Sampling Methods for Uncertainty Quantification in Inverse Problems

John Bardsley

University of Montana

University of Warwick, June 2015

Outline

- Introduce the problem of interest: inverse problems
- Case 1: Linear inverse problems and posterior sampling using optimization
    - Hierarchical models.
- Case 2: Nonlinear inverse problems and posterior sampling using optimization
    - Randomize-then-Optimize (RTO).
- Numerical Tests.

General Statistical Model

Consider the linear, Gaussian statistical model

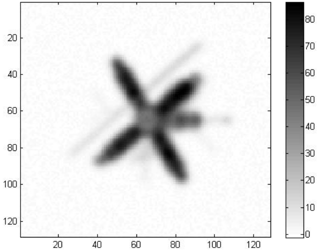
$$\mathbf{y} = \mathbf{A}\mathbf{u} + \boldsymbol{\epsilon},$$

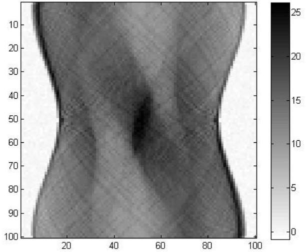
where

- $\mathbf{y} \in \mathbb{R}^m$ is the vector of observations;
- $\mathbf{u} \in \mathbb{R}^n$ is the vector of unknown parameters;
- $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^m$, with $m \geq n$;
- $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$, i.e., $\boldsymbol{\epsilon}$ is i.i.d. Gaussian with mean 0 and variance σ^2.

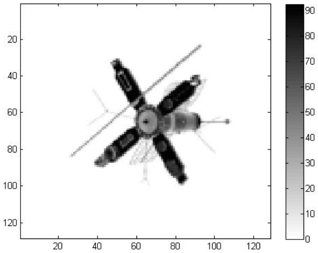
Synthetic Examples

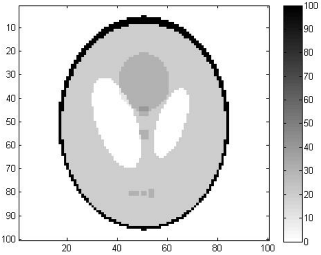
Data $\mathbf{y}$ examples:



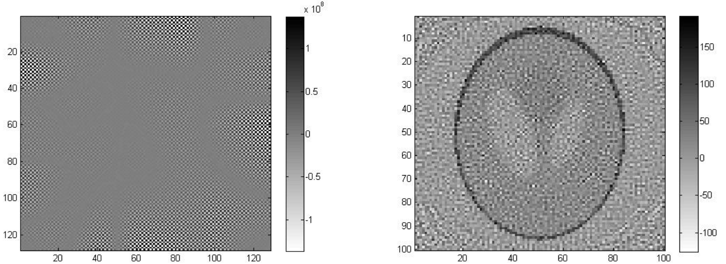


Corresponding true images $\mathbf{u}$:

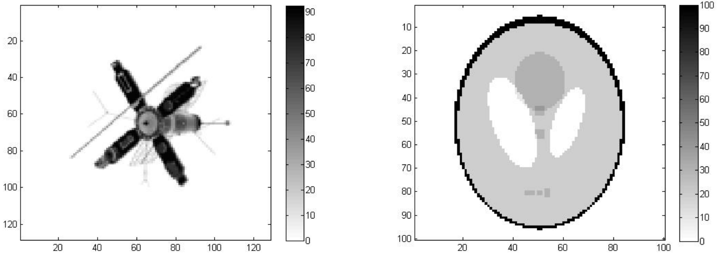




Naive Solutions: $\mathbf{u}_{\text{naive}} = \mathbf{A}^{-1}\mathbf{y}$



Corresponding true images $\mathbf{u}$:



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

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

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T = \sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^T,$$

- the σ_i's get very close to 0 as $i \to n$;
- and 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

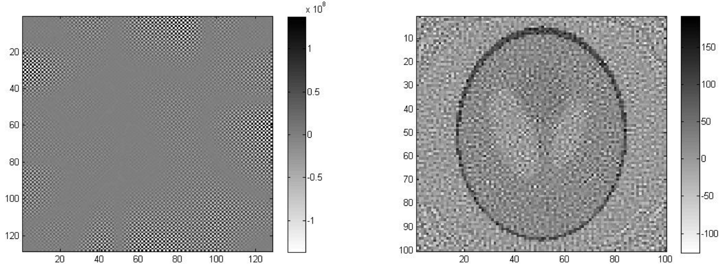
$$\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T = \sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^T,$$

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

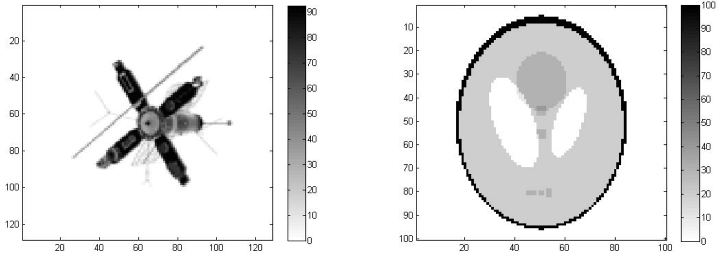
Then the naive solution can then be written

$$\begin{aligned}
\mathbf{A}^{-1}\mathbf{y} &= \mathbf{A}^{-1}(\mathbf{A}\mathbf{u} + \boldsymbol{\epsilon}) \\
&= \mathbf{u} + \mathbf{A}^{-1}\boldsymbol{\epsilon} \\
&= \mathbf{u} + \underbrace{\sum_{i=1}^{n} \left(\frac{\mathbf{u}_i^T \boldsymbol{\epsilon}}{\sigma_i} \right) \mathbf{v}_i}_{\text{large } i \text{ terms dominate}}
\end{aligned}$$

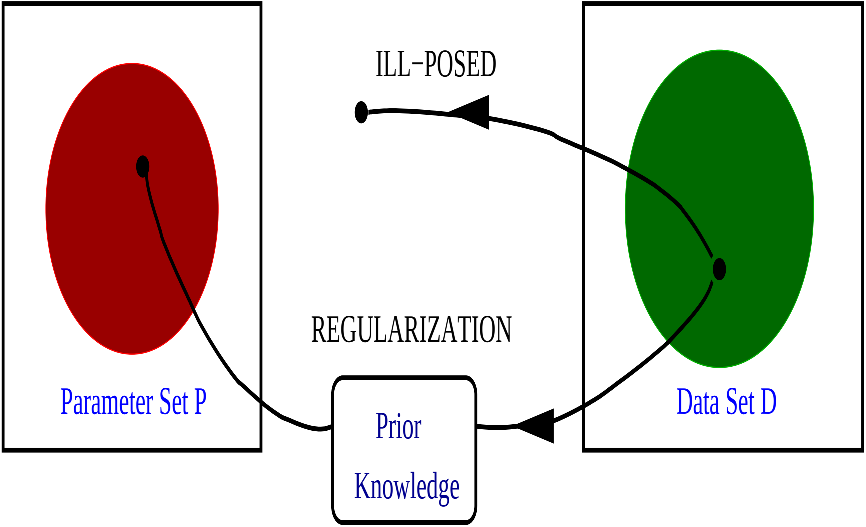
Naive Solutions: $\mathbf{u}_{\text{naive}} = \mathbf{A}^{-1}\mathbf{y}$



Corresponding true images $\mathbf{u}$:



The Fix: Regularization



Bayes Law and Regularization

Bayes' Law:

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

Bayes Law and Regularization

Bayes' Law:

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

For our statistical model, with $\lambda = 1/\sigma^2$,

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

And we assume that the prior has the form

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

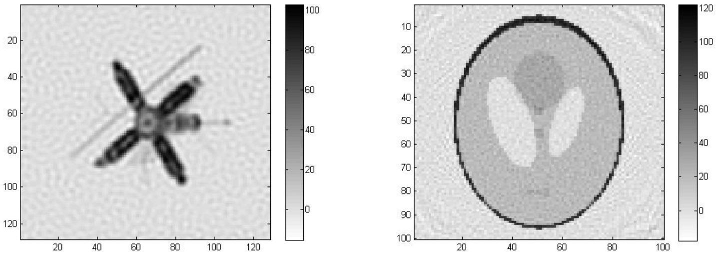
Bayes Law and Regularization

The maximizer of the posterior density is

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

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

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

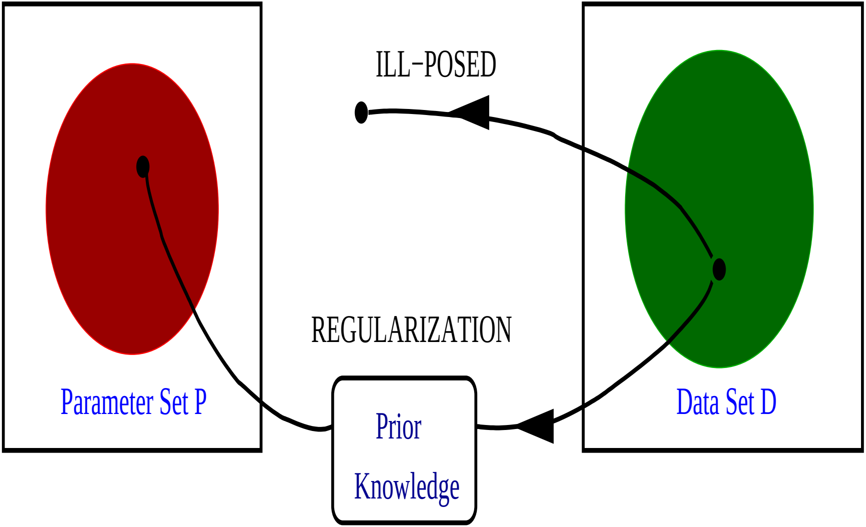


Aside: Develop Your Mathematical Taste

Some Early Inspiration: Algorithms, Numerics & Bayes

1. Vogel
   - *Computational Methods for Inverse Problems*, SIAM 2002.
2. Haario
   - University of Montana Computational Statistics course, Spring 2006.
3. Calvetti & Somersalo
   - *A Gaussian hypermodel to recover blocky objects*, Inverse Problems, 2007.
   - *Intro to Bayesian Scientific Computing*, Springer, 2007.
   - Hierarchical Regularization for Edge-Preserving Reconstruction of PET Images, with D. Calvetti and E. Somersalo, Inverse Problems, **26(3)**, 2010, 035010.
4. Rue & Held
   - *Gaussian Markov Random Fields*, CRC Press, 2005.
   Alternate/equivalent to Calvetti & Somersalo, talk 1.

