DOE/NV/03624–0112

Localization for MCMC: sampling high-dimensional posterior densities with local structure

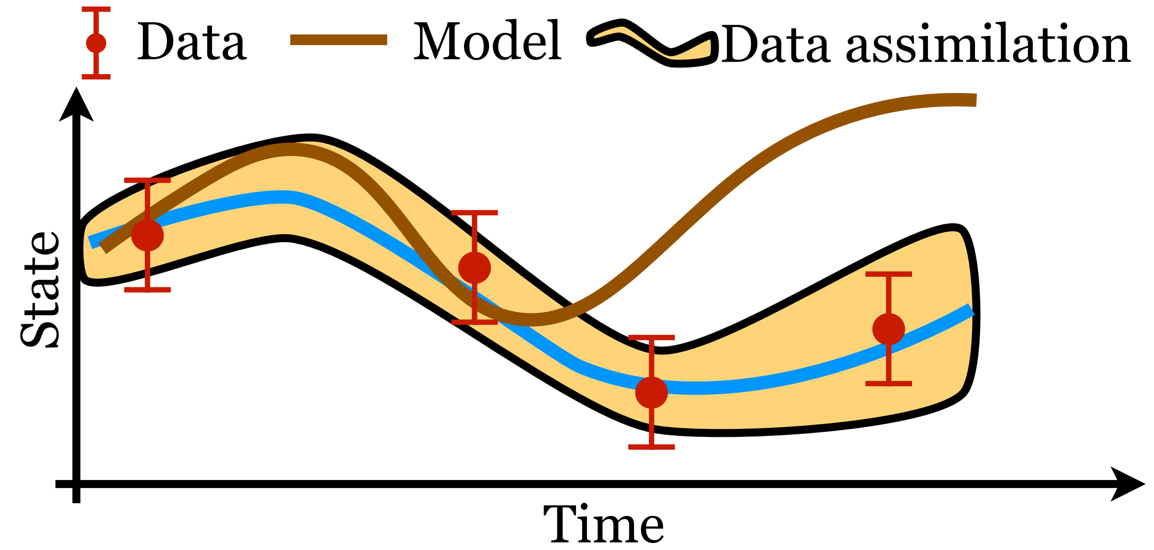
Matthias Morzfeld[o], Xin T. Tong[*], Youssef Marzouk[#]

[o]*University of Arizona*
[*] *National University of Singapore*
[#]*Massachusetts Institute of Technology*

Supported by

This work was partially supported by Mission Support and Test Services, LLC, under Contract No. DE-NA0003624, with the U.S. Department of Energy, and supported by the Site-Directed Research and Development Program

What is localization?



Model:

$$x^k = \mathcal{M}(x^{k-1})$$

Observations:

$$y^k = h(x^k) + \eta^k, \quad \eta^k \sim \mathcal{N}(0, R), \text{ iid}$$

Posterior distribution:

$$p(x^k | y^k) \propto p_0(x^k) p_l(y^k | x^k)$$

EnKF:
- Monte Carlo version of Kalman filter
- Uses ensemble to represent posterior distribution

What is localization?

EnKF

Forecast step:

$$x_i^f = \mathcal{M}(x_i^{k-1})$$
$$P^f = cov(x_i^f)$$

Kalman gain:

$$K = P^f H^T (H P^f H^T + R)^{-1}$$

Analysis ensemble:

$$x_i^k = x_i^f + K(y^k - Hx_i^f + v_i)$$
$$P_a = cov(x_i^k)$$

$$\begin{pmatrix} 2 & -1 & 0.01 & 10^{-3} & 10^{-5} \\ -1 & 2 & -1 & 10^{-3} & 10^{-4} \\ 0.01 & -1 & 2 & -1 & 10^{-4} \\ 10^{-3} & 10^{-3} & -1 & 2 & -1 \\ 10^{-5} & 10^{-4} & 10^{-4} & -1 & 2 \end{pmatrix}$$

$$\downarrow$$

$$\begin{pmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{pmatrix}$$

Dimensions

- Typical number of vars.: 650 million
- Typical number of obs.: 2–10 million
- Typical ensemble size: 50–100

Why should ensemble mean or covariance have any accuracy?

(Part of the) solution: Localization

- Small ensemble size leads to spurious long-range correlations in forecast covariance
- Simple idea: set (small) elements in forecast covariance equal to zero.

Localization is required to make EnKF work.

Localization of inverse problems

	Inverse problem		**Localized inverse problem**
Prior:	$p_0(x) = \mathcal{N}(\mu, C)$	$C \to C_{\mathrm{loc}}$	$p_{0,\mathrm{loc}}(x) = \mathcal{N}(\mu, C_{\mathrm{loc}})$
Observations:	$y = Hx + \eta$ $\eta \sim \mathcal{N}(0, R)$	$H \to H_{\mathrm{loc}}$	$y = H_{\mathrm{loc}}x + \eta$ $\eta \sim \mathcal{N}(0, R)$
Likelihood:	$p_l(y\|x) = \mathcal{N}(Hx, R)$		$p_{l,\mathrm{loc}}(y\|x) = \mathcal{N}(H_{\mathrm{loc}}x, R)$
Posterior:	$p(x\|y) \propto p_0(x)p_l(y\|x)$		$p_{\mathrm{loc}}(x\|y) \propto p_{0,\mathrm{loc}}(x)p_{l,\mathrm{loc}}(y\|x)$

$$\begin{pmatrix} 2 & -1 & 0.01 & 10^{-3} & 10^{-5} \\ -1 & 2 & -1 & 10^{-3} & 10^{-4} \\ 0.01 & -1 & 2 & -1 & 10^{-4} \\ 10^{-3} & 10^{-3} & -1 & 2 & -1 \\ 10^{-5} & 10^{-4} & 10^{-4} & -1 & 2 \end{pmatrix} \qquad \begin{pmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{pmatrix}$$

Localization of inverse problems

	Inverse problem		**Localized inverse problem**
Prior:	$p_0(x) = \mathcal{N}(\mu, C)$	$C \to C_{\text{loc}}$	$p_{0,\text{loc}}(x) = \mathcal{N}(\mu, C_{\text{loc}})$
Observations:	$y = Hx + \eta$ $\eta \sim \mathcal{N}(0, R)$	$H \to H_{\text{loc}}$	$y = H_{\text{loc}}x + \eta$ $\eta \sim \mathcal{N}(0, R)$
Likelihood:	$p_l(y\|x) = \mathcal{N}(Hx, R)$		$p_{l,\text{loc}}(y\|x) = \mathcal{N}(H_{\text{loc}}x, R)$
Posterior:	$p(x\|y) \propto p_0(x)p_l(y\|x)$		$p_{\text{loc}}(x\|y) \propto p_{0,\text{loc}}(x)p_{l,\text{loc}}(y\|x)$

