Sampling Hyperparameters in Hierarchical Models and Fast Inference in Linear-Gaussian Inverse Problems



J. Andrés Christen
Colin Fox, Richard A. Norton

SIAM UQ'18, Garden Grove, April 2018

Topics

- Bayesian hierarchical modeling
- Marginal posterior over hyperparameters
- Two examples:
  - Censored data
  - A linear-Gaussian inverse problem
    - Evaluating *ratio* of determinants
    - Comparison with other samplers and regularization (MTC is fastest)
- Conclusions

Norton Christen Fox *Sampling hyperparameters in hierarchical models* Comm. Stat. 2017

Fox Norton, *Fast sampling in a linear-Gaussian inverse problem* SIAM/ASA JUQ 2016

Bayesian hierarchical modeling

Observed data y depends on latent (field) x via function A

- **First stage:** model the observation y in terms of latent variables x

$$y|x, \theta \sim \pi(y|x, \theta)$$

  with uncertainty in π parameterized by θ.
  E.g. when $y|Ax$ is zero-mean Gaussian, $y|x, \theta \sim \mathrm{N}(Ax, \Sigma(\theta))$. (**likelihood**)

- **Second stage:** model latent variables x

$$x|\theta \sim \pi(x|\theta)$$

  with uncertainty in the model parameterized by θ (**prior**)

- **Third stage:** model unknown hyperparameters θ

$$\theta \sim \pi(\theta)$$

  (**hyperprior**)

Fitting the model to data

Given measured data $\boldsymbol{y}$ determine (the distribution over) unobserved quantities:

- Fit model: full posterior

$$\boldsymbol{x}, \boldsymbol{\theta} | \boldsymbol{y} \sim \pi(\boldsymbol{x}, \boldsymbol{\theta} | \boldsymbol{y}) = \pi(\boldsymbol{y} | \boldsymbol{x}, \boldsymbol{\theta}) \pi(\boldsymbol{x} | \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) / \pi(\boldsymbol{y})$$

- Estimate unknown latent variables: (marginal posterior over latent variables)

$$\boldsymbol{x} | \boldsymbol{y} \sim \int \pi(\boldsymbol{x}, \boldsymbol{\theta} | \boldsymbol{y}) \, d\boldsymbol{\theta}$$

- Or when hyperparameters are of interest (marginal posterior over hyperparameters)

$$\boldsymbol{\theta} | \boldsymbol{y} \sim \int \pi(\boldsymbol{x}, \boldsymbol{\theta} | \boldsymbol{y}) \, d\boldsymbol{x}$$

Samples $\boldsymbol{\theta} | \boldsymbol{y}$ give access to full posterior via

$$\pi(\boldsymbol{x}, \boldsymbol{\theta} | \boldsymbol{y}) = \pi(\boldsymbol{x} | \boldsymbol{\theta}, \boldsymbol{y}) \pi(\boldsymbol{\theta} | \boldsymbol{y})$$

using the *full conditional* for $\boldsymbol{x}$. (MTC)

Marginal-then-conditional sampling

Claim: When the full conditional for the latent variables $\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ has a known form, then the marginal distribution over hyperparameters $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ is available for sampling.

Follows since the θ-dependence of the normalizing constant is known.

Draw iid samples from the full posterior by:

1. Sample from the marginal posterior over $\boldsymbol{\theta}$

$$\boldsymbol{\theta} \overset{iid}{\sim} \pi(\boldsymbol{\theta}|\boldsymbol{y})$$

   usually low-dimensional, so random-walk MCMC has negligible cost e.g., t-walk.

2. Sample from the full conditional over $\boldsymbol{x}$

$$\boldsymbol{x} \sim \pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$$

to give MTC, a.k.a. composition sampling, or two-variate conditional distribution method.

Censored data

y_i is observed with right censoring, i.e., if $y_i > a$ then "observation above a" is recorded.
Let $y_1 < y_2 < \cdots < y_m$ be the uncensored observations, so $n - m$ censored observations.
Introduce latent variables x_i for the unobserved data,

$$y_i = \begin{cases} x_i & \text{if } x_i < a \\ a^+ & \text{if } x_i \geq a \end{cases}$$

Model $x_i \stackrel{iid}{\sim} \mathrm{N}(\mu, \lambda^{-1})$, with $\mu|\lambda \sim \mathrm{N}(\mu_0, k_0\lambda)$ and $\lambda \sim \mathrm{Ga}(\alpha, \beta)$. $k_0 = \alpha = 1$, $\beta = 0.1$

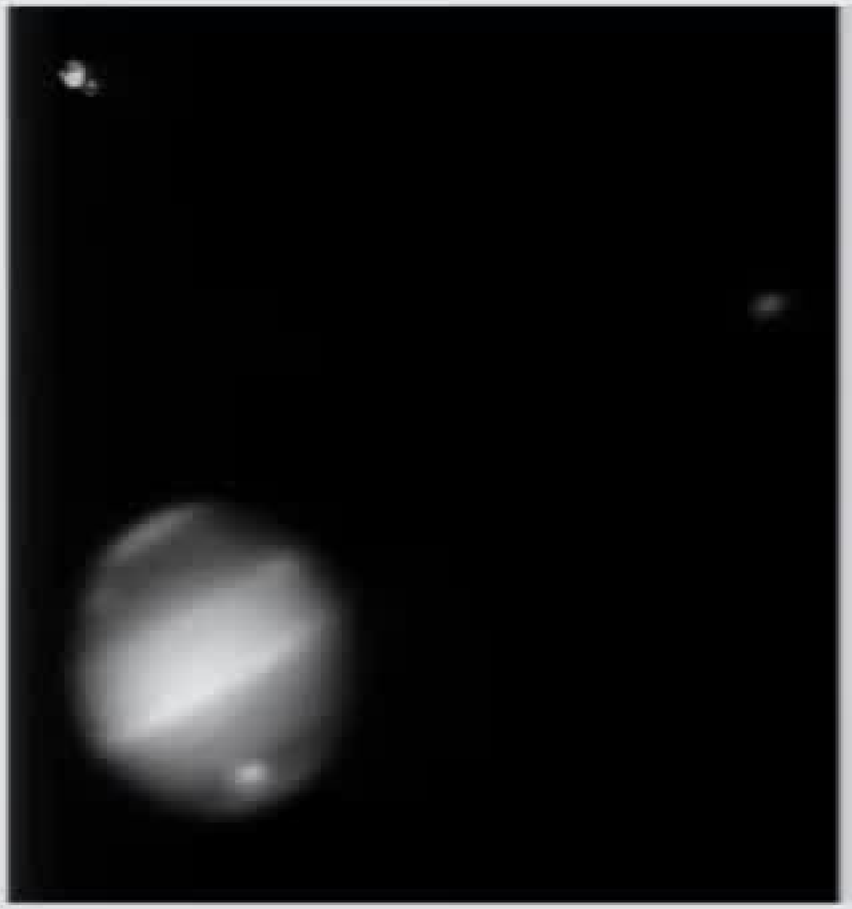
The full conditional $\mu, \lambda|\boldsymbol{y}, \boldsymbol{x}$ is a normal-gamma distribution and each $x_i|\mu, \lambda, \boldsymbol{y}$ full conditional is an iid truncated normal distribution.

Conventional approach is block-Gibbs sampling, has increasing dimension with sample size

> sample $\mu, \lambda|\boldsymbol{y}, \boldsymbol{x}$ by sampling $\lambda|\boldsymbol{y}, \boldsymbol{x} \sim \mathrm{Ga}(\alpha_2, \beta_2)$ then $\mu|\lambda, \boldsymbol{y}, \boldsymbol{x} \sim \mathrm{N}(\mu_2, k_2\lambda)$
>
> sample $x_i|\mu, \lambda, \boldsymbol{y} \sim \mathrm{N}(\mu, \lambda)$ truncated to $x_i \in [a, \infty)$ for $i = 1, \ldots, n - m$.

where $\alpha_2 = \alpha_0 + \frac{n}{2}$, $k_2 = k_0 + n$, $\mu_2 = \frac{1}{k_2}(k_0\mu_0 + \sum_{i=1}^{m} y_i + \sum_{i=1}^{n-m} x_i)$, and $\beta_2 = \beta + \frac{1}{2}(k_0\mu_0^2 - k_2\mu_2^2 + \sum_{i=1}^{m} y_i^2 + \sum_{i=1}^{n-m} x_i^2)$.

A linear Gaussian inverse problem (image deblurring)



Data $\boldsymbol{y}$ is a blurry 256×256 gray-scale photograph of Jupiter in the methane band (780nm).

Estimate the 'true' unblurry image, $\boldsymbol{x}$.

Use the satellite (upper right) as PSF $\boldsymbol{k}$, so *semi-blind* deconvolution.

$$\boldsymbol{y} = \boldsymbol{k} * \boldsymbol{x} + \eta = \boldsymbol{A}\boldsymbol{x} + \eta$$

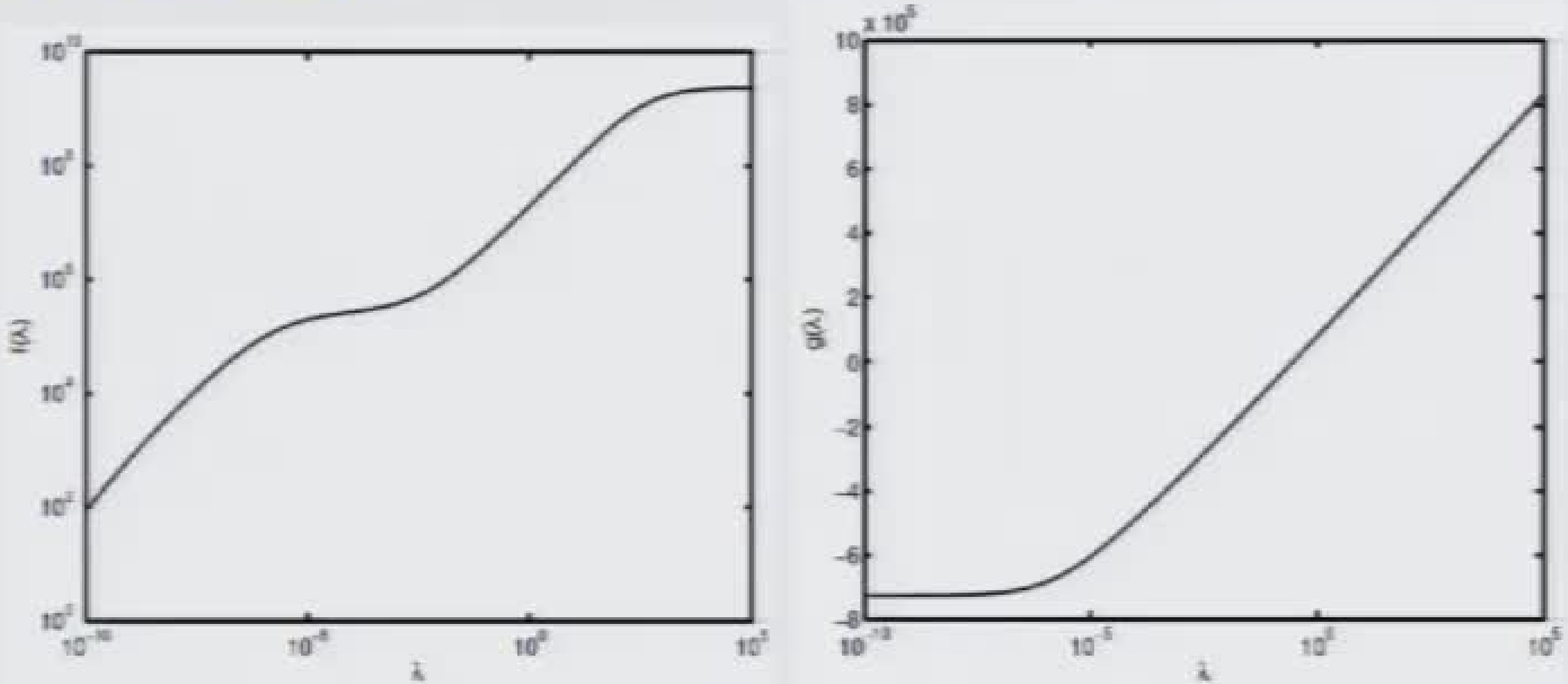
In the continuous setting this is the prototypical ill-posed inverse problem;
$\boldsymbol{k}$ is square integrable $\Rightarrow$ $\boldsymbol{A}$ is Hilbert-Schmidt $\Rightarrow$ compact