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



Gaussian Markov Random field (GMRF) priors

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

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

Gaussian Markov Random field (GMRF) priors

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

$$\mathbf{u}_{\partial_{ij}} = \{u_{i-1,j}, u_{i,j-1}, u_{i+1,j}, u_{i,j+1}\}$$

$$= \begin{bmatrix} & u_{i,j+1} & \\ u_{i-1,j} & u_{ij} & u_{i+1,j} \\ & u_{i,j-1} & \end{bmatrix}.$$

Then we assume

$$u_{i,j}|\mathbf{u}_{\partial_{i,j}} \sim \mathcal{N}\left(\bar{u}_{\partial_{i,j}}, \frac{1}{\delta n_{ij}}\right),$$

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

Gaussian Markov Random field (GMRF) priors

This leads to the prior

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

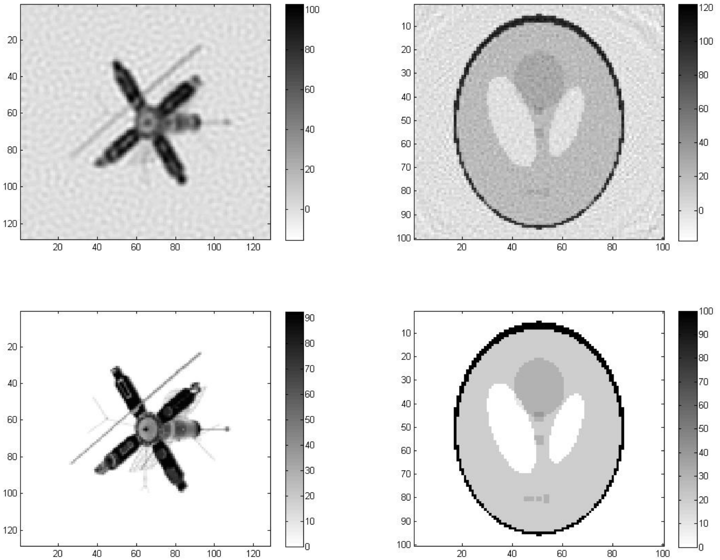
where if $r = (i, j)$ after column-stacking 2D arrays

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

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

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

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



2D Intrinsic GMRF Increment Models (ala Calvetti & Somersalo)

For a 2D signal, suppose

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

2D Intrinsic GMRF Increment Models (ala Calvetti & Somersalo)

For a 2D signal, suppose

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

Then the density function for $\mathbf{u}$ has the form

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

2D Intrinsic GMRF Increment Models (ala Calvetti & Somersalo)

For a 2D signal, suppose

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

Then the density function for $\mathbf{u}$ has the form

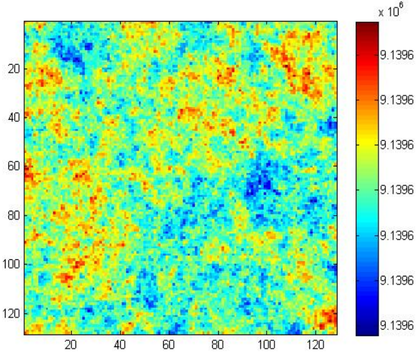
$$
\begin{aligned}
p(\mathbf{u}|\delta) &\propto \delta^{(n^2-1)/2} \exp\left(-\frac{\delta}{2} \sum_{i=1}^{n-1} \sum_{j=1}^{n} w_{ij}^h (u_{i+1,j} - u_{ij})^2\right) \times \\
&\qquad \exp\left(-\frac{\delta}{2} \sum_{i=1}^{n} \sum_{j=1}^{n-1} w_{ij}^v (u_{i,j+1} - u_{ij})^2\right) \\
&= \delta^{(n^2-1)/2}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

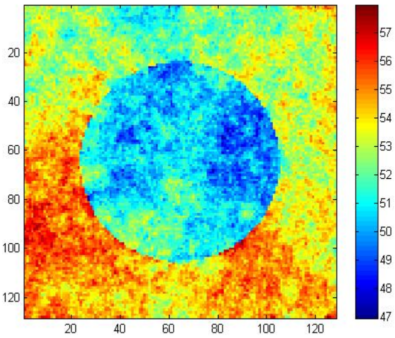
$\mathbf{D}_h = \mathbf{I} \otimes \mathbf{D}$, $\mathbf{D}_v = \mathbf{D} \otimes \mathbf{I}$, $\mathbf{D}$ is a 1D difference matrix,
$\boldsymbol{\Lambda}_h = \text{diag}(\text{vec}(\{w_{ij}^h\}_{ij=1}^{n-1}))$, $\boldsymbol{\Lambda}_v = \text{diag}(\text{vec}(\{w_{ij}^v\}_{ij=1}^{n-1}))$.

2D IGMRF Increment Models

The matrix $\mathbf{D}_h^T \mathbf{\Lambda}_h \mathbf{D}_h + \mathbf{D}_v^T \mathbf{\Lambda}_v \mathbf{D}_v$ is a discretization of

$$-\frac{d}{ds}\left(w_h(s,t)\frac{d}{ds}\right) - \frac{d}{dt}\left(w_v(s,t)\frac{d}{dt}\right)$$





Left: $w_{ij}^h = w_{ij}^v = 1$ for all ij.

Right: $w_{ij}^h = w_{ij}^v = 0.01$ for ij on the circle boundary.

Summary Table

Statistical Assumption	PDE	Reg. Matrix
$u_{ij} \| \mathbf{u}_{\partial_{ij}} \sim \mathcal{N}(\overline{x}_{\partial_{ij}}, (n_{ij} w_{ij})^{-1})$	$-\frac{\partial}{\partial s}\left(w(s,t)\right)\frac{\partial x}{\partial s}\Big) - \frac{\partial}{\partial t}\left(w(s,t)\right)\frac{\partial x}{\partial t}\Big)$	$\mathbf{D}_h^T \boldsymbol{\Lambda} \mathbf{D}_h + \mathbf{D}_v^T \boldsymbol{\Lambda} \mathbf{D}_v$

Caveat:

- These discretizations are grid dependent since I have not scaled $\mathbf{D}$, $\mathbf{D}_h$, and $\mathbf{D}_v$ by step size.
- Even with step size, these break down in the infinite limit, but you could square them to overcome this.

IGMRF Edge-Preserving Reconstruction

0. Set $\mathbf{\Lambda}_h = \mathbf{\Lambda}_v = \mathbf{I}$.
1. Define $\mathbf{L} = \mathbf{D}_h^T \mathbf{\Lambda} \mathbf{D}_h + \mathbf{D}_v^T \mathbf{\Lambda} \mathbf{D}_v$, where
2. Compute the solution $\mathbf{u}_\alpha$ of

$$(\mathbf{A}^T \mathbf{A} + \alpha \mathbf{L})\mathbf{u} = \mathbf{A}^T \mathbf{y}$$

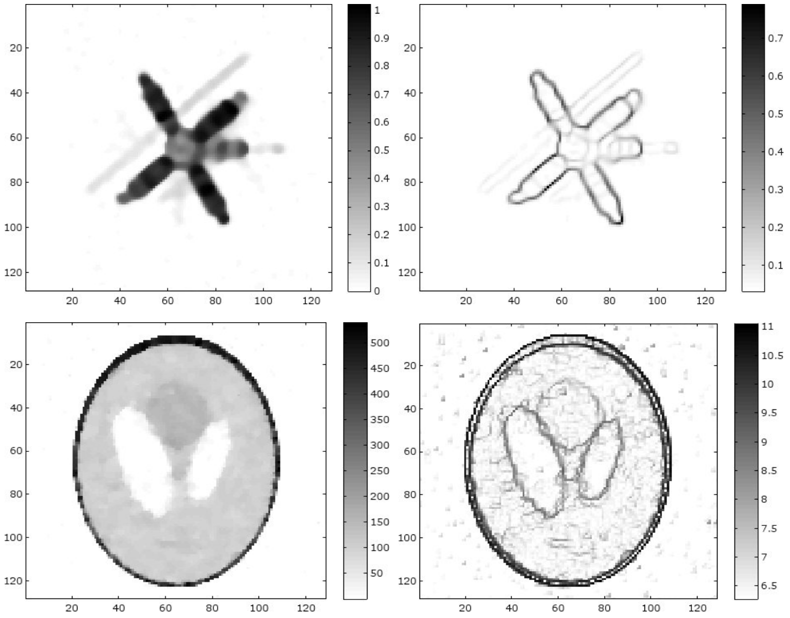
   using PCG with α obtained using GCV.
3. Set

$$\mathbf{\Lambda}(\mathbf{x}_\alpha) = \text{diag}\left(\frac{\mathbf{1}}{\sqrt{(\mathbf{D}_h \mathbf{u}_\alpha)^2 + (\mathbf{D}_v \mathbf{u}_\alpha)^2 + \beta \mathbf{1}}} \right)$$

   and return to Step 1.