- If prior covariance matrix is "nearly" banded

  If each element of *Hx* depends significantly only on a few elements of *x*

  Then: localized posterior mean/covariance are small perturbations of posterior mean/covariance
- Same results hold when prior precision matrix is localized
- Localization is easy to apply in nonlinear problems (see NWP)

Agenda

1. What is localization?
2. **Why should we localize?**
3. Numerical illustrations

High-dimensional local problems

1. State dimension, n, and number of observations, k, is large ($k = O(n)$)
2. Prior covariance matrix *and* prior precision matrix are "nearly" banded
3. Each element of $h(x)$ depends significantly only on a few elements of x
4. R is diagonal

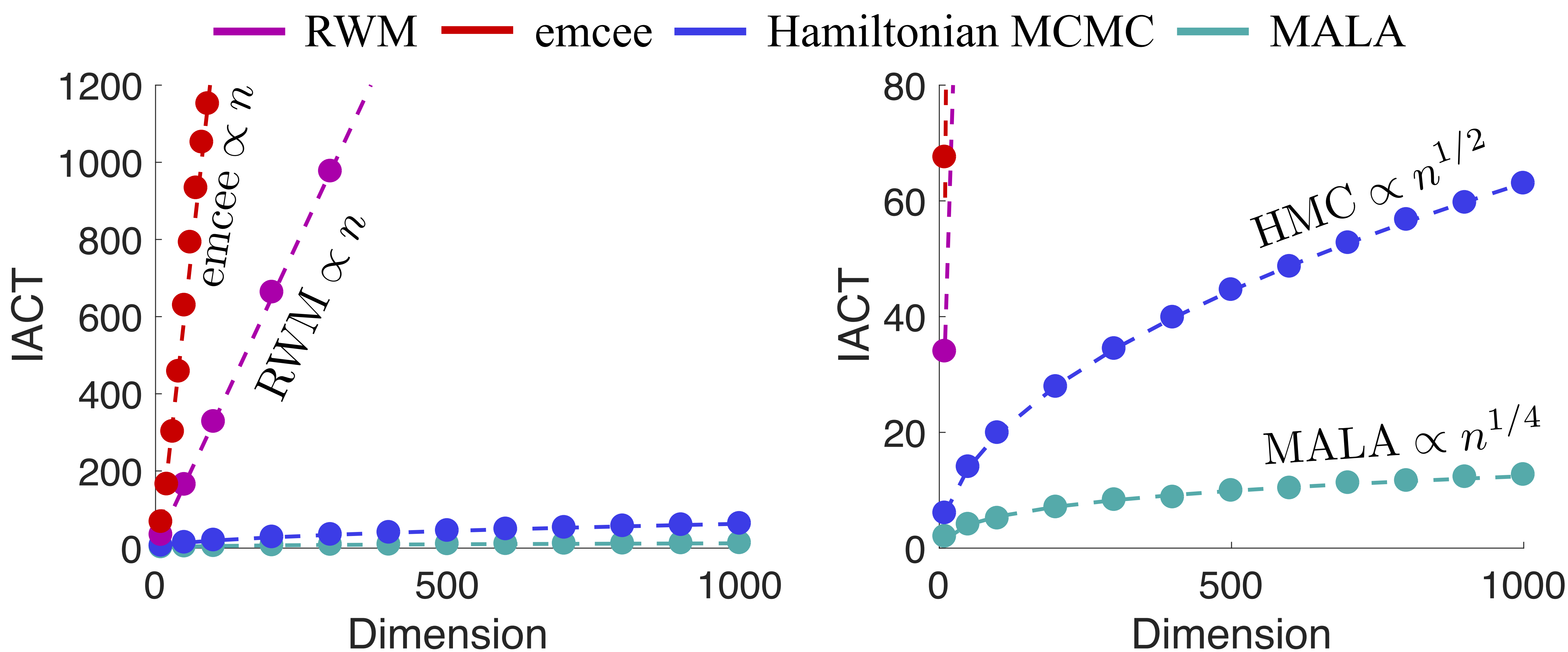
Easy but popular example*: isotropic Gaussian

$$p(x) = \mathcal{N}(0, I)$$

- Extreme example of a local problem
- Has been used to study importance sampling (particle filters) for high-dimensional problems
- How does MCMC do when we sample an isotropic Gaussian and we increase its dimension?

* *Bickel et al.*, 2008, *Bengtsson et al.* 2008, *Snyder et al.* 2008, *Snyder* 2011, Snyder et al. 2015.

What type of MCMC is good for high-dimensional local problems?

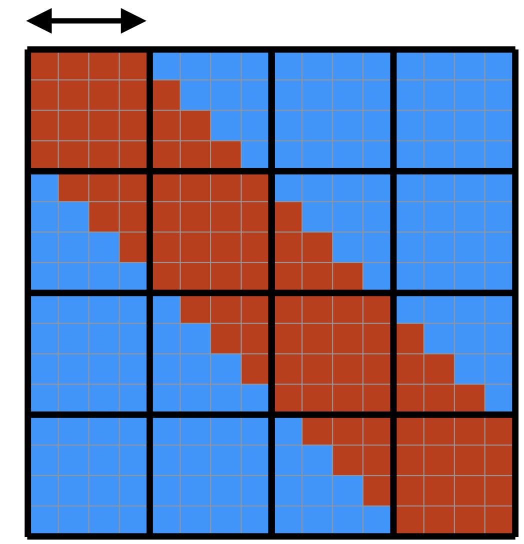


$$N_{e,\text{eff}} = \frac{N_e}{\text{IACT}} \longrightarrow N_e \propto \text{IACT}$$

- Should be trivial (no dependence on dimension).
- Gibbs sampler naturally makes use of local structure and produces independent samples, independently of dimension.
- Can Gibbs sampler "works well" in less trivial local problems?

