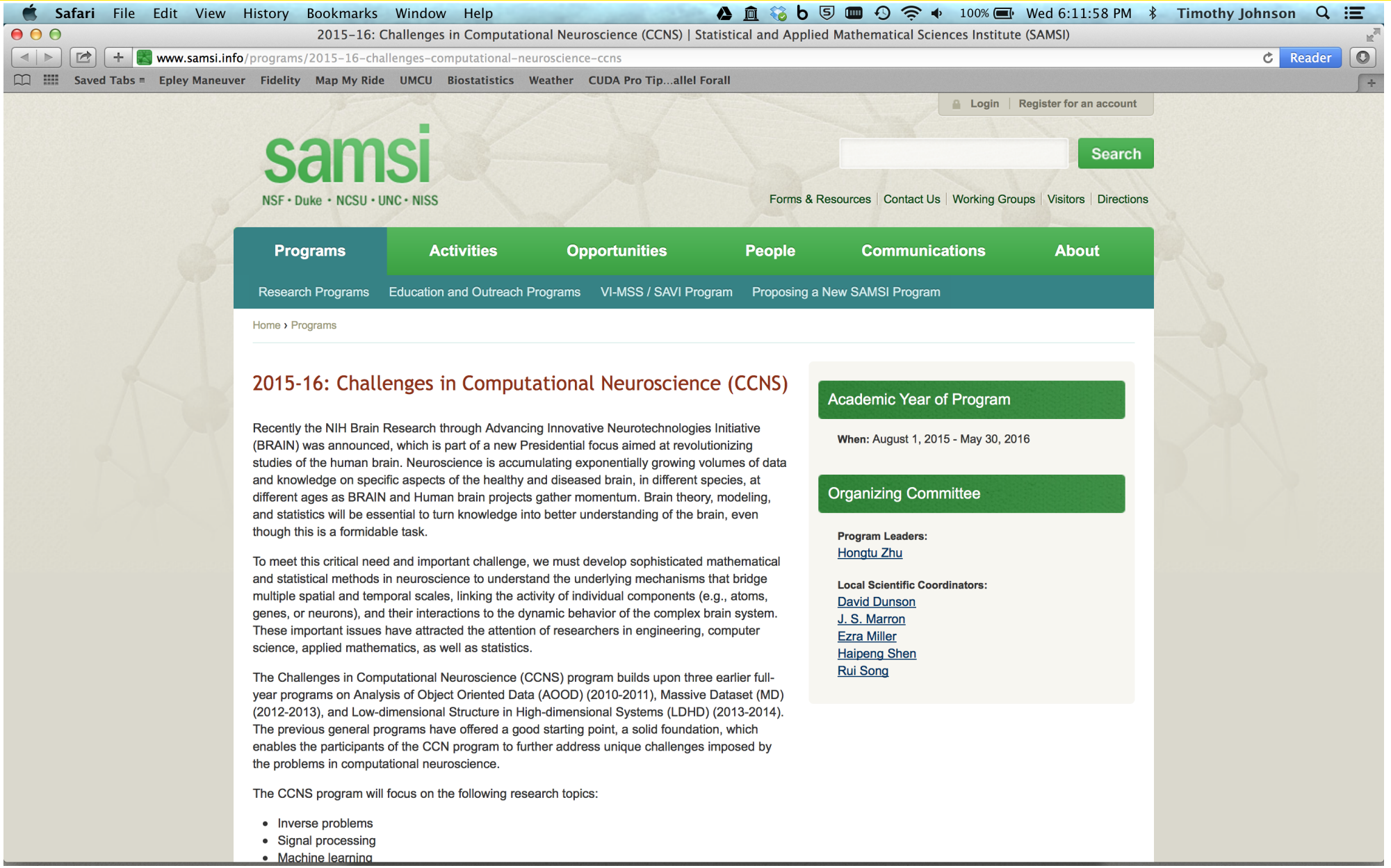
Bayesian Methods in Neuroimaging

Timothy D. Johnson

Department of Biostatistics
University of Michigan

SAMSI Program



Outline

Bayes Theorem

$$\pi(\theta \mid Y) = \frac{\pi(Y \mid \theta)\pi(\theta)}{\pi(Y)}$$

Bayes Theorem

$$\pi(\theta \mid Y) = \frac{\pi(Y \mid \theta)\pi(\theta)}{\pi(Y)}$$

- Y — data

Bayes Theorem

$$\pi(\theta \mid Y) = \frac{\pi(Y \mid \theta)\pi(\theta)}{\pi(Y)}$$

- Y — data
- θ — parameters

Bayes Theorem

$$\pi(\theta \mid Y) = \frac{\pi(Y \mid \theta)\pi(\theta)}{\pi(Y)}$$

- Y — data
- θ — parameters
- $\pi(\theta \mid y)$ — posterior density

Bayes Theorem

$$\pi(\theta \mid Y) = \frac{\pi(Y \mid \theta)\pi(\theta)}{\pi(Y)}$$

- Y — data
- θ — parameters
- $\pi(\theta \mid y)$ — posterior density
- $\pi(Y \mid \theta)$ — likelihood