NOTE: This is just the lagged-diffusivity iteration.

Numerical Results



2D Laplace Increment Models (Anisotropic TV)

For a 2D signal, suppose

$$\begin{aligned} u_{i+1,j} - u_{ij} &\sim \text{Laplace}(0, \delta^{-1}), \quad i = 1, \ldots, n-1, \; j = 1, \ldots, n \\ u_{i,j+1} - u_{ij} &\sim \text{Laplace}(0, \delta^{-1}), \quad i = 1, \ldots, n, \; j = 1, \ldots, n-1 \end{aligned}$$

2D Laplace Increment Models (Anisotropic TV)

For a 2D signal, suppose

$$\begin{aligned} u_{i+1,j} - u_{ij} &\sim \text{Laplace}(0, \delta^{-1}), \quad i = 1, \ldots, n-1,\ j = 1, \ldots, n \\ u_{i,j+1} - u_{ij} &\sim \text{Laplace}(0, \delta^{-1}), \quad i = 1, \ldots, n,\ j = 1, \ldots, n-1 \end{aligned}$$

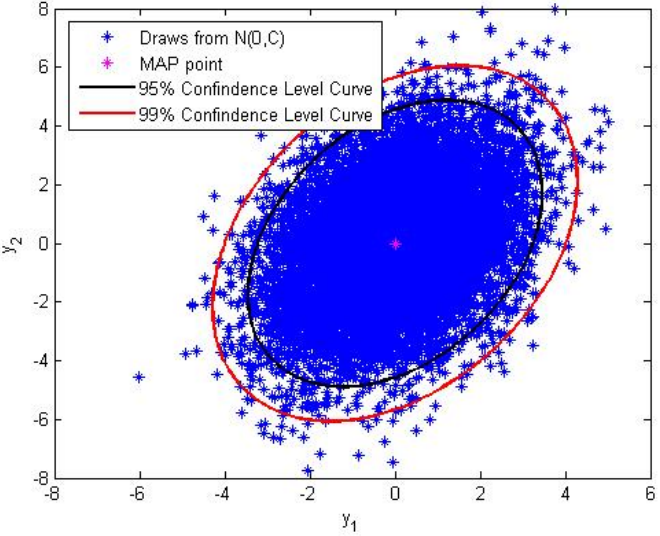
Then the density function for $\mathbf{u}$ has the form

$$\begin{aligned} p(\mathbf{u}|\delta) &\propto \exp\left(-\delta \sum_{i=1}^{n-1} \sum_{j=1}^{n} |u_{i+1,j} - u_{ij}|\right) \times \\ &\qquad \exp\left(-\delta \sum_{i=1}^{n} \sum_{j=1}^{n-1} |u_{i,j+1} - u_{ij}|\right) \\ &= \exp\left(-\delta(\|\mathbf{D}_h \mathbf{u}\|_1 + \|\mathbf{D}_v \mathbf{u}\|_1)\right), \end{aligned}$$

where $\mathbf{D}_h = \mathbf{I} \otimes \mathbf{D}$ and $\mathbf{D}_v = \mathbf{D} \otimes \mathbf{I}$.

Very similar (computationally & theoretically) to TV.

Sampling vs. Computing the MAP



Extending the Bayesian Connection

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

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

is the Bayesian posterior

Extending the Bayesian Connection

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

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

is the Bayesian posterior, where

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

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

$$\begin{aligned} p(\lambda) &\propto \lambda^{\alpha_\lambda - 1} \exp(-\beta_\lambda \lambda) \\ p(\delta) &\propto \delta^{\alpha_\delta - 1} \exp(-\beta_\delta \delta), \end{aligned}$$

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

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

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

The Full Posterior Distribution

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

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

By conjugacy, each conditional distribution lives in the same family as the prior/hyper-prior distribution:

$$\begin{aligned}
\mathbf{u}|\lambda, \delta, \mathbf{y} &\sim N\left((\lambda\mathbf{A}^T\mathbf{A}+\delta\mathbf{L})^{-1}\lambda\mathbf{A}^T\mathbf{y}, (\lambda\mathbf{A}^T\mathbf{A}+\delta\mathbf{L})^{-1}\right),\\
\lambda|\mathbf{u}, \delta, \mathbf{y} &\sim \Gamma\left(n/2+\alpha_\lambda, \frac{1}{2}\|\mathbf{A}\mathbf{u}-\mathbf{y}\|^2+\beta_\lambda\right),\\
\delta|\mathbf{u}, \lambda, \mathbf{y} &\sim \Gamma\left(n/2+\alpha_\delta, \frac{1}{2}\mathbf{u}^T\mathbf{L}\mathbf{u}+\beta_\delta\right).
\end{aligned}$$

An MCMC Method for sampling from $p(\mathbf{u}, \lambda, \delta|\mathbf{y})$

We could compute the MAP, but instead, let's sample from the posterior

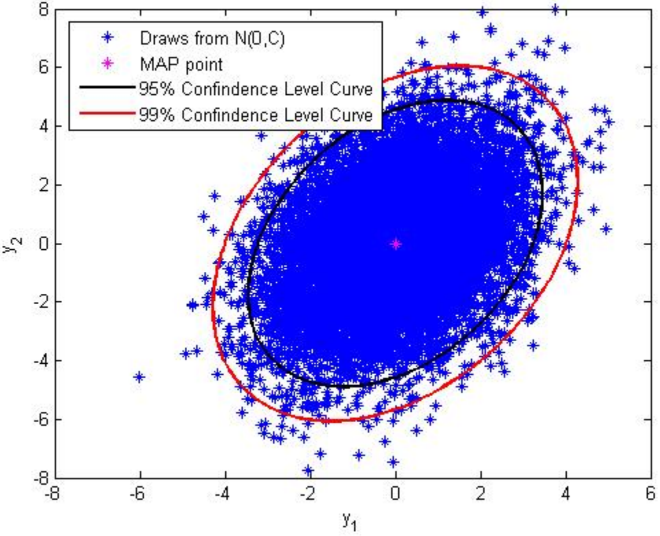
An MCMC Method for Sampling from $p(\mathbf{u}, \delta, \lambda|\mathbf{y})$.

0. δ_0, and λ_0, and set $k = 0$;
1. Sample $\mathbf{u}|\lambda_k, \delta_k, \mathbf{y}$:

$$\mathbf{u}^k \sim N\left((\lambda_k \mathbf{A}^T\mathbf{A} + \delta_k \mathbf{L})^{-1}\lambda_k \mathbf{A}^T\mathbf{y}, (\lambda_k \mathbf{A}^T\mathbf{A} + \delta_k \mathbf{L})^{-1}\right);$$

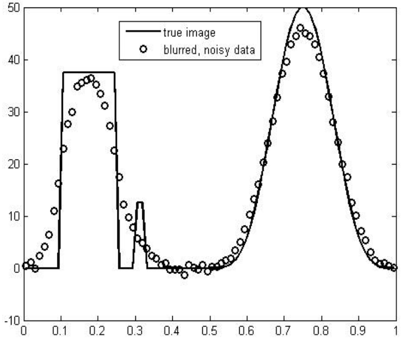
2. Sample $\lambda|\mathbf{u}_k, \delta_k, \mathbf{y}$:
   $\lambda_{k+1} \sim \Gamma\left(n/2 + \alpha_\lambda, \frac{1}{2}\|\mathbf{A}\mathbf{u}^k - \mathbf{y}\|^2 + \beta_\lambda\right)$;
3. Sample $\delta|\mathbf{u}_k, \lambda_k, \mathbf{y}$: $\delta_{k+1} \sim \Gamma\left(n/2 + \alpha_\delta, \frac{1}{2}(\mathbf{u}^k)^T\mathbf{L}\mathbf{u}^k + \beta_\delta\right)$;
4. Set $k = k + 1$ and return to Step 1.