Trace and log determinant

The marginal posterior for $\boldsymbol{\theta}$ can be written

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto \delta^{n/2} \exp\left(-\frac{1}{2}g(\lambda) - \frac{\gamma}{2}f(\lambda)\right)\pi(\boldsymbol{\theta})$$

where $\lambda = \delta/\gamma$, and the functions $f(\lambda) = (\boldsymbol{A}^T\boldsymbol{y})^T((\boldsymbol{A}^T\boldsymbol{A})^{-1} - (\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{L})^{-1})(\boldsymbol{A}^T\boldsymbol{y})$ and $g(\lambda) = \log\det(\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{L})$



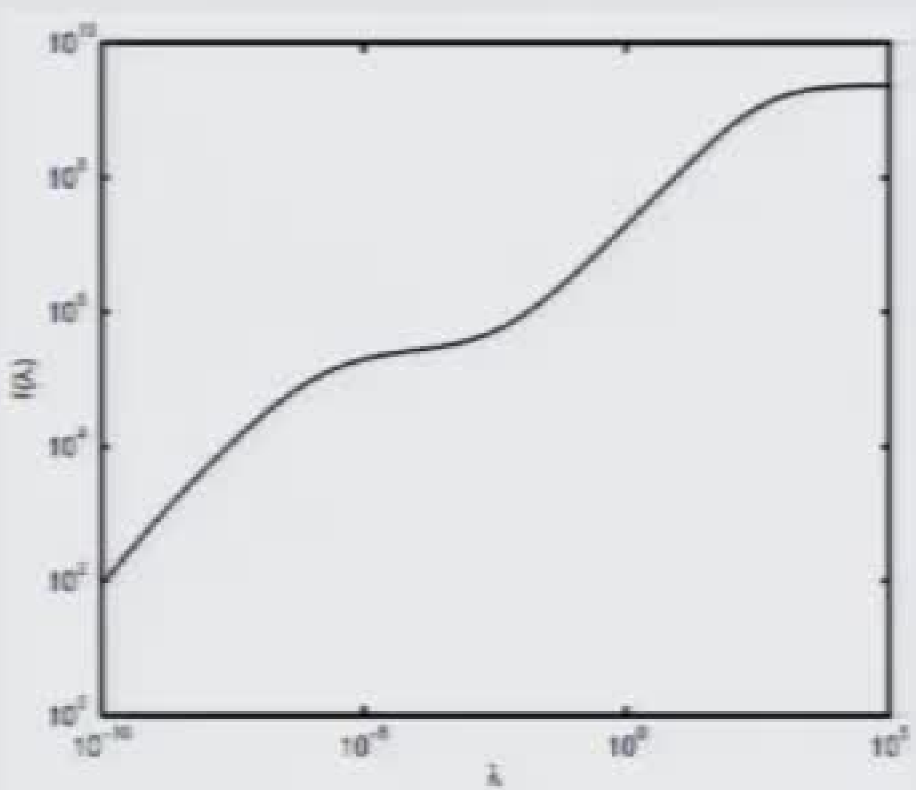
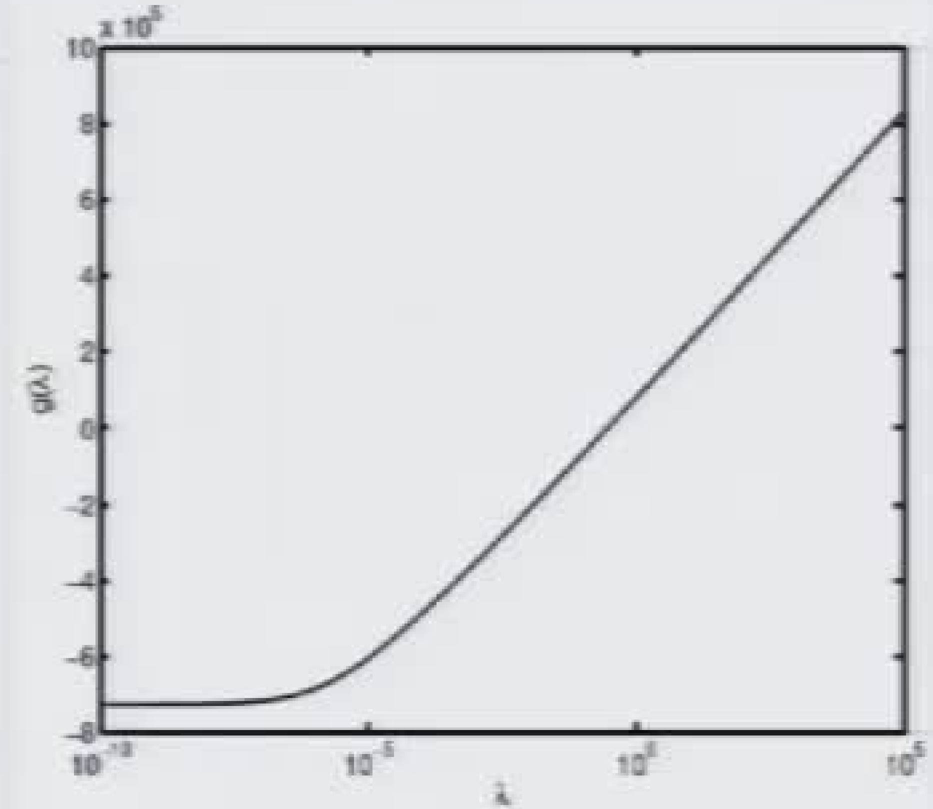
are uni-variate, monotonic, smooth, analytic (periodic case shown)

Trace and log determinant

The marginal posterior for $\boldsymbol{\theta}$ can be written

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto \delta^{n/2} \exp\left(-\frac{1}{2}g(\lambda) - \frac{\gamma}{2}f(\lambda)\right)\pi(\boldsymbol{\theta})$$

where $\lambda = \delta/\gamma$, and the functions $f(\lambda) = (\boldsymbol{A}^T\boldsymbol{y})^T((\boldsymbol{A}^T\boldsymbol{A})^{-1} - (\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{L})^{-1})(\boldsymbol{A}^T\boldsymbol{y})$ and $g(\lambda) = \log\det(\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{L})$

are uni-variate, monotonic, smooth, analytic (periodic case shown)

Sampling the full conditional for x

For the example

$$\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y} \sim \mathrm{N}\left((\boldsymbol{A}^T\boldsymbol{A}+(\delta/\gamma)\boldsymbol{L})^{-1}\boldsymbol{A}^T\boldsymbol{y},(\gamma\boldsymbol{A}^T\boldsymbol{A}+\delta\boldsymbol{L})^{-1}\right)$$

Independent $\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y}$ computed by RTO (randomize then optimize), i.e. solving the generalized deconvolution eqns with random RHS

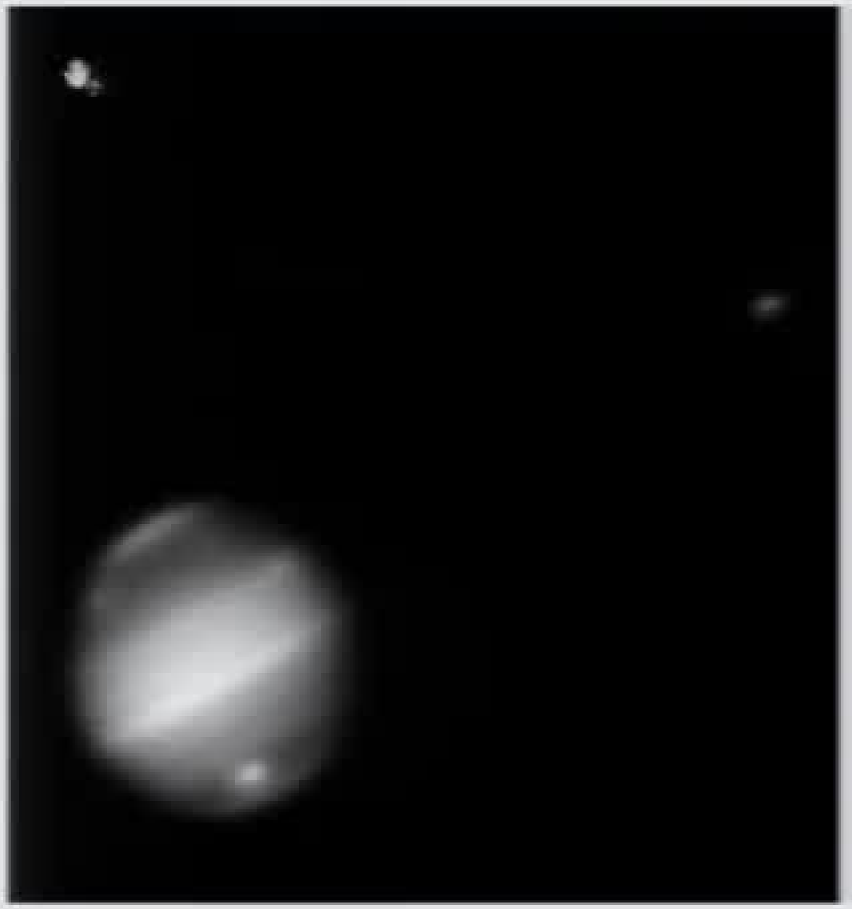
$$\left(\gamma\boldsymbol{A}^\mathsf{T}\boldsymbol{A}+\delta\boldsymbol{L}\right)\boldsymbol{x}=\gamma\boldsymbol{A}^\mathsf{T}\boldsymbol{y}+w$$

where $w = v_1 + v_2$ with independent $v_1 \sim \mathrm{N}\left(0,\gamma\boldsymbol{A}^\mathsf{T}\boldsymbol{A}\right)$ and $v_2 \sim \mathrm{N}\left(0,\delta\boldsymbol{L}\right)$

Requires one linear solve

Oliver He Reynolds 1996, Wikle Milliff Nychka Berliner 2001, Lalanne Prévost Chavel 2001, Tan Li Stoica 2010, Orieux Féron Giovannelli 2012, Bardsley 2012

A linear Gaussian inverse problem (image deblurring)



Data $\boldsymbol{y}$ is a blurry 256×256 gray-scale photograph of Jupiter in the methane band (780nm).

Estimate the 'true' unblurry image, $\boldsymbol{x}$.