Gibbs sampler for linear problems with banded structure

q (block-size)



Assumptions:

1. Target density is Gaussian $p(x) = \mathcal{N}(\mu, C)$
2. C is q-block-tridiagonal, condition number is $\mathcal{C}$

$$[C]_{i,j} = 0 \quad \text{for} \quad (i,j) \notin \{(i,i),(i,i+1),(i,i-1), i=1,\cdots,m\}.$$

Result:

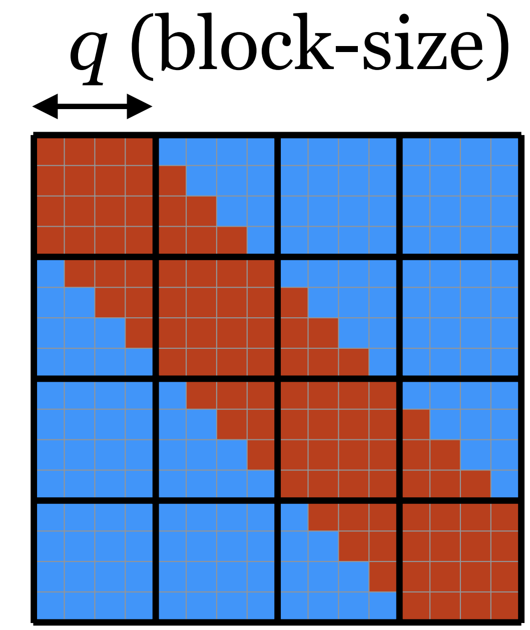
Let x^k be the samples of the Gibbs sampler (block-size q).
The distribution of x^k converges to p geometrically fast in all coordinates, and we can couple x^k and a sample $z \sim \mathcal{N}(\mu, C)$

$$E\|C^{-1/2}(x^k - z)\|^2 \leq \beta^k n(1 + \|C^{-1/2}(x^0 - \mu)\|^2),$$

$$\beta \leq \frac{2(1-\mathcal{C}^{-1})^2\mathcal{C}^4}{1+2(1-\mathcal{C}^{-1})^2\mathcal{C}^4},$$

1. **Convergence *rate* independent of dimension**
2. **Convergence if quick if condition number is “small”**

Gibbs sampler for linear problems with banded structure



Assumptions:

$p(x) = \mathcal{N}(\mu, \Omega^{-1})$

1. Target density is Gaussian $\cancel{p(x) = \mathcal{N}(\mu, C)}$
2. Ω $\cancel{C}$ is q-block-tridiagonal, condition number is $\mathcal{C}$

$$[\Omega]_{i,j} = 0 \quad \text{for} \quad (i,j) \notin \{(i,i),(i,i+1),(i,i-1), i = 1, \cdots, m\}.$$

Result:

Let x^k be the samples of the Gibbs sampler (block-size q).
The distribution of x^k converges to p geometrically fast in all coordinates, and we can couple x^k and a sample $\cancel{z \sim \mathcal{N}(\mu, C)}$ $z \sim \mathcal{N}(\mu, \Omega^{-1})$

$$E\|C^{-1/2}(x^k - z)\|^2 \leq \beta^k n(1 + \|C^{-1/2}(x^0 - \mu)\|^2),$$

$$\cancel{\beta \leq \frac{2(1-\mathcal{C}^{-1})^2\mathcal{C}^4}{1+2(1-\mathcal{C}^{-1})^2\mathcal{C}^4}}, \quad \beta \leq \frac{\mathcal{C}(1-\mathcal{C}^{-1})^2}{1+\mathcal{C}(1-\mathcal{C}^{-1})^2},$$

1. Convergence *rate* independent of dimension
2. Convergence if quick if condition number is “small”

Summary: Gibbs sampler

Convergence rate is independent of dimension if:

Theorem by Bickel & Lindner (2012)

1. Target density is Gaussian
2. Covariance matrix is q-block-tridiagonal, condition number is small

   or

3. Precision matrix is q-block-tridiagonal, condition number is small

Gibbs sampler useful for high-dimensional Gaussians with local statistical interactions (correlation & conditional dependence)

Metropolis-within-Gibbs sampler for nonlinear local problems

Metropolis-within-Gibbs:

1. Propose local move using block-Gibbs sampler for prior

$$x_j' \sim p(x_j | x_1^{k+1}, \dots x_{j-1}^{k+1}, x_{j+1}^k, \dots, x_n^k)$$

2. Accept local move with probability

$$a = \min\left\{1, \frac{\exp\left(-0.5||R^{-1/2}(y - h(x'))||^2\right)}{\exp\left(-0.5||R^{-1/2}(y - h(x))||^2\right)}\right\}$$

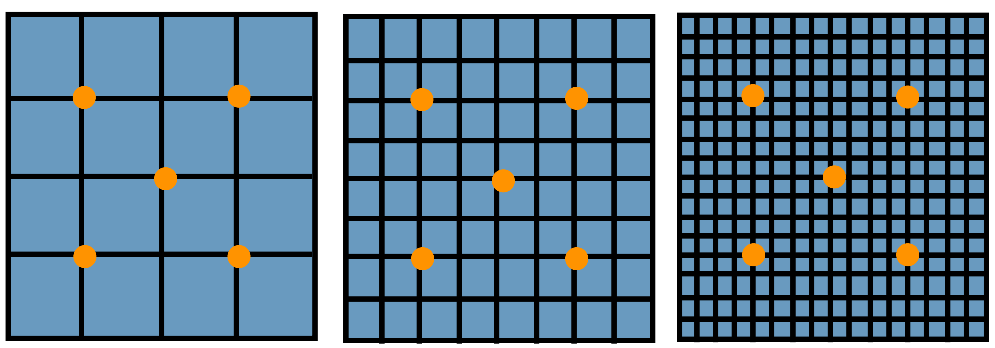
- *If precision matrix is given and banded: condition on neighbors only*
- *Linear algebra for matrices of size of the blocks, not overall dimension*
- *Proposal covariance independent of dimension.*

- *Acceptance high because change is local*
- *Acceptance rate independent of dimension?*

Connections with function space MCMC/dimension independence

Function space MCMC:

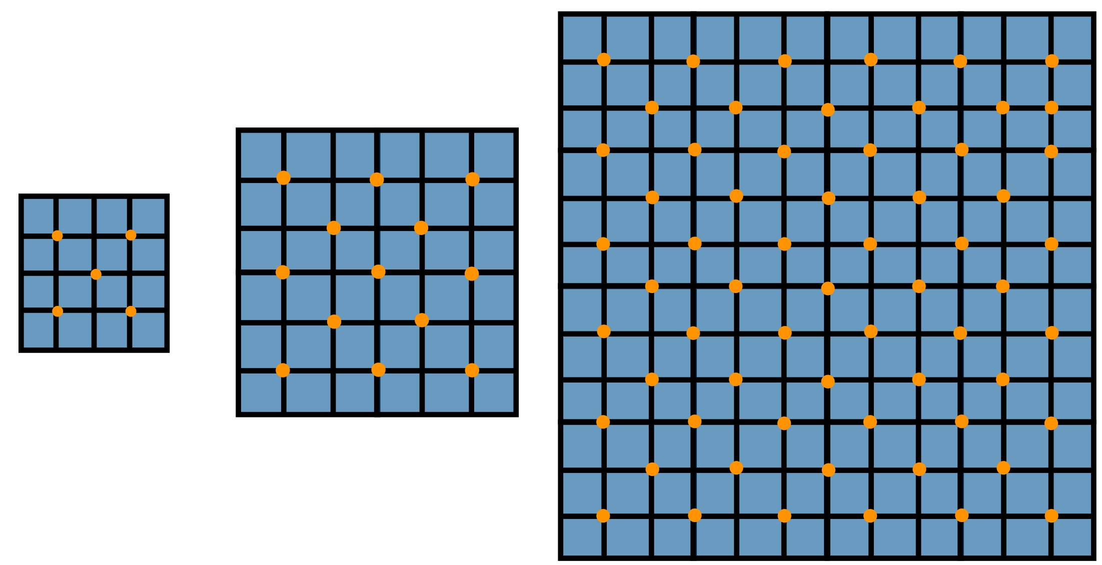
1. Discretization refines
2. Dimension increases
3. Number of obs. const.
4. “Effective dimension” const.
5. MCMC is dimension invariant



