,$$

and hence

$$p(x|\mathbf{y},\lambda,\delta) \quad \propto \quad \exp\left(-\frac{\lambda}{2}\|\mathcal{A}_m x-\mathbf{y}\|^2-\frac{\delta}{2}\langle x,\mathcal{L}x\rangle\right).$$

The Infinite Dimnensional Limit

$$\lim_{n\to\infty}\left\{\frac{\lambda}{2}\|\mathbf{A}\mathbf{x}-\mathbf{y}\|^2+\frac{\delta c(n)}{2}\mathbf{x}^T\mathbf{L}\mathbf{x}\right\}=\frac{\lambda}{2}\|\mathcal{A}_m x-\mathbf{y}\|^2+\frac{\delta}{2}\langle x,\mathcal{L}x\rangle,$$

and hence

$$p(x|\mathbf{y},\lambda,\delta) \quad \propto \quad \exp\left(-\frac{\lambda}{2}\|\mathcal{A}_m x-\mathbf{y}\|^2-\frac{\delta}{2}\langle x,\mathcal{L}x\rangle\right).$$

Question: When is the resulting probability measure well-defined:

$$\mu^{\text{post}}(A)\stackrel{\text{def}}{=}\int_A p(x|\mathbf{y},\lambda,\delta)\,dx?$$

The Infinite Dimnensional Limit

$$\lim_{n\to\infty}\left\{\frac{\lambda}{2}\|\mathbf{A}\mathbf{x}-\mathbf{y}\|^2+\frac{\delta c(n)}{2}\mathbf{x}^T\mathbf{L}\mathbf{x}\right\}=\frac{\lambda}{2}\|\mathcal{A}_m x-\mathbf{y}\|^2+\frac{\delta}{2}\langle x,\mathcal{L}x\rangle,$$

and hence

$$p(x|\mathbf{y},\lambda,\delta) \quad\propto\quad \exp\left(-\frac{\lambda}{2}\|\mathcal{A}_m x-\mathbf{y}\|^2-\frac{\delta}{2}\langle x,\mathcal{L}x\rangle\right).$$

Question: When is the resulting probability measure well-defined:

$$\mu^{\text{post}}(A)\stackrel{\text{def}}{=}\int_A p(x|\mathbf{y},\lambda,\delta)\,dx?$$

Answer [Stuart]: When $\mathcal{L}^{-1}$ is a trace-class operator on $L^2(\Omega)$, i.e., when $\sum_{i=1}^{\infty}\langle\phi_i,\mathcal{L}^{-1}\phi_i\rangle<\infty$ for any o.n. basis $\{\phi_i\}$ of $L^2(\Omega)$.

The Infinite Dimnensional Limit

In one-dimension, if

$$\mathcal{L} = -\frac{d}{ds}\left(w(s)\frac{d}{ds}\right), \quad 0 < s < 1,$$

then $\mathcal{L}^{-1}$ is trace class.

The Infinite Dimnensional Limit

In one-dimension, if

$$\mathcal{L} = -\frac{d}{ds}\left(w(s)\frac{d}{ds}\right), \quad 0 < s < 1,$$

then $\mathcal{L}^{-1}$ is trace class.

In two-dimensions, if

$$\mathcal{L} = -\frac{\partial}{\partial s}\left(w_s(s,t)\frac{\partial}{\partial s}\right) - \frac{\partial}{\partial t}\left(w_t(s,t)\frac{\partial}{\partial t}\right), \quad 0 < s, t < 1,$$

then $\underline{\mathcal{L}^{-1} \text{ is } \textit{not} \text{ trace-class}}$.

The Infinite Dimnensional Limit

In one-dimension, if

$$\mathcal{L} = -\frac{d}{ds}\left(w(s)\frac{d}{ds}\right), \quad 0 < s < 1,$$

then $\mathcal{L}^{-1}$ is trace class.

In two-dimensions, if

$$\mathcal{L} = -\frac{\partial}{\partial s}\left(w_s(s,t)\frac{\partial}{\partial s}\right) - \frac{\partial}{\partial t}\left(w_t(s,t)\frac{\partial}{\partial t}\right), \quad 0 < s, t < 1,$$

then $\underline{\mathcal{L}^{-1} \text{ is } \textit{not} \text{ trace-class}}$.

FIX: in two-dimensions, use $\mathcal{L} \stackrel{\text{def}}{=} \mathcal{L}^2$, which is trace class; note that if $w_s = w_t = 1$ above, this is called the *biharmonic operator*.

An Alternative: Second-Order GMRFs

For a 2D signal, suppose

$$\begin{aligned} x_{i-1,j} - 2x_{ij} + x_{i+1,j} &\sim \mathcal{N}(0, (w_{ij}^h \delta)^{-1}), \quad i, j = 1, \ldots, \sqrt{n} \\ x_{i,j-1} - 2x_{ij} + x_{i,j+1} &\sim \mathcal{N}(0, (w_{ij}^v \delta)^{-1}), \quad i, j = 1, \ldots, \sqrt{n}. \end{aligned}$$

An Alternative: Second-Order GMRFs

For a 2D signal, suppose

$$\begin{aligned} x_{i-1,j} - 2x_{ij} + x_{i+1,j} &\sim \mathcal{N}(0, (w^h_{ij}\delta)^{-1}), \quad i,j = 1,\dots,\sqrt{n} \\ x_{i,j-1} - 2x_{ij} + x_{i,j+1} &\sim \mathcal{N}(0, (w^v_{ij}\delta)^{-1}), \quad i,j = 1,\dots,\sqrt{n}. \end{aligned}$$

Assuming independence, the density function for $\mathbf{x}$ has the form

$$\begin{aligned} p(\mathbf{x}|\delta) \quad \propto \quad & \delta^{n/2} \exp\left(-\frac{\delta}{2} \sum_{i,j=1}^{n} w^h_{ij}(x_{i-1,j} - 2x_{ij} + x_{i+1,j})^2 \right) \times \\ & \exp\left(-\frac{\delta}{2} \sum_{i,j=1}^{n} w^v_{ij}(x_{i,j-1} - 2x_{ij} + x_{i,j+1})^2 \right) \end{aligned}$$

An Alternative: Second-Order GMRFs

For a 2D signal, suppose

$$\begin{aligned} x_{i-1,j} - 2x_{ij} + x_{i+1,j} &\sim \mathcal{N}(0, (w_{ij}^h \delta)^{-1}), \quad i,j = 1, \ldots, \sqrt{n} \\ x_{i,j-1} - 2x_{ij} + x_{i,j+1} &\sim \mathcal{N}(0, (w_{ij}^v \delta)^{-1}), \quad i,j = 1, \ldots, \sqrt{n}. \end{aligned}$$

Assuming independence, the density function for $\mathbf{x}$ has the form

$$\begin{aligned} p(\mathbf{x}|\delta) &\propto \delta^{n/2} \exp\left(-\frac{\delta}{2} \sum_{i,j=1}^{n} w_{ij}^h (x_{i-1,j} - 2x_{ij} + x_{i+1,j})^2 \right) \times \\ &\qquad \exp\left(-\frac{\delta}{2} \sum_{i,j=1}^{n} w_{ij}^v (x_{i,j-1} - 2x_{ij} + x_{i,j+1})^2 \right) \\ &= \delta^{n/2} \exp\left(-\frac{\delta}{2} \mathbf{x}^T (\mathbf{L}_h^T \mathbf{\Lambda}_h \mathbf{L}_h + \mathbf{L}_v^T \mathbf{\Lambda}_v \mathbf{L}_v) \mathbf{x} \right), \end{aligned}$$

- $\mathbf{L}_v = \mathbf{L} \otimes \mathbf{I}$, $\mathbf{L}_h = \mathbf{I} \otimes \mathbf{L}$, $\mathbf{L}$ = 1D discrete neg-Laplacian;
- $\mathbf{\Lambda}_h = \text{diag}(\text{vec}(\{w_{ij}^h\}_{ij=1}^{\sqrt{n}}))$, $\mathbf{\Lambda}_v = \text{diag}(\text{vec}(\{w_{ij}^v\}_{ij=1}^{\sqrt{n}}))$.