Use the satellite (upper right) as PSF $\boldsymbol{k}$, so *semi-blind* deconvolution.

$$\boldsymbol{y} = \boldsymbol{k} * \boldsymbol{x} + \eta = \boldsymbol{A}\boldsymbol{x} + \eta$$

In the continuous setting this is the prototypical ill-posed inverse problem;
$\boldsymbol{k}$ is square integrable $\Rightarrow$ $\boldsymbol{A}$ is Hilbert-Schmidt $\Rightarrow$ compact

Fitting the model to data

Given measured data $\boldsymbol{y}$ determine (the distribution over) unobserved quantities:

- Fit model: full posterior

$$\boldsymbol{x}, \boldsymbol{\theta} | \boldsymbol{y} \sim \pi(\boldsymbol{x}, \boldsymbol{\theta} | \boldsymbol{y}) = \pi(\boldsymbol{y} | \boldsymbol{x}, \boldsymbol{\theta}) \pi(\boldsymbol{x} | \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) / \pi(\boldsymbol{y})$$

- Estimate unknown latent variables: (marginal posterior over latent variables)

$$\boldsymbol{x} | \boldsymbol{y} \sim \int \pi(\boldsymbol{x}, \boldsymbol{\theta} | \boldsymbol{y}) \, d\boldsymbol{\theta}$$

- Or when hyperparameters are of interest (marginal posterior over hyperparameters)

$$\boldsymbol{\theta} | \boldsymbol{y} \sim \int \pi(\boldsymbol{x}, \boldsymbol{\theta} | \boldsymbol{y}) \, d\boldsymbol{x}$$

Samples $\boldsymbol{\theta} | \boldsymbol{y}$ give access to full posterior via

$$\pi(\boldsymbol{x}, \boldsymbol{\theta} | \boldsymbol{y}) = \pi(\boldsymbol{x} | \boldsymbol{\theta}, \boldsymbol{y}) \pi(\boldsymbol{\theta} | \boldsymbol{y})$$

using the *full conditional* for $\boldsymbol{x}$. (MTC)

Bayesian hierarchical modeling

Observed data y depends on latent (field) x via function A

- **First stage:** model the observation y in terms of latent variables x

$$y|x,\theta \sim \pi(y|x,\theta)$$

  with uncertainty in π parameterized by θ.
  E.g. when $y|Ax$ is zero-mean Gaussian, $y|x,\theta \sim \mathrm{N}(Ax,\Sigma(\theta))$. (**likelihood**)

- **Second stage:** model latent variables x

$$x|\theta \sim \pi(x|\theta)$$

  with uncertainty in the model parameterized by θ (**prior**)

- **Third stage:** model unknown hyperparameters θ

$$\theta \sim \pi(\theta)$$

  (**hyperprior**)

Fitting the model to data

Given measured data $\boldsymbol{y}$ determine (the distribution over) unobserved quantities:

- Fit model: full posterior

$$\boldsymbol{x}, \boldsymbol{\theta} | \boldsymbol{y} \sim \pi(\boldsymbol{x}, \boldsymbol{\theta} | \boldsymbol{y}) = \pi(\boldsymbol{y} | \boldsymbol{x}, \boldsymbol{\theta}) \pi(\boldsymbol{x} | \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) / \pi(\boldsymbol{y})$$

- Estimate unknown latent variables: (marginal posterior over latent variables)

$$\boldsymbol{x} | \boldsymbol{y} \sim \int \pi(\boldsymbol{x}, \boldsymbol{\theta} | \boldsymbol{y}) \, d\boldsymbol{\theta}$$

- Or when hyperparameters are of interest (marginal posterior over hyperparameters)

$$\boldsymbol{\theta} | \boldsymbol{y} \sim \int \pi(\boldsymbol{x}, \boldsymbol{\theta} | \boldsymbol{y}) \, d\boldsymbol{x}$$

Samples $\boldsymbol{\theta} | \boldsymbol{y}$ give access to full posterior via

$$\pi(\boldsymbol{x}, \boldsymbol{\theta} | \boldsymbol{y}) = \pi(\boldsymbol{x} | \boldsymbol{\theta}, \boldsymbol{y}) \pi(\boldsymbol{\theta} | \boldsymbol{y})$$

using the *full conditional* for $\boldsymbol{x}$. (MTC)

Bayesian hierarchical modeling

Observed data y depends on latent (field) x via function A

- **First stage:** model the observation y in terms of latent variables x

$$y|x,\theta \sim \pi(y|x,\theta)$$

  with uncertainty in π parameterized by θ.
  E.g. when $y|Ax$ is zero-mean Gaussian, $y|x,\theta \sim \mathrm{N}(Ax,\Sigma(\theta))$. (**likelihood**)

- **Second stage:** model latent variables x

$$x|\theta \sim \pi(x|\theta)$$

  with uncertainty in the model parameterized by θ (**prior**)

- **Third stage:** model unknown hyperparameters θ

$$\theta \sim \pi(\theta)$$

  (**hyperprior**)

Fitting the model to data

Given measured data $\boldsymbol{y}$ determine (the distribution over) unobserved quantities:

- Fit model: full posterior

$$\boldsymbol{x}, \boldsymbol{\theta} | \boldsymbol{y} \sim \pi(\boldsymbol{x}, \boldsymbol{\theta} | \boldsymbol{y}) = \pi(\boldsymbol{y} | \boldsymbol{x}, \boldsymbol{\theta}) \pi(\boldsymbol{x} | \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) / \pi(\boldsymbol{y})$$

- Estimate unknown latent variables: (marginal posterior over latent variables)

$$\boldsymbol{x} | \boldsymbol{y} \sim \int \pi(\boldsymbol{x}, \boldsymbol{\theta} | \boldsymbol{y}) \, d\boldsymbol{\theta}$$

- Or when hyperparameters are of interest (marginal posterior over hyperparameters)

$$\boldsymbol{\theta} | \boldsymbol{y} \sim \int \pi(\boldsymbol{x}, \boldsymbol{\theta} | \boldsymbol{y}) \, d\boldsymbol{x}$$

Samples $\boldsymbol{\theta} | \boldsymbol{y}$ give access to full posterior via

$$\pi(\boldsymbol{x}, \boldsymbol{\theta} | \boldsymbol{y}) = \pi(\boldsymbol{x} | \boldsymbol{\theta}, \boldsymbol{y}) \pi(\boldsymbol{\theta} | \boldsymbol{y})$$

using the *full conditional* for $\boldsymbol{x}$. (MTC)

Marginal-then-conditional sampling

Claim: When the full conditional for the latent variables $\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ has a known form, then the marginal distribution over hyperparameters $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ is available for sampling.

Follows since the θ-dependence of the normalizing constant is known.

Draw iid samples from the full posterior by:

1. Sample from the marginal posterior over $\boldsymbol{\theta}$

$$\boldsymbol{\theta} \overset{iid}{\sim} \pi(\boldsymbol{\theta}|\boldsymbol{y})$$

   usually low-dimensional, so random-walk MCMC has negligible cost e.g., t-walk.

2. Sample from the full conditional over $\boldsymbol{x}$

$$\boldsymbol{x} \sim \pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$$

to give MTC, a.k.a. composition sampling, or two-variate conditional distribution method.

Censored data

y_i is observed with right censoring, i.e., if $y_i > a$ then "observation above a" is recorded.

Let $y_1 < y_2 < \cdots < y_m$ be the uncensored observations, so $n - m$ censored observations.

Introduce latent variables x_i for the unobserved data,

$$y_i = \begin{cases} x_i & \text{if } x_i < a \\ a^+ & \text{if } x_i \geq a \end{cases}$$

Model $x_i \overset{iid}{\sim} \mathrm{N}(\mu, \lambda^{-1})$, with $\mu|\lambda \sim \mathrm{N}(\mu_0, k_0\lambda)$ and $\lambda \sim \mathrm{Ga}(\alpha, \beta)$. $k_0 = \alpha = 1,\ \beta = 0.1$

The full conditional $\mu, \lambda|\boldsymbol{y}, \boldsymbol{x}$ is a normal-gamma distribution and each $x_i|\mu, \lambda, \boldsymbol{y}$ full conditional is an iid truncated normal distribution.

Conventional approach is block-Gibbs sampling, has increasing dimension with sample size

> sample $\mu, \lambda|\boldsymbol{y}, \boldsymbol{x}$ by sampling $\lambda|\boldsymbol{y}, \boldsymbol{x} \sim \mathrm{Ga}(\alpha_2, \beta_2)$ then $\mu|\lambda, \boldsymbol{y}, \boldsymbol{x} \sim \mathrm{N}(\mu_2, k_2\lambda)$
>
> sample $x_i|\mu, \lambda, \boldsymbol{y} \sim \mathrm{N}(\mu, \lambda)$ truncated to $x_i \in [a, \infty)$ for $i = 1, \ldots, n - m$.

where $\alpha_2 = \alpha_0 + \frac{n}{2}$, $k_2 = k_0 + n$, $\mu_2 = \frac{1}{k_2}(k_0\mu_0 + \sum_{i=1}^{m} y_i + \sum_{i=1}^{n-m} x_i)$, and $\beta_2 = \beta + \frac{1}{2}(k_0\mu_0^2 - k_2\mu_2^2 + \sum_{i=1}^{m} y_i^2 + \sum_{i=1}^{n-m} x_i^2)$.

Censored data

y_i is observed with right censoring, i.e., if $y_i > a$ then "observation above a" is recorded.
Let $y_1 < y_2 < \cdots < y_m$ be the uncensored observations, so $n - m$ censored observations.
Introduce latent variables x_i for the unobserved data,

$$y_i = \begin{cases} x_i & \text{if } x_i < a \\ a^+ & \text{if } x_i \geq a \end{cases}$$

Model $x_i \overset{iid}{\sim} \mathrm{N}(\mu, \lambda^{-1})$, with $\mu|\lambda \sim \mathrm{N}(\mu_0, k_0\lambda)$ and $\lambda \sim \mathrm{Ga}(\alpha, \beta)$. $k_0 = \alpha = 1,\ \beta = 0.1$

The full conditional $\mu, \lambda|\boldsymbol{y}, \boldsymbol{x}$ is a normal-gamma distribution and each $x_i|\mu, \lambda, \boldsymbol{y}$ full conditional is an iid truncated normal distribution.

Conventional approach is block-Gibbs sampling, has increasing dimension with sample size

sample $\mu, \lambda|\boldsymbol{y}, \boldsymbol{x}$ by sampling $\lambda|\boldsymbol{y}, \boldsymbol{x} \sim \mathrm{Ga}(\alpha_2, \beta_2)$ then $\mu|\lambda, \boldsymbol{y}, \boldsymbol{x} \sim \mathrm{N}(\mu_2, k_2\lambda)$

sample $x_i|\mu, \lambda, \boldsymbol{y} \sim \mathrm{N}(\mu, \lambda)$ truncated to $x_i \in [a, \infty)$ for $i = 1, \ldots, n - m$.

where $\alpha_2 = \alpha_0 + \frac{n}{2}$, $k_2 = k_0 + n$, $\mu_2 = \frac{1}{k_2}(k_0\mu_0 + \sum_{i=1}^{m} y_i + \sum_{i=1}^{n-m} x_i)$, and $\beta_2 = \beta + \frac{1}{2}(k_0\mu_0^2 - k_2\mu_2^2 + \sum_{i=1}^{m} y_i^2 + \sum_{i=1}^{n-m} x_i^2)$.