- ***Low-rank priors***
- ***Small number of obs.***
- ***Low-rank prior to posterior updates***
- ***Low effective dimension***

MCMC for local problems:

1. Discretization is const.
2. Dimension increases
3. Number of obs. increases.
4. “Effective dimension” increases.
5. MCMC is dimension invariant



- ***High-rank but sparse priors***
- ***Large number of obs.***
- ***High-rank prior to posterior updates***
- ***Large effective dimension***

Agenda

1. What is localization?
2. Why should we localize?
3. **Numerical illustrations**

Image deblurring — Gibbs sampler

Problem formulation

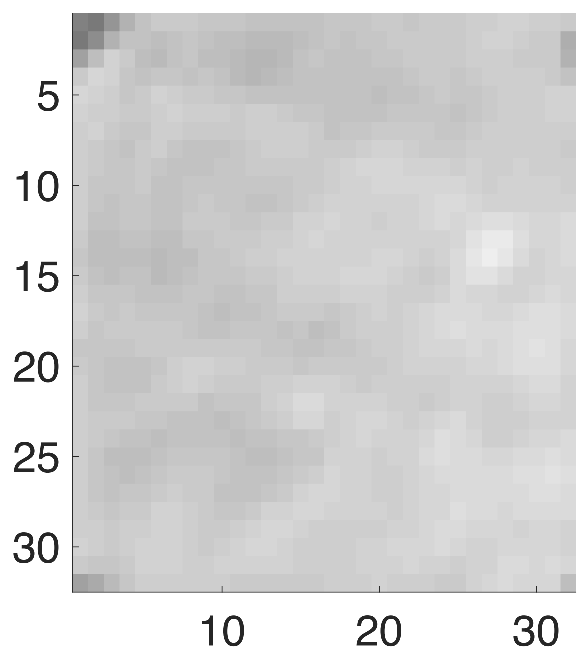
- Gaussian prior (Laplacian as precision)
- Linear problem (convolution)
- Dimension is large ~10^4
- Effective dimension huge

$$p(x) = \mathcal{N}(0, \delta^{-1} L^{-1})$$

$$y = Hx + \eta, \quad \eta \sim \mathcal{N}(0, \lambda^{-1} I)$$

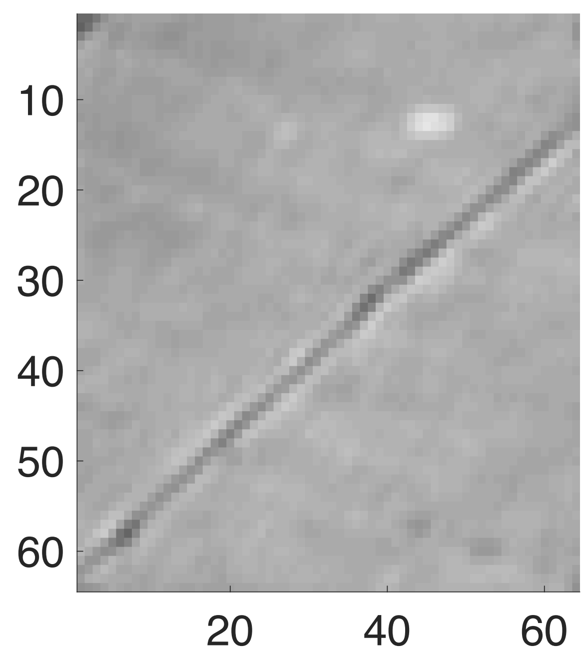
$$p(x|y) \propto \exp\left(-\frac{\lambda}{2} ||Hx - y||^2 - \frac{\delta}{2} ||Lx||^2 \right)$$