The Infinite Dimensional Limit

Let $\mathbf{L} = \mathbf{L}_h^T \boldsymbol{\Lambda}_h \mathbf{L}_h + \mathbf{L}_v^T \boldsymbol{\Lambda}_v \mathbf{L}_v$, then

$$\lim_{n \to \infty} c(n) \mathbf{x}^T \mathbf{L} \mathbf{x} = \langle x, \mathcal{L} x \rangle,$$

where

$$\mathcal{L} = -\frac{\partial^2}{\partial s^2} \left(w_s(s,t) \frac{\partial^2}{\partial s^2} \right) - \frac{\partial^2}{\partial t^2} \left(w_t(s,t) \frac{\partial^2}{\partial t^2} \right), \quad 0 < s, t < 1,$$

The Infinite Dimensional Limit

Let $\mathbf{L} = \mathbf{L}_h^T \mathbf{\Lambda}_h \mathbf{L}_h + \mathbf{L}_v^T \mathbf{\Lambda}_v \mathbf{L}_v$, then

$$\lim_{n\to\infty} c(n)\mathbf{x}^T \mathbf{L}\mathbf{x} = \langle x, \mathcal{L}x \rangle,$$

where

$$\mathcal{L} = -\frac{\partial^2}{\partial s^2}\left(w_s(s,t)\frac{\partial^2}{\partial s^2}\right) - \frac{\partial^2}{\partial t^2}\left(w_t(s,t)\frac{\partial^2}{\partial t^2}\right), \quad 0 < s,t < 1,$$

NOTE: $\mathcal{L}^{-1}$ is trace-class, and the posterior density function

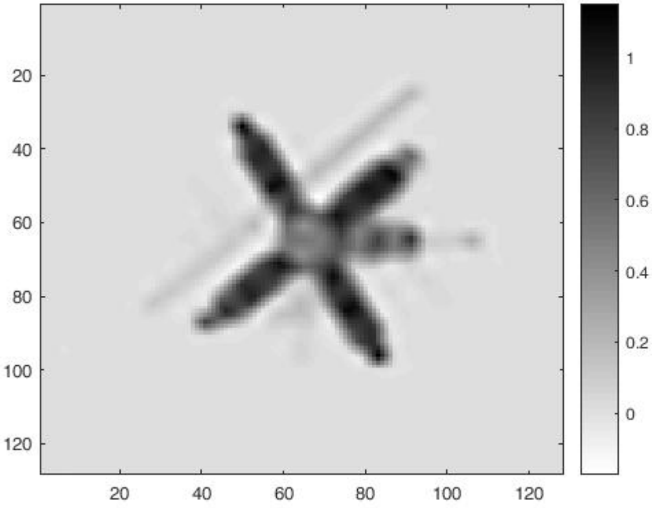
$$p(x|\mathbf{y},\lambda,\delta) \quad \propto \quad \exp\left(-\frac{\lambda}{2}\|\mathcal{A}_M x - \mathbf{y}\|^2 - \frac{\delta}{2}\langle x, \mathcal{L}x\rangle\right)$$

is a well-defined probability measure on $Ł^2(\Omega)$.

Higher-Order GMRF, Edge-Preserving Reconstruction

0. Set $\boldsymbol{\Lambda}_h = \boldsymbol{\Lambda}_v = \mathbf{I}$.
1. Define $\mathbf{L} = \mathbf{L}_h^T \boldsymbol{\Lambda}_h \mathbf{L}_h + \mathbf{L}_v^T \boldsymbol{\Lambda}_v \mathbf{L}_v$, where
2. Compute
$$\mathbf{x}_\alpha = (\mathbf{A}^T\mathbf{A} + \alpha\mathbf{L})^{-1}\mathbf{A}^T\mathbf{y}$$
using PCG with $\underline{\alpha \text{ obtained using GCV}}$.
3. Set
$$\begin{aligned}
\boldsymbol{\Lambda}_h(\mathbf{x}_\alpha) &= \operatorname{diag}\left(\frac{\mathbf{1}}{\sqrt{(\mathbf{L}_h\mathbf{x}_\alpha)^2 + \beta\mathbf{1}}}\right) \\
\boldsymbol{\Lambda}_v(\mathbf{x}_\alpha) &= \operatorname{diag}\left(\frac{\mathbf{1}}{\sqrt{(\mathbf{L}_v\mathbf{x}_\alpha)^2 + \beta\mathbf{1}}}\right)
\end{aligned}$$
$0 < \beta \ll 1$, and return to Step 1.

Plot after 10 iterations



Hierarchical Bayes: Assume Hyper-Priors on λ and δ

Uncertainty in λ and δ: $\lambda \sim p(\lambda)$ and $\delta \sim p(\delta)$. Then

$$p(\mathbf{x}, \lambda, \delta | \mathbf{y}) \propto p(\mathbf{y}|\mathbf{x}, \lambda) p(\lambda) \, p(\mathbf{x}|\delta) p(\delta),$$

is the Bayesian posterior

Hierarchical Bayes: Assume Hyper-Priors on λ and δ

Uncertainty in λ and δ: $\lambda \sim p(\lambda)$ and $\delta \sim p(\delta)$. Then

$$p(\mathbf{x}, \lambda, \delta|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x}, \lambda)p(\lambda)\, p(\mathbf{x}|\delta)p(\delta),$$

is the Bayesian posterior, where

$$p(\mathbf{y}|\mathbf{x}, \lambda) \propto \lambda^{m/2} \exp\left(-\frac{\lambda}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2\right),$$

and we choose a GMRF prior and Gamma hyper-priors:

$$\begin{aligned} p(\mathbf{x}|\delta) &\propto \delta^{n/2} \exp\left(-\frac{\delta}{2}\mathbf{x}^T\mathbf{L}\mathbf{x}\right), \\ p(\lambda) &\propto \lambda^{\alpha_\lambda - 1} \exp(-\beta_\lambda \lambda), \\ p(\delta) &\propto \delta^{\alpha_\delta - 1} \exp(-\beta_\delta \delta), \end{aligned}$$

where $\alpha_\lambda = \alpha_\delta = 1$ and $\beta_\lambda = \beta_\delta = 10^{-4}$, and hence

$$\text{mean} = \alpha/\beta = 10^4, \quad \text{var} = \alpha/\beta^2 = 10^8.$$

The Full Posterior Distribution

$$p(\mathbf{x}, \lambda, \delta | \mathbf{y}) \propto \text{the posterior}$$

$$\lambda^{m/2+\alpha_\lambda-1}\delta^{n/2+\alpha_\delta-1}\exp\left(-\frac{\lambda}{2}\|\mathbf{A}\mathbf{x}-\mathbf{y}\|^2 - \frac{\delta}{2}\mathbf{x}^T\mathbf{L}\mathbf{x} - \beta_\lambda\lambda - \beta_\delta\delta\right).$$

The Full Posterior Distribution

$$p(\mathbf{x}, \lambda, \delta | \mathbf{y}) \propto \text{the posterior}$$

$$\lambda^{m/2+\alpha_\lambda - 1} \delta^{n/2+\alpha_\delta - 1} \exp\left(-\frac{\lambda}{2} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2 - \frac{\delta}{2} \mathbf{x}^T \mathbf{L} \mathbf{x} - \beta_\lambda \lambda - \beta_\delta \delta \right).$$

By conjugacy, the full conditionals are in the same family as the prior/hyper-prior distribution:

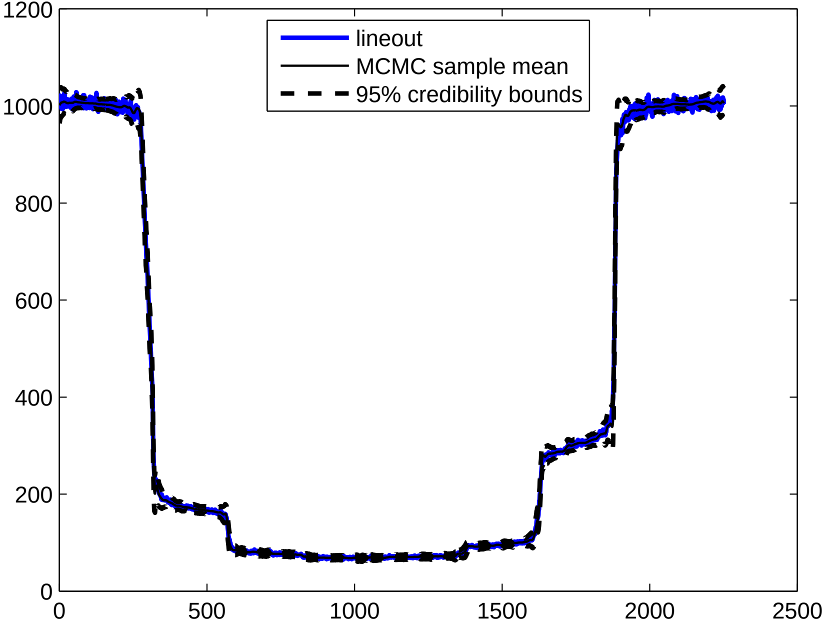
1. $p(\lambda | \mathbf{x}, \delta, \mathbf{y}) \propto \lambda^{m/2+\alpha_\lambda - 1} \exp\left(-\left(\frac{1}{2} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2 + \beta_\lambda \right) \lambda \right)$;
2. $p(\delta | \mathbf{x}, \lambda, \mathbf{y}) \propto \delta^{n/2+\alpha_\delta - 1} \exp\left(-\left(\frac{1}{2} \mathbf{x}^T \mathbf{L} \mathbf{x} + \beta_\delta \right) \delta \right)$;
3. $p(\mathbf{x} | \lambda, \delta, \mathbf{y}) \propto \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T (\lambda \mathbf{A}^T \mathbf{A} + \delta \mathbf{L}) (\mathbf{x} - \boldsymbol{\mu}) \right)$,
   where $\boldsymbol{\mu} = (\lambda \mathbf{A}^T \mathbf{A} + \delta \mathbf{L})^{-1} \lambda \mathbf{A}^T \mathbf{y}$.

The Gibbs sampler for sampling from $p(\mathbf{x}, \lambda, \delta | \mathbf{y})$

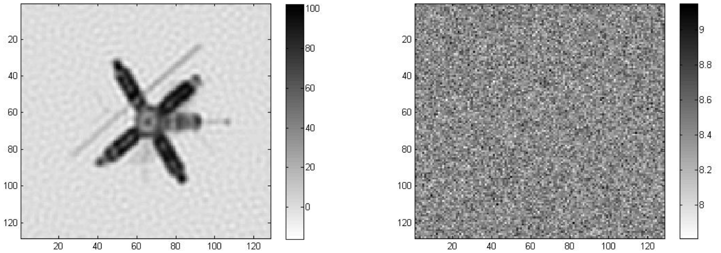
0. Choose $\mathbf{x}^0$, and set $k = 0$;
1. Compute $\lambda_{k+1} \sim \Gamma\left(m/2 + \alpha_\lambda, \frac{1}{2}\|\mathbf{A}\mathbf{x}^k - \mathbf{y}\|^2 + \beta_\lambda\right)$;
2. Compute $\delta_{k+1} \sim \Gamma\left(n/2 + \alpha_\delta, \frac{1}{2}(\mathbf{x}^k)^T \mathbf{L}\mathbf{x}^k + \beta_\delta\right)$;
3. Compute $\mathbf{x}^{k+1} \sim$ $\mathcal{N}\left((\lambda_{k+1}\mathbf{A}^T\mathbf{A} + \delta_{k+1}\mathbf{L})^{-1}\lambda_{k+1}\mathbf{A}^T\mathbf{y}, (\lambda_{k+1}\mathbf{A}^T\mathbf{A} + \delta_{k+1}\mathbf{L})^{-1}\right)$;
4. Set $k = k + 1$ and return to Step 1.

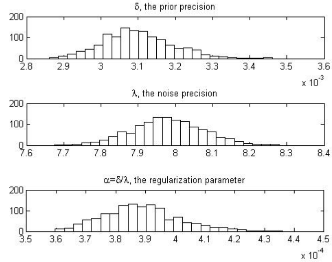
NOTE: the Markov chain $\{(\mathbf{x}_k, \lambda_k, \delta_k)\}$ generated by this Gibbs sampler is guaranteed to converge in distribution to the posetior density function $p(\mathbf{x}, \lambda, \delta | \mathbf{y})$.

An example from X-ray Radiography (w/ Luttman)

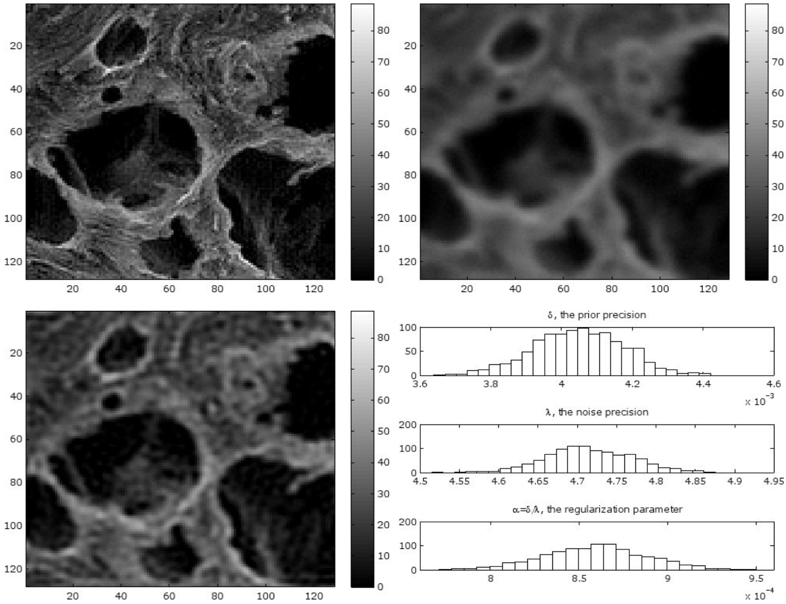


Deblurring with periodic boundary conditions





Deblurring with Neumann boundary conditions (w/ Nagy)



Computational bottleneck: Step 3. Compute $\mathbf{x}^k \sim N\left((\lambda_k \mathbf{A}^T\mathbf{A} + \delta_k \mathbf{L})^{-1}\lambda_k \mathbf{A}^T\mathbf{y}, (\lambda_k \mathbf{A}^T\mathbf{A} + \delta_k \mathbf{L})^{-1}\right)$

The conditional density for $\mathbf{x}|\mathbf{y}, \lambda, \delta$, dropping k for simplicity, is

$$p(\mathbf{x}|\lambda, \delta, \mathbf{y}) \quad \propto \quad \exp\left(-\frac{1}{2}\left\|\begin{bmatrix} \lambda^{1/2}\mathbf{A} \\ (\delta\mathbf{L})^{1/2} \end{bmatrix}\mathbf{x} - \begin{bmatrix} \lambda^{1/2}\mathbf{y} \\ \mathbf{0} \end{bmatrix}\right\|^2\right).$$

Computational bottleneck: Step 3. Compute $\mathbf{x}^k \sim N\left((\lambda_k \mathbf{A}^T\mathbf{A} + \delta_k \mathbf{L})^{-1}\lambda_k \mathbf{A}^T\mathbf{y}, (\lambda_k \mathbf{A}^T\mathbf{A} + \delta_k \mathbf{L})^{-1}\right)$

The conditional density for $\mathbf{x}|\mathbf{y}, \lambda, \delta$, dropping k for simplicity, is

$$p(\mathbf{x}|\lambda, \delta, \mathbf{y}) \quad \propto \quad \exp\left(-\frac{1}{2}\left\|\begin{bmatrix} \lambda^{1/2}\mathbf{A} \\ (\delta\mathbf{L})^{1/2} \end{bmatrix}\mathbf{x} - \begin{bmatrix} \lambda^{1/2}\mathbf{y} \\ \mathbf{0} \end{bmatrix}\right\|^2\right).$$

From here on out, we define:

$$\bar{\mathbf{A}} = \begin{bmatrix} \lambda^{1/2}\mathbf{A} \\ (\delta\mathbf{L})^{1/2} \end{bmatrix}, \quad \text{and} \quad \bar{\mathbf{y}} = \begin{bmatrix} \lambda^{1/2}\mathbf{y} \\ \mathbf{0} \end{bmatrix}.$$

