Efficient MCMC Sampling for Hierarchical Bayesian Inverse Problems

Andrew Brown\textsuperscript{1,2}, Arvind Saibaba\textsuperscript{3}, Sarah Vallélian\textsuperscript{2,3}

CCNS Transition Workshop
SAMSI
May 5, 2016

Supported by SAMSI Visiting Research Fellowship

\textsuperscript{1} Department of Mathematical Sciences, Clemson University, Clemson, SC 29634, USA
\textsuperscript{2} Statistical and Applied Mathematical Sciences Institute, Research Triangle Park, NC 27709, USA
\textsuperscript{3} Department of Mathematics, North Carolina State University, Raleigh, NC 27695, USA
Interdisciplinary work...

- Things I’m somewhat competent about
- Things I’m going to talk about

Questions I can answer
We are concerned with cases in which this problem isn’t ‘well behaved’.
In general,

\[
\hat{x} = \arg \min_x \|Ax - d\|_2^2 + \lambda \|L(x - x_0)\|_2^2
\]

\[
= (A^T A + \lambda^2 Q)^{-1}(A^T d + \lambda^2 Qx_0),
\]

where \(Q = L^T L\) and \(x_0\) is a “default” solution.

Note:

\[
\hat{x} = \arg \max_x k_1 \exp \left( -\frac{1}{2} \|Ax - d\|_2^2 \right) \left( -\frac{\lambda}{2} \|L(x - x_0)\|_2^2 \right) \equiv f(d|x)
\]

\[
\equiv \pi(x|\lambda)
\]

Inverse problem admits **Bayesian interpretation**

The Bayesian Machinery

- Given prior information and a data generating model, goal = update information about the parameters of interest, given the observed data via Bayes’ rule:
  \[
  \text{Posterior} \propto \text{Likelihood} \times \text{Prior}
  \]

- MAP estimator = posterior mode
  \[
  \arg \max_x \pi(x \mid d, \lambda) = \arg \max_x f(d \mid x) \pi(x \mid \lambda)
  \]

- The posterior distribution facilitates more “complete” inferences
  - Other point estimators (posterior mean, posterior median, etc.)

- In particular, it allows quantification of uncertainty about the estimators

Suppose $d \mid x, \mu \sim N(Ax, \mu^{-1}I)$, $x \mid \sigma \sim N(0, \sigma^{-1}\Gamma)$.

With $\mu$ and $\sigma$ fixed, the posterior $x \mid d, \sigma, \mu \sim N(m^*, \Sigma^*)$, where

$$
\Sigma^* = (\mu A^T A + \sigma \Gamma^{-1})^{-1}
$$
$$
m^* = \Sigma^* \mu A^T d
$$

MAP = posterior mode = posterior mean

$$
\hat{x} = (\mu A^T A + \sigma \Gamma^{-1})^{-1} \mu A^T d
$$
$$
\equiv (A^T A + \lambda L^T L)^{-1} A^T d,
$$

where $\lambda = \sigma / \mu$ and $\Gamma^{-1} = L^T L$.

Lindley and Smith (1972)
Hierarchical Model

\[ d \mid x, \mu \sim N_m(Ax, \mu^{-1}I) \]
\[ x \mid \sigma \sim N_n(0, \sigma^{-1}\Gamma) \]
\[ \mu \sim \text{Ga}(a_\mu, b_\mu) \]
\[ \sigma \sim \text{Ga}(a_\sigma, b_\sigma) \]

- Full conditional distributions for Gibbs sampling:

\[ x \mid \sigma, \mu, d \sim N_n \left( \left( \mu A^T A + \sigma \Gamma^{-1} \right)^{-1} \mu A^T d, \left( \mu A^T A + \sigma \Gamma^{-1} \right)^{-1} \right) \]

\[ \mu \mid x, \sigma, d \sim \text{Ga} \left( \frac{m}{2} + a_\mu, \frac{1}{2} \|Ax - d\|_2^2 + b_\mu \right) \]

\[ \sigma \mid x, \mu \sim \text{Ga} \left( \frac{n}{2} + a_\sigma, \frac{1}{2} \|Lx\|_2^2 + b_\sigma \right) \]

where \( L^T L = \Gamma^{-1} \).