Sampling vs. Computing the MAP

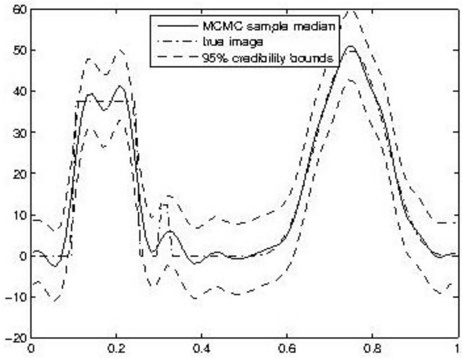


A One-dimensional example

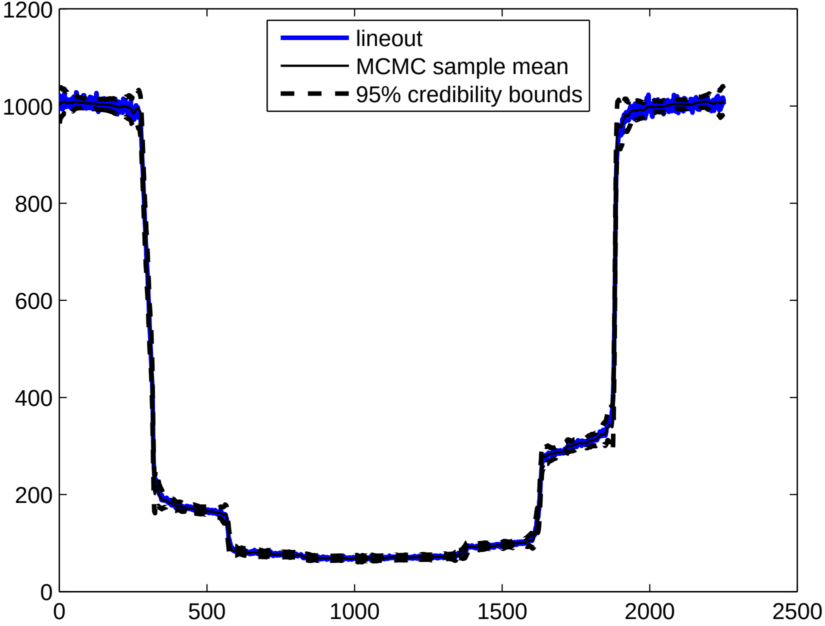
True Image and Blurred, Noisy Data



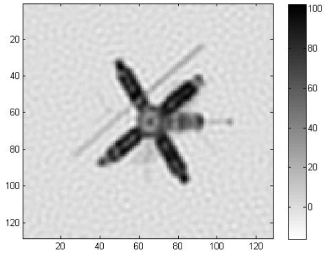
Mean and 95% Confidence Images

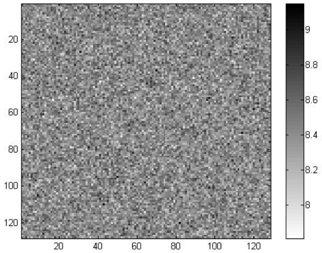


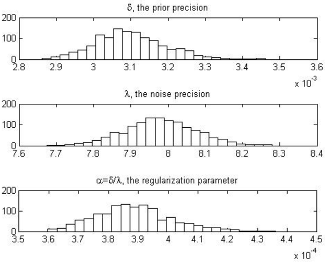
An example from X-ray Radiography (w/ Luttman)



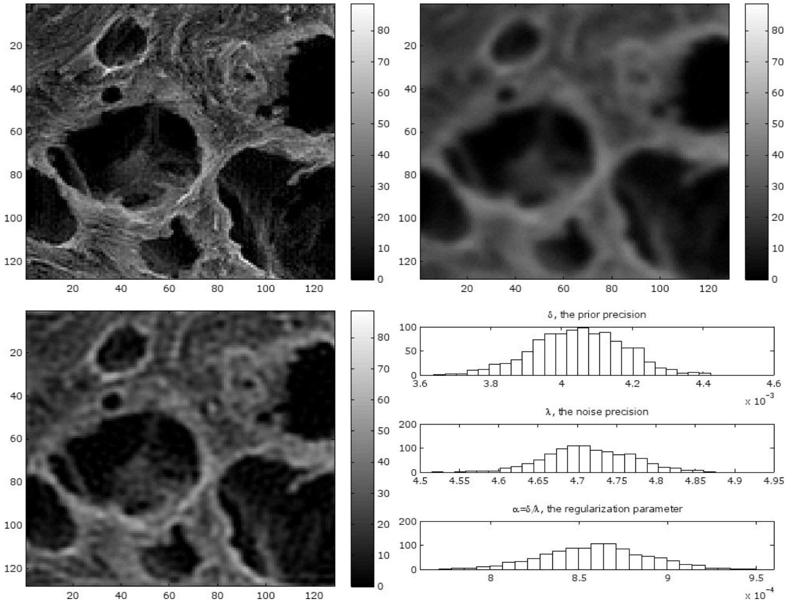
Deblurring with periodic boundary conditions







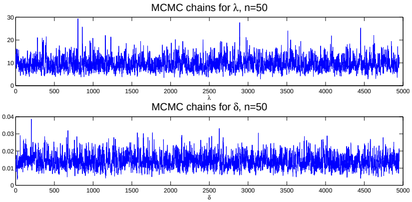
Deblurring with Neumann boundary conditions

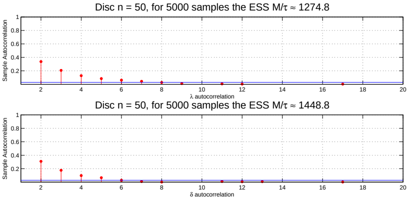


But wait! Correlation in the δ-chain increases as $n \to \infty$

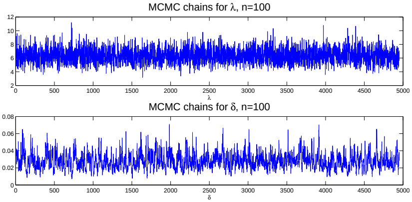
Work with S. Agapiou, O. Papaspiliopoulos, & Andrew Stuart

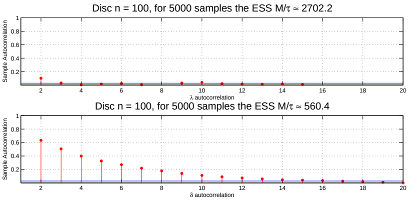
$$n = 50$$





$$n = 100$$

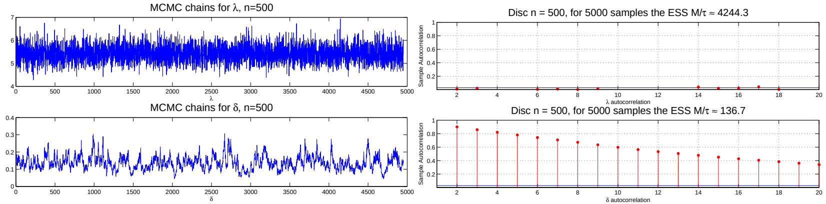




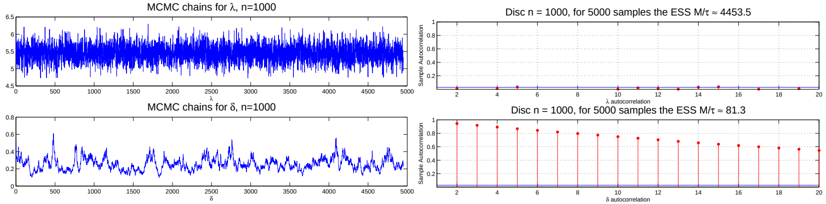
But wait! Correlation in the δ-chain increases as $n \to \infty$

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

$$n = 500$$



$$n = 1000$$



One fix, Marginalization: $p(\lambda, \delta|\mathbf{y}) = \int_{\mathbb{R}^n} p(\mathbf{u}, \lambda, \delta|\mathbf{y}) d\mathbf{u}$

Thanks Sergios!

$$
\begin{aligned}
p(\mathbf{u}, \lambda, \delta|\mathbf{y}) \quad \propto \quad & \exp\left(-\frac{1}{2}(\mathbf{u}-\mathbf{m})^T(\lambda \mathbf{A}^T\mathbf{A} + \delta \mathbf{L})(\mathbf{u}-\mathbf{m})\right) \times \\
& \lambda^{n/2+\alpha}\delta^{n/2+\alpha} \exp\left(-\beta\lambda - \beta\delta\right) \times \\
& \exp\left(-\frac{1}{2}\mathbf{y}^T(\lambda \mathbf{I} - \lambda^2 \mathbf{A}(\lambda \mathbf{A}^T\mathbf{A} + \delta \mathbf{L})^{-1}\mathbf{A}^T)\mathbf{y}\right).
\end{aligned}
$$

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

One fix, Marginalization: $p(\lambda, \delta|\mathbf{y}) = \int_{\mathbb{R}^n} p(\mathbf{u}, \lambda, \delta|\mathbf{y}) d\mathbf{u}$

Thanks Sergios!

$$\begin{aligned} p(\mathbf{u}, \lambda, \delta|\mathbf{y}) \quad \propto \quad & \exp\left(-\frac{1}{2}(\mathbf{u}-\mathbf{m})^T(\lambda\mathbf{A}^T\mathbf{A}+\delta\mathbf{L})(\mathbf{u}-\mathbf{m})\right) \times \\ & \lambda^{n/2+\alpha}\delta^{n/2+\alpha}\exp\left(-\beta\lambda-\beta\delta\right) \times \\ & \exp\left(-\frac{1}{2}\mathbf{y}^T(\lambda\mathbf{I}-\lambda^2\mathbf{A}(\lambda\mathbf{A}^T\mathbf{A}+\delta\mathbf{L})^{-1}\mathbf{A}^T)\mathbf{y}\right). \end{aligned}$$

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

$$\begin{aligned} p(\lambda, \delta|\mathbf{y}) \quad \propto \quad & |\det(\lambda\mathbf{A}^T\mathbf{A}+\delta\mathbf{L})|^{1/2}\lambda^{n/2+\alpha}\delta^{n/2+\alpha}\exp\left(-\beta\lambda-\beta\delta\right) \times \\ & \exp\left(-\frac{1}{2}\mathbf{y}^T(\lambda\mathbf{I}-\lambda^2\mathbf{A}(\lambda\mathbf{A}^T\mathbf{A}+\delta\mathbf{L})^{-1}\mathbf{A}^T)\mathbf{y}\right), \end{aligned}$$

One fix, Marginalization: $p(\lambda, \delta|\mathbf{y}) = \int_{\mathbb{R}^n} p(\mathbf{u}, \lambda, \delta|\mathbf{y}) d\mathbf{u}$

Thanks Sergios!