Note then that

$$\frac{1}{2}\|\bar{\mathbf{A}}\mathbf{x} - \bar{\mathbf{y}}\|^2 = \frac{\lambda}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2 + \frac{\delta}{2}\mathbf{x}^T\mathbf{L}\mathbf{x}.$$

Computational bottleneck: Step 3. Compute $\mathbf{x} \sim N\left((\lambda \mathbf{A}^T\mathbf{A} + \delta \mathbf{L})^{-1}\lambda \mathbf{A}^T\mathbf{y}, (\lambda \mathbf{A}^T\mathbf{A} + \delta \mathbf{L})^{-1}\right)$

For large-scale problems, you can use optimization:

$$\mathbf{x} = \arg\min_{\boldsymbol{\psi}} \|\bar{\mathbf{A}}\boldsymbol{\psi} - (\bar{\mathbf{y}} + \boldsymbol{\epsilon})\|^2, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

Computational bottleneck: Step 3. Compute $\mathbf{x} \sim N\left((\lambda \mathbf{A}^T\mathbf{A} + \delta \mathbf{L})^{-1}\lambda \mathbf{A}^T\mathbf{y}, (\lambda \mathbf{A}^T\mathbf{A} + \delta \mathbf{L})^{-1}\right)$

For large-scale problems, you can use optimization:

$$\mathbf{x} = \arg\min_{\boldsymbol{\psi}} \|\bar{\mathbf{A}}\boldsymbol{\psi} - (\bar{\mathbf{y}} + \boldsymbol{\epsilon})\|^2, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

Note $\mathbf{x}$ is a random variable defined by

$$\mathbf{x} = (\bar{\mathbf{A}}^T\bar{\mathbf{A}})^{-1}\bar{\mathbf{A}}^T(\bar{\mathbf{y}} + \boldsymbol{\epsilon}), \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

Computational bottleneck: Step 3. Compute $\mathbf{x} \sim N\left((\lambda \mathbf{A}^T\mathbf{A} + \delta \mathbf{L})^{-1}\lambda \mathbf{A}^T\mathbf{y}, (\lambda \mathbf{A}^T\mathbf{A} + \delta \mathbf{L})^{-1}\right)$

For large-scale problems, you can use optimization:

$$\mathbf{x} = \arg\min_{\boldsymbol{\psi}} \|\bar{\mathbf{A}}\boldsymbol{\psi} - (\bar{\mathbf{y}} + \boldsymbol{\epsilon})\|^2, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

Note $\mathbf{x}$ is a random variable defined by

$$\mathbf{x} = (\bar{\mathbf{A}}^T\bar{\mathbf{A}})^{-1}\bar{\mathbf{A}}^T(\bar{\mathbf{y}} + \boldsymbol{\epsilon}), \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

QR-rewrite: if $\bar{\mathbf{A}} = \mathbf{QR}$, with $\mathbf{Q} \in \mathbb{R}^{m\times n}$, $\mathbf{R} \in \mathbb{R}^{n\times n}$, then

$$\begin{aligned}\mathbf{x} &= \mathbf{R}^{-1}\underbrace{\mathbf{Q}^T(\bar{\mathbf{y}} + \boldsymbol{\epsilon})}_{\stackrel{\text{def}}{=}\mathbf{v}}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \\ &\stackrel{\text{def}}{=} \mathbf{F}^{-1}(\mathbf{v}), \quad \mathbf{v} \sim \mathcal{N}(\mathbf{Q}^T\bar{\mathbf{y}}, \mathbf{I}).\end{aligned}$$

Proof that $\mathbf{x} \sim N\left((\lambda \mathbf{A}^T\mathbf{A} + \delta \mathbf{L})^{-1}\lambda \mathbf{A}^T\mathbf{y}, (\lambda \mathbf{A}^T\mathbf{A} + \delta \mathbf{L})^{-1}\right)$

What we know:

- $\mathbf{v} \sim \mathcal{N}(\mathbf{Q}^T\bar{\mathbf{y}}, \mathbf{I}) \Longrightarrow p_{\mathbf{v}}(\mathbf{v}) \propto \exp\left(-\frac{1}{2}\|\mathbf{v} - \mathbf{Q}^T\bar{\mathbf{y}}\|^2\right)$;
- $\mathbf{F}(\mathbf{x}) = \mathbf{R}\mathbf{x}$.

Proof that $\mathbf{x} \sim N\left((\lambda \mathbf{A}^T\mathbf{A} + \delta\mathbf{L})^{-1}\lambda\mathbf{A}^T\mathbf{y}, (\lambda\mathbf{A}^T\mathbf{A} + \delta\mathbf{L})^{-1}\right)$

What we know:

- $\mathbf{v} \sim \mathcal{N}(\mathbf{Q}^T\bar{\mathbf{y}}, \mathbf{I}) \Longrightarrow p_{\mathbf{v}}(\mathbf{v}) \propto \exp\left(-\frac{1}{2}\|\mathbf{v} - \mathbf{Q}^T\bar{\mathbf{y}}\|^2\right)$;
- $\mathbf{F}(\mathbf{x}) = \mathbf{R}\mathbf{x}$.

$$\begin{aligned} p(\mathbf{x}) &= \overbrace{(2\pi)^{-n/2}\left|\det\left(\mathbf{R}\right)\right|\exp\left(-\frac{1}{2}\|\mathbf{R}\mathbf{x} - \mathbf{Q}^T\bar{\mathbf{y}}\|^2\right)}^{\mathbf{x}=\mathbf{F}^{-1}(\mathbf{v}) \quad \Rightarrow \quad p(\mathbf{x})=\left|\det\left(\frac{d}{d\mathbf{x}}\mathbf{F}(\mathbf{x})\right)\right|p_{\mathbf{v}}(\mathbf{F}(\mathbf{x}))} \\ &= (2\pi)^{-n/2}\left|\det\left(\bar{\mathbf{A}}^T\bar{\mathbf{A}}\right)\right|^{1/2}\exp\left(-\frac{1}{2}\|\bar{\mathbf{A}}\mathbf{x} - \bar{\mathbf{y}}\|^2\right) \end{aligned}$$

Proof that $\mathbf{x} \sim N\left((\lambda \mathbf{A}^T\mathbf{A} + \delta \mathbf{L})^{-1}\lambda \mathbf{A}^T\mathbf{y}, (\lambda \mathbf{A}^T\mathbf{A} + \delta \mathbf{L})^{-1}\right)$

What we know:

- $\mathbf{v} \sim \mathcal{N}(\mathbf{Q}^T\bar{\mathbf{y}}, \mathbf{I}) \implies p_{\mathbf{v}}(\mathbf{v}) \propto \exp\left(-\frac{1}{2}\|\mathbf{v} - \mathbf{Q}^T\bar{\mathbf{y}}\|^2\right)$;
- $\mathbf{F}(\mathbf{x}) = \mathbf{R}\mathbf{x}$.

$$
\begin{aligned}
p(\mathbf{x}) &= \overbrace{(2\pi)^{-n/2} \left|\det(\mathbf{R})\right| \exp\left(-\frac{1}{2}\|\mathbf{R}\mathbf{x} - \mathbf{Q}^T\bar{\mathbf{y}}\|^2\right)}^{\mathbf{x}=\mathbf{F}^{-1}(\mathbf{v}) \quad \Rightarrow \quad p(\mathbf{x}) = \left|\det\left(\frac{d}{d\mathbf{x}}\mathbf{F}(\mathbf{x})\right)\right| p_{\mathbf{v}}(\mathbf{F}(\mathbf{x}))} \\
&= (2\pi)^{-n/2} \left|\det\left(\bar{\mathbf{A}}^T\bar{\mathbf{A}}\right)\right|^{1/2} \exp\left(-\frac{1}{2}\|\bar{\mathbf{A}}\mathbf{x} - \bar{\mathbf{y}}\|^2\right)
\end{aligned}
$$

Thus as desired, we have

$$
\begin{aligned}
\mathbf{x} &\sim \mathcal{N}\left((\bar{\mathbf{A}}^T\bar{\mathbf{A}})^{-1}\bar{\mathbf{A}}^T\bar{\mathbf{y}}, (\bar{\mathbf{A}}^T\bar{\mathbf{A}})^{-1}\right) \\
&\overset{'dist'}{=} \mathcal{N}\left((\lambda \mathbf{A}^T\mathbf{A} + \delta \mathbf{L})^{-1}\lambda \mathbf{A}^T\mathbf{y}, (\lambda \mathbf{A}^T\mathbf{A} + \delta \mathbf{L})^{-1}\right).
\end{aligned}
$$

MCMC Chain Diagnostics

Question: How correlated is the MCMC chain $\{(\mathbf{x}^k, \lambda_k, \delta_k)\}_{k=1}^N$?

