# Efficient MCMC Sampling for Hierarchical Bayesian Inverse Problems

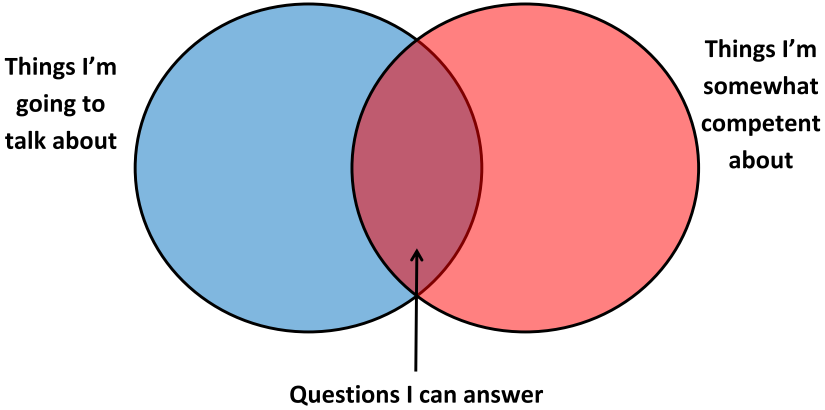
Andrew Brown[1,2], Arvind Saibaba[3], Sarah Vallélian[2,3]

CCNS Transition Workshop
SAMSI
May 5, 2016

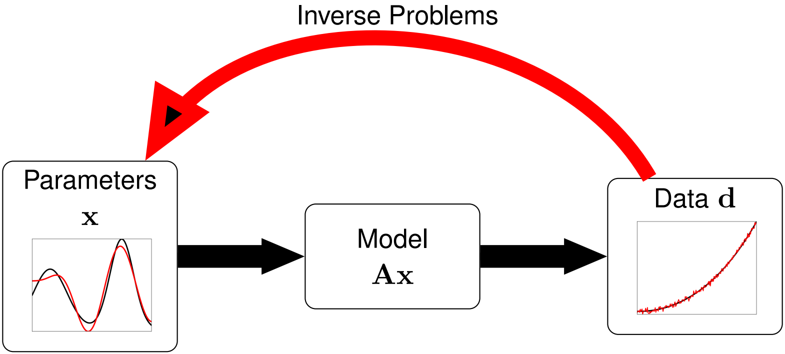
Supported by SAMSI Visiting Research Fellowship

[1] Department of Mathematical Sciences, Clemson University, Clemson, SC 29634, USA
[2] Statistical and Applied Mathematical Sciences Institute, Research Triangle Park, NC 27709, USA
[3] Department of Mathematics, North Carolina State University, Raleigh, NC 27695, USA

# Interdisciplinary work...



Courtesy of A. Saibaba




We are concerned with cases in which this problem isn't 'well behaved'.

# Regularized Inversion

- In general,

$$\hat{\mathbf{x}} = \arg\min_{x} \|\mathbf{A}\mathbf{x} - \mathbf{d}\|_2^2 + \lambda\|\mathbf{L}(\mathbf{x} - \mathbf{x}_0)\|_2^2$$

$$= (\mathbf{A}^T\mathbf{A} + \lambda^2\mathbf{Q})^{-1}(\mathbf{A}^T\mathbf{d} + \lambda^2\mathbf{Q}\mathbf{x}_0),$$

where $\mathbf{Q} = \mathbf{L}^T\mathbf{L}$ and $\mathbf{x}_0$ is a “default” solution.

- Note:

$$\hat{\mathbf{x}} = \arg\max_{x} \underbrace{k_1 \exp\left(-\frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{d}\|_2^2\right)}_{\equiv f(\mathbf{d}|\mathbf{x})} \underbrace{k_2 \exp\left(-\frac{\lambda}{2}\|\mathbf{L}(\mathbf{x} - \mathbf{x}_0)\|_2^2\right)}_{\equiv \pi(\mathbf{x}|\lambda)}$$

- Inverse problem admits **Bayesian interpretation**

Hoerl and Kinnard (1970), Tikhonov and Arsenin (1977), Press et al. (2007), Fox et al. (2013)

# 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

$$\operatorname*{arg\,max}_x \pi(\mathbf{x} \mid \mathbf{d}, \lambda) = \operatorname*{arg\,max}_x f(\mathbf{d} \mid \mathbf{x})\pi(\mathbf{x} \mid \lambda)$$

- 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

Berger (1985), Bernardo and Smith (1994), Robert (2007), Carlin and Louis (2009), Gelman et al. (2013), . . .