Censored data

y_i is observed with right censoring, i.e., if $y_i > a$ then "observation above a" is recorded.

Let $y_1 < y_2 < \cdots < y_m$ be the uncensored observations, so $n - m$ censored observations.

Introduce latent variables x_i for the unobserved data,

$$y_i = \begin{cases} x_i & \text{if } x_i < a \\ a^+ & \text{if } x_i \geq a \end{cases}$$

Model $x_i \overset{iid}{\sim} \mathrm{N}(\mu, \lambda^{-1})$, with $\mu|\lambda \sim \mathrm{N}(\mu_0, k_0\lambda)$ and $\lambda \sim \mathrm{Ga}(\alpha, \beta)$. $k_0 = \alpha = 1,\ \beta = 0.1$

The full conditional $\mu, \lambda | \boldsymbol{y}, \boldsymbol{x}$ is a normal-gamma distribution and each $x_i | \mu, \lambda, \boldsymbol{y}$ full conditional is an iid truncated normal distribution.

Conventional approach is block-Gibbs sampling, has increasing dimension with sample size

> sample $\mu, \lambda | \boldsymbol{y}, \boldsymbol{x}$ by sampling $\lambda | \boldsymbol{y}, \boldsymbol{x} \sim \mathrm{Ga}(\alpha_2, \beta_2)$ then $\mu | \lambda, \boldsymbol{y}, \boldsymbol{x} \sim \mathrm{N}(\mu_2, k_2\lambda)$
>
> sample $x_i | \mu, \lambda, \boldsymbol{y} \sim \mathrm{N}(\mu, \lambda)$ truncated to $x_i \in [a, \infty)$ for $i = 1, \ldots, n - m$.

where $\alpha_2 = \alpha_0 + \frac{n}{2}$, $k_2 = k_0 + n$, $\mu_2 = \frac{1}{k_2}(k_0\mu_0 + \sum_{i=1}^{m} y_i + \sum_{i=1}^{n-m} x_i)$, and $\beta_2 = \beta + \frac{1}{2}(k_0\mu_0^2 - k_2\mu_2^2 + \sum_{i=1}^{m} y_i^2 + \sum_{i=1}^{n-m} x_i^2)$.

Censored data :: MTC

Since the full conditional for latent field is tractable, the marginal posterior for hyperparameters μ, λ is available, and is

$$f(\mu, \lambda | \boldsymbol{y}) \propto \lambda^{\alpha_1 - 1/2} \exp\left(-\lambda\left(\frac{k_1}{2}(\mu - \mu_1)^2 + \beta_1\right)\right)\left(1 - \Phi(\sqrt{\lambda}(a - \mu))\right)^{n-m}$$

where $\alpha_1 = \alpha + \frac{m}{2}$, $k_1 = k_0 + m$, $\mu_1 = \frac{1}{k_1}(k_0\mu_0 + \sum_{i=1}^m y_i)$, $\beta_1 = \beta + \frac{1}{2}(k_0\mu_0^2 - k_1\mu_1^2 + \sum_{i=1}^m y_i^2)$ depend only on the uncensored data.

This is a 2-dim distribution so computational cost of MCMC is independent of sample size, once m, $\sum_{i=1}^m y_i$ and $\sum_{i=1}^m y_i^2$ evaluated.

Can sample from this distribution using the t-walk and the computational cost will remain almost constant with sample size. Moreover, if IACT remains constant with sample size then CCES (computing cost per effectively-independent sample) also remains constant for increasing n.

A linear Gaussian inverse problem (image deblurring)

Data $\boldsymbol{y}$ is a blurry 256×256 gray-scale photograph of Jupiter in the methane band (780nm).

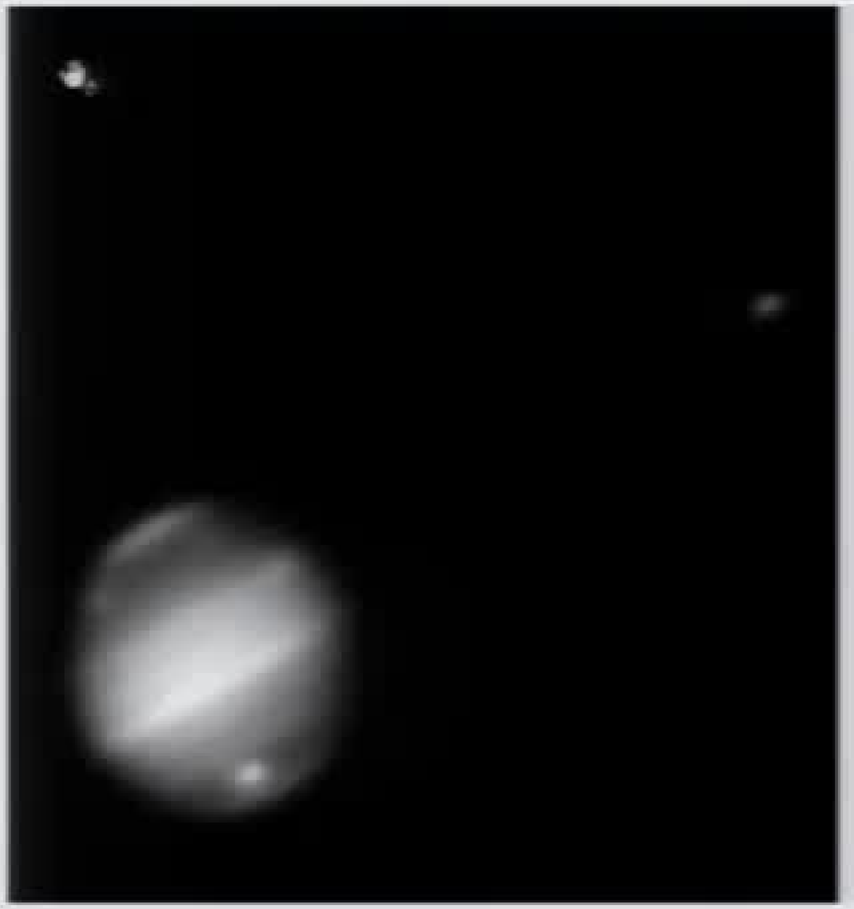
Estimate the 'true' unblurry image, $\boldsymbol{x}$.

Use the satellite (upper right) as PSF $\boldsymbol{k}$, so *semi-blind* deconvolution.

$$\boldsymbol{y} = \boldsymbol{k} * \boldsymbol{x} + \eta = \boldsymbol{A}\boldsymbol{x} + \eta$$

In the continuous setting this is the prototypical ill-posed inverse problem;
$\boldsymbol{k}$ is square integrable $\Rightarrow$ $\boldsymbol{A}$ is Hilbert-Schmidt $\Rightarrow$ compact

A linear Gaussian inverse problem (image deblurring)



Data $\boldsymbol{y}$ is a blurry 256×256 gray-scale photograph of Jupiter in the methane band (780nm).

Estimate the 'true' unblurry image, $\boldsymbol{x}$.

Use the satellite (upper right) as PSF $\boldsymbol{k}$, so *semi-blind* deconvolution.

$$\boldsymbol{y} = \boldsymbol{k} * \boldsymbol{x} + \eta = \boldsymbol{A}\boldsymbol{x} + \eta$$

In the continuous setting this is the prototypical ill-posed inverse problem;
$\boldsymbol{k}$ is square integrable $\Rightarrow$ $\boldsymbol{A}$ is Hilbert-Schmidt $\Rightarrow$ compact

Bayesian hierarchical model

Linear forward map, Gaussian noise and prior

$$y|x,\theta \sim \mathrm{N}\left(Ax, (\gamma I)^{-1}\right) \quad \text{(likelihood)}$$

$$x|\theta \sim \mathrm{N}\left(\mu, (\delta L)^{-1}\right) \quad \text{(prior)}$$

$$\theta = (\gamma, \delta) \sim \pi(\theta) \quad \text{(hyperprior)}$$

where γ is precision of measurements, δ is lumping constant in true image.

Common model in spatial statistics

Since

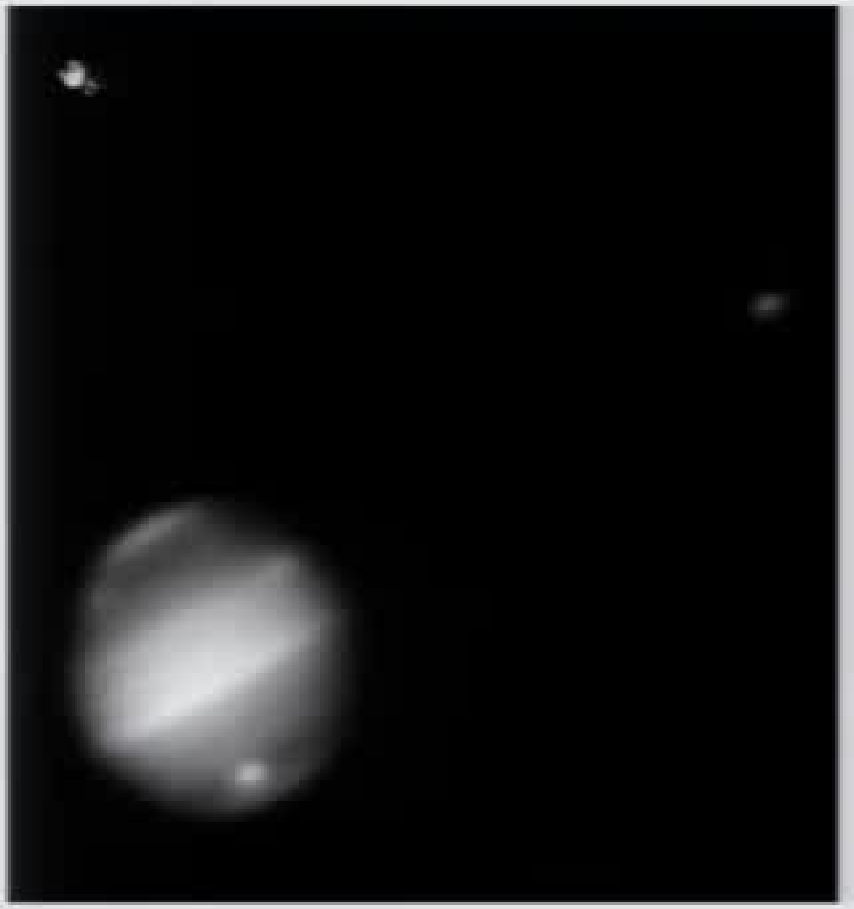
$$\pi(y|x,\theta) = \frac{\gamma^{n/2}}{\sqrt{2\pi}} \exp\left\{-\frac{\gamma}{2}\|Ax - y\|^2\right\} \quad \text{and } \pi(x|\theta) = \frac{\delta^{n/2}\sqrt{\det L}}{\sqrt{2\pi}} \exp\left\{-\frac{\delta}{2} x^\mathsf{T} L x\right\}$$

by conditional Bayes rule, the full conditional over the latent field

$$\pi(x|y,\theta) \propto \exp\left\{-\frac{\gamma}{2}\left(\|Ax - y\|^2 - \frac{\delta}{\gamma} x^\mathsf{T} L x\right)\right\}$$

is normal

A linear Gaussian inverse problem (image deblurring)



Data $\boldsymbol{y}$ is a blurry 256×256 gray-scale photograph of Jupiter in the methane band (780nm).

