Localization for high dimensional MCMC

Xin T Tong

National University of Singapore
Joint work with Matthias Morzfeld and Youssef Marzouk

Bayesian Computations for High-Dimensional statistical models.
IMS, Singapore 2018
- Curse of dimensionality for MCMC.
- Localization technique from numerical weather forecast.
- Localization for MCMC.
- Numerical examples.
Bayesian inverse problem

- Suppose $\mathbf{x} \sim p_0 = \mathcal{N}(\mathbf{m}, \mathbf{C})$, we observe

$$
\mathbf{y} = h(\mathbf{x}) + \mathbf{v}, \quad \mathbf{v} \sim \mathcal{N}(0, R).
$$

Try to recover the value and uncertainty of $\mathbf{x}$.

- Possible applications:
  - $\mathbf{x}$ is the real image, $h$ defocus map.
  - $\mathbf{x}$ initial condition, $h$ forward map of a PDE.
  - $\mathbf{x}$ model parameters, $h$ gives model outcome.

Often $\mathbf{x}$ is high dimension.

- Bayesian approach: try to sample the posterior

$$
p(\mathbf{x}|\mathbf{y}) \propto p_0(\mathbf{x})p_l(\mathbf{y}|\mathbf{x}).
$$

$$
p_l(\mathbf{y}|\mathbf{x}) = \mathcal{N}(h(\mathbf{x}), R).
$$
Bayesian inverse problem

- Suppose $\mathbf{x} \sim p_0 = \mathcal{N}(\mathbf{m}, \mathbf{C})$, we observe
  \[ \mathbf{y} = h(\mathbf{x}) + \mathbf{v}, \quad \mathbf{v} \sim \mathcal{N}(0, \mathbf{R}). \]
  Try to recover the value and uncertainty of $\mathbf{x}$.

- Possible applications:
  - $\mathbf{x}$ is the real image, $h$ defocus map.
  - $\mathbf{x}$ initial condition, $h$ forward map of a PDE.
  - $\mathbf{x}$ model parameters, $h$ gives model outcome.
  Often $\mathbf{x}$ is high dimension.

- Bayesian approach: try to sample the posterior
  \[ p(\mathbf{x}|\mathbf{y}) \propto p_0(\mathbf{x})p_l(\mathbf{y}|\mathbf{x}). \]
  
  \[ p_l(\mathbf{y}|\mathbf{x}) = \mathcal{N}(h(\mathbf{x}), \mathbf{R}). \]
Bayesian inverse problem

- Suppose $\mathbf{x} \sim p_0 = \mathcal{N}(\mathbf{m}, \mathbf{C})$, we observe

$$\mathbf{y} = h(\mathbf{x}) + \mathbf{v}, \quad \mathbf{v} \sim \mathcal{N}(0, R).$$

Try to recover the value and uncertainty of $\mathbf{x}$.

- Possible applications:
  - $\mathbf{x}$ is the real image, $h$ defocus map.
  - $\mathbf{x}$ initial condition, $h$ forward map of a PDE.
  - $\mathbf{x}$ model parameters, $h$ gives model outcome.

Often $\mathbf{x}$ is high dimension.

- Bayesian approach: try to sample the posterior

$$p(\mathbf{x}|\mathbf{y}) \propto p_0(\mathbf{x})p_l(\mathbf{y}|\mathbf{x}).$$

$$p_l(\mathbf{y}|\mathbf{x}) = \mathcal{N}(h(\mathbf{x}), R).$$
MCMC sampler

- Given a target distribution \( p(x) \), generate a sequence of samples:
  \[ x^{(1)}, x^{(2)}, \ldots, x^{(N)} \]

  Use sample statistics to approximate population ones.

- Standard MCMC steps
  - Generate proposals \( x' \sim q(x^{(k)}, x') \)
  - Accept with prob. \( \alpha(x, x') = \min\{1, p(x')q(x', x^{(k)})/q(x^{(k)}, x')p(x)\} \)

- Popular choices of proposals \( \xi_k \sim \mathcal{N}(0, I_n) \).
  - RWM: \( x' = x_k + \sigma \xi_k \)
  - MALA: \( x' = x_k + \frac{\sigma^2}{2} \nabla \log p(x_k) + \sigma \xi_k \).
  - pCN: \( \Delta x'_{k+1} = \sqrt{1 - \beta^2} \Delta x_k + \beta \xi_k \).