Bayes Theorem

$$\pi(\theta \mid Y) = \frac{\pi(Y \mid \theta)\pi(\theta)}{\pi(Y)}$$

- Y — data
- θ — parameters
- $\pi(\theta \mid y)$ — posterior density
- $\pi(Y \mid \theta)$ — likelihood
- $\pi(\theta)$ — prior density

Bayes Theorem

$$\pi(\theta \mid Y) = \frac{\pi(Y \mid \theta)\pi(\theta)}{\pi(Y)}$$

- Y — data
- θ — parameters
- $\pi(\theta \mid y)$ — posterior density
- $\pi(Y \mid \theta)$ — likelihood
- $\pi(\theta)$ — prior density
- $\pi(Y)$ — marginal density

Bayes Theorem

$$\pi(\theta \mid Y) = \frac{\pi(Y \mid \theta)\pi(\theta)}{\pi(Y)}$$

- Y — data
- θ — parameters
- $\pi(\theta \mid y)$ — posterior density
- $\pi(Y \mid \theta)$ — likelihood
- $\pi(\theta)$ — prior density
- $\pi(Y)$ — marginal density

Simple, yet profound

Example—Gaussian Data

- $Y = (Y_1, \ldots, Y_n)$, where each Y_i is Gaussian dist:

$$[Y_i \mid \mu] \sim \mathsf{N}(\mu, \sigma^2)$$

Example—Gaussian Data

- $Y = (Y_1, \ldots, Y_n)$, where each Y_i is Gaussian dist:

$$[Y_i \mid \mu] \sim \mathsf{N}(\mu, \sigma^2)$$

- $\mu \sim \mathsf{N}(\nu, \phi^2)$

Example—Gaussian Data

- $Y = (Y_1, \ldots, Y_n)$, where each Y_i is Gaussian dist:

$$[Y_i \mid \mu] \sim \mathsf{N}(\mu, \sigma^2)$$

- $\mu \sim \mathsf{N}(\nu, \phi^2)$
- ν, ϕ^2 and σ^2 known constants

Example—Gaussian Data

- $Y = (Y_1, \ldots, Y_n)$, where each Y_i is Gaussian dist:

$$[Y_i \mid \mu] \sim \mathsf{N}(\mu, \sigma^2)$$

- $\mu \sim \mathsf{N}(\nu, \phi^2)$
- ν, ϕ^2 and σ^2 known constants
- Then

$$[\mu \mid Y] \sim \mathsf{N}(m, v)$$

where

$$\begin{aligned} v &= \frac{\sigma^2 \phi^2}{n\phi^2 + \sigma^2} \\ m &= \left(\frac{\sum_i^n y_i}{\sigma^2} + \frac{\nu}{\phi^2} \right) / v \end{aligned}$$

Example (cont.)

- This example is a simple ”toy example” with a simple posterior distribution

Example (cont.)

- This example is a simple ”toy example” with a simple posterior distribution
- Most models today are much more complicated

Example (cont.)

- This example is a simple ”toy example” with a simple posterior distribution
- Most models today are much more complicated
- Including Neuroimaging examples

Example (cont.)

- This example is a simple ”toy example” with a simple posterior distribution
- Most models today are much more complicated
- Including Neuroimaging examples
  - Very high dimensional Hierarchical models
  - Posteriors have no closed form

Example (cont.)

- This example is a simple ”toy example” with a simple posterior distribution
- Most models today are much more complicated
- Including Neuroimaging examples
  - Very high dimensional Hierarchical models
  - Posteriors have no closed form

Must rely on Monte Carlo simulation techniques

Monte Carlo Simulation

- Markov Chain Monte Carlo

Monte Carlo Simulation

- Markov Chain Monte Carlo
  - Metropolis algorithm

Monte Carlo Simulation

- Markov Chain Monte Carlo
  - Metropolis algorithm
  - Metropolis-Hastings algorithm

Monte Carlo Simulation

- Markov Chain Monte Carlo
  - Metropolis algorithm
  - Metropolis-Hastings algorithm
  - Gibbs sampling

Monte Carlo Simulation

- Markov Chain Monte Carlo
  - Metropolis algorithm
  - Metropolis-Hastings algorithm
  - Gibbs sampling
  - Metropolis-within-Gibbs

Monte Carlo Simulation

- Markov Chain Monte Carlo
  - Metropolis algorithm
  - Metropolis-Hastings algorithm
  - Gibbs sampling
  - Metropolis-within-Gibbs
- Particle filtering

...to name a few

Monte Carlo Simulation

- Markov Chain Monte Carlo
  - Metropolis algorithm
  - Metropolis-Hastings algorithm
  - Gibbs sampling
  - Metropolis-within-Gibbs
- Particle filtering

...to name a few

- Theory guarantees these all converge to the posterior distribution

Monte Carlo Simulation

- Markov Chain Monte Carlo
  - Metropolis algorithm
  - Metropolis-Hastings algorithm
  - Gibbs sampling
  - Metropolis-within-Gibbs
- Particle filtering