Estimate the 'true' unblurry image, $\boldsymbol{x}$.

Use the satellite (upper right) as PSF $\boldsymbol{k}$, so *semi-blind* deconvolution.

$$\boldsymbol{y} = \boldsymbol{k} * \boldsymbol{x} + \eta = \boldsymbol{A}\boldsymbol{x} + \eta$$

In the continuous setting this is the prototypical ill-posed inverse problem;
$\boldsymbol{k}$ is square integrable $\Rightarrow$ $\boldsymbol{A}$ is Hilbert-Schmidt $\Rightarrow$ compact

Bayesian hierarchical model

Linear forward map, Gaussian noise and prior

$$\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta} \sim \mathrm{N}\left(\boldsymbol{A}\boldsymbol{x}, (\gamma I)^{-1}\right) \qquad \text{(likelihood)}$$

$$\boldsymbol{x}|\boldsymbol{\theta} \sim \mathrm{N}\left(\mu, (\delta \boldsymbol{L})^{-1}\right) \qquad \text{(prior)}$$

$$\boldsymbol{\theta} = (\gamma, \delta) \sim \pi(\boldsymbol{\theta}) \qquad \text{(hyperprior)}$$

where γ is precision of measurements, δ is lumping constant in true image.

Common model in spatial statistics

Since

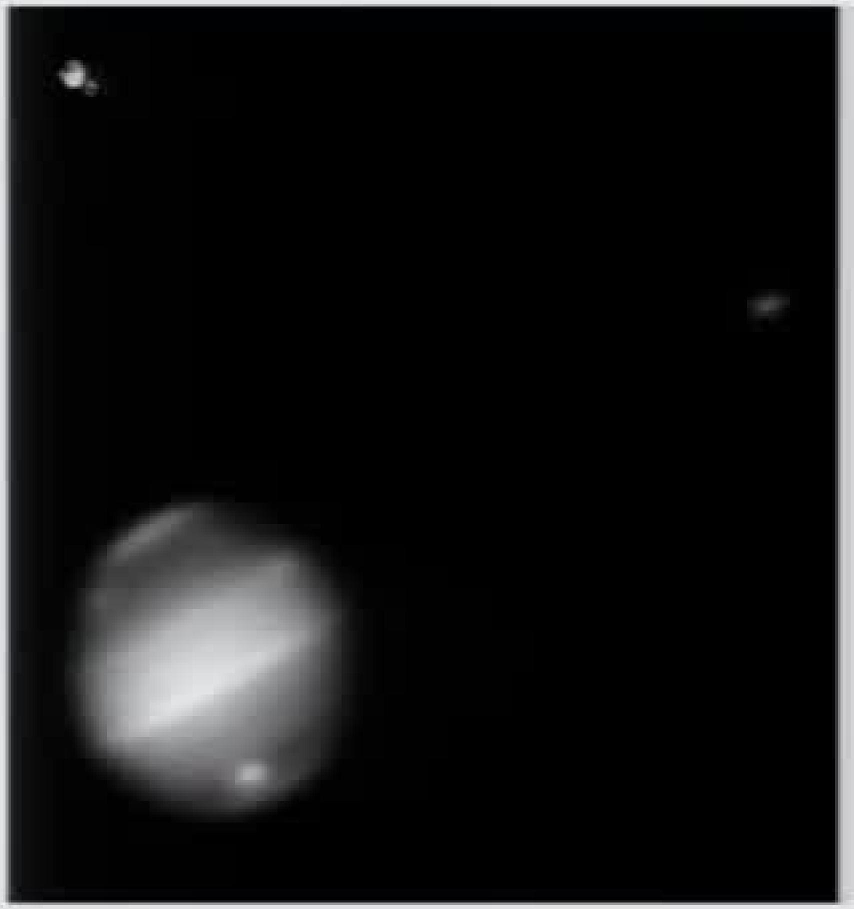
$$\pi(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}) = \frac{\gamma^{n/2}}{\sqrt{2\pi}} \exp\left\{-\frac{\gamma}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|^2\right\} \quad \text{and } \pi(\boldsymbol{x}|\boldsymbol{\theta}) = \frac{\delta^{n/2}\sqrt{\det \boldsymbol{L}}}{\sqrt{2\pi}} \exp\left\{-\frac{\delta}{2}\boldsymbol{x}^\mathsf{T}\boldsymbol{L}\boldsymbol{x}\right\}$$

by conditional Bayes rule, the full conditional over the latent field

$$\pi(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}) \propto \exp\left\{-\frac{\gamma}{2}\left(\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|^2 - \frac{\delta}{\gamma}\boldsymbol{x}^\mathsf{T}\boldsymbol{L}\boldsymbol{x}\right)\right\}$$

is normal

A linear Gaussian inverse problem (image deblurring)



Data $\boldsymbol{y}$ is a blurry 256×256 gray-scale photograph of Jupiter in the methane band (780nm).

Estimate the 'true' unblurry image, $\boldsymbol{x}$.

Use the satellite (upper right) as PSF $\boldsymbol{k}$, so *semi-blind* deconvolution.

$$\boldsymbol{y} = \boldsymbol{k} * \boldsymbol{x} + \eta = \boldsymbol{A}\boldsymbol{x} + \eta$$

In the continuous setting this is the prototypical ill-posed inverse problem;
$\boldsymbol{k}$ is square integrable $\Rightarrow$ $\boldsymbol{A}$ is Hilbert-Schmidt $\Rightarrow$ compact

Marginal posterior for θ

Lemma 1

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\pi(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{x})\,\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})\,\pi(\boldsymbol{y})}$$

Proof. $\pi(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}) = \pi(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})\,\pi(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ and $\pi(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}) = \pi(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$.
Since $\pi(\boldsymbol{y}) \neq 0$, the result follows by Bayes rule $\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \pi(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})/\pi(\boldsymbol{y})$. ■

Since $\pi(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})$ has known form, $\boldsymbol{x}$-dependence of RHS can be eliminated (i.e., an algebraic route to integrating over $\boldsymbol{x}$.)

For general Gaussian-linear model : Σ = noise covariance, Q = prior precision

$$\pi(\theta|y) \propto \sqrt{\frac{\det(\Sigma^{-1})\det(Q)}{\det(Q + A^T\Sigma^{-1}A)}}$$

$$\exp\left\{-\frac{1}{2}(y - A\mu)^T\Sigma^{-1}A\left[(A^T\Sigma^{-1}A)^{-1} - (A^T\Sigma^{-1}A + Q)^{-1}\right]A^T\Sigma^{-1}(y - A\mu)\right\}\pi(\theta)$$

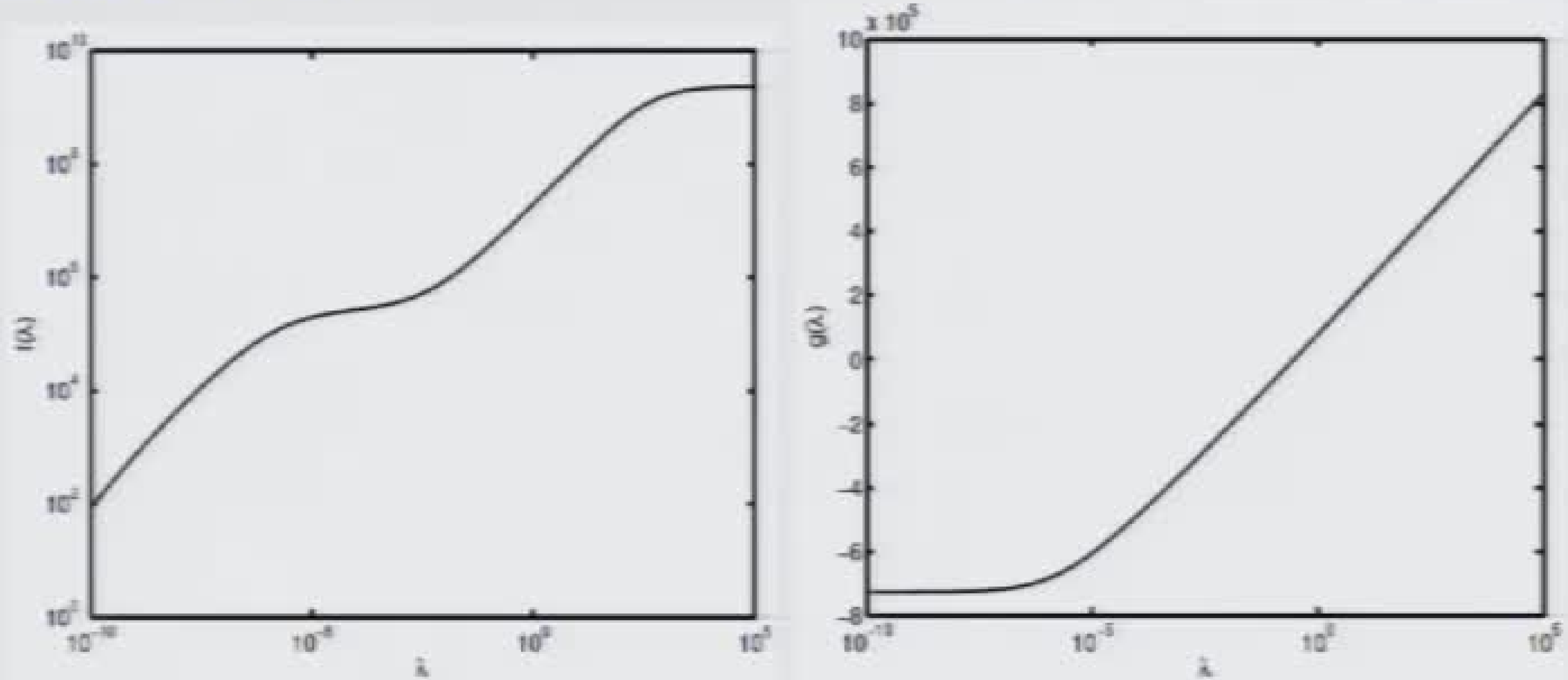
Traditional difficulty:: MCMC requires ratios of determinants of Σ^{-1}, Q and $Q + A^T\Sigma^{-1}A$, and differences of arguments of the exponential.