Also emcee and Hamiltonian MCMC.
\( \sigma, \beta \) are tuning parameters.
How does MCMC work in high dim?

Sample isotropic Gaussian $p = \mathcal{N}(0, I_n)$.

Measurement of efficiency: integrated auto-correlation time (IACT)
Measure how many iterations to get an “uncorrelated” sample.

Increases with dimension
Let’s look at RWM, assume $x_k = 0$.

- Propose $x' = \sigma \xi_k$
- Accept with probability $\exp(-\frac{1}{2} \sigma^2 \|\xi_k\|^2) \sim \exp(-\frac{1}{2} \sigma^2 n)$.

If we keep $\sigma = 1$, “never” accept if $n > 20$.

If we want acceptance at a constant rate, $\sigma = n^{-\frac{1}{2}}$. But then $x' = x_k + \sigma \xi_k$ is highly correlated with $x_k$.

Similar for MALA, $\sigma = n^{-1/3}$. Hamiltonian MCMC, $\sigma = n^{-1/4}$.

Is it possible to break this curse of dimensionality?

Is high dimensionality an issue in other related fields?
Where is the problem?

Let’s look at RWM, assume $x_k = 0$.

- Propose $x' = \sigma \xi_k$
- Accept with probability $\exp(-\frac{1}{2} \sigma^2 \|\xi_k\|^2) \sim \exp(-\frac{1}{2} \sigma^2 n)$.

If we keep $\sigma = 1$, “never” accept if $n > 20$.

If we want acceptance at a constant rate, $\sigma = n^{-1/2}$. But then $x' = x_k + \sigma \xi_k$ is highly correlated with $x_k$.

Similar for MALA, $\sigma = n^{-1/3}$. Hamiltonian MCMC, $\sigma = n^{-1/4}$.

Is it possible to break this curse of dimensionality?

Is high dimensionality an issue in other related fields?
Let’s look at RWM, assume $x_k = 0$.

- Propose $x' = \sigma \xi_k$
- Accept with probability $\exp\left(-\frac{1}{2} \sigma^2 \|\xi_k\|^2\right) \sim \exp\left(-\frac{1}{2} \sigma^2 n\right)$.

If we keep $\sigma = 1$, “never” accept if $n > 20$.

If we want acceptance at a constant rate, $\sigma = n^{-\frac{1}{2}}$.
But then $x' = x_k + \sigma \xi_k$ is highly correlated with $x_k$.

Similar for MALA, $\sigma = n^{-1/3}$. Hamiltonian MCMC, $\sigma = n^{-1/4}$.

Is it possible to break this curse of dimensionality?

Is high dimensionality an issue in other related fields?
Weather forecast:

Formulation: $y = h(\Psi(x)) + \nu$.

- $x$: atmosphere and ocean state of yesterday
- $\Psi$: PDE solver maps yesterday to today.
- $h$: partial observation made by satellites

Main challenge: high dimension, $n \sim 10^6 - 10^8$. 
EnKF is a popular tool

- Generate samples $\mathbf{x}^i$, $i = 1, \cdots, K$.
- Compute $\mathbf{y}^i = h(\Psi(\mathbf{x})) + \mathbf{v}^i$.
- Interpret the distribution of $(\mathbf{x}, \mathbf{y})$ as a Gaussian, with mean and cov based on $\{\mathbf{x}^i, \mathbf{y}^i\}$.
- Apply the Kalman formula to get the posterior.
- (Update $\mathbf{x}^i = \Psi(\mathbf{x}^i)$ and repeat).