MCMC Chain Diagnostics

Question: How correlated is the MCMC chain $\{(\mathbf{x}^k, \lambda_k, \delta_k)\}_{k=1}^N$?

Answer: First, note that

$$p(\mathbf{x}, \lambda, \delta | \mathbf{y}) \propto p(\mathbf{x} | \mathbf{y}, \lambda, \delta) p(\lambda, \delta | \mathbf{y}).$$

Thus one can compute an independent sample $(\mathbf{x}', \lambda', \delta')$ from $p(\mathbf{x}', \lambda', \delta' | \mathbf{y})$ via:

1. $(\lambda', \delta') \sim p(\lambda', \delta' | \mathbf{y})$,
2. $\mathbf{x}' \sim p(\mathbf{x}' | \mathbf{y}, \lambda', \delta')$.

Key Observation: the correlation in the (λ, δ)-chain drives the correlation in the $(\mathbf{x}, \lambda, \delta)$-chain.

Chain Correlation: How correlated is $\{\delta_i\}_{k=1}^{K}$?

The *autocorrelation function* is defined

$$\hat{\rho}(k) = C(k)/C(0),$$

where

$$C(k) = \frac{1}{K-|k|} \sum_{i=1}^{K-|k|} (\delta_i - \bar{\delta})(\delta_{i+|k|} - \bar{\delta}), \quad \text{where} \quad \bar{\delta} = \frac{1}{K} \sum_{i=1}^{k} \delta_i.$$

Chain Correlation: How correlated is $\{\delta_i\}_{k=1}^{K}$?

The *autocorrelation function* is defined

$$\hat{\rho}(k) = C(k)/C(0),$$

where

$$C(k) = \frac{1}{K-|k|}\sum_{i=1}^{K-|k|}(\delta_i - \bar{\delta})(\delta_{i+|k|} - \bar{\delta}), \quad \text{where} \quad \bar{\delta} = \frac{1}{K}\sum_{i=1}^{k}\delta_i.$$

The *integrated autocorrelation time* is defined

$$\hat{\tau}_{\text{int}} = \sum_{k=-\bar{K}}^{\bar{K}} \hat{\rho}(k),$$

where $\bar{K}$ is the smallest integer such that $\bar{K} \geq 3\hat{\tau}_{\text{int}}$,

Chain Correlation: How correlated is $\{\delta_i\}_{k=1}^{K}$?

The *autocorrelation function* is defined

$$\hat{\rho}(k) = C(k)/C(0),$$

where

$$C(k) = \frac{1}{K-|k|}\sum_{i=1}^{K-|k|}(\delta_i - \bar{\delta})(\delta_{i+|k|} - \bar{\delta}), \quad \text{where} \quad \bar{\delta} = \frac{1}{K}\sum_{i=1}^{k}\delta_i.$$

The *integrated autocorrelation time* is defined

$$\hat{\tau}_{\text{int}} = \sum_{k=-\bar{K}}^{\bar{K}} \hat{\rho}(k),$$

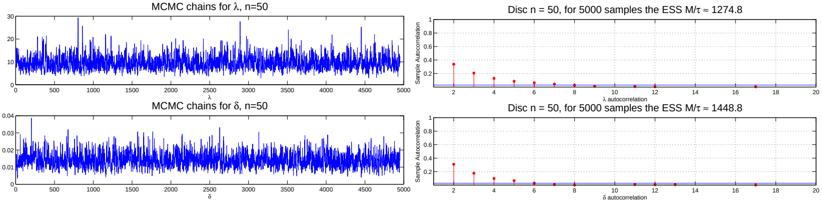
where $\bar{K}$ is the smallest integer such that $\bar{K} \geq 3\hat{\tau}_{\text{int}}$, and

$$\#\,\text{independent samples in}\,\{\delta_i\}_{k=1}^{K} \approx K/\hat{\tau}_{\text{int}}.$$

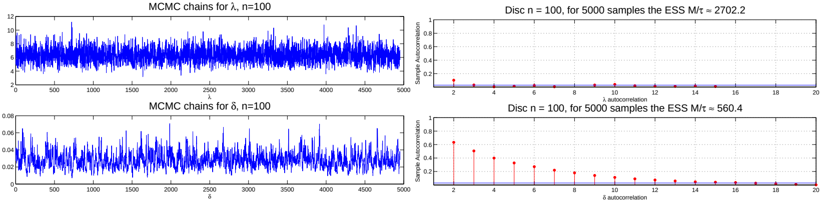
As $n \to \infty$, correlation in λ/δ-chains disappears/increases

(Work with S. Agapiou, O. Papaspiliopoulos, & Andrew Stuart)

$$n = 50$$



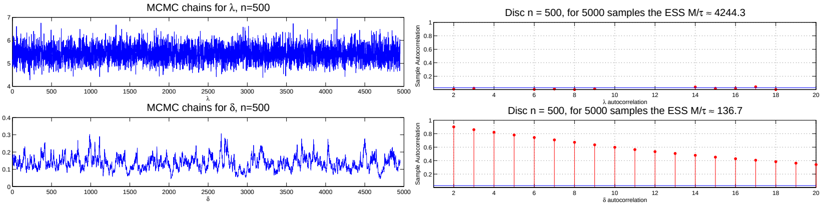
$$n = 100$$



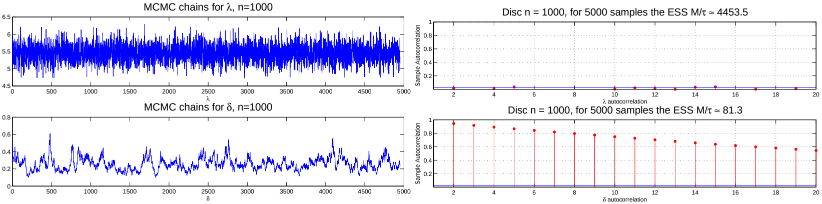
As $n \to \infty$, correlation in λ/δ-chains disappears/increases

(Work with S. Agapiou, O. Papaspiliopoulos, & Andrew Stuart)

$$n = 500$$



$$n = 1000$$



To overcome this issue, we use marginalization

First note that

$$\begin{aligned}\frac{\lambda}{2}\|\mathbf{A}\mathbf{x}-\mathbf{y}\|^2+\frac{\delta}{2}\mathbf{x}^T\mathbf{L}\mathbf{x} &= \frac{1}{2}\underbrace{(\lambda\|\mathbf{y}\|^2-\boldsymbol{\mu}^T(\lambda\mathbf{A}^T\mathbf{A}+\delta\mathbf{L})\boldsymbol{\mu})}_{U(\lambda,\delta)}+\\ &\quad \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T(\lambda\mathbf{A}^T\mathbf{A}+\delta\mathbf{L})(\mathbf{x}-\boldsymbol{\mu}),\end{aligned}$$

where $\boldsymbol{\mu}=(\lambda\mathbf{A}^T\mathbf{A}+\delta\mathbf{L})^{-1}\lambda\mathbf{A}^T\mathbf{y}$.