Trace and log determinant

The marginal posterior for $\boldsymbol{\theta}$ can be written

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto \delta^{n/2} \exp\left(-\frac{1}{2}g(\lambda) - \frac{\gamma}{2}f(\lambda)\right)\pi(\boldsymbol{\theta})$$

where $\lambda = \delta/\gamma$, and the functions $f(\lambda) = (\boldsymbol{A}^T\boldsymbol{y})^T((\boldsymbol{A}^T\boldsymbol{A})^{-1} - (\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{L})^{-1})(\boldsymbol{A}^T\boldsymbol{y})$ and $g(\lambda) = \log\det(\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{L})$



are uni-variate, monotonic, smooth, analytic (periodic case shown)

Evaluation of (ratio of) determinants for MCMC

- Periodic boundary conditions (diagonalize matrices by FFT):
  - Option 1: $\mathcal{O}(n)$ calculation: $\det(\boldsymbol{A}^T\boldsymbol{A}+\lambda\boldsymbol{L}) = \Pi_{i=1}^{n}(Ki^2+\lambda L_i)$. RWM over $\pi(\boldsymbol{\theta}|\boldsymbol{y})$
  - Option 2: $\mathcal{O}(1)$ series expansion of f and g. MWG with bespoke directions over $\pi(\boldsymbol{\theta}|\boldsymbol{y})$
- General case: Write $B = \boldsymbol{A}^T\boldsymbol{A}+\lambda\boldsymbol{L}$

$$f^{(r)}(\lambda) = (-1)^{r+1}k!(\boldsymbol{A}^T\boldsymbol{y})^T(B^{-1}\boldsymbol{L})^r B^{-1}(\boldsymbol{A}^T\boldsymbol{y}), \qquad r = 1, 2, \ldots.$$

Using the identity (Gohberg Goldberg Krupnik 2000)

$$\log(\det(I+tF)) = \sum_{r=1}^{\infty}\frac{(-1)^{r+1}}{r!}\operatorname{tr}(F^r)t^r$$

the derivatives of g are

$$g^{(r)}(\lambda) = (-1)^{r+1}\operatorname{tr}((B^{-1}\boldsymbol{L})^r), \qquad r = 1, 2, \ldots.$$

Estimate traces via $\operatorname{tr}((B^{-1}\boldsymbol{L})^r) = \mathrm{E}[z^T(B^{-1}\boldsymbol{L})^r z]$, $z_i \overset{\text{iid}}{\sim} \text{Unif}(\{-1, 1\})$ (Meurant2009)
No determinants need be evaluated!

Comparing algorithms

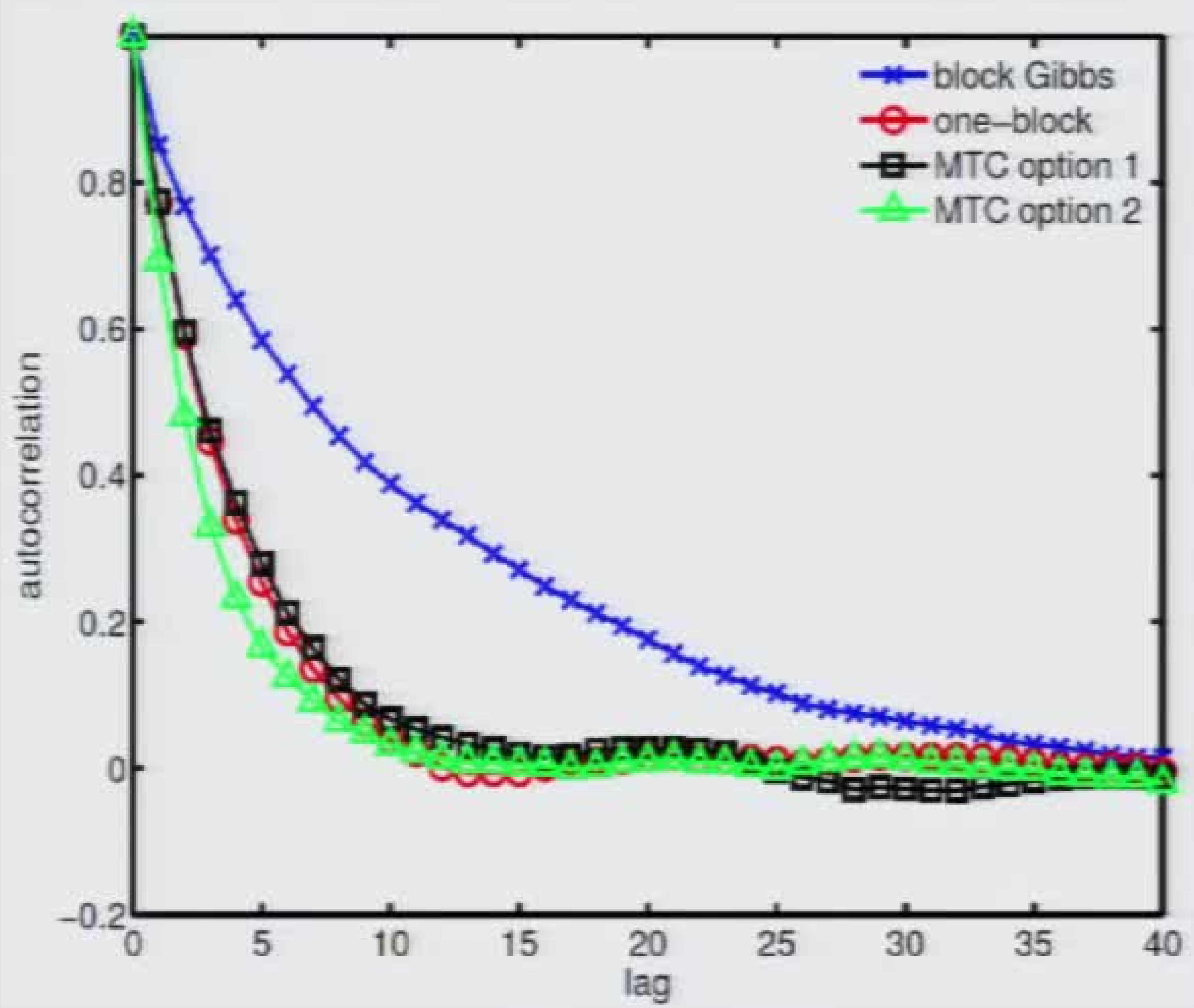
MTC: First draw (quasi) independent samples from the marginal posterior over $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \int \pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{x}$, then from full conditional over $\boldsymbol{x}$

Block Gibbs: Gibbs sweep $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}|\boldsymbol{y})$ then $\boldsymbol{x} \sim \pi(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})$ in sequence, repeatedly

One-block: Draw $\boldsymbol{\theta}' \sim \pi(\boldsymbol{\theta}|\boldsymbol{y})$ then $\boldsymbol{x}' \sim \pi(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}')$ and put $(\boldsymbol{x}', \boldsymbol{\theta}')$ as proposal in MH-MCMC

Regularized inversion: Estimate $\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} \pi(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})$ with $\lambda = \delta/\gamma$ chosen according to L-curve criterion.

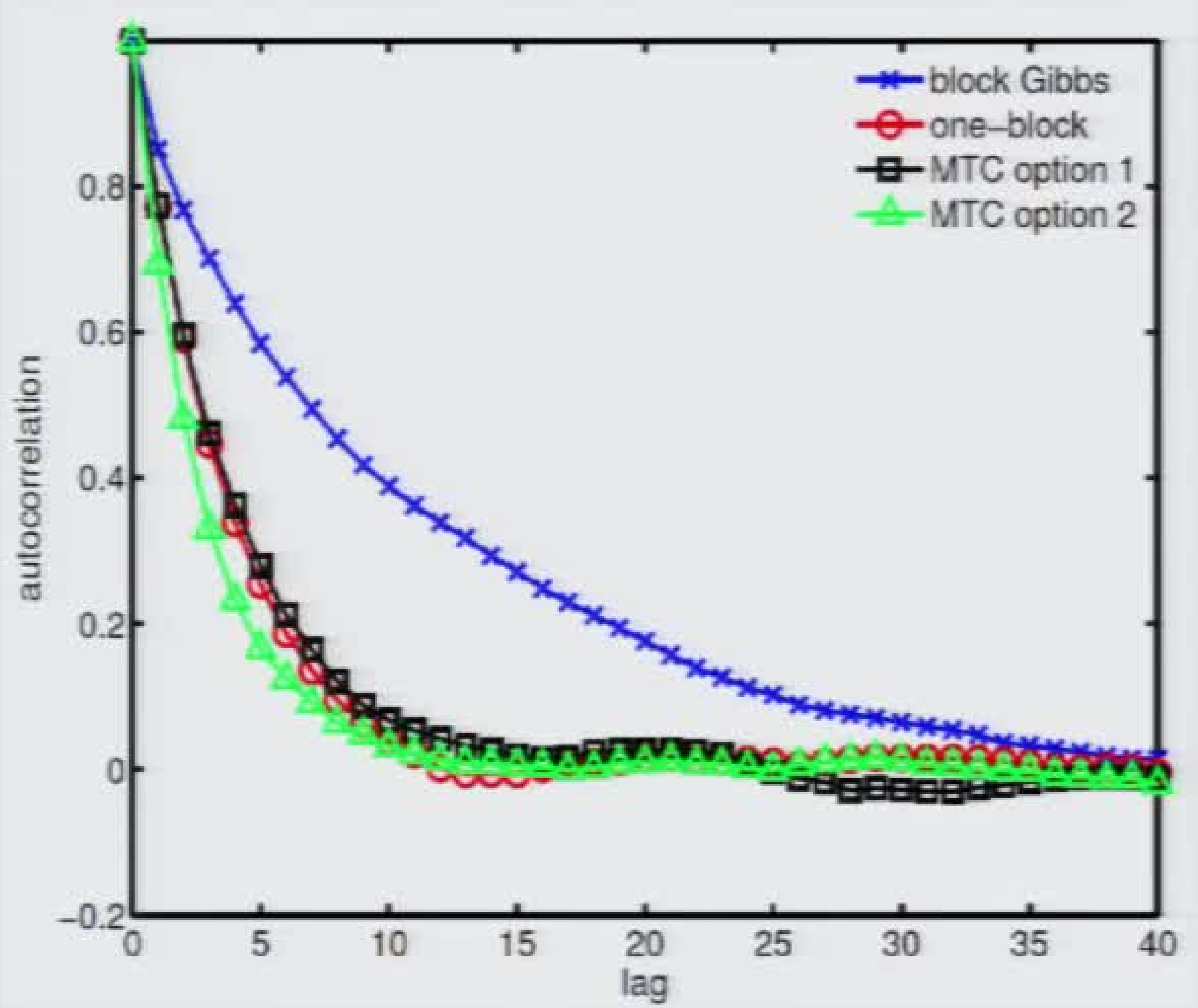
Autocorrelation of $\lambda = \delta/\gamma$ (periodic BC)



Gibbs slowest :: MTC opt. 2 cost per iteration is $O(1)$, all others 1 linear solve per iteration

Dimension-independent re-parametrization of Gibbs improves IACT to that of one-block, but increases cost per iteration to 3 linear solves, hence CCES unchanged.

Autocorrelation of $\lambda = \delta/\gamma$ (periodic BC)



Gibbs slowest :: MTC opt. 2 cost per iteration is $O(1)$, all others 1 linear solve per iteration

Dimension-independent re-parametrization of Gibbs improves IACT to that of one-block, but increases cost per iteration to 3 linear solves, hence CCES unchanged.

Posterior expectation

$$\mathrm{E}_{\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}}\left[h\left(\boldsymbol{x}\right)\right] = \mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{y}}\left[\mathrm{E}_{\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y}}\left[h\left(\boldsymbol{x}\right)\right]\right]$$

which is a weighted sum in $\boldsymbol{\theta}$ of expectations over full conditionals in $\boldsymbol{x}$.

In the linear Gaussian problem any moment may be evaluated this way, i.e. for polynomial h.

The mean further simplifies to

$$\mathrm{E}[\boldsymbol{x}|\boldsymbol{y}] = \int (\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{L})^{-1}\boldsymbol{A}^T\boldsymbol{y}\,\pi(\lambda|\boldsymbol{y})\,\mathrm{d}\lambda$$

Weights for the numerical integration given by the marginal posterior histogram for λ.