...to name a few

- Theory guarantees these all converge to the posterior distribution
- No guarantee how long it will take

Monte Carlo Simulation

- For complex problems, including those in Neuroimaging,
  - these (MC)MC simulations methods computationally intense
  - "behave poorly"—samples highly correlated (called slow mixing)

Monte Carlo Simulation

- For complex problems, including those in Neuroimaging,
  - these (MC)MC simulations methods computationally intense
  - "behave poorly"—samples highly correlated (called slow mixing)
- Must run the simulation a very long time to obtain good estimates of the posterior
  - Weeks to months

Back to this latter

Pre-surgical fMRI

Liu, Z., Berrocal, V. J., Bartsch, A. J., Johnson, T. D. (2014) Pre-Surgical fMRI data analysis using a spatially adaptive conditionally autoregressive model. Submitted to *Bayesian Analysis*.

- Standard fMRI methods have too strict control of false positives

Pre-surgical fMRI

Liu, Z., Berrocal, V. J., Bartsch, A. J., Johnson, T. D. (2014) Pre-Surgical fMRI data analysis using a spatially adaptive conditionally autoregressive model. Submitted to *Bayesian Analysis*.

- Standard fMRI methods have too strict control of false positives
- For pre-surgical fMRI, control of false negatives is vital
  - Don't want to cut out functionally eloquent regions by mistake

Pre-surgical fMRI

Liu, Z., Berrocal, V. J., Bartsch, A. J., Johnson, T. D. (2014) Pre-Surgical fMRI data analysis using a spatially adaptive conditionally autoregressive model. Submitted to *Bayesian Analysis*.

- Standard fMRI methods have too strict control of false positives
- For pre-surgical fMRI, control of false negatives is vital
  - Don't want to cut out functionally eloquent regions by mistake
- As is control of smoothing between boundaries of high and low signal intensity
  - Want to smooth where signal changes slowly
  - Don't want to smooth where signal is rapidly changing

This motivates our approach

Pre-surgical fMRI—Our Approach

- At voxel i model the signal indep. with mean μ_i and var. σ_i^2
- Place a spatially adaptive CAR model on the μ_i
  - Spatially correlates the means
  - Spatially adapts smoothness to the image

$$\begin{aligned}
\left[Y_i \mid \mu_i, \sigma_i^2\right] &\sim \mathrm{N}(\mu_i, \sigma_i^2) \\
\left[\mu_i \mid \mu_{-i}, \sigma_i^2\right] &\sim \mathrm{N}\left(\sum_{j\sim i} \mu_j / N_i, c_i \sigma_i^2\right) \\
\left[\ln(\sigma_i^2) \mid \ln(\sigma_{-i}^2), \phi^2\right] &\sim \mathrm{N}\left(\sum_{j\sim i} \ln(\sigma_j^2)/N_i, \phi^2/N_i\right) \\
c_i = p_i/(1-p_i), &\qquad p_i \sim \mathrm{Beta}(\alpha, \beta)
\end{aligned}$$

p_i controls the amount of smoothing in the full conditional of μ_i

Pre-surgical fMRI—Our Approach

- Bayesian decision theory: loss function penalizes false positives and false negatives asymmetrically.

Pre-surgical fMRI—Our Approach

- Bayesian decision theory: loss function penalizes false positives and false negatives asymmetrically.
- With the guidance of a subject area expert
  - We penalize false negatives 11 times more heavily than false positives