To overcome this issue, we use marginalization

First note that

$$\frac{\lambda}{2}\|\mathbf{A}\mathbf{x}-\mathbf{y}\|^2+\frac{\delta}{2}\mathbf{x}^T\mathbf{L}\mathbf{x} \;=\; \frac{1}{2}\underbrace{(\lambda\|\mathbf{y}\|^2-\boldsymbol{\mu}^T(\lambda\mathbf{A}^T\mathbf{A}+\delta\mathbf{L})\boldsymbol{\mu})}_{U(\lambda,\delta)}+ \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T(\lambda\mathbf{A}^T\mathbf{A}+\delta\mathbf{L})(\mathbf{x}-\boldsymbol{\mu}),$$

where $\boldsymbol{\mu}=(\lambda\mathbf{A}^T\mathbf{A}+\delta\mathbf{L})^{-1}\lambda\mathbf{A}^T\mathbf{y}$. Then

$$p(\mathbf{x},\lambda,\delta|\mathbf{y}) \;\propto\; p(\lambda)p(\delta)\exp\left(-\frac{1}{2}U(\lambda,\delta)\right)\times \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T(\lambda\mathbf{A}^T\mathbf{A}+\delta\mathbf{L})(\mathbf{x}-\boldsymbol{\mu})\right)$$

To overcome this issue, we use marginalization

$$
\begin{aligned}
p(\lambda, \delta|\mathbf{y}) &\propto \int_{\mathbb{R}^n} p(\mathbf{x}, \lambda, \delta|\mathbf{y}) d\mathbf{x} \\
&\propto p(\lambda)p(\delta) \exp\left(-\frac{1}{2}U(\lambda, \delta)\right) \times \\
&\quad \underbrace{\int_{\mathbb{R}^n} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T(\lambda\mathbf{A}^T\mathbf{A}+\delta\mathbf{L})(\mathbf{x}-\boldsymbol{\mu})\right) d\mathbf{x}}_{(2\pi)^{n/2}\det(\lambda\mathbf{A}^T\mathbf{A}+\delta\mathbf{L})^{-1/2}}
\end{aligned}
$$

To overcome this issue, we use marginalization

$$
\begin{aligned}
p(\lambda, \delta|\mathbf{y}) &\propto \int_{\mathbb{R}^n} p(\mathbf{x}, \lambda, \delta|\mathbf{y}) d\mathbf{x} \\
&\propto p(\lambda)p(\delta) \exp\left(-\frac{1}{2}U(\lambda, \delta)\right) \times \\
&\quad \underbrace{\int_{\mathbb{R}^n} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T(\lambda\mathbf{A}^T\mathbf{A}+\delta\mathbf{L})(\mathbf{x}-\boldsymbol{\mu})\right) d\mathbf{x}}_{(2\pi)^{n/2}\det(\lambda\mathbf{A}^T\mathbf{A}+\delta\mathbf{L})^{-1/2}} \\
&\propto p(\lambda)p(\delta) \exp\left(-\frac{1}{2}U(\lambda, \delta) - \frac{1}{2}\underbrace{\ln\det(\lambda\mathbf{A}^T\mathbf{A}+\delta\mathbf{L})}_{c(\lambda,\delta)}\right).
\end{aligned}
$$

To overcome this issue, we use marginalization

Thus we have the *marginal density*

$$p(\lambda, \delta|\mathbf{y}) \propto p(\lambda)p(\delta) \exp\left(-\frac{1}{2}U(\lambda, \delta) - \frac{1}{2}c(\lambda, \delta)\right),$$

where

$$\begin{aligned} U(\lambda, \delta) &= \mathbf{y}^T(\lambda \mathbf{I} - \lambda^2 \mathbf{A}(\lambda \mathbf{A}^T \mathbf{A} + \delta \mathbf{L})^{-1} \mathbf{A}^T)\mathbf{y} \\ c(\lambda, \delta) &= \ln \det(\lambda \mathbf{A}^T \mathbf{A} + \delta \mathbf{L}). \end{aligned}$$

Partially Collapsed Gibbs: Step 1, Reduce Conditioning

First ‘reduce the conditioning’ in the problematic step 2 of the Gibbs sampler, which is guaranteed to improve performance.

Reduce Conditioning in the Gibbs Sampler

0. Choose $\mathbf{x}^0$, and set $k = 0$;
1. Compute $\lambda_{k+1} \sim \Gamma\left(m/2 + \alpha_\lambda, \frac{1}{2}\|\mathbf{A}\mathbf{x}^k - \mathbf{y}\|^2 + \beta_\lambda\right)$;
2. Compute $\underbrace{(\hat{\mathbf{x}}, \delta_{k+1}) \sim p(\hat{\mathbf{x}}, \delta_{k+1}|\mathbf{y}, \lambda_{k+1})}_{\text{previously: } \delta_{k+1} \sim p(\delta_{k+1}|\mathbf{y}, \mathbf{x}^k, \lambda_{k+1})}$;
3. Compute $\mathbf{x}^{k+1} \sim$ $N\left((\lambda_{k+1}\mathbf{A}^T\mathbf{A} + \delta_{k+1}\mathbf{L})^{-1}\lambda_{k+1}\mathbf{A}^T\mathbf{y}, (\lambda_{k+1}\mathbf{A}^T\mathbf{A} + \delta_{k+1}\mathbf{L})^{-1}\right)$;
4. Set $k = k + 1$ and return to Step 1.

Partially Collapsed Gibbs: Step 2, Collapse/Marginalize

In step 2, $\hat{\mathbf{x}}$ is redundant, so we can integrate in out, to obtain

$$\begin{aligned}
\delta_{k+1} &\sim \int_{\mathbb{R}^n} p(\hat{\mathbf{x}}, \delta_{k+1}|\mathbf{y}, \lambda_{k+1}) d\hat{\mathbf{x}} \\
&\stackrel{'d'}{=} p(\delta_{k+1}|\mathbf{y}, \lambda_{k+1}) \\
&\propto p(\delta_{k+1}) \exp\left(-\frac{1}{2}U(\lambda_{k+1}, \delta_{k+1}) - \frac{1}{2}c(\lambda_{k+1}, \delta_{k+1})\right).
\end{aligned}$$

Partially Collapsed Gibbs: Step 2, Collapse/Marginalize

In step 2, $\hat{\mathbf{x}}$ is redundant, so we can integrate in out, to obtain

$$\begin{aligned} \delta_{k+1} &\sim \int_{\mathbb{R}^n} p(\hat{\mathbf{x}}, \delta_{k+1} | \mathbf{y}, \lambda_{k+1}) d\hat{\mathbf{x}} \\ &\overset{'d'}{=} p(\delta_{k+1} | \mathbf{y}, \lambda_{k+1}) \\ &\propto p(\delta_{k+1}) \exp\left(-\frac{1}{2} U(\lambda_{k+1}, \delta_{k+1}) - \frac{1}{2} c(\lambda_{k+1}, \delta_{k+1}) \right). \end{aligned}$$

Partially Collapsed Gibbs Sampler for $p(\mathbf{x}, \delta, \lambda | \mathbf{y})$.

0. Choose $\mathbf{x}^0$, and set $k = 0$;
1. Compute $\lambda_{k+1} \sim \Gamma\left(m/2 + \alpha_\lambda, \frac{1}{2}\|\mathbf{A}\mathbf{x}^k - \mathbf{y}\|^2 + \beta_\lambda\right)$;
2. Compute $\delta_{k+1} \sim p(\delta_{k+1} | \mathbf{y}, \lambda_{k+1})$;
3. Compute $\mathbf{x}^{k+1} \sim$ $N\left((\lambda_{k+1}\mathbf{A}^T\mathbf{A} + \delta_{k+1}\mathbf{L})^{-1}\lambda_{k+1}\mathbf{A}^T\mathbf{y}, (\lambda_{k+1}\mathbf{A}^T\mathbf{A} + \delta_{k+1}\mathbf{L})^{-1}\right)$;
4. Set $k = k + 1$ and return to Step 1.

2. Compute $\delta_{k+1} \sim p(\delta_{k+1}|\mathbf{y}, \lambda_{k+1})$

n_{MH} *adaptive Metropolis steps:* Set $\delta_{k,0} = \delta_k$ and $j = 1$.

(i) Compute $\rho^* \sim \mathcal{N}(\ln(\delta_{k,j-1}), \sigma_k^2)$ and set $\delta_{k*} = \exp(\rho^*)$. With probability

$$\alpha = \min\left\{1, \frac{p(\delta_{k*}|\mathbf{y}, \lambda_k)}{p(\delta_{k,j-1}|\mathbf{y}, \lambda_k)}\right\},$$

set $\delta_{kj} = \delta_{k*}$, else $\delta_{k,j} = \delta_{k,j-1}$.

(ii) If $j < n_{\text{MH}}$, set $j = j + 1$ and return to step 2(i).