$$\begin{aligned} p(\mathbf{u}, \lambda, \delta|\mathbf{y}) \quad \propto \quad & \exp\left(-\frac{1}{2}(\mathbf{u}-\mathbf{m})^T(\lambda \mathbf{A}^T\mathbf{A} + \delta \mathbf{L})(\mathbf{u}-\mathbf{m})\right) \times \\ & \lambda^{n/2+\alpha}\delta^{n/2+\alpha} \exp\left(-\beta\lambda - \beta\delta\right) \times \\ & \exp\left(-\frac{1}{2}\mathbf{y}^T(\lambda \mathbf{I} - \lambda^2 \mathbf{A}(\lambda \mathbf{A}^T\mathbf{A} + \delta \mathbf{L})^{-1}\mathbf{A}^T)\mathbf{y}\right). \end{aligned}$$

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

$$\begin{aligned} p(\lambda, \delta|\mathbf{y}) \quad \propto \quad & |\det(\lambda \mathbf{A}^T\mathbf{A} + \delta \mathbf{L})|^{1/2}\lambda^{n/2+\alpha}\delta^{n/2+\alpha} \exp\left(-\beta\lambda - \beta\delta\right) \times \\ & \exp\left(-\frac{1}{2}\mathbf{y}^T(\lambda \mathbf{I} - \lambda^2 \mathbf{A}(\lambda \mathbf{A}^T\mathbf{A} + \delta \mathbf{L})^{-1}\mathbf{A}^T)\mathbf{y}\right), \end{aligned}$$

and hence,

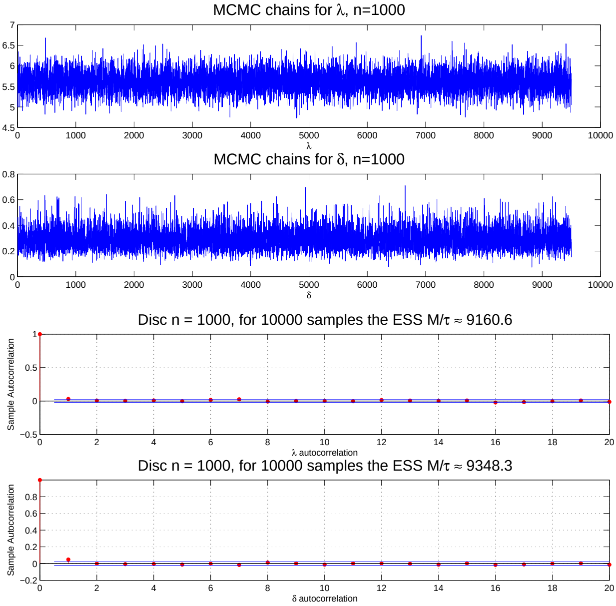
$$\begin{aligned} p(\delta|\mathbf{y}, \lambda) \quad \propto \quad & |\det(\lambda \mathbf{A}^T\mathbf{A} + \delta \mathbf{L})|^{1/2}\delta^{n/2+\alpha} \times \\ & \exp\left(\frac{\lambda^2}{2}\mathbf{y}^T\mathbf{A}(\lambda \mathbf{A}^T\mathbf{A} + \delta \mathbf{L})^{-1}\mathbf{A}^T\mathbf{y} - \beta\delta\right). \end{aligned}$$

MCMC Method with Marginalization (w/ Kevin Joyce)

Another MCMC Method for Sampling from $p(\mathbf{u}, \delta, \lambda | \mathbf{y})$.

0. δ_0, and λ_0, and set $k = 0$;
1. Sample $\mathbf{u}|\lambda_k, \delta_k, \mathbf{y}$:
$$\mathbf{u}^k \sim N\left((\lambda_k \mathbf{A}^T\mathbf{A} + \delta_k \mathbf{L})^{-1}\lambda_k \mathbf{A}^T\mathbf{y}, (\lambda_k \mathbf{A}^T\mathbf{A} + \delta_k \mathbf{L})^{-1}\right);$$
2. Sample $\lambda|\mathbf{u}_k, \delta_k, \mathbf{y}$:
$\lambda_{k+1} \sim \Gamma\left(n/2 + \alpha_\lambda, \frac{1}{2}\|\mathbf{A}\mathbf{u}^k - \mathbf{y}\|^2 + \beta_\lambda\right)$;
3. Sample $\delta|\lambda_k, \mathbf{y}$ (e.g., using Metropolis-Hastings):
$$\begin{aligned} p(\delta|\mathbf{y}, \lambda_k) \quad &\propto \quad |\det(\lambda_k \mathbf{A}^T\mathbf{A} + \delta \mathbf{L})|^{1/2}\delta^{n/2+\alpha}\times \\ &\exp\left(\frac{\lambda_k^2}{2}\mathbf{y}^T\mathbf{A}(\lambda_k \mathbf{A}^T\mathbf{A} + \delta \mathbf{L})^{-1}\mathbf{A}^T\mathbf{y} - \beta\delta\right); \end{aligned}$$
4. Set $k = k + 1$ and return to Step 1.

Chain and auto-correlation plots of λ and δ



Computational bottleneck: computing $\mathbf{u}^k \sim N\left((\lambda_k \mathbf{A}^T\mathbf{A} + \delta_k \mathbf{L})^{-1}\lambda_k \mathbf{A}^T\mathbf{y}, (\lambda_k \mathbf{A}^T\mathbf{A} + \delta_k \mathbf{L})^{-1}\right)$

The conditional density for $\mathbf{u}|\lambda_k, \delta_k, \mathbf{y}$ can be written in the form

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

Computational bottleneck: computing $\mathbf{u}^k \sim N\left((\lambda_k \mathbf{A}^T\mathbf{A} + \delta_k \mathbf{L})^{-1}\lambda_k \mathbf{A}^T\mathbf{y}, (\lambda_k \mathbf{A}^T\mathbf{A} + \delta_k \mathbf{L})^{-1}\right)$

The conditional density for $\mathbf{u}|\lambda_k, \delta_k, \mathbf{y}$ can be written in the form

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

From here on out, we define:

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

Note then that

$$\frac{1}{2}\|\mathbf{J}_k\mathbf{u} - \bar{\mathbf{y}}_k\|^2 = \frac{\lambda_k}{2}\|\mathbf{A}\mathbf{u} - \mathbf{y}\|^2 + \frac{\delta_k}{2}\mathbf{u}^T\mathbf{L}\mathbf{u}.$$

Computational bottleneck: computing $\mathbf{u}^k \sim N\left((\lambda_k \mathbf{A}^T\mathbf{A} + \delta_k \mathbf{L})^{-1}\lambda_k \mathbf{A}^T\mathbf{y}, (\lambda_k \mathbf{A}^T\mathbf{A} + \delta_k \mathbf{L})^{-1}\right)$

For large-scale problems, you can use optimization

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

Computational bottleneck: computing $\mathbf{u}^k \sim N\left((\lambda_k \mathbf{A}^T\mathbf{A} + \delta_k \mathbf{L})^{-1}\lambda_k \mathbf{A}^T\mathbf{y}, (\lambda_k \mathbf{A}^T\mathbf{A} + \delta_k \mathbf{L})^{-1}\right)$

For large-scale problems, you can use optimization

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

Note $\mathbf{u}^k$ is a random variable defined by

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

Computational bottleneck: computing $\mathbf{u}^k \sim N\left((\lambda_k \mathbf{A}^T\mathbf{A} + \delta_k \mathbf{L})^{-1}\lambda_k \mathbf{A}^T\mathbf{y}, (\lambda_k \mathbf{A}^T\mathbf{A} + \delta_k \mathbf{L})^{-1}\right)$

For large-scale problems, you can use optimization

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

Note $\mathbf{u}^k$ is a random variable defined by

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

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

then

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

Proof that $\mathbf{u}^k \sim N\left((\lambda_k \mathbf{A}^T\mathbf{A} + \delta_k \mathbf{L})^{-1}\lambda_k \mathbf{A}^T\mathbf{y}, (\lambda_k \mathbf{A}^T\mathbf{A} + \delta_k \mathbf{L})^{-1}\right)$

What we know:

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

Proof that $\mathbf{u}^k \sim N\left((\lambda_k \mathbf{A}^T\mathbf{A} + \delta_k \mathbf{L})^{-1}\lambda_k \mathbf{A}^T\mathbf{y}, (\lambda_k \mathbf{A}^T\mathbf{A} + \delta_k \mathbf{L})^{-1}\right)$

What we know:

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

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

Proof that $\mathbf{u}^k \sim N\left((\lambda_k \mathbf{A}^T\mathbf{A} + \delta_k \mathbf{L})^{-1}\lambda_k \mathbf{A}^T\mathbf{y}, (\lambda_k \mathbf{A}^T\mathbf{A} + \delta_k \mathbf{L})^{-1}\right)$

What we know:

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

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

Thus as desired, we have

$$
\begin{aligned}
\mathbf{u}^k &\sim \mathcal{N}\left((\mathbf{J}_k^T \mathbf{J}_k)^{-1} \mathbf{J}_k^T \bar{\mathbf{y}}_k, (\mathbf{J}_k^T \mathbf{J}_k)^{-1}\right) \\
&\overset{\text{dist}}{=} \mathcal{N}\left((\lambda_k \mathbf{A}^T\mathbf{A} + \delta_k \mathbf{L})^{-1}\lambda_k \mathbf{A}^T\mathbf{y}, (\lambda_k \mathbf{A}^T\mathbf{A} + \delta_k \mathbf{L})^{-1}\right).
\end{aligned}
$$

Now Consider a Nonlinear Statistical Model

Now assume the non-linear, Gaussian statistical model

$$\mathbf{y} = \mathbf{A}(\mathbf{u}) + \boldsymbol{\epsilon},$$

where

- $\mathbf{y} \in \mathbb{R}^m$ is the vector of observations;
- $\mathbf{u} \in \mathbb{R}^n$ is the vector of unknown parameters;
- $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^m$, with $m \geq n$, is nonlinear;
- $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \lambda^{-1}\mathbf{I})$, i.e., $\boldsymbol{\epsilon}$ is i.i.d. Gaussian with mean 0 and variance $\lambda^{-1} = \sigma^2$.