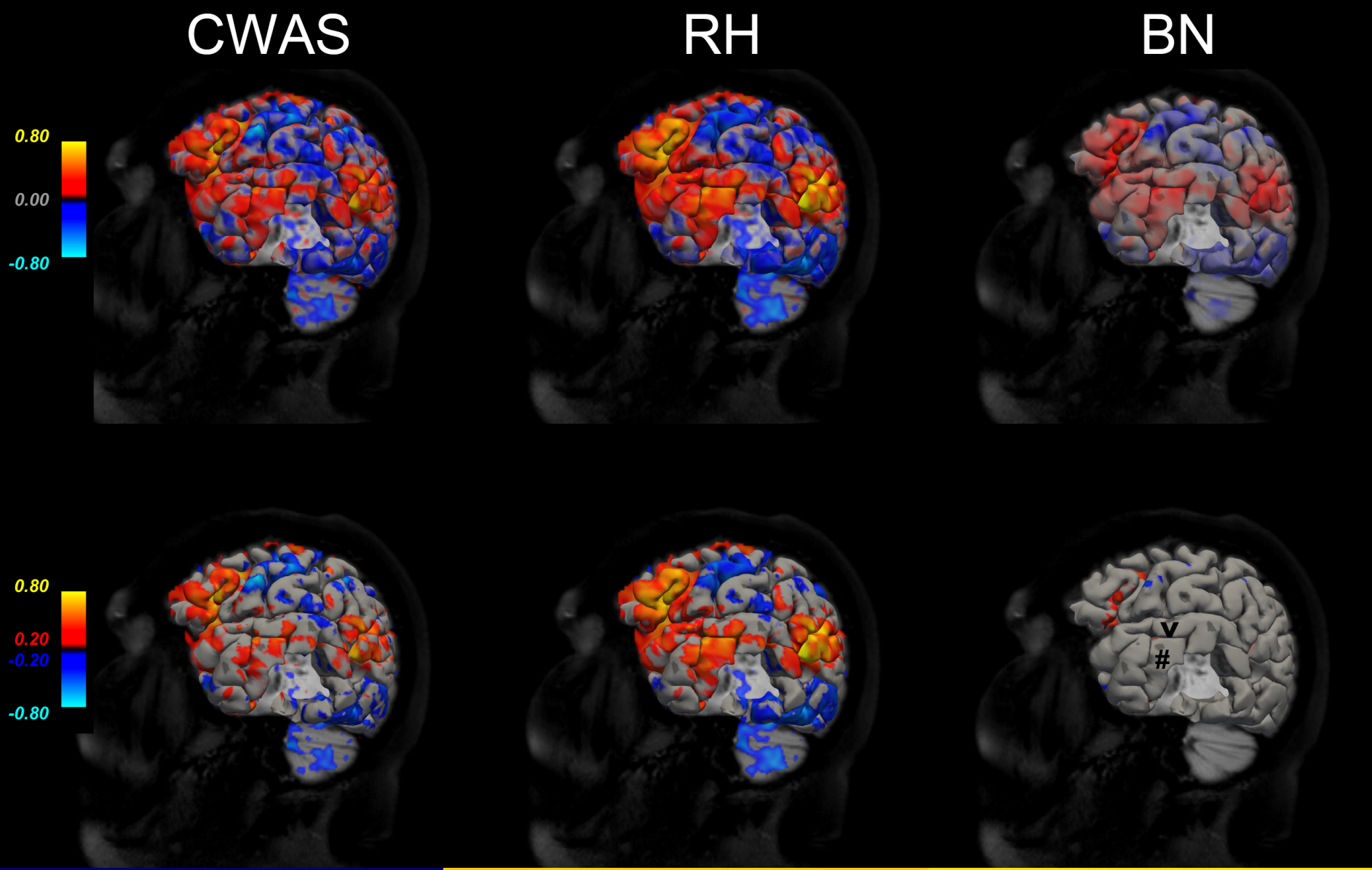
Pre-surgical fMRI—Our Approach

- Bayesian decision theory: loss function penalizes false positives and false negatives asymmetrically.
- With the guidance of a subject area expert
  - We penalize false negatives 11 times more heavily than false positives

We compare results with those from
two other spatially adaptive CAR models

Speed: Fast, about 1 hour

Pre-surgical fMRI—Results



Group Level fMRI Analysis

Xu, L., Johnson, T. D., Nichols, T. E., Nee, D. (2009) Modeling inter-subject variability in fMRI activation location: a Bayesian hierarchical spatial model. *Biometrics* **65** 1041–1051.

Study of Proactive Interference Resolution

- Proactive interference occurs when current information is lost because it is mixed up with previously learned, similar, information

Group Level fMRI Analysis

Xu, L., Johnson, T. D., Nichols, T. E., Nee, D. (2009) Modeling inter-subject variability in fMRI activation location: a Bayesian hierarchical spatial model. *Biometrics* **65** 1041–1051.

Study of Proactive Interference Resolution

- Proactive interference occurs when current information is lost because it is mixed up with previously learned, similar, information
  - One's ability to resolve proactive interference is key to in determining how much information one can store in short term memory

Group Level fMRI Analysis

- Recent probes task
  - Subject given a small set of items to remember (a list of 6 letters, say) in a short time period (*target probe*)

Group Level fMRI Analysis

- Recent probes task
  - Subject given a small set of items to remember (a list of 6 letters, say) in a short time period (*target probe*)
  - The subject then given a short list of items (*recognition probe*) that may (*positive probe*) or may not (*negative probe*) be a subset of the target probe

Group Level fMRI Analysis

- Recent probes task
  - Subject given a small set of items to remember (a list of 6 letters, say) in a short time period (*target probe*)
  - The subject then given a short list of items (*recognition probe*) that may (*positive probe*) or may not (*negative probe*) be a subset of the target probe
  - The recognition probe could further be a member of a previous trial (*recent probe*) or not (*non-recent probe*)

Group Level fMRI Analysis

- Recent probes task
  - Subject given a small set of items to remember (a list of 6 letters, say) in a short time period (*target probe*)
  - The subject then given a short list of items (*recognition probe*) that may (*positive probe*) or may not (*negative probe*) be a subset of the target probe
  - The recognition probe could further be a member of a previous trial (*recent probe*) or not (*non-recent probe*)
  - Subjects show slower reaction time and increased error rates when rejecting *recent negative probes* compared to *non-recent negative probes*

Group Level fMRI Analysis

- Recent probes task
  - Subject given a small set of items to remember (a list of 6 letters, say) in a short time period (*target probe*)
  - The subject then given a short list of items (*recognition probe*) that may (*positive probe*) or may not (*negative probe*) be a subset of the target probe
  - The recognition probe could further be a member of a previous trial (*recent probe*) or not (*non-recent probe*)
  - Subjects show slower reaction time and increased error rates when rejecting *recent negative probes* compared to *non-recent negative probes*
  - Performance decrease a marker of proactive interference