Thank You

Thank You

Take-home messages

- Don't do Gibbs unless you have a very good reason
- If the full conditional over x has known form, do MTC
- No restriction on prior
- For censored data example, sampling is independent of data size
- For the linear-Gaussian inverse problem ...
  - One linear solve per independent sample is optimal ...
  - ... independent of image dimension
  - Faster than Gibbs (including dimension-independent parametrization), one-block, regularization
  - Does not require trace-class prior covariance, nor consistent discretization

Take-home messages

- Don't do Gibbs unless you have a very good reason
- If the full conditional over x has known form, do MTC
- No restriction on prior
- For censored data example, sampling is independent of data size
- For the linear-Gaussian inverse problem ...
  - One linear solve per independent sample is optimal ...
  - ... independent of image dimension
  - Faster than Gibbs (including dimension-independent parametrization), one-block, regularization
  - Does not require trace-class prior covariance, nor consistent discretization

Posterior expectation

$$\mathrm{E}_{\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}}\left[h\left(\boldsymbol{x}\right)\right] = \mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{y}}\left[\mathrm{E}_{\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y}}\left[h\left(\boldsymbol{x}\right)\right]\right]$$

which is a weighted sum in $\boldsymbol{\theta}$ of expectations over full conditionals in $\boldsymbol{x}$.

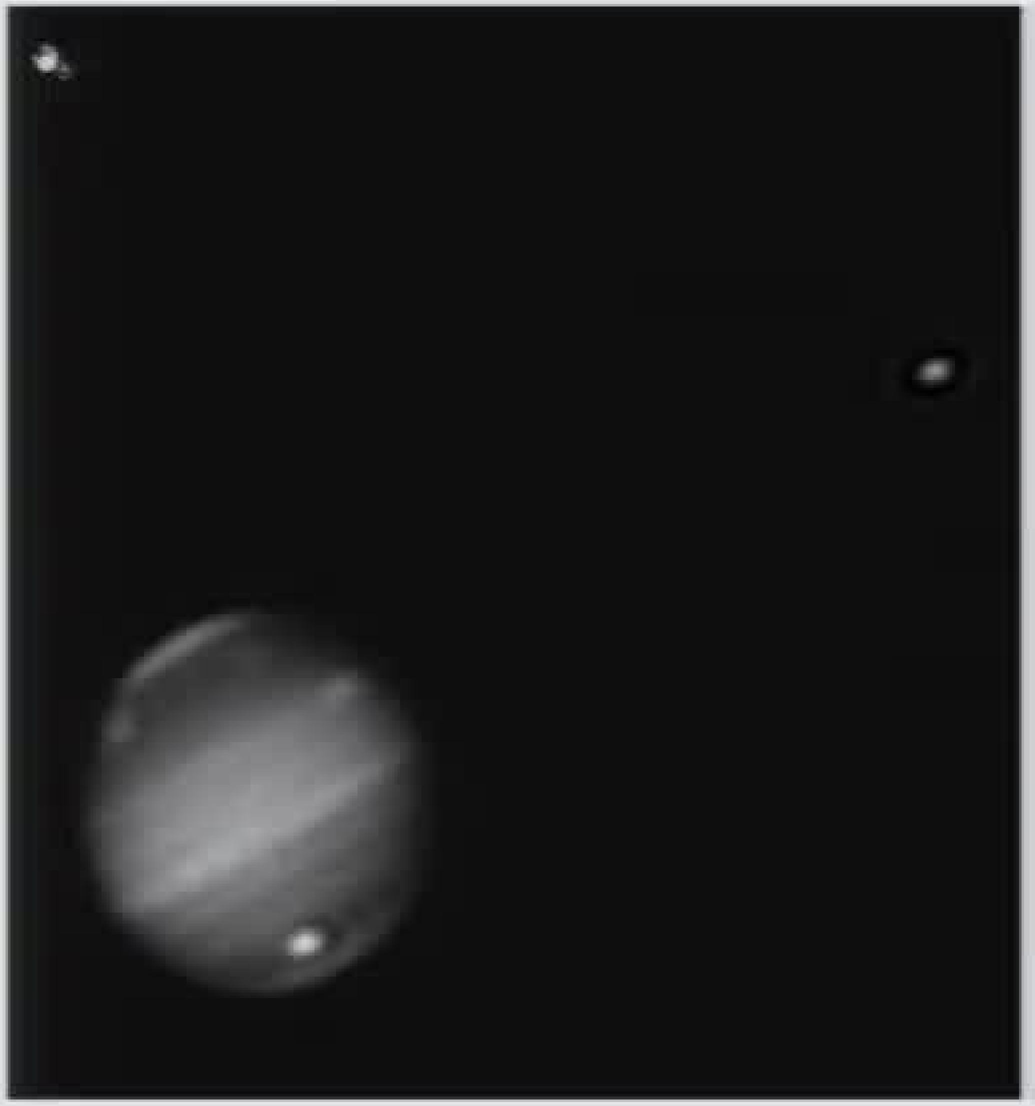
In the linear Gaussian problem any moment may be evaluated this way, i.e. for polynomial h.

The mean further simplifies to

$$\mathrm{E}[\boldsymbol{x}|\boldsymbol{y}] = \int (\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{L})^{-1}\boldsymbol{A}^T\boldsymbol{y}\,\pi(\lambda|\boldsymbol{y})\,\mathrm{d}\lambda$$

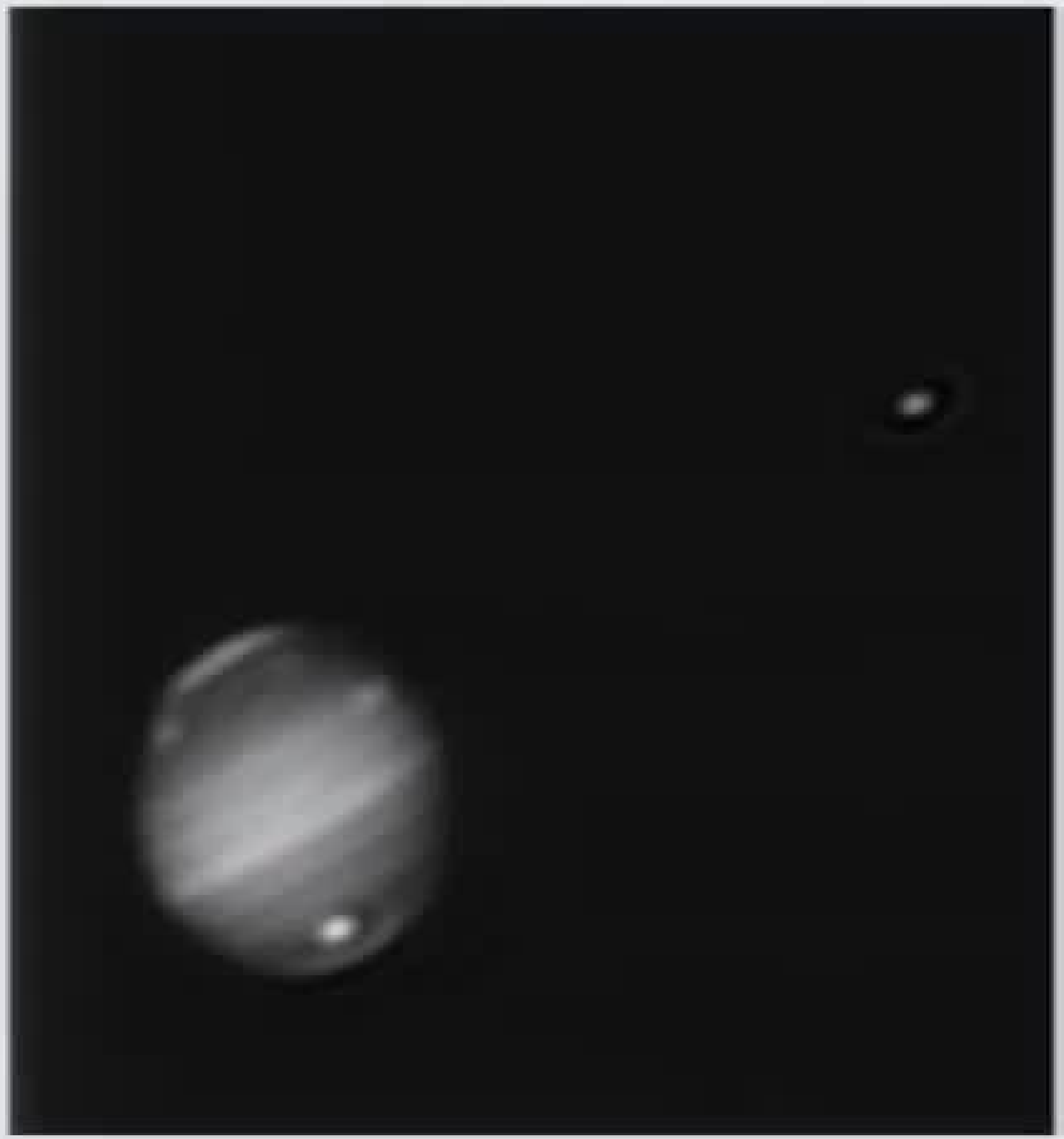
Weights for the numerical integration given by the marginal posterior histogram for λ.

Regularized solution



Total of 201 solves

Bayesian mean image



Total of 84 solves

Dirichlet BC outside border of nuisance pixels, mean image integrates over nuisance pixels

Gibbs requires $\gtrsim 2100$ solves, even when dimension-independent form is available

Take-home messages

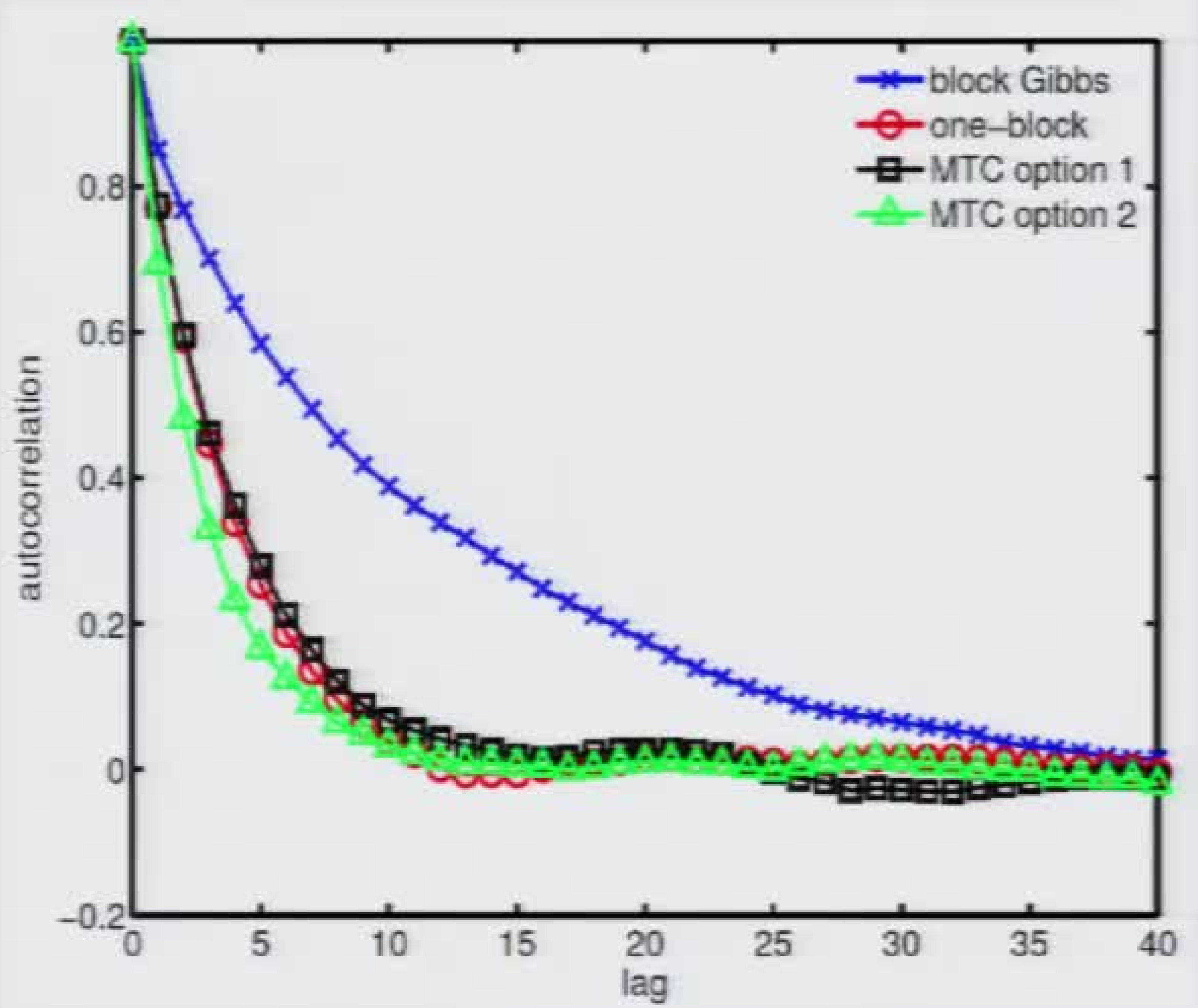
- Don't do Gibbs unless you have a very good reason
- If the full conditional over x has known form, do MTC
- No restriction on prior
- For censored data example, sampling is independent of data size
- For the linear-Gaussian inverse problem ...
  - One linear solve per independent sample is optimal ...
  - ... independent of image dimension
  - Faster than Gibbs (including dimension-independent parametrization), one-block, regularization
  - Does not require trace-class prior covariance, nor consistent discretization