32 x 32



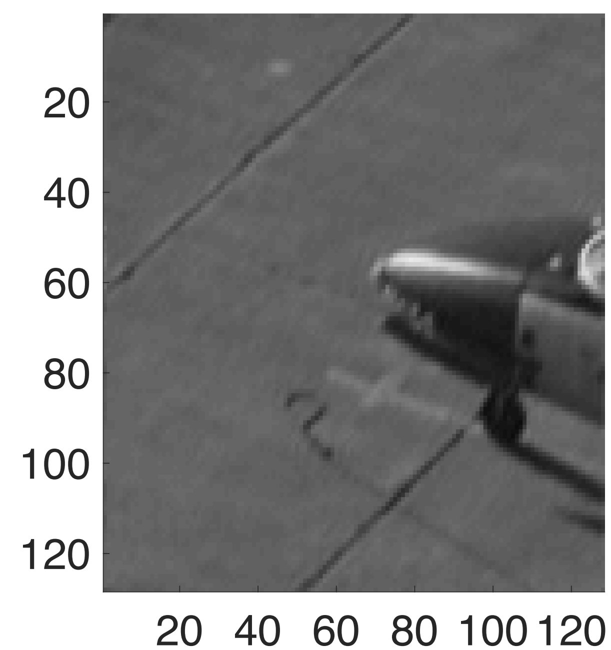
$n = 1,024$

64 x 64



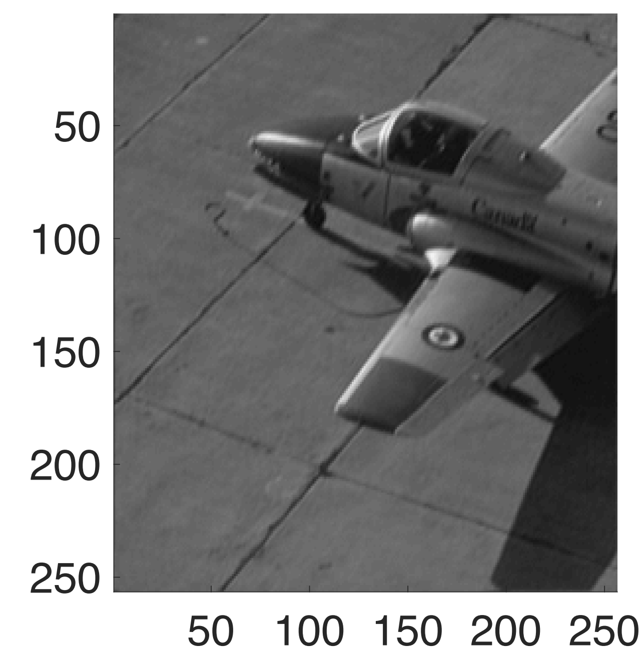
$n = 4,096$

128 x 128



$n = 16,384$

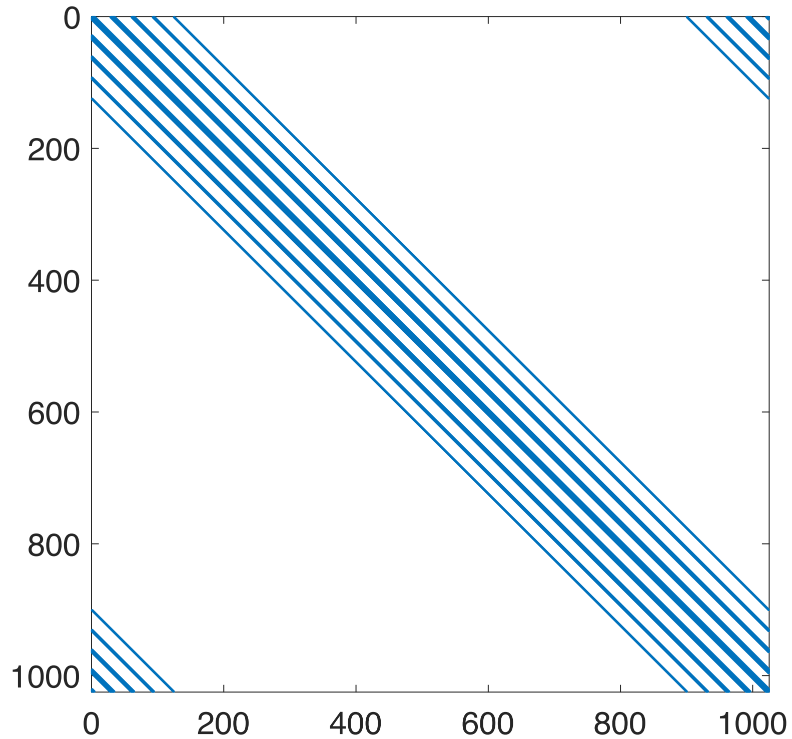
256 x 256

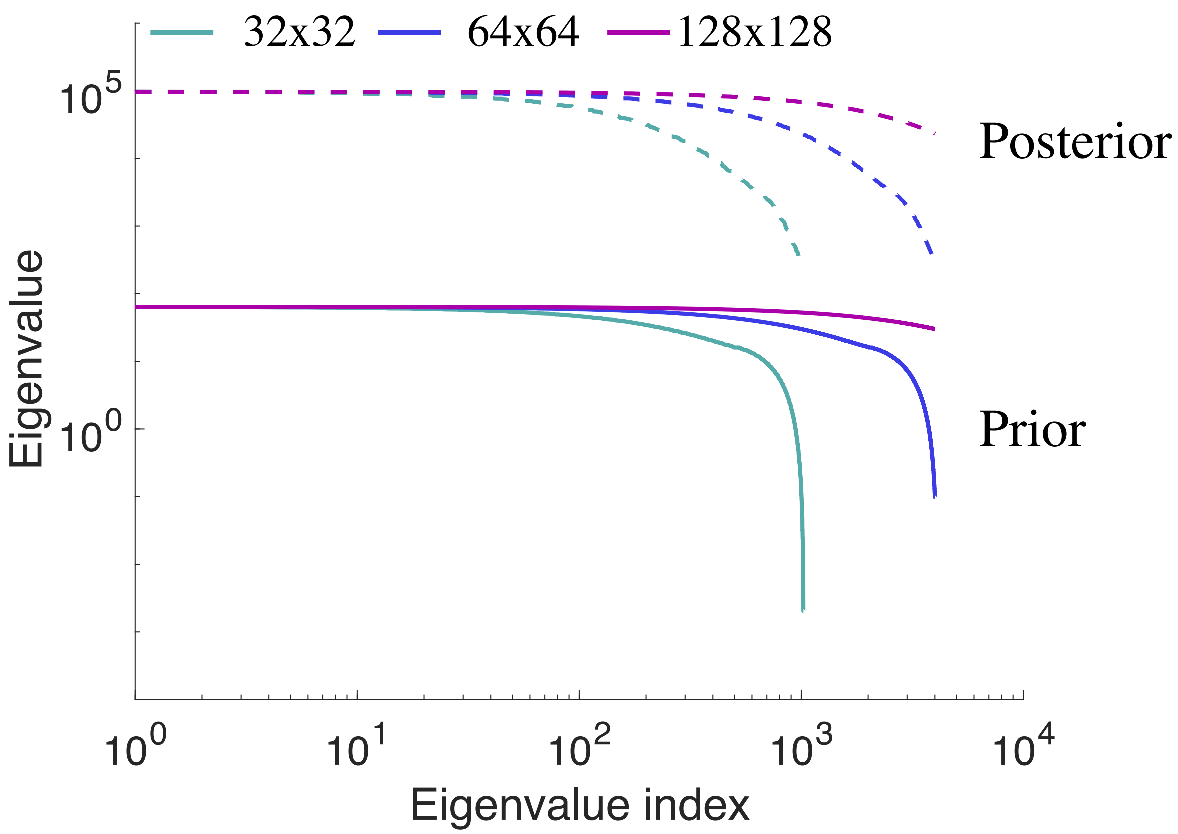


$n = 65,536$

Image deblurring — Gibbs sampler

$$\Omega = \lambda H^T H + \delta L x$$





- Only nearby pixels are blurred
- *H* is banded
- Prior precision is banded
- Posterior precision is banded

- Number of “large” eigenvalues increases with image size (dimension)
- Effective dimension increases with image size (dimension)