Assume λ (and later δ) are known,
so we will only discuss sampling $\mathbf{u}$!

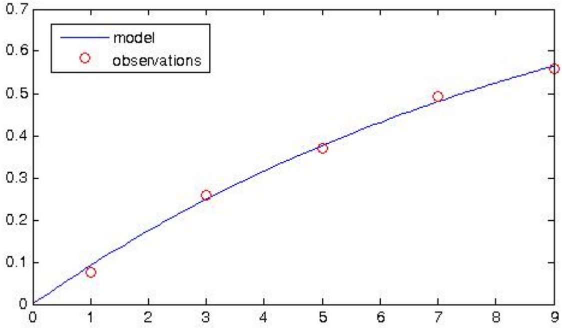
Toy example

Consider the following nonlinear, two-parameter **pre-whitened** model.

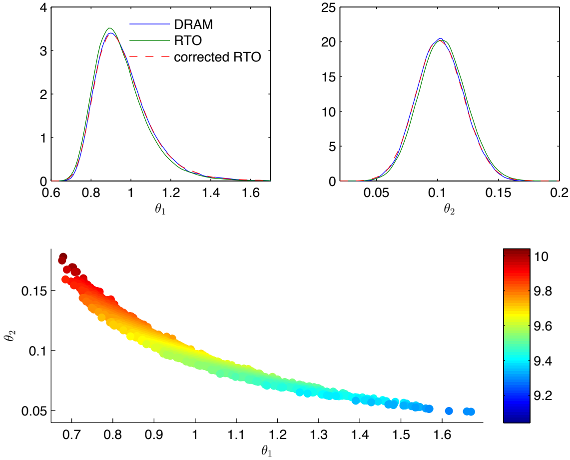
$$y_i = u_1(1 - \exp(-u_2 x_i)) + \epsilon, \quad \epsilon \sim N(0, \sigma^2), \quad i = 1, 2, 3, 4, 5,$$

with $x_i = 2i - 1$, $\sigma = 0.0136$, and $\mathbf{y} = [.076, .258, .369, .492, .559]$.

GOAL: estimate a probability distribution for $\mathbf{u} = (u_1, u_2)$.



Toy example continued: the Bayesian posterior $p(\theta_1, \theta_2 | \mathbf{y})$



What is the Posterior Density Function $p(\mathbf{u}|\mathbf{y})$?

The likelihood function has the form

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

Given the prior $p(\mathbf{u})$, the posterior density then has the form

$$p(\mathbf{u}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{u})p(\mathbf{u}).$$

What is the Posterior Density Function $p(\mathbf{u}|\mathbf{y})$?

The likelihood function has the form

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

Given the prior $p(\mathbf{u})$, the posterior density then has the form

$$p(\mathbf{u}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{u})p(\mathbf{u}).$$

Assumption: we assume that the posterior has least squares form

$$p(\mathbf{u}|\mathbf{y}) \propto \exp\left(-\frac{1}{2}\|\bar{\mathbf{A}}(\mathbf{u}) - \bar{\mathbf{y}}\|^2\right).$$

1. If $p(\mathbf{u})$ is uniform, $\bar{\mathbf{A}} = \lambda^{1/2}\mathbf{A}$ and $\bar{\mathbf{y}} = \lambda^{1/2}\mathbf{y}$.

What is the Posterior Density Function $p(\mathbf{u}|\mathbf{y})$?

The likelihood function has the form

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

Given the prior $p(\mathbf{u})$, the posterior density then has the form

$$p(\mathbf{u}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{u})p(\mathbf{u}).$$

Assumption: we assume that the posterior has least squares form

$$p(\mathbf{u}|\mathbf{y}) \propto \exp\left(-\frac{1}{2}\|\bar{\mathbf{A}}(\mathbf{u})-\bar{\mathbf{y}}\|^2\right).$$

1. If $p(\mathbf{u})$ is uniform, $\bar{\mathbf{A}} = \lambda^{1/2}\mathbf{A}$ and $\bar{\mathbf{y}} = \lambda^{1/2}\mathbf{y}$.
2. If $p(\mathbf{u}) \propto \exp\left(-\frac{\delta}{2}(\mathbf{u}-\mathbf{u}_0)^T\mathbf{L}(\mathbf{u}-\mathbf{u}_0)\right)$, then
   $\bar{\mathbf{A}}(\mathbf{u}) = \begin{bmatrix} \lambda^{1/2}\mathbf{A}(\mathbf{u}) \\ (\delta\mathbf{L})^{1/2}\mathbf{u} \end{bmatrix}$ and $\bar{\mathbf{y}} = \begin{bmatrix} \lambda^{1/2}\mathbf{y} \\ (\delta\mathbf{L})^{1/2}\mathbf{u}_0 \end{bmatrix}$.

Extension of optimization-based approach to nonlinear problems: Randomized maximum likelihood

An obvious extension to nonlinear problems is as follows:

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

Comment: This is *randomized maximum likelihood*.

Problem: It is an open question what the probability of $\mathbf{u}$ is.

One solution: Tan Bui-Thanh, Ghattas, et. al., "A Randomized MAP Algorithm for Large-Scale Bayesian Inverse Problems."

Extension to nonlinear problems

As in the linear case, we create a nonlinear mapping

$$\mathbf{u} = \mathbf{F}^{-1}(\mathbf{v}), \quad \mathbf{v} \sim \mathcal{N}(\mathbf{Q}^T \bar{\mathbf{y}}, \mathbf{I}).$$

Extension to nonlinear problems

As in the linear case, we create a nonlinear mapping

$$\mathbf{u} = \mathbf{F}^{-1}(\mathbf{v}), \quad \mathbf{v} \sim \mathcal{N}(\mathbf{Q}^T \bar{\mathbf{y}}, \mathbf{I}).$$

First, define

$$\mathbf{u}_{\mathrm{MAP}} = \arg\min_{\mathbf{u}} \|\bar{\mathbf{A}}(\mathbf{u}) - \bar{\mathbf{y}}\|^2,$$

then first-order optimality yields

$$\mathbf{J}(\mathbf{u}_{\mathrm{MAP}})^T (\bar{\mathbf{A}}(\mathbf{u}_{\mathrm{MAP}}) - \bar{\mathbf{y}}) = \mathbf{0}.$$

Nonlinear Mapping, Cont. (w/ Solonen)

$\mathbf{u}_{\mathrm{MAP}}$ is a solution of the nonlinear equation

$$\mathbf{J}(\mathbf{u}_{\mathrm{MAP}})^T \bar{\mathbf{A}}(\mathbf{u}) = \mathbf{J}(\mathbf{u}_{\mathrm{MAP}})^T \bar{\mathbf{y}}.$$

Nonlinear Mapping, Cont. (w/ Solonen)

$\mathbf{u}_{\text{MAP}}$ is a solution of the nonlinear equation

$$\mathbf{J}(\mathbf{u}_{\text{MAP}})^T \bar{\mathbf{A}}(\mathbf{u}) = \mathbf{J}(\mathbf{u}_{\text{MAP}})^T \bar{\mathbf{y}}.$$

QR-rewrite: this equation can be equivalently expressed

$$\mathbf{Q}^T \bar{\mathbf{A}}(\mathbf{u}) = \mathbf{Q}^T \bar{\mathbf{y}},$$

where $\mathbf{J}(\mathbf{u}_{\text{MAP}}) = [\mathbf{Q}, \bar{\mathbf{Q}}]\mathbf{R}$, with $\mathbf{Q} \in \mathbb{R}^{m \times n}$, $\bar{\mathbf{Q}} \in \mathbb{R}^{m \times (m-n)}$, and $\mathbf{R} \in \mathbb{R}^{m \times n}$ upper-triangular.

Nonlinear Mapping, Cont. (w/ Solonen)

$\mathbf{u}_{\mathrm{MAP}}$ is a solution of the nonlinear equation

$$\mathbf{J}(\mathbf{u}_{\mathrm{MAP}})^T \bar{\mathbf{A}}(\mathbf{u}) = \mathbf{J}(\mathbf{u}_{\mathrm{MAP}})^T \bar{\mathbf{y}}.$$

QR-rewrite: this equation can be equivalently expressed

$$\mathbf{Q}^T \bar{\mathbf{A}}(\mathbf{u}) = \mathbf{Q}^T \bar{\mathbf{y}},$$

where $\mathbf{J}(\mathbf{u}_{\mathrm{MAP}}) = [\mathbf{Q}, \bar{\mathbf{Q}}]\mathbf{R}$, with $\mathbf{Q} \in \mathbb{R}^{m\times n}$, $\bar{\mathbf{Q}} \in \mathbb{R}^{m\times(m-n)}$, and $\mathbf{R} \in \mathbb{R}^{m\times n}$ upper-triangular.