Group Level fMRI Analysis

- Recent probes task
  - Subject given a small set of items to remember (a list of 6 letters, say) in a short time period (*target probe*)
  - The subject then given a short list of items (*recognition probe*) that may (*positive probe*) or may not (*negative probe*) be a subset of the target probe
  - The recognition probe could further be a member of a previous trial (*recent probe*) or not (*non-recent probe*)
  - Subjects show slower reaction time and increased error rates when rejecting *recent negative probes* compared to *non-recent negative probes*
  - Performance decrease a marker of proactive interference
- The left lateral prefrontal cortex is a region linked to proactive interference resolution

Group Level fMRI Analysis—Overview of Our Approach

A Bayesian Spatial Hierarchical Model

Group Level fMRI Analysis—Overview of Our Approach

A Bayesian Spatial Hierarchical Model

- Level 1: subject level data
  - Unsmoothed Z-stat image modeled as a mixture distribution
  - Spatial correlation accounted for in the mixing weights

Group Level fMRI Analysis—Overview of Our Approach

A Bayesian Spatial Hierarchical Model

- Level 1: subject level data
  - Unsmoothed Z-stat image modeled as a mixture distribution
  - Spatial correlation accounted for in the mixing weights
- Level 2: subject level data
  - Cluster mixing wt means about “activation centers”

Group Level fMRI Analysis—Overview of Our Approach

A Bayesian Spatial Hierarchical Model

- Level 1: subject level data
  - Unsmoothed Z-stat image modeled as a mixture distribution
  - Spatial correlation accounted for in the mixing weights
- Level 2: subject level data
  - Cluster mixing wt means about “activation centers”
- Level 3: population level data
  - Activation centers cluster around pop level centers

Group Level fMRI Analysis—Overview of Our Approach

A Bayesian Spatial Hierarchical Model

- Level 1: subject level data
  - Unsmoothed Z-stat image modeled as a mixture distribution
  - Spatial correlation accounted for in the mixing weights
- Level 2: subject level data
  - Cluster mixing wt means about “activation centers”
- Level 3: population level data
  - Activation centers cluster around pop level centers
- Level 4: Dirichlet process prior
  - Population parameters modeled as a Dirichlet process

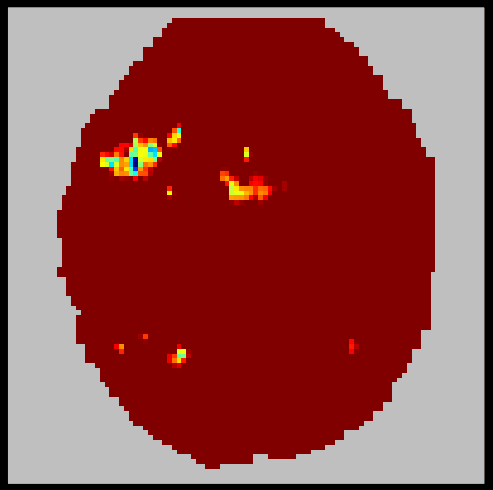
Group Level fMRI Analysis—Overview of Our Approach

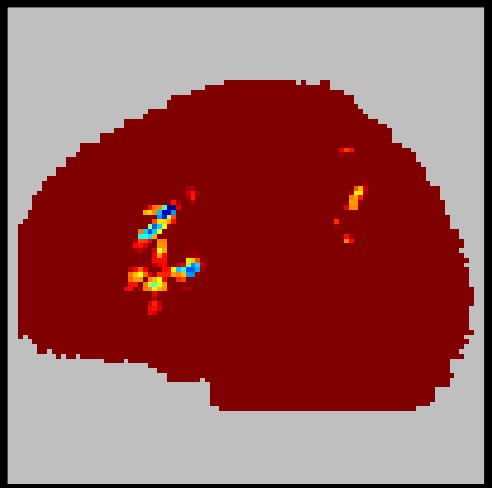
A Bayesian Spatial Hierarchical Model

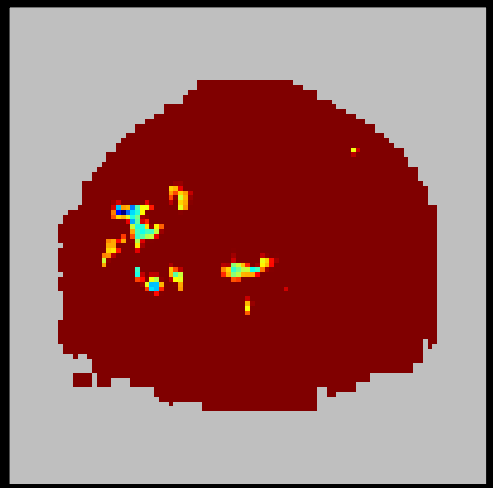
- Level 1: subject level data
  - Unsmoothed Z-stat image modeled as a mixture distribution
  - Spatial correlation accounted for in the mixing weights
- Level 2: subject level data
  - Cluster mixing wt means about “activation centers”
- Level 3: population level data
  - Activation centers cluster around pop level centers
- Level 4: Dirichlet process prior
  - Population parameters modeled as a Dirichlet process