Image deblurring — Gibbs sampler

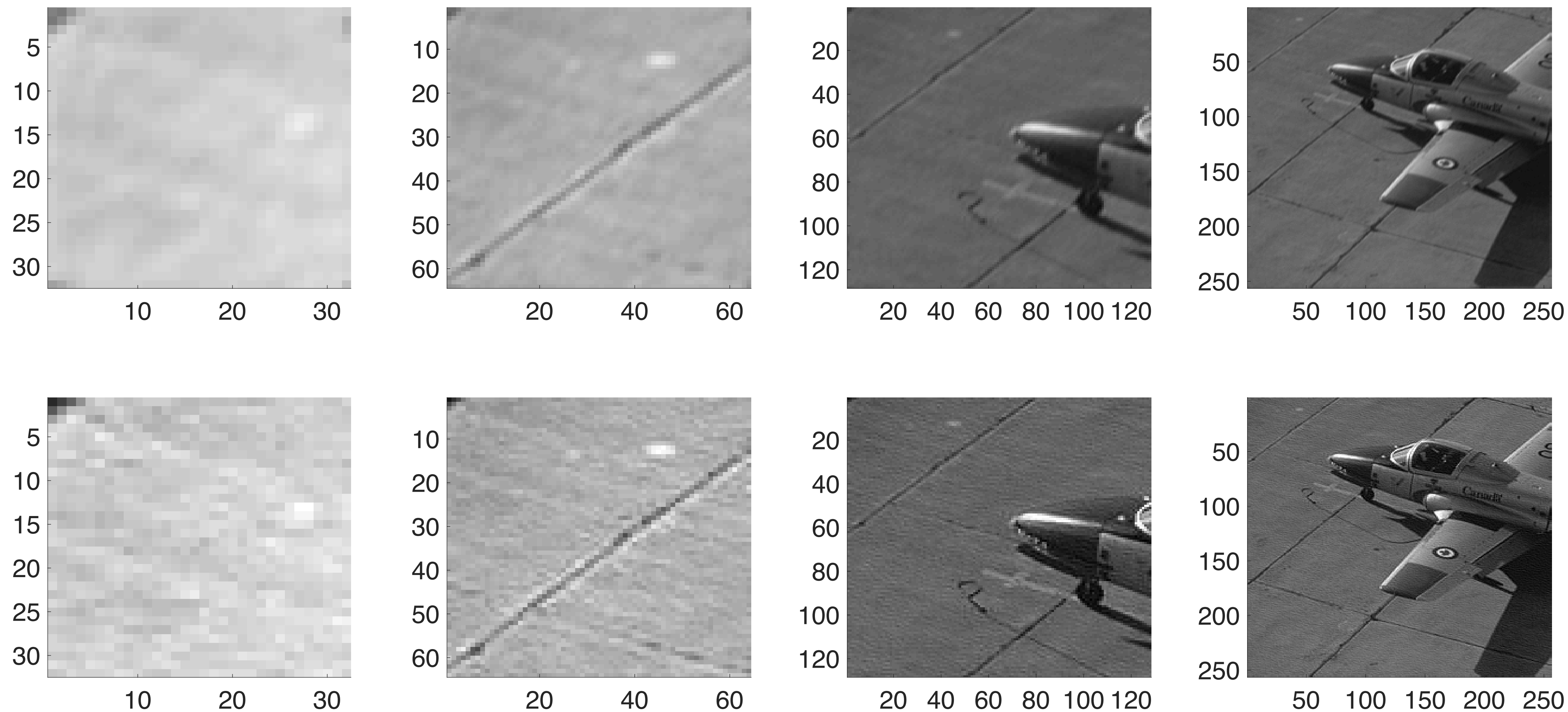
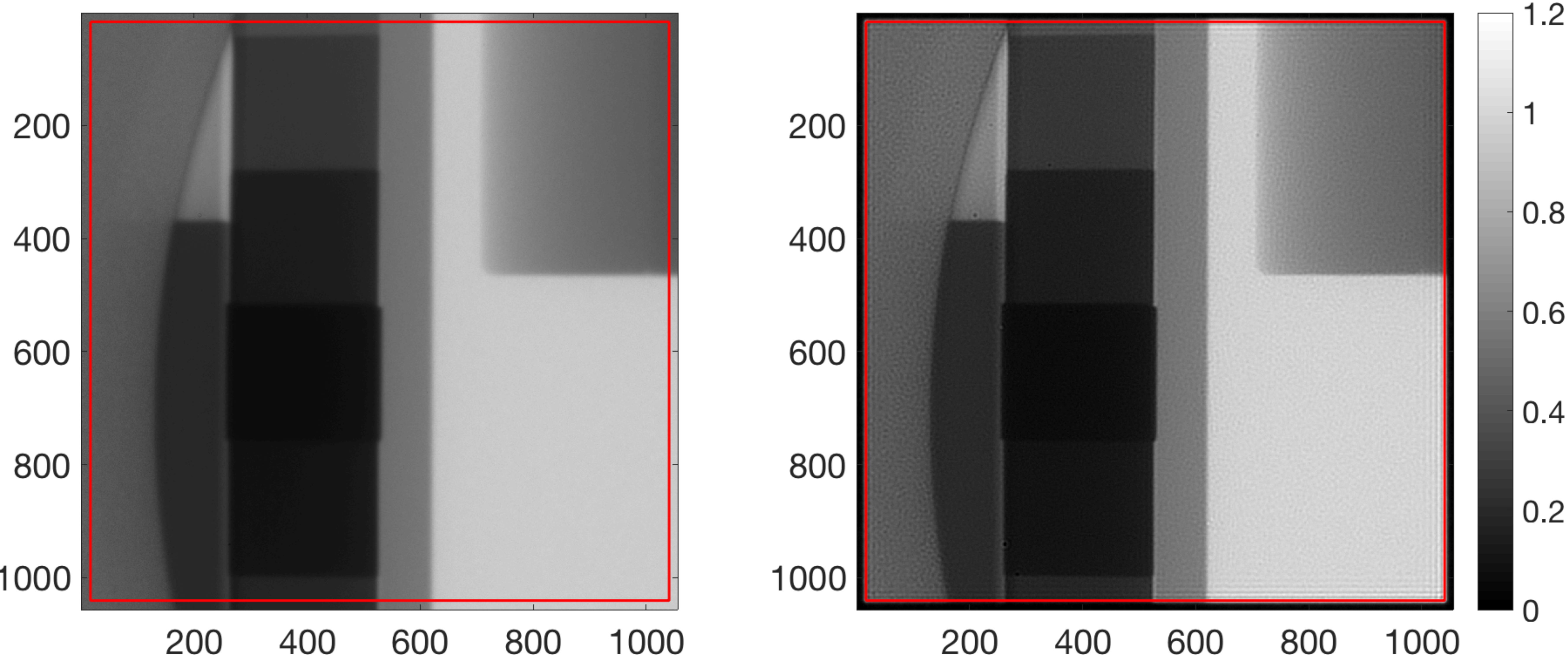


Image size	32 x 32	64 x 64	128 x 128	256 x 256
Dimension	1,024	4,096	16,348	16,536
Eff. Dimension	$4.8 \cdot 10^{8}$	$7.4 \cdot 10^{9}$	$1.2 \cdot 10^{11}$	-
IACT (Gibbs)	2.92	2.97	1.74	1.11
Blocksize (Gibbs)	16	16	32	64