- Suppose $\mathbf{d} \mid \mathbf{x}, \mu \sim N(\mathbf{A}\mathbf{x}, \mu^{-1}\mathbf{I}), \;\; \mathbf{x} \mid \sigma \sim N(\mathbf{0}, \sigma^{-1}\mathbf{\Gamma})$
- **With** $\mu$ **and** $\sigma$ **fixed**, the posterior $\mathbf{x} \mid \mathbf{d}, \sigma, \mu \sim N\left(\mathbf{m}^*, \mathbf{\Sigma}^*\right)$, where

$$\begin{aligned} \mathbf{\Sigma}^* &= \left(\mu\mathbf{A}^T\mathbf{A} + \sigma\mathbf{\Gamma}^{-1}\right)^{-1} \\ \mathbf{m}^* &= \mathbf{\Sigma}^*\mu\mathbf{A}^T\mathbf{d} \end{aligned}$$

- MAP = posterior mode = posterior mean

$$\begin{aligned} \hat{\mathbf{x}} &= \left(\mu\mathbf{A}^T\mathbf{A} + \sigma\mathbf{\Gamma}^{-1}\right)^{-1} \mu\mathbf{A}^T\mathbf{d} \\ &\equiv \left(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{L}^T\mathbf{L}\right)^{-1} \mathbf{A}^T\mathbf{d}, \end{aligned}$$

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

Lindley and Smith (1972)

# Hierarchical Model

$$
\begin{aligned}
\mathbf{d} \mid \mathbf{x}, \mu &\sim N_m(\mathbf{A}\mathbf{x}, \mu^{-1}\mathbf{I}) \\
\mathbf{x} \mid \sigma &\sim N_n(\mathbf{0}, \sigma^{-1}\mathbf{\Gamma}) \\
\mu &\sim \text{Ga}(a_\mu, b_\mu) \\
\sigma &\sim \text{Ga}(a_\sigma, b_\sigma)
\end{aligned}
$$

- Full conditional distributions for Gibbs sampling:

$$
\begin{aligned}
\mathbf{x} \mid \sigma, \mu, \mathbf{d} &\sim N_n\left( \left(\mu\mathbf{A}^T\mathbf{A} + \sigma\mathbf{\Gamma}^{-1}\right)^{-1} \mu\mathbf{A}^T\mathbf{d}, \left(\mu\mathbf{A}^T\mathbf{A} + \sigma\mathbf{\Gamma}^{-1}\right)^{-1} \right) \\
\mu \mid \mathbf{x}, \sigma, \mathbf{d} &\sim \text{Ga}\left( \frac{m}{2} + a_\mu, \ \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{d}\|_2^2 + b_\mu \right) \\
\sigma \mid \mathbf{x}, \mu &\sim \text{Ga}\left( \frac{n}{2} + a_\sigma, \ \frac{1}{2}\|\mathbf{L}\mathbf{x}\|_2^2 + b_\sigma \right)
\end{aligned}
$$

where $\mathbf{L}^T\mathbf{L} = \mathbf{\Gamma}^{-1}$.

Geman and Geman (1984), Gelfand and Smith (1990), Gelman and Rubin (1992), Brooks and Gelman (1998), Carlin and Louis (2009)

# Approximate Sampling from Conditional Distributions

- Sampling from the full conditional of $\mathbf{x}$ requires $\left(\mu \mathbf{A}^T \mathbf{A} + \sigma \boldsymbol{\Gamma}^{-1}\right)^{-1/2} \mathbf{z}, \;\; \mathbf{z} \sim N(\mathbf{0}, \mathbf{I})$.
- When $\mathbf{x}$ 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.

Metropolis et al. (1953), Hastings (1970), Tierney (1994), Rosenthal (2011), Gelman et al. (2013)

- Note:

$$\begin{aligned}\mathbf{\Sigma}^* &:= (\mu\mathbf{A}^T\mathbf{A} + \sigma\mathbf{L}^T\mathbf{L})^{-1} \\ &= \mathbf{L}^{-1}(\mu\mathbf{L}^{-T}\mathbf{A}^T\mathbf{A}\mathbf{L}^{-1} + \sigma\mathbf{I})^{-1}\mathbf{L}^{-T}\end{aligned}$$

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

$$\mathbf{L}^{-T}\mathbf{A}^T\mathbf{A}\mathbf{L}^{-1} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T \approx \mathbf{V}_k\mathbf{\Lambda}_k\mathbf{V}_k^T$$

- $\mathbf{L}^{-T}\mathbf{A}^T\mathbf{A}\mathbf{L}^{-1}$ does not need to be explicitly computed so that the $k$ largest eigenvalues can be found relatively quickly