Approximate Sampling from Conditional Distributions

Sampling from the full conditional of \( x \) requires
\[
(\mu A^T A + \sigma \Gamma^{-1})^{-1/2} z, \quad z \sim N(0, I).
\]

When \( x \) is high-dimensional, this is **computationally expensive**

For difficult conditional distributions, common to use Metropolis-Hastings with simpler proposal distributions

Proposed alternative: Find a computationally cheap approximation, and correct for the approximation using M-H.

Note:

\[ \Sigma^* := (\mu A^T A + \sigma L^T L)^{-1} \]
\[ = L^{-1}(\mu L^{-T} A^T A L^{-1} + \sigma I)^{-1} L^{-T} \]

When \( A \) is poorly conditioned, we expect the spectrum of \( L^{-T} A^T A L^{-1} \) to decay quickly. I.e.,

\[ L^{-T} A^T A L^{-1} = V \Lambda V^T \approx V_k \Lambda_k V_k^T \]

\( L^{-T} A^T A L^{-1} \) does not need to be explicitly computed so that the \( k \) largest eigenvalues can be found relatively quickly.
Some algebra and Woodbury formula yields

\[
L^{-1}(\mu L^{-T} A^T A L^{-1} + \sigma I)^{-1} L^{-T} \approx \sigma^{-1} L^{-1} \left( I + \frac{\mu}{\sigma} V_k \Lambda_k V_k^T \right)^{-1} L^{-T}
\]

\[
= \sigma^{-1} L^{-1} (I - V_k D V_k^T) L^{-T}
\]

\[
=: \tilde{\Sigma}
\]

where

\[
D = \text{diag} \left( \frac{\mu \lambda_j}{\mu \lambda_j + \sigma} \right).
\]

Similarly, we can factor \( \tilde{\Sigma} = GG^T \) with

\[
G = \sigma^{-1/2} L^{-1} (I - V_k \tilde{D} V_k^T),
\]

where

\[
\tilde{D} = \text{diag} \left( 1 \pm \sqrt{1 - (D)_{jj}} \right).
\]
Suggests a Gaussian proposal distribution with an easy-to-compute covariance matrix and factorization

\[ x^* \mid \mu, \sigma, d \sim N \left( \tilde{\Sigma}(\mu A^T d), \tilde{\Sigma} \right) \]

**Idea**: Inside the block-Gibbs sampler, use the cheap proposal as an approximation to the target (full conditional) distribution of \( x \) in a Hastings independence sampler

- Fast to sample from this distribution
- Fast to evaluate the likelihood function associated with this distribution
Target density:

\[ h(x) = \exp \left\{ -\frac{1}{2}(x - \mu \Sigma A^T d)^T \Sigma^{-1}(x - \mu \Sigma A^T d) \right\} \]

Proposal density:

\[ q(x) = \exp \left\{ -\frac{1}{2}(x - \mu \tilde{\Sigma} A^T d)^T \tilde{\Sigma}^{-1}(x - \mu \tilde{\Sigma} A^T d) \right\} \]

Acceptance ratio:

\[
\frac{h(x^*)}{q(x^*)} \cdot \frac{h(x)}{q(x)} = \exp \left\{ -\frac{1}{2}x^*,T \left( \Sigma^{-1} - \tilde{\Sigma}^{-1} \right)x^* + \frac{1}{2}x^T \left( \Sigma^{-1} - \tilde{\Sigma}^{-1} \right)x \right\} = w(x^*)/w(x),
\]

We can show that if the remaining eigenvalues from the low-rank approximation are sufficiently small, \( w(x) \approx 1 \Rightarrow \text{very high acceptance rate} \)
Simulated EEG Data

- Model:
  \[ \mathbf{d} = \mathbf{Ax} + \text{noise} \]

- \( \mathbf{d} \in \mathbb{R}^m \) represents the electrode measurements at different locations along the scalp.

- \( \mathbf{x} \in \mathbb{R}^n \) represents the current sources on a discretized grid in the brain.

- \( \mathbf{A} \in \mathbb{R}^{m \times n}, m < n \), is the **leadfield matrix** determined by conductivity and geometry of the head.

- Simulate randomly-oriented dipoles located on a intracerebral source grid.