SLOW—days to converge

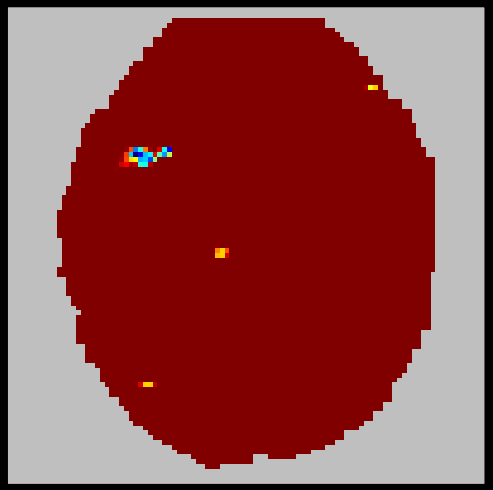
Group Level fMRI Analysis—Patient Level Results (Sbj 4)

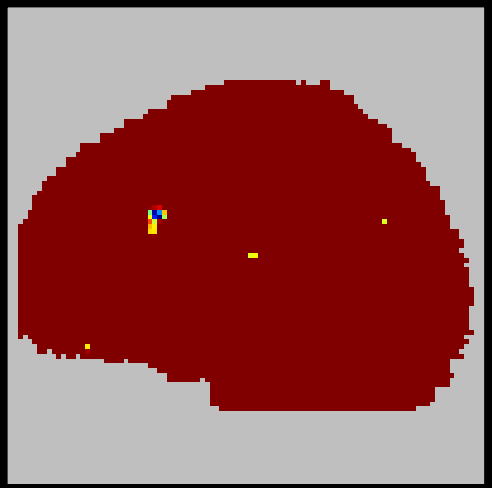


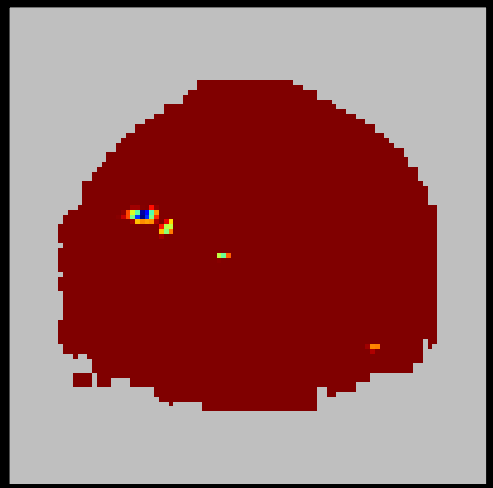




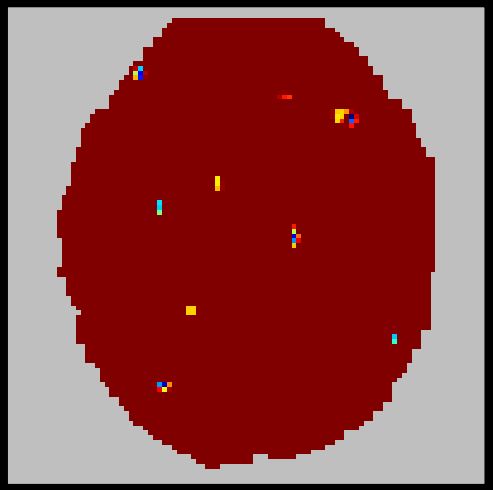
Group Level fMRI Analysis—Patient Level Results (Sbj 6)