- Some algebra and Woodbury formula yields

$$\begin{aligned}\mathbf{L}^{-1}(\mu\mathbf{L}^{-T}\mathbf{A}^T\mathbf{A}\mathbf{L}^{-1}+\sigma\mathbf{I})^{-1}\mathbf{L}^{-T} &\approx \sigma^{-1}\mathbf{L}^{-1}\left(\mathbf{I}+\frac{\mu}{\sigma}\mathbf{V}_k\boldsymbol{\Lambda}_k\mathbf{V}_k^T\right)^{-1}\mathbf{L}^{-T}\\ &= \sigma^{-1}\mathbf{L}^{-1}(\mathbf{I}-\mathbf{V}_k\mathbf{D}\mathbf{V}_k^T)\mathbf{L}^{-T}\\ &=: \widetilde{\boldsymbol{\Sigma}}\end{aligned}$$

where

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

- Similarly, we can factor $\widetilde{\boldsymbol{\Sigma}}=\mathbf{G}\mathbf{G}^T$ with

$$\mathbf{G}=\sigma^{-1/2}\mathbf{L}^{-1}(\mathbf{I}-\mathbf{V}_k\widetilde{\mathbf{D}}\mathbf{V}_k^T),$$

where

$$\widetilde{\mathbf{D}}=\text{diag}\left(1\pm\sqrt{1-(\mathbf{D})_{jj}}\right)$$

- Suggests a Gaussian proposal distribution with an easy-to-compute covariance matrix and factorization

$$\mathbf{x}^* \mid \mu, \sigma, \mathbf{d} \sim N\left(\widetilde{\boldsymbol{\Sigma}}(\mu \mathbf{A}^T \mathbf{d}), \widetilde{\boldsymbol{\Sigma}}\right)$$

- **Idea**: Inside the block-Gibbs sampler, use the cheap proposal as an approximation to the target (full conditional) distribution of $\mathbf{x}$ in a Hastings independence sampler
  - Fast to sample from this distribution
  - Fast to evaluate the likelihood function associated with this distribution

- Target density:

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

- Proposal density:

$$q(\mathbf{x}) = \exp\left\{-\frac{1}{2}(\mathbf{x} - \mu\widetilde{\mathbf{\Sigma}}\mathbf{A}^T\mathbf{d})^T\widetilde{\mathbf{\Sigma}}^{-1}(\mathbf{x} - \mu\widetilde{\mathbf{\Sigma}}\mathbf{A}^T\mathbf{d})\right\}$$

- Acceptance ratio:

$$\begin{aligned}\frac{h(\mathbf{x}^*)/q(\mathbf{x}^*)}{h(\mathbf{x})/q(\mathbf{x})} &= \exp\left\{-\frac{1}{2}\mathbf{x}^{*,T}\left(\mathbf{\Sigma}^{-1} - \widetilde{\mathbf{\Sigma}}^{-1}\right)\mathbf{x}^* + \frac{1}{2}\mathbf{x}^T\left(\mathbf{\Sigma}^{-1} - \widetilde{\mathbf{\Sigma}}^{-1}\right)\mathbf{x}\right\} \\ &\equiv \frac{w(\mathbf{x}^*)}{w(\mathbf{x})},\end{aligned}$$

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

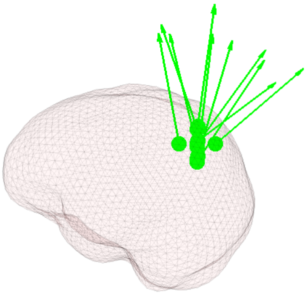
# Simulated EEG Data

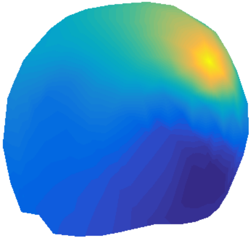
- Model:

$$\mathbf{d} = \mathbf{A}\mathbf{x} + \text{noise}$$

- $\mathbf{d} \in \mathbb{R}^m$ 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:
  - German Gomez-Herrero EEG Tutorial: `http://germangh.github.io/tutorials/dipoles/tutorial_dipoles.htm`
  - Fieldtrip MATLAB Toolbox for EEG: `http://www.fieldtriptoolbox.org`

Oostendrop and Oosterom (1989), Hauk (2004)