- Software:
  - Fieldtrip MATLAB Toolbox for EEG: [http://www.fieldtriptoolbox.org](http://www.fieldtriptoolbox.org)

Here, $\text{dim}(d) = 257$ and $\text{dim}(x) = 1261$

**Figure**: First 300 eigenvalues of $A^TA$
Figure: Solutions to the EEG inverse problem using block Gibbs (left panel), Hastings-within-Gibbs (middle panel), and MAP (right panel)
## Efficiency

<table>
<thead>
<tr>
<th>Parameter</th>
<th>PSRF (Gibbs)</th>
<th>PSRF (HwG)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>1.104</td>
<td>1.041</td>
</tr>
<tr>
<td>$x$</td>
<td>1.571</td>
<td>1.577</td>
</tr>
</tbody>
</table>

Table: Potential scale reduction factors from Gibbs sampling and Hastings within Gibbs sampling

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Wall Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block Gibbs</td>
<td>1979.764 s</td>
</tr>
<tr>
<td>Hastings-within-Gibbs</td>
<td><strong>219.17 s</strong></td>
</tr>
</tbody>
</table>

- Acceptance rate for Hastings sampler = 100% for all three chains
Simulated Computed Tomography Data

- X-ray is passed through a body from a source \( s = 0 \) to a sensor \( s = S \) along a line determined by angle and distance with respect to a fixed origin.

- Use Shepp-Logan phantom as the true image.

- Target image is discretized so that \( \dim(x) = 128 \times 128 = 16384 \).

- Simulate observed data (Radon transform model) over discretized lines and angles so that \( \dim(z) = 5000 \).

- Data generating model:

\[
z = Ax + e,
\]

where \( e \sim N(0, \mu^{-1}I) \), \( \mu^{-1/2} = 0.01 \|Ax\|_\infty \).

- MATLAB code: [http://www.math.umt.edu/bardsley/codes.html](http://www.math.umt.edu/bardsley/codes.html)

Kaipio and Somersalo (2005), Bardsley (2011)
Figure: First 8000 eigenvalues of $A^T A$
Figure: Posterior mean (left panel) and MAP estimate (right panel)
<table>
<thead>
<tr>
<th>Estimate</th>
<th>Relative Error</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>0.4411</td>
<td>0.1081</td>
</tr>
<tr>
<td>Posterior Mean</td>
<td>0.4440</td>
<td>0.1081</td>
</tr>
</tbody>
</table>

- **Wall time for HwG sampler = 3270.35 s (≪ 1 hr.)**

- Posterior provides access to almost any estimator we want and quantifies the associated uncertainty
Features of the Approach

- Spectral decomposition does **not** require explicitly computing $L^{-T}A^TAL^{-1}$, but only matrix-vector products.
  - Forming $A$ itself is often challenging.

- Finding the eigenvalues is a precomputation **before** iterating. Once found, the proposal is cheap for any given $\mu$ and $\sigma$.

- For ill-posed problems, the acceptance probability is close to one.
  - Can accept every proposed draw as an approximation to avoid evaluating the likelihood.

- We are still modeling the full dimension of $x$, **not** projecting onto a lower-dimensional subspace.
  - Exploiting the nature of the forward model

- This approach allows incorporation of strong or vague prior information about the solution through specification of the prior covariance (precision) matrix
  - Prior information determined from fMRI can help to solve the EEG problem
  - Prior smoothness assumptions through a GP prior or Laplacian

Dale et al. (2000), Banerjee et al. (2008), Higdon et al. (2008), Banerjee et al. (2012)
Thoughts About Future Directions

- Application to real data
- Approximations based on Krylov spaces
  - Covariance factorization is not necessary in this case
- Allow estimation of hyperparameters in the prior covariance
  - Prior distributions on matrices with special structure
  - Parameters estimated via, e.g., empirical Bayes and kept fixed
- Exploration of other penalties in the prior
  - Many types of regression penalties ("shrinkage priors") can be expressed as scale mixtures of Normal distributions
- Incorporation into MCMC algorithms with very computationally intense forward models
  - E.g., delayed acceptance algorithms.

Christen and Fox (2005), Park and Casella (2008), Qian and Wu (2008), Polson and Scott (2010), Parker and Fox (2012), Fox et al. (2013)
Thank you!