Dimension curse comes from another angle

- The covariance of $\mathbf{x}$ estimated using the sample covariance

\[
\tilde{\mathbf{x}} = \frac{1}{K} \sum \mathbf{x}^i, \quad \hat{\mathbf{C}} = \frac{1}{K-1} \sum (\mathbf{x}^i - \tilde{\mathbf{x}})(\mathbf{x}^i - \tilde{\mathbf{x}})^T.
\]

- If $\mathbf{x}^i \sim \mathcal{N}(0, \Sigma)$, $\|\hat{\mathbf{C}} - \Sigma\| = O(\sqrt{n/K})$. 
The ensemble Kalman filter

EnKF is a popular tool

- Generate samples $x^i$, $i = 1, \cdots, K$.
- Compute $y^i = h(\Psi(x)) + v^i$.
- Interpret the distribution of $(x, y)$ as a Gaussian, with mean and cov based on $\{x^i, y^i\}$.
- Apply the Kalman formula to get the posterior.
- (Update $x^i = \Psi(x^i)$ and repeat).

Dimension curse comes from another angle

- The covariance of $x$ estimated using the sample covariance

$$\bar{x} = \frac{1}{K} \sum x^i, \quad \hat{C} = \frac{1}{K - 1} \sum (x^i - \bar{x})(x^i - \bar{x})^T.$$  

- If $x^i \sim \mathcal{N}(0, \Sigma)$, $||\hat{C} - \Sigma|| = O(\sqrt{n/K})$. 
High dimension often comes from dense grids.

Interaction often is local: PDE discretization.

Example: Lorenz 96 model

\[
\dot{x}_i(t) = (x_{i+1} - x_{i-2})x_{i-1} - x_i dt + F, \quad i = 1, \ldots, n
\]

Information travels along interaction, and is dissipated.
Sparsity: local covariance

- Correlation depends on information propagation.
- Correlation decays quickly with the distance.
- Covariance is localized with a structure $\Phi$, e.g. $\Phi(x) = \rho^x$

\[
[\hat{C}]_{i,j} \propto \Phi(|i - j|)
\]

$\Phi(x) \in [0, 1]$ is decreasing. Distance can be general.

Correlation of Lorenz 96
Covariance Localization

- Localization: use only local information.
- Implementation: Schur product with a mask

\[ \left[ \hat{C} \circ D_L \right]_{i,j} = \left[ \hat{C} \right]_{i,j} \cdot \left[ D_L \right]_{i,j} \]

Use \( \hat{C} \circ D_L \) to describe uncertainty

\[ [D_L]_{i,j} = \phi(|i - j|), \text{ with a radius } L. \]

Gaspari-Cohn matrix: \( \phi(x) = \exp(-4x^2/L^2)1_{|i-j| \leq L} \).

Cutoff matrix: \( \phi(x) = 1_{|i-j| \leq L} \).

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

\[
\begin{pmatrix}
2 & -1 & 0 & 0 & 0 \\
-1 & 2 & -1 & 0 & 0 \\
0 & -1 & 2 & -1 & 0 \\
0 & 0 & -1 & 2 & -1 \\
0 & 0 & 0 & -1 & 2
\end{pmatrix}
\]
Advantage with localization

Intuitively, ignoring the long distance covariance terms, reduces the sampling difficulty, and necessary sampling size.

Theorem (Bickel, Levina 08)

\[ X^{(1)}, \ldots, X^{(K)} \sim \mathcal{N}(0, \Sigma), \text{ sample covariance matrix} \]

\[ C = \frac{1}{K} \sum_{k=1}^{K} X^{(k)} \otimes X^{(k)}. \text{ There is a constant } c, \text{ and for any } t > 0 \]

\[ \mathbb{P}(\|C \circ D_L - \Sigma \circ D_L\| > \|D_L\|_1 t) \leq 8 \exp(2 \log n - cK \min\{t, t^2\}) \]

This indicates that \( K \propto \|D_L\|_1^2 \log n \) is the necessary sample size.

\[ \|D_L\|_1 = \max_i \sum_j |D_L|_{i,j} \text{ is independent of } d, \text{ e.g, the cut-off/branding matrix. } [D_{cut}^L]_{i,j} = 1_{|i-j| \leq L}, \|D_{cut}^L\|_1 \approx 2L. \]

Theorem (T. 18)

If EnKF has a stable localized covariance structure, the sample size is of order log n.
Intuitively, ignoring the long distance covariance terms, reduces the sampling difficulty, and necessary sampling size.