(iii) Else once $j = n_{\text{MH}}$, define $\delta_{k+1} = \delta_{n_{\text{MH}}}$,

$$\sigma_{k+1}^2 = \text{var}([\ln \delta_1, \ldots, \ln \delta_{k+1}]),$$

and go to step 3.

2. Compute $\delta_{k+1} \sim p(\delta_{k+1}|\mathbf{y}, \lambda_{k+1})$

n_{MH} *adaptive Metropolis steps:* Set $\delta_{k,0} = \delta_k$ and $j = 1$.

(i) Compute $\rho^* \sim \mathcal{N}(\ln(\delta_{k,j-1}), \sigma_k^2)$ and set $\delta_{k*} = \exp(\rho^*)$. With probability

$$\alpha = \min\left\{1, \frac{p(\delta_{k*}|\mathbf{y}, \lambda_k)}{p(\delta_{k,j-1}|\mathbf{y}, \lambda_k)}\right\},$$

set $\delta_{kj} = \delta_{k*}$, else $\delta_{k,j} = \delta_{k,j-1}$.

(ii) If $j < n_{\mathrm{MH}}$, set $j = j + 1$ and return to step 2(i).

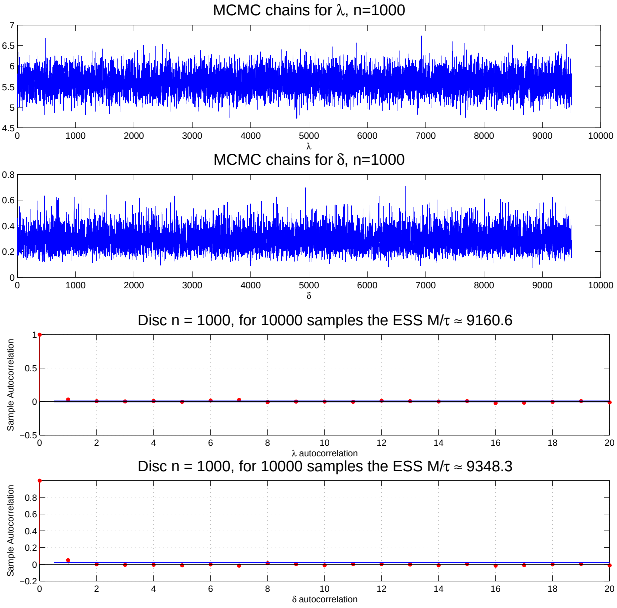
(iii) Else once $j = n_{\mathrm{MH}}$, define $\delta_{k+1} = \delta_{n_{\mathrm{MH}}}$,

$$\sigma_{k+1}^2 = \mathrm{var}([\ln \delta_1, \ldots, \ln \delta_{k+1}]),$$

and go to step 3.

NOTE: if n_{MH} is sufficiently large, this will yield essentially independent samples from $p(\delta|\mathbf{y}, \lambda_{k+1})$.

Chain auto-correlation plots for Partially Collapsed Gibbs



Some observations regarding $p(\lambda, \delta|\mathbf{y})$

Note that

$$p(\mathbf{x}, \lambda, \delta|\mathbf{y}) \propto p(\mathbf{x}|\mathbf{y}, \lambda, \delta)p(\lambda, \delta|\mathbf{y}).$$

Thus one can sample from $p(\mathbf{x}, \lambda, \delta|\mathbf{y})$ via

1. $(\lambda', \delta') \sim p(\lambda', \delta'|\mathbf{y})$,
2. $\mathbf{x}' \sim p(\mathbf{x}'|\mathbf{y}, \lambda', \delta')$.

Fox and Norton call this marginal-then-conditional sampling.

Some observations regarding $p(\lambda, \delta|\mathbf{y})$

Note that

$$p(\mathbf{x}, \lambda, \delta|\mathbf{y}) \propto p(\mathbf{x}|\mathbf{y}, \lambda, \delta)p(\lambda, \delta|\mathbf{y}).$$

Thus one can sample from $p(\mathbf{x}, \lambda, \delta|\mathbf{y})$ via

1. $(\lambda', \delta') \sim p(\lambda', \delta'|\mathbf{y})$,
2. $\mathbf{x}' \sim p(\mathbf{x}'|\mathbf{y}, \lambda', \delta')$.

Fox and Norton call this marginal-then-conditional sampling.

Observation: one can use MCMC to compute a (λ, δ)-chain for $p(\lambda, \delta|\mathbf{y})$ in step 1, then afterwards compute $\mathbf{x}$-samples in step 2.

Sample directly from $p(\lambda, \delta|\mathbf{y})$ using adaptive Metropolis

0. Initialize λ_0, δ_0, and $\mathbf{C}_0 \in \mathbb{R}^{2\times 2}$. Set $k = 1$. Define k_{total}.
1. Compute $\begin{bmatrix} \rho^* \\ \gamma^* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \ln(\lambda_{k-1}) \\ \ln(\delta_{k-1}) \end{bmatrix}, \mathbf{C}_{k-1}\right)$ and set $[\lambda^*, \delta^*]^T = [\exp(\rho^*), \exp(\gamma^*)]^T$. With probability

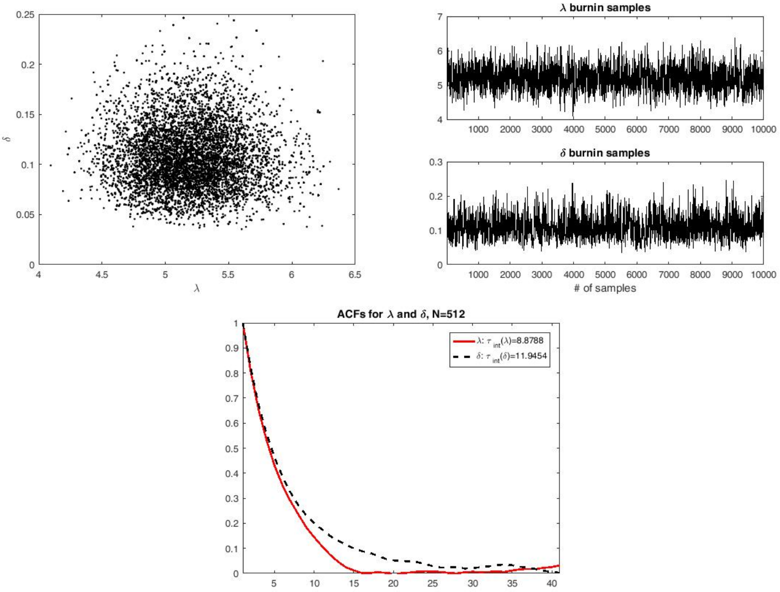
$$\alpha = \min\left\{1, \frac{p(\lambda^*, \delta^*|\mathbf{y})}{p(\lambda_{k-1}, \delta_{k-1}|\mathbf{y})}\right\},$$

   set $[\lambda_k, \delta_k]^T = [\lambda^*, \delta^*]^T$, else set $[\lambda_k, \delta_k]^T = [\lambda_{k-1}, \delta_{k-1}]^T$.
2. Update the proposal covariance:

$$\mathbf{C}_k = \text{cov}\left(\begin{bmatrix} \ln(\lambda_0) & \ln(\delta_0) \\ \vdots & \vdots \\ \ln(\lambda_k) & \ln(\delta_k) \end{bmatrix}\right) + \epsilon\mathbf{I}, \quad 0 < \epsilon \ll 1.$$

3. If $k = k_{\text{total}}$ stop, else set $k = k + 1$ and return to Step 1.

Chain diagnostics for AM applied to $p(\lambda, \delta|\mathbf{y})$



Computational Bottlekneck

Evaluating

$$p(\lambda, \delta|\mathbf{y}) \propto p(\lambda)p(\delta) \exp\left(-\frac{1}{2}U(\lambda,\delta) - \frac{1}{2}c(\lambda,\delta)\right),$$

requires

$$\begin{aligned} U(\lambda,\delta) &= \mathbf{y}^T(\lambda\mathbf{I} - \lambda^2\mathbf{A}(\lambda\mathbf{A}^T\mathbf{A} + \delta\mathbf{L})^{-1}\mathbf{A}^T)\mathbf{y} \\ c(\lambda,\delta) &= \ln\det(\lambda\mathbf{A}^T\mathbf{A} + \delta\mathbf{L}), \end{aligned}$$

which in turn requires

- computing $\mathbf{x}_{\text{MAP}} = (\lambda\mathbf{A}^T\mathbf{A} + \delta\mathbf{L})^{-1}\lambda\mathbf{A}^T\mathbf{y}$;
- computing $\ln\det(\lambda\mathbf{A}^T\mathbf{A} + \delta\mathbf{L})$.