Thank You

Take-home messages

- Don't do Gibbs unless you have a very good reason
- If the full conditional over x has known form, do MTC
- No restriction on prior
- For censored data example, sampling is independent of data size
- For the linear-Gaussian inverse problem ...
  - One linear solve per independent sample is optimal ...
  - ... independent of image dimension
  - Faster than Gibbs (including dimension-independent parametrization), one-block, regularization
  - Does not require trace-class prior covariance, nor consistent discretization

Autocorrelation of $\lambda = \delta/\gamma$ (periodic BC)



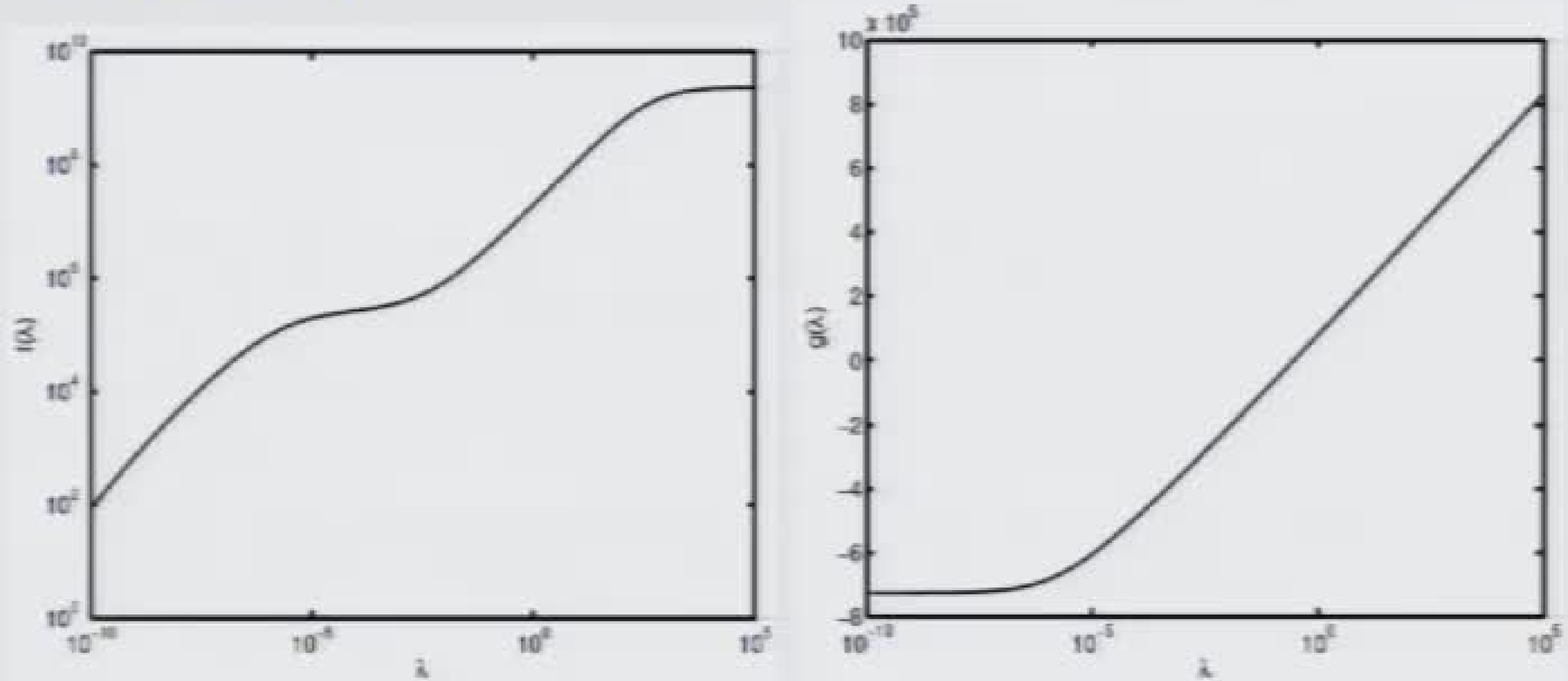
Gibbs slowest :: MTC opt. 2 cost per iteration is $O(1)$, all others 1 linear solve per iteration
Dimension-independent re-parametrization of Gibbs improves IACT to that of one-block, but increases cost per iteration to 3 linear solves, hence CCES unchanged.

Trace and log determinant

The marginal posterior for $\boldsymbol{\theta}$ can be written

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto \delta^{n/2} \exp\left(-\frac{1}{2} g(\lambda) - \frac{\gamma}{2} f(\lambda)\right) \pi(\boldsymbol{\theta})$$

where $\lambda = \delta/\gamma$, and the functions $f(\lambda) = (\boldsymbol{A}^T \boldsymbol{y})^T ((\boldsymbol{A}^T \boldsymbol{A})^{-1} - (\boldsymbol{A}^T \boldsymbol{A} + \lambda \boldsymbol{L})^{-1})(\boldsymbol{A}^T \boldsymbol{y})$ and $g(\lambda) = \log\det(\boldsymbol{A}^T \boldsymbol{A} + \lambda \boldsymbol{L})$



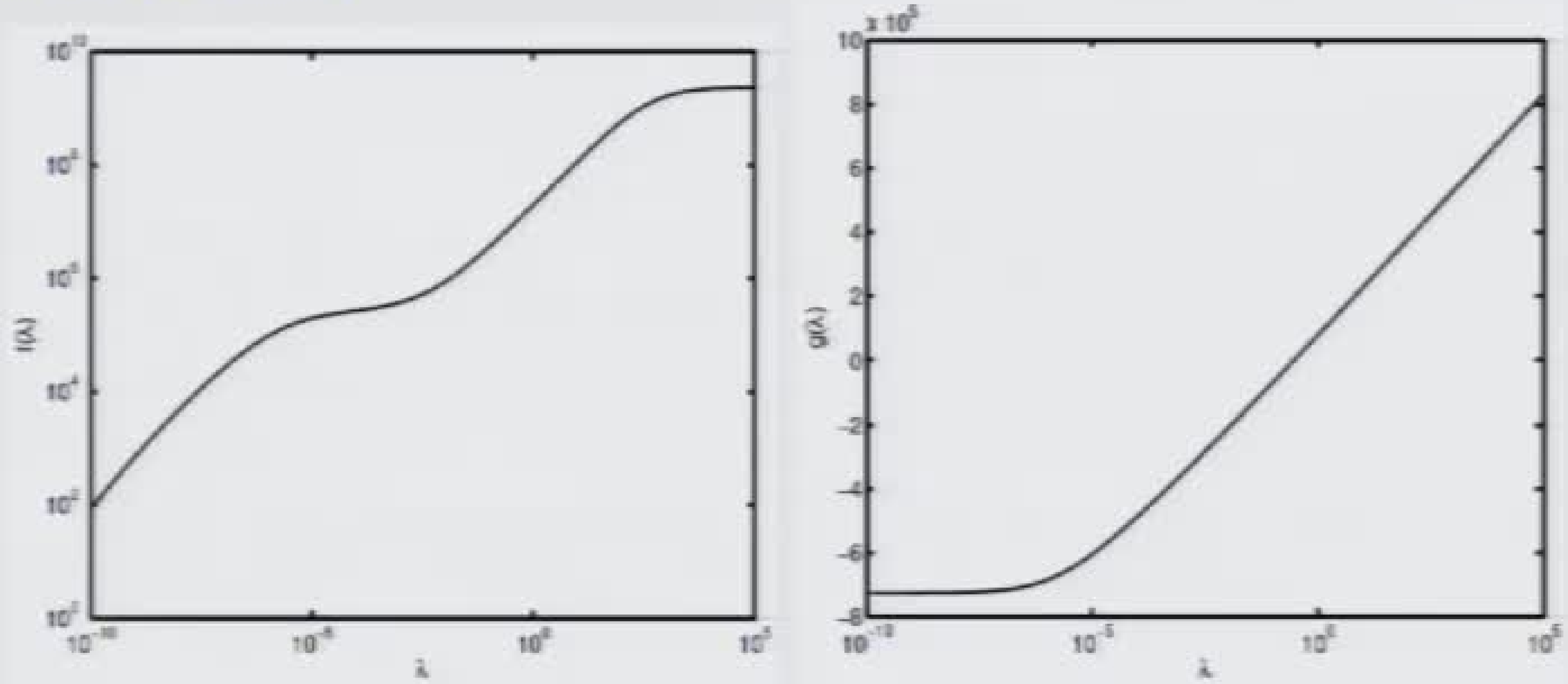
are uni-variate, monotonic, smooth, analytic (periodic case shown)

Trace and log determinant

The marginal posterior for $\boldsymbol{\theta}$ can be written

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto \delta^{n/2} \exp\left(-\frac{1}{2} g(\lambda) - \frac{\gamma}{2} f(\lambda)\right) \pi(\boldsymbol{\theta})$$

where $\lambda = \delta/\gamma$, and the functions $f(\lambda) = (\boldsymbol{A}^T\boldsymbol{y})^T((\boldsymbol{A}^T\boldsymbol{A})^{-1} - (\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{L})^{-1})(\boldsymbol{A}^T\boldsymbol{y})$ and $g(\lambda) = \log\det(\boldsymbol{A}^T\boldsymbol{A} + \lambda\boldsymbol{L})$



are uni-variate, monotonic, smooth, analytic (periodic case shown)