**Theorem (Bickel, Levina 08)**

\[ X^{(1)}, \ldots, X^{(K)} \sim \mathcal{N}(0, \Sigma), \text{ sample covariance matrix} \]

\[ C = \frac{1}{K} \sum_{k=1}^{K} X^{(k)} \otimes X^{(k)}. \text{ There is a constant } c, \text{ and for any } t > 0 \]

\[ \mathbb{P}(\|C \circ D_L - \Sigma \circ D_L\| > \|D_L\|_1 t) \leq 8 \exp(2 \log n - cK \min\{t, t^2\}) \]

*This indicates that* \( K \propto \|D_L\|_1^2 \log n \) *is the necessary sample size.*

\[ \|D_L\|_1 = \max_i \sum_j |D_L|_{i,j} \text{ is independent of } d, \text{ e.g, the cut-off/branding matrix. } [D_{cut}^L]_{i,j} = 1_{|i-j| \leq L}, \|D_{cut}^L\|_1 \approx 2L. \]

**Theorem (T. 18)**

*If EnKF has a stable localized covariance structure, the sample size is of order* \( \log n \).
From EnKF to MCMC?

- Localization for EnKF:
  - Update only in small local blocks.
  - Works when covariance have local structure.

How to apply localization to MCMC?

How to update $x^i$ component by component?

- Gibbs sampling implements this idea exactly!
  - Write $x^i = [x^i_1, x^i_2, \ldots, x^i_m]$.
  - $x^i_k$ can be of dimension $q$, then $n = qm$.
  - Generate $x^i_1 + 1 \sim p(x_1 | x^i_2, x^i_3 \ldots, x^i_m)$.
  - Generate $x^i_2 + 1 \sim p(x_2 | x^i_1 + 1, x^i_3 \ldots, x^i_m)$.
  - \ldots
  - Generate $x^i_m + 1 \sim p(x_m | x^i_1 + 1, x^i_2 + 1 \ldots, x^i_{m-1})$. 
From EnKF to MCMC?

- Localization for EnKF:
  - Update only in small local blocks.
  - Works when covariance have local structure.

How to apply localization to MCMC?

- How to update $x^i$ component by component?

Gibbs sampling implements this idea exactly!

- Write $x^i = [x^i_1, x^i_2, \ldots, x^i_m]$.
- $x^i_k$ can be of dimension $q$, then $n = qm$.
- Generate $x^i_1 + 1 \sim p(x_1 | x^i_2, x^i_3, \ldots, x^i_m)$.
- Generate $x^i_2 + 1 \sim p(x_2 | x^i_1 + 1, x^i_3, \ldots, x^i_m)$.
- \ldots
- Generate $x^i_m + 1 \sim p(x_m | x^i_1 + 1, x^i_2 + 1, \ldots, x^i_{m-1})$.  


How efficient is Gibbs?

First just test with $p = \mathcal{N}(0, I_n)$

- Generate $x_1^{i+1} \sim p(x_1 | x_2^i, x_3^i, \ldots, x_m^i) = \mathcal{N}(0, I_q)$.
- Generate $x_2^{i+1} \sim p(x_2 | x_1^{i+1}, x_3^i, \ldots, x_m^i) = \mathcal{N}(0, I_q)$.
- \ldots
- Generate $x_m^{i+1} \sim p(x_m | x_1^{i+1}, x_2^{i+1}, \ldots, x_{m-1}^{i+1}) = \mathcal{N}(0, I_q)$.

Gibbs naturally exploits the component independence. It works efficiently against the dimension. How about component with sparse/local independence?
How efficient is Gibbs?