Image deblurring — Gibbs sampler

- Scales well to larger problems (10^6 — 10^7)
- No need to assemble matrices — assemble required blocks on the fly
- See Jesse Adam's talk (yesterday)



Lorenz ‘96

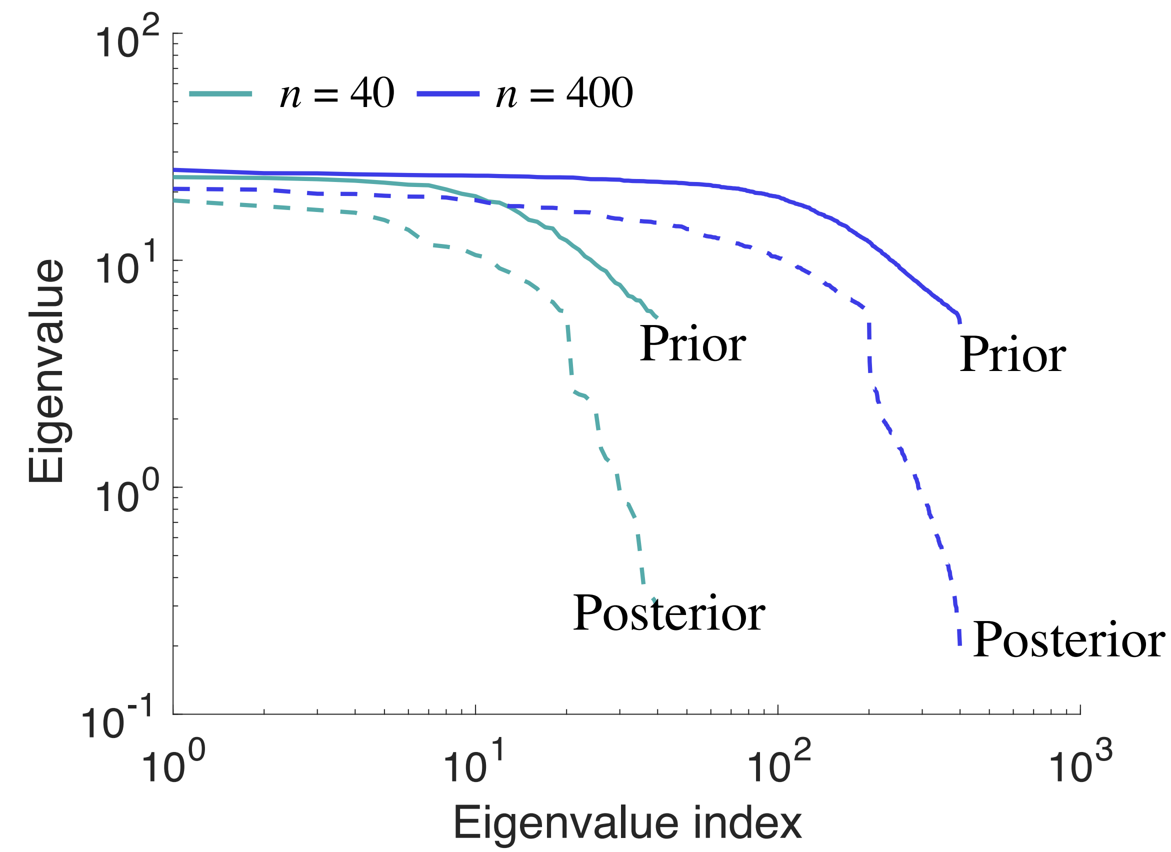
Goal: Estimate initial conditions of L96, given noisy observations at time T

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = (x_{i+1} - x_{i-2})x_{i-1} - x_i + 8$$

$$y = H\mathcal{M}_{0\to T}(x_0) + \eta, \quad \eta \sim \mathcal{N}(0, I)$$

Problem formulation

- Gaussian prior (“Climatology”)
- L96 dynamics (RK4)
- Dimension $n = 40$ or $n = 400$
- $k = n/2$ observations at time $T = 0.2$
- Eff. dim $n_{\text{eff}} = 18$ or $n_{\text{eff}} = 181$
- Number of “large” evals increases with n