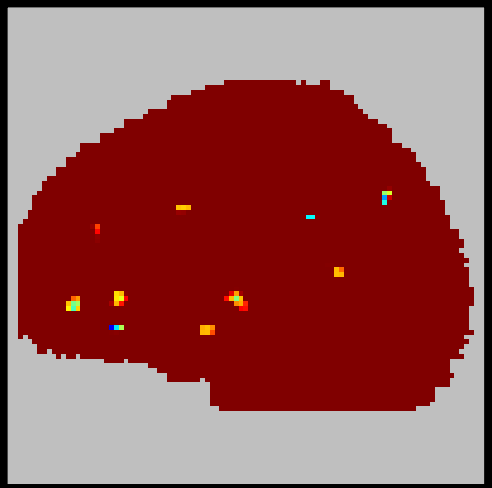


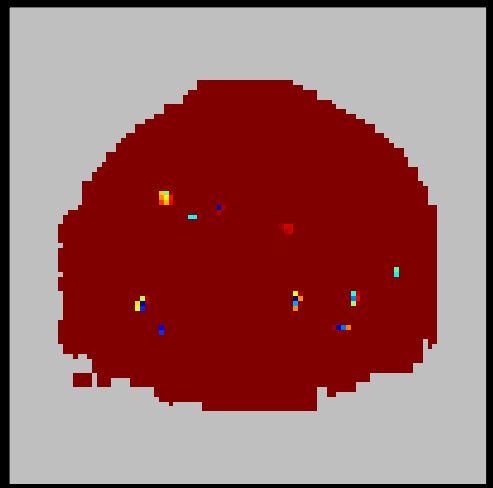




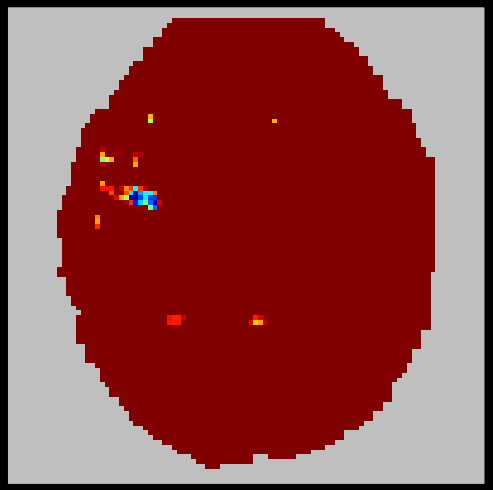
Group Level fMRI Analysis—Patient Level Results (Sbj 13)

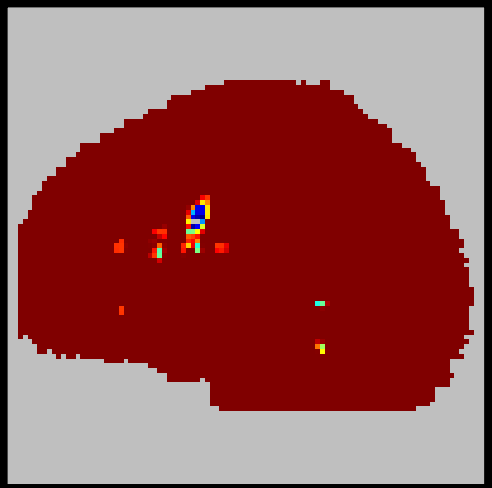


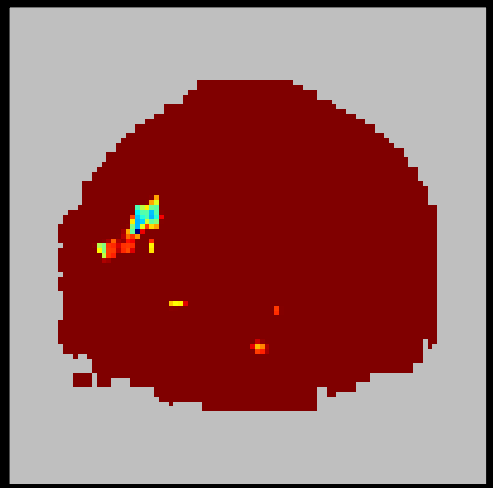




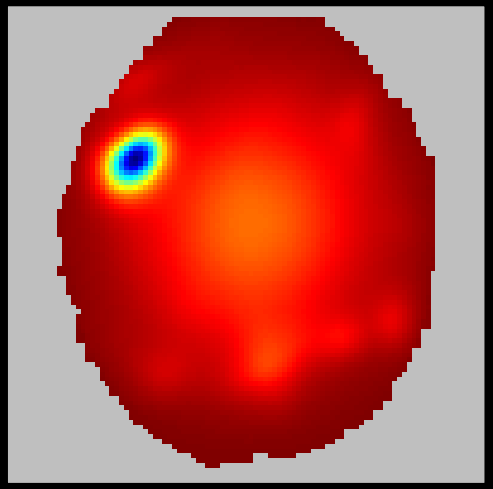
Group Level fMRI Analysis—Patient Level Results (Sbj 15)