First just test with $p = \mathcal{N}(0, I_n)$

- Generate $x_1^{i+1} \sim p(x_1 | x_2^i, x_3^i, \ldots, x_m^i) = \mathcal{N}(0, I_q)$.
- Generate $x_2^{i+1} \sim p(x_2 | x_1^{i+1}, x_3^i, \ldots, x_m^i) = \mathcal{N}(0, I_q)$.
- \ldots
- Generate $x_m^{i+1} \sim p(x_m | x_1^{i+1}, x_2^{i+1}, \ldots, x_{m-1}^{i+1}) = \mathcal{N}(0, I_q)$.

Gibbs naturally exploits the component independence. It works efficiently against the dimension. How about component with sparse/local independence?
- Local covariance matrix \( \mathbf{C} \):
  \( [\mathbf{C}]_{i,j} \) decays to zero quickly when \( |i - j| \) becomes large.

- Localized covariance matrix \( \mathbf{C} \):
  \( [\mathbf{C}]_{i,j} = 0 \) when \( |i - j| > L \). \( \mathbf{C} \) has a bandwidth \( 2L \).

- We will see ”local” is a perturbation of ”localized”

- We can choose \( q = L \) in \( \mathbf{x}^i = [\mathbf{x}_1^i, \mathbf{x}_2^i, \ldots, \mathbf{x}_m^i] \),
  Then \( \mathbf{C} \) is block tridiagonal.
Theorem (Morzfeld, T., Marzouk)

Apply Gibbs sampler with block-size \( q \) to \( p = \mathcal{N}(\mathbf{m}, \mathbf{C}) \). Suppose \( \mathbf{C} \) is \( q \)-block-tridiagonal. Then the distribution of \( \mathbf{x}^k \) converges to \( p \) geometrically fast in all coordinates, and we can couple \( \mathbf{x}^k \) and a sample \( \mathbf{z} \sim \mathcal{N}(\mathbf{m}, \mathbf{C}) \) such that

\[
\mathbb{E} \| \mathbf{C}^{-1/2}(\mathbf{x}^k - \mathbf{z}) \|^2 \leq \beta^k n(1 + \| \mathbf{C}^{-1/2}(\mathbf{x}^0 - \mathbf{m}) \|^2),
\]

where

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

with \( C \) being the condition number of \( \mathbf{C} \).

Localized covariance + mild condition \( \Rightarrow \) dimension free convergence.
Localized precision

Theorem (Morzfeld, T., Marzouk)

Apply Gibbs sampler with block-size $q$ to $p = \mathcal{N}(\mathbf{m}, \mathbf{C})$. Suppose $\Sigma = \mathbf{C}^{-1}$ is $q$-block-tridiagonal. Then the distribution of $\mathbf{x}^k$ converges to $p$ geometrically fast in all coordinates, and we can couple $\mathbf{x}^k$ and a sample $\mathbf{z} \sim \mathcal{N}(\mathbf{m}, \mathbf{C})$ such that

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

where

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

with $C$ being the condition number of $\mathbf{C}$.

Localized precision+mild condition $\Rightarrow$ dimension free covergence.
Why both localized covariance and precision? 
- A lemma in Bickle & Lindner 2012. 
- Localized covariance+mild condition $\Rightarrow$ local precision. 
- Localized precision+mild condition $\Rightarrow$ local covariance. 
- We will see ”local” is a perturbation of ”localized” 

For computation of Gibbs sampler, localized precision is superior: 

$$x_{j}^{k+1} \sim \mathcal{N} \left( m_j - \sum_{i<j} \Omega_{j,j}^{-1} \Omega_{j,i} (x_{i}^{k+1} - m_i) - \sum_{i>j} \Omega_{j,j}^{-1} \Omega_{j,i} (x_{i}^{k} - m_i), \Omega_{j,j}^{-1} \right).$$ 

When $\Omega$ is sparse, meaning fast computation.
Covariance v.s. Precision

Why both localized covariance and precision?