Nonlinear mapping: define $\mathbf{F} \stackrel{\text{def}}{=} \mathbf{Q}^T \bar{\mathbf{A}}$ and

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

PDF for $\mathbf{u} = \mathbf{F}^{-1}(\mathbf{v})$, $\mathbf{v} \sim \mathcal{N}(\mathbf{Q}^T\bar{\mathbf{y}}, \mathbf{I})$

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

PDF for $\mathbf{u} = \mathbf{F}^{-1}(\mathbf{v})$, $\mathbf{v} \sim \mathcal{N}(\mathbf{Q}^T\bar{\mathbf{y}}, \mathbf{I})$

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

Next we need $\frac{d}{d\mathbf{u}}\mathbf{F}(\mathbf{u}) \in \mathbb{R}^{n\times n}$ to be invertible. Then

$$\begin{aligned} p_{\mathbf{u}_{\mathrm{MAP}}}(\mathbf{u}) &\propto \left|\det\left(\frac{d}{d\mathbf{u}}\mathbf{F}(\mathbf{u})\right)\right| p_{\mathbf{v}}(\mathbf{F}(\mathbf{u})) \\ &= \left|\det\left(\mathbf{Q}^T\mathbf{J}(\mathbf{u})\right)\right| \exp\left(-\frac{1}{2}\|\mathbf{Q}^T(\bar{\mathbf{A}}(\mathbf{u}) - \bar{\mathbf{y}})\|^2\right) \end{aligned}$$

PDF for $\mathbf{u} = \mathbf{F}^{-1}(\mathbf{v})$, $\mathbf{v} \sim \mathcal{N}(\mathbf{Q}^T\bar{\mathbf{y}}, \mathbf{I})$

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

Next we need $\frac{d}{d\mathbf{u}}\mathbf{F}(\mathbf{u}) \in \mathbb{R}^{n\times n}$ to be invertible. Then

$$\begin{aligned}
p_{\mathbf{u}_{\mathrm{MAP}}}(\mathbf{u}) &\propto \left|\det\left(\frac{d}{d\mathbf{u}}\mathbf{F}(\mathbf{u})\right)\right| p_{\mathbf{v}}(\mathbf{F}(\mathbf{u})) \\
&= \left|\det\left(\mathbf{Q}^T\mathbf{J}(\mathbf{u})\right)\right| \exp\left(-\frac{1}{2}\|\mathbf{Q}^T(\bar{\mathbf{A}}(\mathbf{u}) - \bar{\mathbf{y}})\|^2\right) \\
&= \left|\det\left(\mathbf{Q}^T\mathbf{J}(\mathbf{u})\right)\right| \exp\left(\frac{1}{2}\|\bar{\mathbf{Q}}^T(\bar{\mathbf{A}}(\mathbf{u}) - \bar{\mathbf{y}})\|^2\right) \\
&\qquad \exp\left(-\frac{1}{2}\|\bar{\mathbf{A}}(\mathbf{u}) - \bar{\mathbf{y}}\|^2\right) \\
&= c(\mathbf{u})p(\mathbf{u}|\mathbf{y}).
\end{aligned}$$

Theorem (RTO probability density)

Let $\bar{\mathbf{A}} : \mathbb{R}^n \to \mathbb{R}^m$, $\bar{\mathbf{y}} \in \mathbb{R}^m$, *and assume*

- $\bar{\mathbf{A}}$ *is continuously differentiable;*
- $\mathbf{J}(\mathbf{u}) \in \mathbb{R}^{m\times n}$ *is rank* n *for every* $\mathbf{u}$*;*
- $\mathbf{J}(\mathbf{u}_{\mathrm{MAP}}) = \mathbf{QR}$ *and* $\mathbf{F} \stackrel{\mathrm{def}}{=} \mathbf{Q}^T \bar{\mathbf{A}}$*;*
- $\mathbf{Q}^T\mathbf{J}(\mathbf{u})$ *is invertible for all relevant* $\mathbf{u}$.

Then the random variable

$$\mathbf{u} = \mathbf{F}^{-1}(\mathbf{v}), \quad \mathbf{v} \sim \mathcal{N}(\mathbf{Q}^T\bar{\mathbf{y}}, \mathbf{I}),$$

has probability density function

$$p_{\mathbf{u}_{\mathrm{MAP}}}(\mathbf{u}) \propto c(\mathbf{u})p(\mathbf{u}|\mathbf{y}),$$

where

$$c(\mathbf{u}) = \left|\det(\mathbf{Q}^T\mathbf{J}(\mathbf{u}))\right| \exp\left(\frac{1}{2}\|\bar{\mathbf{Q}}^T(\bar{\mathbf{y}} - \bar{\mathbf{A}}(\mathbf{u}))\|^2\right).$$

Interpretation

Note that

$$\underbrace{e^{-\frac{1}{2}\|\bar{\mathbf{f}}(\boldsymbol{\theta})-\bar{\mathbf{y}}\|^2}}_{\text{the posterior density}} = \underbrace{e^{-\frac{1}{2}\|\mathbf{Q}^T(\bar{\mathbf{f}}(\boldsymbol{\theta})-\bar{\mathbf{y}})\|^2}}_{\text{the part we capture}} \times \underbrace{e^{-\frac{1}{2}\|\bar{\mathbf{Q}}^T(\bar{\mathbf{f}}(\boldsymbol{\theta})-\bar{\mathbf{y}})\|^2}}_{\text{the part we miss}},$$

and the RTO density looks like

$$p_{\boldsymbol{\theta}_{\mathrm{MAP}}}(\boldsymbol{\theta}) \propto \underbrace{\left|\mathbf{Q}^T\mathbf{J}(\boldsymbol{\theta})\right| \exp\left(\frac{1}{2}\|\bar{\mathbf{Q}}^T\bar{\mathbf{y}} - \bar{\mathbf{Q}}^T\bar{\mathbf{f}}(\boldsymbol{\theta})\|^2\right)}_{\text{captures the deviation from the posterior}} \underbrace{p(\boldsymbol{\theta}|\mathbf{y})}_{\text{posterior}}.$$

Optimization Comes In During Implementation

Randomize-the-Optimize for computing $\mathbf{u} = \mathbf{F}^{-1}(\mathbf{v})$

1. Compute a random draw from $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.
2. Solve

$$\mathbf{u} \;=\; \arg\min_{\boldsymbol{\psi}} \left\{ \ell(\boldsymbol{\psi}, \boldsymbol{\epsilon}) \stackrel{\text{def}}{=} \|\mathbf{Q}^T(\mathbf{A}(\boldsymbol{\psi}) - (\bar{\mathbf{y}} + \boldsymbol{\epsilon}))\|^2 \right\}.$$

3. Reject $\mathbf{u}$ when $\mathbf{Q}^T(\bar{\mathbf{y}} + \boldsymbol{\epsilon})$ is not in the range of $\mathbf{Q}^T\mathbf{A}(\cdot)$.

NOTE the presence of Q in the least squares!

Metropolis-Hastings using RTO

Given $\mathbf{u}^{k-1}$ and proposal $\mathbf{u}^* \sim p_{\mathbf{u}_{\text{MAP}}}(\mathbf{u})$, accept with probability

$$\begin{aligned} r &= \min\left(1, \frac{p(\mathbf{u}^*|\mathbf{y})p_{\mathbf{u}_{\text{MAP}}}(\mathbf{u}^{k-1})}{p(\mathbf{u}^{k-1}|\mathbf{y})p_{\mathbf{u}_{\text{MAP}}}(\mathbf{u}^*)}\right) \\ &= \min\left(1, \frac{p(\mathbf{u}^*|\mathbf{y})c(\mathbf{u}^{k-1})p(\mathbf{u}^{k-1}|\mathbf{y})}{p(\mathbf{u}^{k-1}|\mathbf{y})c(\mathbf{u}^*)p(\mathbf{u}^*|\mathbf{y})}\right) \\ &= \min\left(1, \frac{c(\mathbf{u}^{k-1})}{c(\mathbf{u}^*)}\right). \end{aligned}$$

For numerical reasons, use

$$r = \min\left(1, \exp\left(\ln c(\mathbf{u}^{k-1}) - \ln c(\mathbf{u}^*)\right)\right).$$

Metropolis-Hastings using RTO, Cont.

The RTO Metropolis-Hastings Algorithm

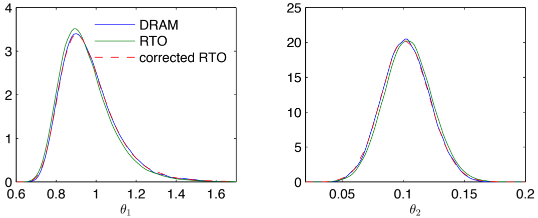
1. Choose $\mathbf{u}^0$ and number of samples N. Set $k = 1$.
2. Compute an RTO sample $\mathbf{u}^* \sim p_{\mathbf{u}_{\mathrm{MAP}}}(\mathbf{u})$.
3. Compute the acceptance probability

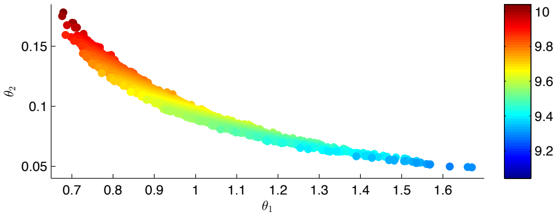
$$r = \min\left(1, \frac{c(\mathbf{u}^{k-1})}{c(\mathbf{u}^*)}\right).$$

4. Draw $u \sim U(0,1)$. If $u < r$, $\mathbf{u}^k = \mathbf{u}^*$, else $\mathbf{u}^k = \mathbf{u}^{k-1}$.
5. If $k < N$, set $k = k + 1$ and return to Step 2.