Algorithms tested

1. pCN
2. MALA
3. l-MwG

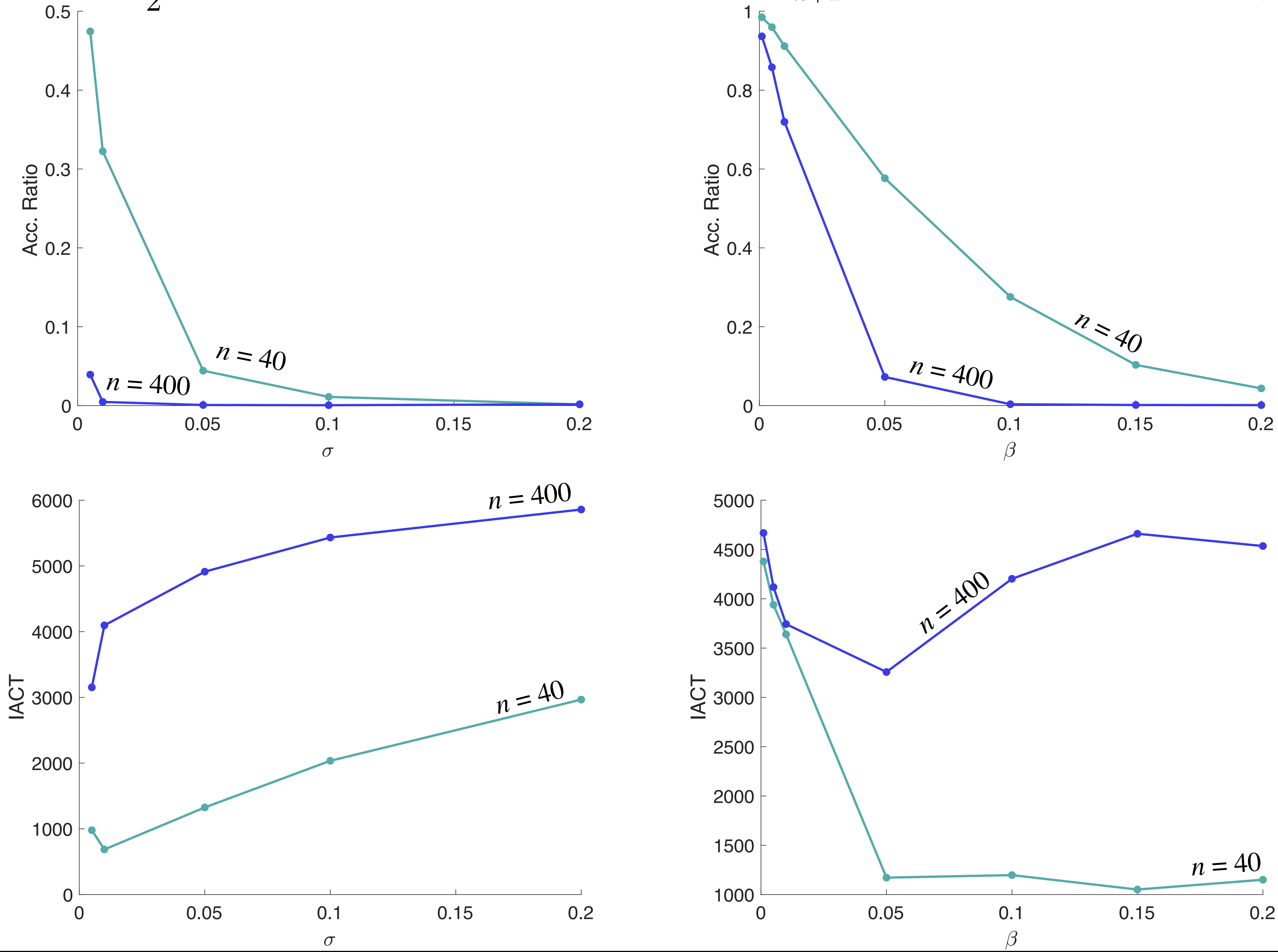
Lorenz ’96: tuning of MALA and pCN

MALA

$$x'_{k+1} = x_k + \frac{\sigma^2}{2}\nabla \log p(x_k) + \sigma\,\xi \quad \xi \sim \mathcal{N}(0, C_{\mathrm{MALA}})$$

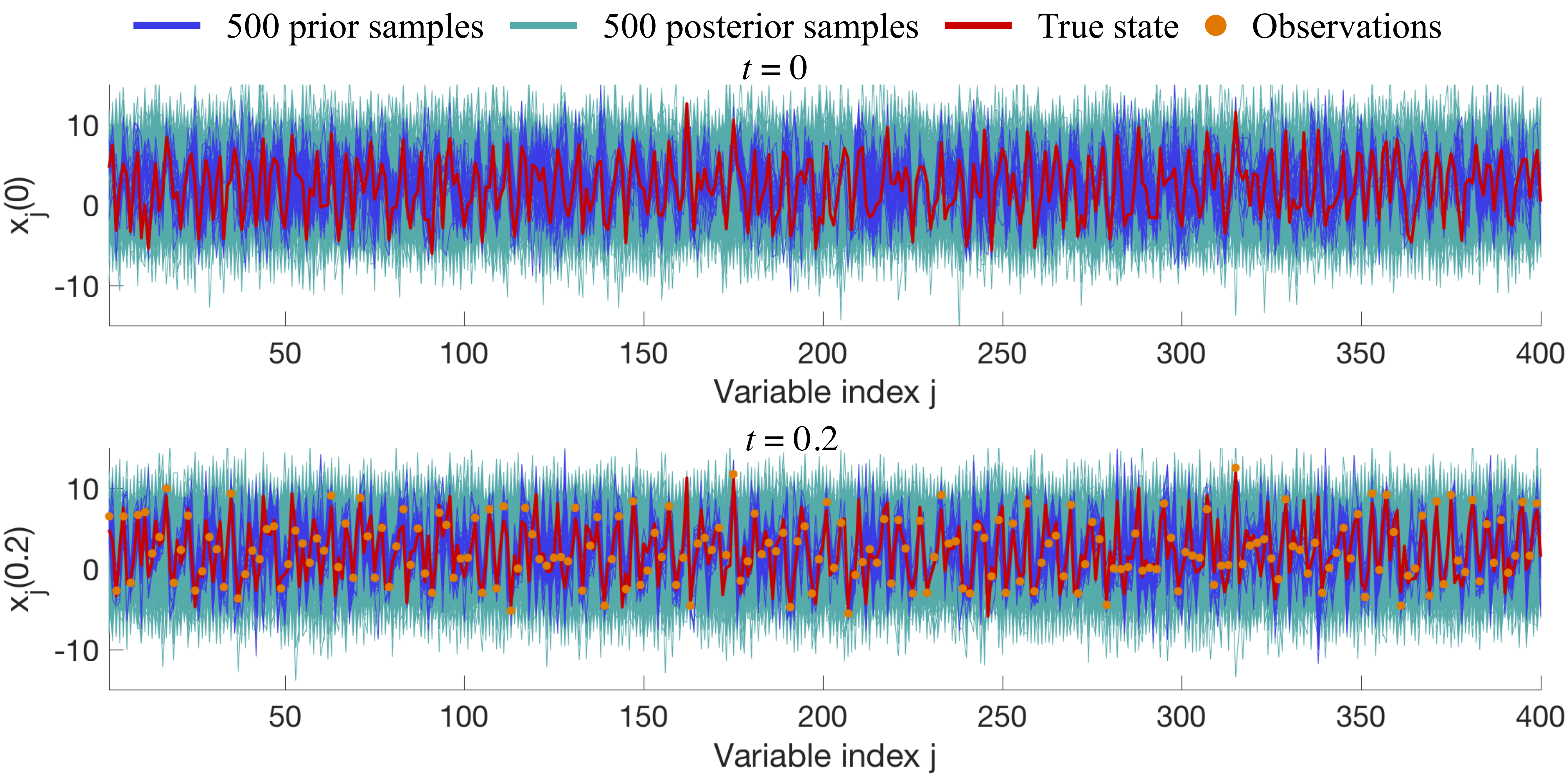
pCN

$$\Delta x'_{k+1} = \sqrt{1-\beta^2}\Delta x_k + \beta\,\xi, \xi \sim \mathcal{N}(0, C)$$



Lorenz '96: MALA, pCN and MwG

	MALA	pCN	l-MwG-B2	l-MwG-B4	l-MwG-B8
$n = 40$	686	1051	55	60	266
$n = 400$	3,153	3,257	43	81	257



Localization

- *NWP*: delete spurious correlations and restrict influence of observations to a neighborhood
- *Inverse problems*: enforce prior statistical interactions (correlations/ conditional dependencies) to be local and restrict influence of an observation to its neighborhood
- Localization introduces small errors if the problem is “local”

MCMC for localized problems

- Some MCMC algorithms struggle on local problems
- MCMC based on Gibbs sampler may be promising
- Gibbs sampler prototypical algorithm for exploring local structure
- Notion of “high-dimension” different from function space MCMC

Thank you.

NEVADA NATIONAL NNSS SECURITY SITE
Alfred P. Sloan FOUNDATION
U.S. NAVAL RESEARCH LABORATORY