- A lemma in Bickle & Lindner 2012.
- Localized covariance+mild condition ⇒ local precision.
- Localized precision+mild condition ⇒ local covariance.
- We will see ”local” is a perturbation of ”localized”

For computation of Gibbs sampler, localized precision is superior:

\[
x_{j}^{k+1} \sim \mathcal{N} \left( m_{j} - \sum_{i<j} \Omega_{j,j}^{-1} \Omega_{j,i} (x_{i}^{k+1} - m_{i}) - \sum_{i>j} \Omega_{j,j}^{-1} \Omega_{j,i} (x_{i}^{k} - m_{i}), \Omega_{j,j}^{-1} \right).
\]

When \( \Omega \) is sparse, meaning fast computation.
Gibbs works for Gaussian sampling, with localized covariance or precision.

How about Bayesian inverse problem?

\[ y = h(x) + v, \quad v \sim \mathcal{N}(0, R), \quad x \sim p_0 = \mathcal{N}(m, C). \]

If \( h \) is linear, \( p \) is also Gaussian, Gibbs is directly applicable.

What to do when \( C \) e.t.c. are not localized but local?
Add in Metropolis steps to incorporate information

- Generate \( x'_1 \sim p_0(x_1 | x^i_2, x^i_3, \ldots, x^i_m) \)
- Accept as \( x^{i+1}_1 \) with \( \alpha_1(x^i_1, x'_1, x^i_2:m) \)

\[
\alpha_1(x^i_1, x'_1, x^i_2:m) = \min \left\{ 1, \frac{\exp\left(-\frac{1}{2} \| y - h(x') \|^2_R \right)}{\exp\left(-\frac{1}{2} \| y - h(x^i) \|^2_R \right)} \right\},
\]

where \( x' = (x'_1, x^i_2:m) \).

- Repeat for all 2, ..., \( m \) blocks

When \( h \) has a dimension free Lipschitz constant, \( \| y - h(x') \|^2_R - \| y - h(x^i) \|^2_R \) is independent of \( n \).

- Dimension independent acceptance rate.
- Should have fast convergence, though proof is unclear.
Theoretical discussion

- Does not follow the usual spatial mixing condition

\[ x_{j}^{k+1} \sim \mathcal{N} \left( m_j - \sum_{i<j} \Omega_{j,j}^{-1} \Omega_{j,i} (x_{i}^{k+1} - m_i) - \sum_{i>j} \Omega_{j,j}^{-1} \Omega_{j,i} (x_{i}^{k} - m_i), \Omega_{j,j}^{-1} \right) \]

- \( p(x_j|x_{j-}) \) has no bound from below.
- Total variation norm is not very useful in this high dimensional analysis.
- Wasserstein norm is more appropriate
- For this linear case, becomes a numerical algebra problem, related to Gauss-Seidel’s method
- Difficult to generalize.
- No existing literature for high dimensional Gibbs sampler? Especially when M-H step is involved?
Localization

Often $\Omega$ and $h$ are local

- $[\Omega]_{i,j}$ decays to zero quickly when $|i - j|$ increases.
- $[h(x)]_j$ depends significantly only over a few $x_i$.

Fast sparse computation is possible with localized parameters

- $[\Omega]_{i,j}$ decays to zero quickly when $|i - j|$ increases.
- $[h(x)]_j$ depends significantly only over a few $x_i$.

Localization: truncate the near zero terms, $\Omega \rightarrow \Omega^L$, $h \rightarrow h^L$.

We call MwG with localization as l-MwG.

Theorem (Morzfeld, T., Marzouk 2018)

The perturbation to the inverse problem is of order

$$\max \left\{ \| \Omega - \Omega^L \|_1, \sqrt{\| (H - H^L)(H - H^L)^T \|_1} \right\}.$$ 

$$\| A \|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^{n} |A_{i,j}|.$$
Comparison with function space MCMC

Function space MCMC:
- Discretization refines, domain const.
- Number of obs. const.
- Effective dimension const.
- Low-rank priors.
- Low-rank prior to posterior update.

Solved by dimension reduction.

MCMC for local problems:
- Domain size increases, discretization is const.
- Number of obs. increases.
- Effective dimension increases.
- High-rank, sparse priors.
- High-rank prior to posterior update.