BOD, Good: $\mathbf{A}(\mathbf{u}) = u_1(1 - \exp(-u_2\mathbf{x}))$

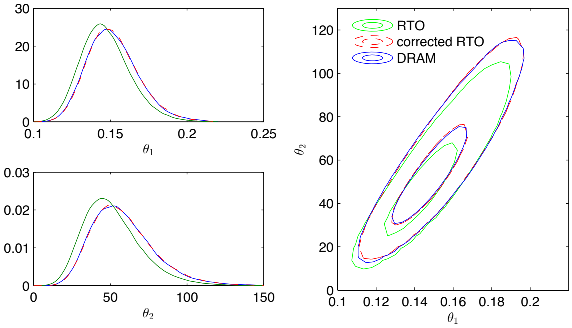
- $\mathbf{x}$ = 20 linearly spaced observations in $1 \leq x \leq 9$;
- $\mathbf{y} = \mathbf{A}(\mathbf{u}) + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ with $\sigma = 0.01$;
- $\mathbf{u} = (1, 0.1)$.



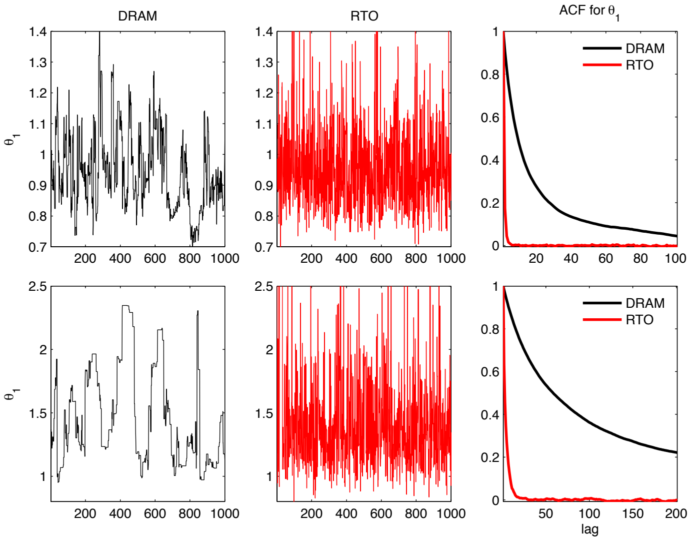


MONOD: $\mathbf{A}(\mathbf{u}) = u_1\mathbf{x}/(u_2+\mathbf{x})$ Solonen, Laine, Haario

$$
\begin{aligned}
\mathbf{x} &= (28, 55, 83, 110, 138, 225, 375) \\
\mathbf{y} &= (0.053, 0.060, 0.112, 0.105, 0.099, 0.122, 0.125).
\end{aligned}
$$



Autocorrelation plots for θ_1 for good and bad BOD



Electrical Impedance Tomography Seppänen, Solonen, Haario, Kaipio

$$
\begin{aligned}
\nabla \cdot (\theta \nabla \varphi) = 0, \quad & \vec{r} \in \Omega \\
\varphi + z_\ell \theta \frac{\partial \varphi}{\partial \vec{n}} = y_\ell, \quad & \vec{r} \in e_\ell,\ \ell = 1, \ldots, L \\
\int_{e_\ell} \theta \frac{\partial \varphi}{\partial \vec{n}} \mathrm{d}S = I_\ell, \quad & \ell = 1, \ldots, L \\
\theta \frac{\partial \varphi}{\partial \vec{n}} = 0, \quad & \vec{r} \in \partial\Omega \setminus \cup_{\ell=1}^{L} e_\ell
\end{aligned}
$$

- $\theta = \theta(\vec{r})$ & $\varphi = \varphi(\vec{r})$: electrical conductivity & potential.
- $\vec{r} \in \Omega$: spatial coordinate.
- e_ℓ: area under the ℓth electrode.
- z_ℓ: contact impedance between ℓth electrode and object.
- y_ℓ & I_ℓ: amplitudes of the electrode potential and current.
- $\vec{n}$: outward unit normal
- L: number of electrodes.

EIT, Forward/Inverse Problem

Forward Problem: Use finite elements to discretize the PDE, and compute the potential distribution φ and the electrode potential $\mathbf{y} = (y_1, \ldots, y_L)$ given the discrete conductivity distribution $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)$, the contact impedances z_ℓ, and the current pattern I_ℓ. This defines the forward mapping

$$\mathbf{y} = \mathbf{f}(\boldsymbol{\theta}).$$

Inverse Problem: Given electrode potential measurements $\mathbf{y}$, construct the posterior density

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{f}(\boldsymbol{\theta})\|^2 + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \mathbf{L}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\right).$$

Gaussian process prior

$$p(\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T \mathbf{L}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\right)$$

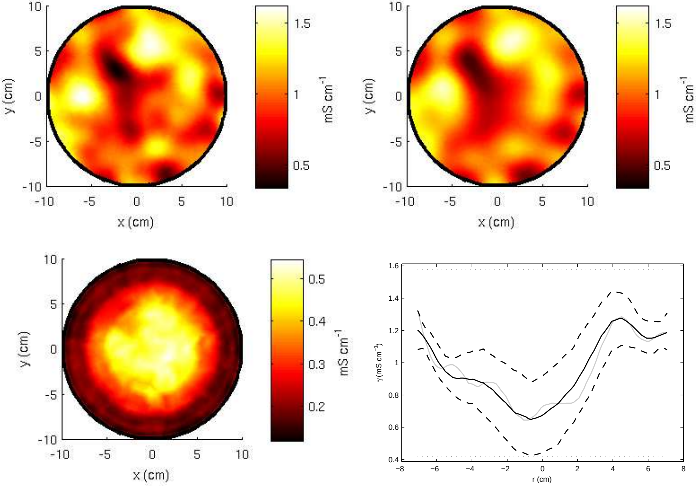
$$[\mathbf{L}^{-1}]_{ij} = a \exp\left\{-\frac{\|\vec{r}_i - \vec{r}_j\|_2^2}{2b^2}\right\} + c\delta_{ij}$$

- $\vec{r}_i = (x_i, y_i)$ is the spatial coordinate: $\theta_i = \theta(\vec{r}_i)$.
- a, b and c are parameters controlling the variance and spatial smoothness of $\boldsymbol{\theta}$.
- the *correlation length* ℓ – the distance for which $\mathbf{L}^{-1}(i,j)$ drops to 1% of the $\mathbf{L}^{-1}(i,i) = \mathrm{var}(\gamma_i)$ – is given by

$$\ell = b\sqrt{2\ln(100) - 2\ln\left(1 + c/a\right)}.$$

RTO Metropolis-Hastings applied to EIT example

True Conductivity = Realization from Smoothness Prior

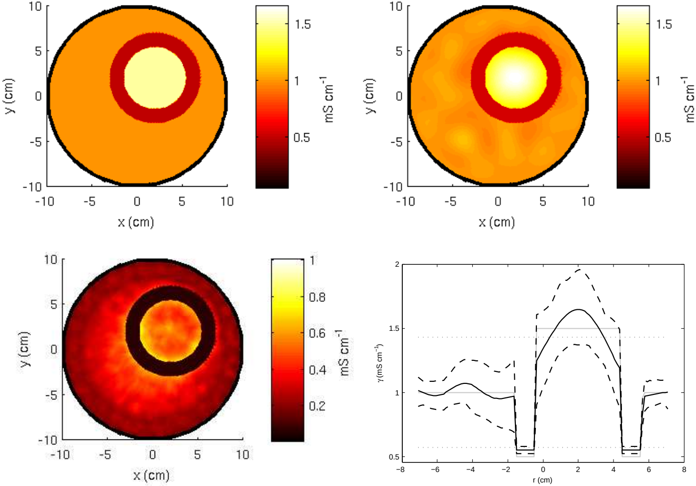


Upper images: truth & conditional mean.

Lower images: 99% c.i.'s & profiles of all of the above.

RTO Metropolis-Hastings applied to EIT example

True Conductivity = Internal Structure #1

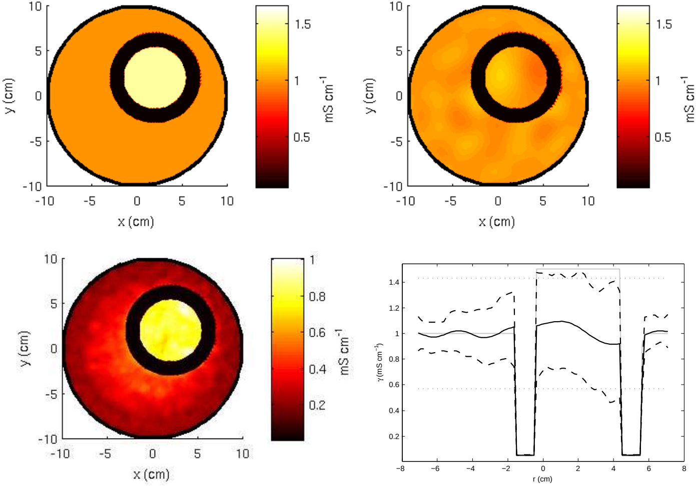


Upper images: truth & conditional mean.

Lower images: 99% c.i.'s & profiles of all of the above.

RTO Metropolis-Hastings applied to EIT example

True Conductivity = Internal Structure #2



Upper images: truth & conditional mean.

Lower images: 99% c.i.'s & profiles of all of the above.

Conclusions

OPEN QUESTIONS:

- Are we using posterior samples to measure uncertainty in the best possible way?
- How do we do uncertainty quantification for TV?
- How does RTO behave in the infinite limit?

Conclusions

OPEN QUESTIONS:

- Are we using posterior samples to measure uncertainty in the best possible way?
- How do we do uncertainty quantification for TV?
- How does RTO behave in the infinite limit?

HAVE FUN!