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

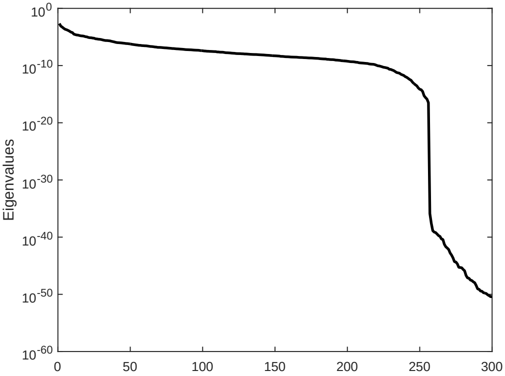



Figure : First 300 eigenvalues of $\mathbf{A}^T\mathbf{A}$

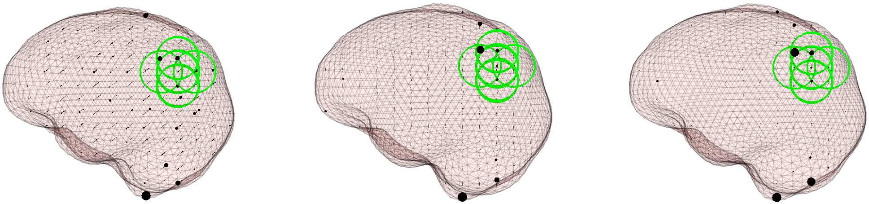


Figure : Solutions to the EEG inverse problem using block Gibbs (left panel), Hastings-within-Gibbs (middle panel), and MAP (right panel)

# Efficiency

| Parameter | PSRF (Gibbs) | PSRF (HwG) |
|---|---|---|
| $\mu$ | 1.000 | 1.000 |
| $\sigma$ | 1.104 | 1.041 |
| $\mathbf{x}$ | 1.571 | 1.577 |

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

| Algorithm | Wall Time |
|---|---|
| Block Gibbs | 1979.764 s |
| Hastings-within-Gibbs | **219.17 s** |

- Acceptance rate for Hastings sampler = 100% for all three chains

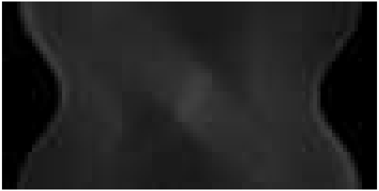
# Simulated Computed Tomography Data

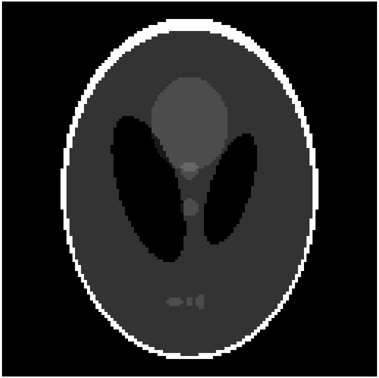
- X-ray is passed through a body from a source ($s = 0$) to a sensor ($s = S$) along a line determined by angle and distance with respect to a fixed origin
- Use Shepp-Logan phantom as the true image
- Target image is discretized so that $\dim(\mathbf{x}) = 128 \times 128 = 16384$
- Simulate observed data (Radon transform model) over discretized lines and angles so that $\dim(\mathbf{z}) = 5000$
- Data generating model:

$$\mathbf{z} = \mathbf{A}\mathbf{x} + \mathbf{e},$$

  where $\mathbf{e} \sim N(\mathbf{0}, \mu^{-1}\mathbf{I})$, $\mu^{-1/2} = 0.01\|\mathbf{A}\mathbf{x}\|_\infty$.
- MATLAB code: `http://www.math.umt.edu/bardsley/codes.html`

Kaipio and Somersalo (2005), Bardsley (2011)





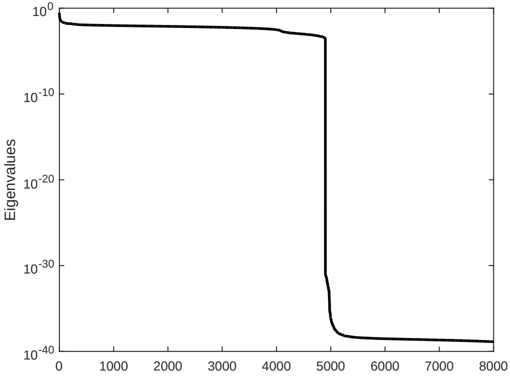



Figure : First 8000 eigenvalues of $\mathbf{A}^T\mathbf{A}$

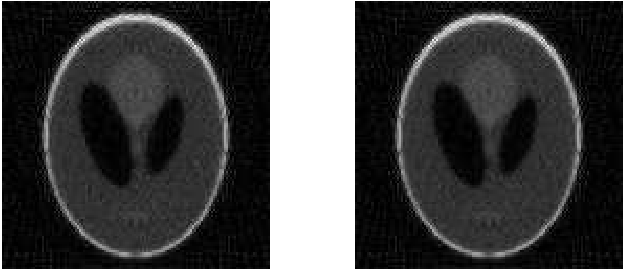


Figure : Posterior mean (left panel) and MAP estimate (right panel)

| Estimate | Relative Error | RMSE |
|---|---|---|
| MAP | 0.4411 | 0.1081 |
| Posterior Mean | 0.4440 | 0.1081 |

- **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 $\mathbf{L}^{-T}\mathbf{A}^T\mathbf{A}\mathbf{L}^{-1}$, but only matrix-vector products.
  - Forming $\mathbf{A}$ 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 $\mathbf{x}$, *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!