Solved by localization.
Example I: image deblurring

- Truth $\mathbf{x} \sim \mathcal{N}(0, \delta^{-1} L^{-2})$, $L$ is Laplacian.
- Defocus obs: $\mathbf{y} = \mathbf{A}\mathbf{x} + \eta$, $\eta \sim \mathcal{N}(0, \lambda^{-1} \mathbf{I})$.
- Dimension is large $O(10^4)$. 
Example I: image deblurring

\[ \Omega = \lambda A^T A + \delta L x \]

Precision is sparse. Effective dimension is large.

<table>
<thead>
<tr>
<th>Image size</th>
<th>32 x 32</th>
<th>64 x 64</th>
<th>128 x 128</th>
<th>256 x 256</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimension</td>
<td>1,024</td>
<td>4,096</td>
<td>16,348</td>
<td>16,536</td>
</tr>
<tr>
<td>Eff. Dimension</td>
<td>4.8 \times 10^8</td>
<td>7.4 \times 10^9</td>
<td>1.2 \times 10^{11}</td>
<td>-</td>
</tr>
<tr>
<td>IACT (Gibbs)</td>
<td>2.92</td>
<td>2.97</td>
<td>1.74</td>
<td>1.11</td>
</tr>
<tr>
<td>Blocksize (Gibbs)</td>
<td>16</td>
<td>16</td>
<td>32</td>
<td>64</td>
</tr>
</tbody>
</table>
Example II: Lorenz 96 inverse

- Truth $\mathbf{x}_0 \sim p_0$, $p_0$ is Gaussian Climatology.
- $\Psi_t : \mathbf{x}_0 \mapsto \mathbf{x}_t : d\mathbf{x}_i = (\mathbf{x}_{i+1} - \mathbf{x}_{i-2})\mathbf{x}_{i-1} - \mathbf{x}_i + 8$
- Observe every other $\mathbf{x}_t$, $y = H(\Psi_t(\mathbf{x}_0)) + \xi$.

<table>
<thead>
<tr>
<th></th>
<th>MALA</th>
<th>pCN</th>
<th>l-MwG-B2</th>
<th>l-MwG-B4</th>
<th>l-MwG-B8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 40$</td>
<td>686</td>
<td>1051</td>
<td>55</td>
<td>60</td>
<td>266</td>
</tr>
<tr>
<td>$n = 400$</td>
<td>3,153</td>
<td>3,257</td>
<td>43</td>
<td>81</td>
<td>257</td>
</tr>
</tbody>
</table>

500 prior samples  500 posterior samples  True state  Observations

$t = 0$

$t = 0.2$
Most MCMC suffers from high dimensionality due to degenerate acceptance.

Localization technique in EnKF significantly reduces sampling complexity.

Gibbs sampler has dimension free convergence sampling local Gaussian dist.

Local proposals help MCMC has dimension free acceptance.

Different setting comparing with functional space MCMC.

Successful applications with image deblurring and Lorenz inverse problem.

Efficient MCMC is possible for high dimensional problem with structures.
Future direction: l-MALA

Applied to other MCMC algorithms, e.g. MALA.

Do local proposals

$$\tilde{x}_j^k = x_j^k + \tau v_j(x^k) + \sqrt{2\tau} \xi_j^k, \quad v_j(x) = \nabla_{x_j} \log \pi(x).$$

Accept local proposals

$$\alpha_j(x, \tilde{x}) = \min \left\{ 1, \frac{\pi(\tilde{x}) \exp \left( -\frac{1}{4\tau} \| x_j - \tilde{x}_j - \tau v_j(\tilde{x}) \|^2 \right)}{\pi(x) \exp \left( -\frac{1}{4\tau} \| \tilde{x}_j - x_j - \tau v_j(x) \|^2 \right)} \right\}. $$

- Dimension free step size+acceptance rate.
- Dimension free convergence, assuming log-concavity
- Compared with ULA: bias free!
- Better than l-MwG: use posterior directly
- Working well on “local” problems.
Reference

- Localization for MCMC: sampling high-dimensional posterior distributions with local structure. arXiv:1710.07747

Links and slides can be found at www.math.nus.edu.sg/~mattxin.

Thank you!