Computational Bottlekneck

Evaluating

$$p(\lambda, \delta|\mathbf{y}) \propto p(\lambda)p(\delta) \exp\left(-\frac{1}{2}U(\lambda, \delta) - \frac{1}{2}c(\lambda, \delta)\right),$$

requires

$$\begin{aligned} U(\lambda, \delta) &= \mathbf{y}^T(\lambda\mathbf{I} - \lambda^2\mathbf{A}(\lambda\mathbf{A}^T\mathbf{A} + \delta\mathbf{L})^{-1}\mathbf{A}^T)\mathbf{y} \\ c(\lambda, \delta) &= \ln\det(\lambda\mathbf{A}^T\mathbf{A} + \delta\mathbf{L}), \end{aligned}$$

which in turn requires

- computing $\mathbf{x}_{\text{MAP}} = (\lambda\mathbf{A}^T\mathbf{A} + \delta\mathbf{L})^{-1}\lambda\mathbf{A}^T\mathbf{y}$;
- computing $\ln\det(\lambda\mathbf{A}^T\mathbf{A} + \delta\mathbf{L})$.

NOTE: For the CT test case, these can only computed approximately.

The Gibbs sampler for sampling from $p(\mathbf{x}, \lambda, \delta|\mathbf{y})$

0. δ_0, and λ_0, and set $k = 0$;
1. Compute $\lambda_{k+1} \sim \Gamma\left(n/2 + \alpha_\lambda, \frac{1}{2}\|\mathbf{A}\mathbf{x}^k - \mathbf{y}\|^2 + \beta_\lambda\right)$;
2. Compute $\delta_{k+1} \sim \Gamma\left(n/2 + \alpha_\delta, \frac{1}{2}(\mathbf{x}^k)^T\mathbf{L}\mathbf{x}^k + \beta_\delta\right)$;
3. Compute $\mathbf{x}^{k+1} \sim$
   $\mathcal{N}\left((\lambda_{k+1}\mathbf{A}^T\mathbf{A} + \delta_{k+1}\mathbf{L})^{-1}\lambda_{k+1}\mathbf{A}^T\mathbf{y}, (\lambda_{k+1}\mathbf{A}^T\mathbf{A} + \delta_{k+1}\mathbf{L})^{-1}\right)$;
4. Set $k = k + 1$ and return to Step 1.

NOTE: step 3 is the computational bottleneck for many large-scale problems, such as computed tomography.

Gradient Scan Gibbs Sampler

Replace step 3 of the Gibbs sampler

$$\mathbf{x}^{k+1} \sim \mathcal{N}\left((\lambda_{k+1}\mathbf{A}^T\mathbf{A} + \delta_{k+1}\mathbf{L})^{-1}\lambda_{k+1}\mathbf{A}^T\mathbf{y}, (\lambda_{k+1}\mathbf{A}^T\mathbf{A} + \delta_{k+1}\mathbf{L})^{-1}\right)$$

with j_k CG iterations applied to

$$(\lambda_k\mathbf{A}^T\mathbf{A}+\delta_k\mathbf{L})(\mathbf{x}^{k-1}+\mathbf{p}) = \lambda_k\mathbf{A}^T\mathbf{y}+\lambda_k^{1/2}\mathbf{A}^T\boldsymbol{\epsilon}_1+\delta_k^{1/2}\tilde{\mathbf{D}}^T\boldsymbol{\epsilon}_2, \quad (1)$$

where $\boldsymbol{\epsilon}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$ and $\boldsymbol{\epsilon}_2 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$, and then define

$$\mathbf{x}^k = \mathbf{x}^{k-1} + \mathbf{p}^{j_k},$$

where $\mathbf{p}^{j_k}$ is the final CG iterate.

NOTE: in exact arithmetic, if $j_k = n$, using (1) is equivalent to the orginal Gibbs sampler.

Gradient Scan Gibbs Sampler

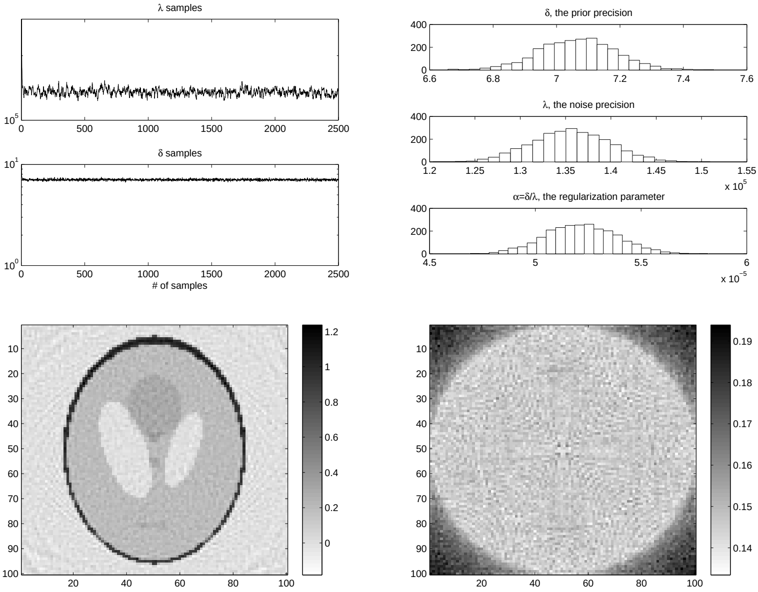
0. Initialize λ_0 and δ_0, set $k = 1$ and define k_{total}. Define $\mathbf{x}^0$ to be the j_0^{th} iterate of CG applied to $(\lambda_0 \mathbf{A}^* \mathbf{A} + \delta_0 \mathbf{L})\mathbf{x} = \lambda_0 \mathbf{A}^* \mathbf{y}$;
1. Compute $\lambda_k \sim \Gamma\left(M/2 + \alpha_\lambda, \frac{1}{2}\|\mathbf{A}\mathbf{x}^{k-1} - \mathbf{y}\|^2 + \beta_\lambda\right)$;
2. Compute $\delta_k \sim \Gamma\left(\overline{N}/2 + \alpha_\delta, \frac{1}{2}(\mathbf{x}^{k-1})^T \mathbf{L}\mathbf{x}^{k-1} + \beta_\delta\right)$;
3. Apply j_k iterations of CG to

   $$(\lambda_k \mathbf{A}^T \mathbf{A} + \delta_k \mathbf{L})(\mathbf{x}^{k-1} + \mathbf{p}) = \lambda_k \mathbf{A}^T \mathbf{y} + \lambda_k^{1/2} \mathbf{A}^T \boldsymbol{\epsilon}_1 + \delta_k^{1/2} \tilde{\mathbf{D}}^T \boldsymbol{\epsilon}_2,$$

   and define $\mathbf{x}^k = \mathbf{x}^{k-1} + \mathbf{p}^{j_k}$, where $\mathbf{p}^{j_k}$ is the j_k^{th} CG iterate.
4. If $k = k_{\text{total}}$ stop, otherwise, set $k = k + 1$ and return to Step 1.

NOTE: the smaller is j_k, the more correlated will be the $\mathbf{x}$-chain.

Grad Scan Gibbs Numerical Test: $j_k = 20$, $n = 128^2$.



Conclusions/Takeaways

- Inverse problems have unique characteristics, making the use of Bayesian methods for their solution practical, challenging, and interesting.
- GMRFs provide a way of modelling the prior from pixel-level assumptions. However, not all GMRFs yield a well-defined posterior density in the infinite dimensional limit.
- Placing probability densities on λ and δ yields a hierarchical posterior density $p(\mathbf{x}, \lambda, \delta|\mathbf{y})$.
- We provided MCMC methods for sampling from the posterior $p(\mathbf{x}, \lambda, \delta|\mathbf{y})$ and the marginal density $p(\lambda, \delta|\mathbf{y})$.