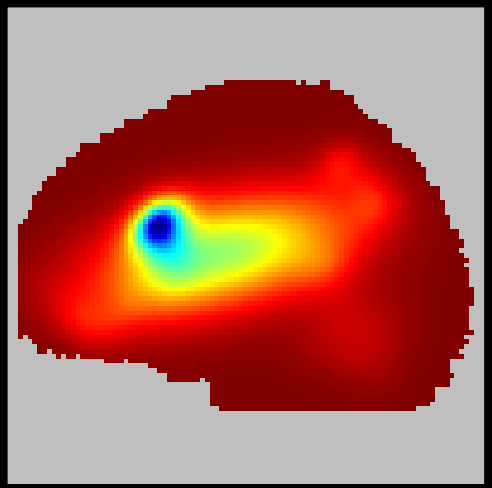


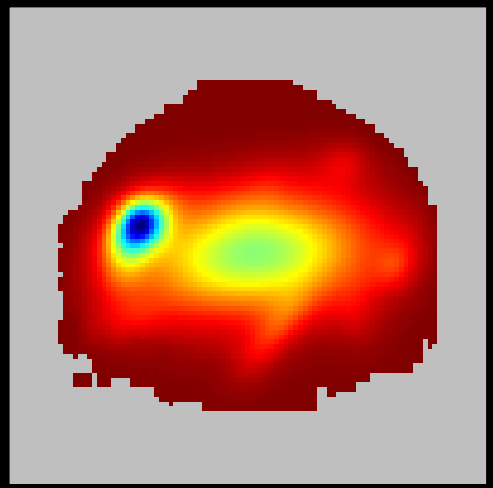




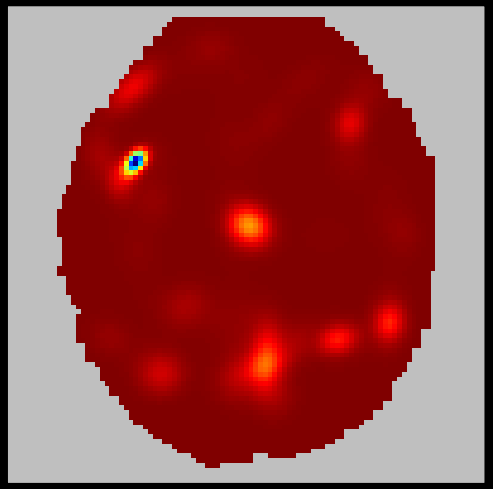
Group Level fMRI Analysis—Marginal PPD of Ind Centers

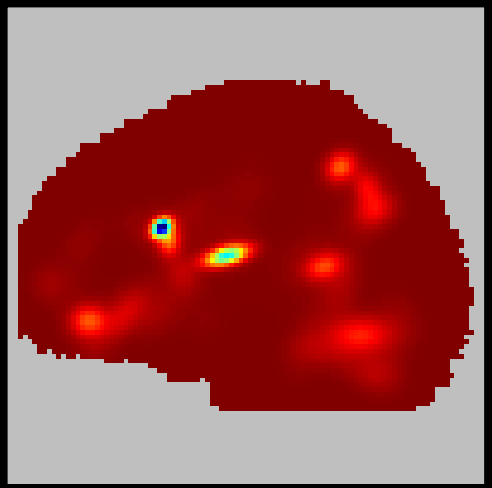


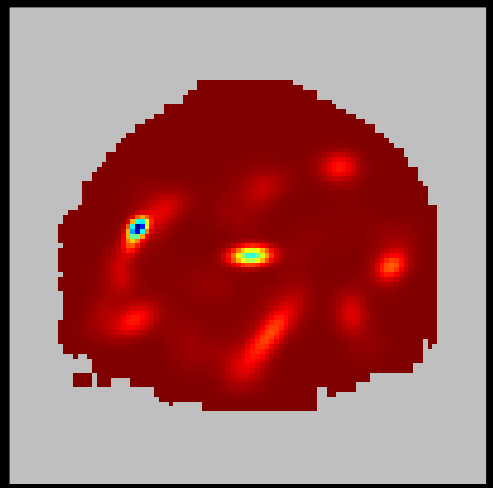




Group Level fMRI Analysis—Marginal PPD of Population Centers







Other Areas

For every imaging problem there is a Bayesian solution

- Review paper:
  - Zhang, L., Guindani, M., Vannucci M. (2014) Bayesian Models for fMRI Data Analysis, *WIRES: Computational Statistics* (to appear)
- Particle Filtering:
  - Aston, J. A. D., Johansen, A. D. (2014) Bayesian Inference on the Brain: Bayesian Solutions to Selected Problems in Neuroimaging, To appear in *Proceedings of the IWBCTA 2013*, Varanasi, India..

Algorithms

Approximation Algorithms

Stochastic

- Hamiltonian Monte Carlo (HMC)
- Reimannian Manifold HMC (RMHMC)

Deterministic

- Variational Bayes (VB)
- Integrated Nested Laplacian Approximation (INLA)

Algorithms

Approximation Algorithms

Stochastic

- Hamiltonian Monte Carlo (HMC)
- Reimannian Manifold HMC (RMHMC)

Theory guarantees approx. error can be made arbitrarily small

Deterministic

- Variational Bayes (VB)
- Integrated Nested Laplacian Approximation (INLA)

Algorithms

Approximation Algorithms

Stochastic

- Hamiltonian Monte Carlo (HMC)
- Reimannian Manifold HMC (RMHMC)

Theory guarantees approx. error can be made arbitrarily small

Should assess sensitivity to priors

Deterministic

- Variational Bayes (VB)
- Integrated Nested Laplacian Approximation (INLA)

Algorithms

Approximation Algorithms

Stochastic

- Hamiltonian Monte Carlo (HMC)
- Reimannian Manifold HMC (RMHMC)

Theory guarantees approx. error can be made arbitrarily small

Should assess sensitivity to priors

Deterministic

- Variational Bayes (VB)
- Integrated Nested Laplacian Approximation (INLA)

No such theory

Algorithms

Approximation Algorithms

Stochastic

- Hamiltonian Monte Carlo (HMC)
- Reimannian Manifold HMC (RMHMC)

Theory guarantees approx. error can be made arbitrarily small

Should assess sensitivity to priors

Deterministic

- Variational Bayes (VB)
- Integrated Nested Laplacian Approximation (INLA)

No such theory

Should conduct thorough sim. studies to assess approx. error

Algorithms

Approximation Algorithms

Stochastic

- Hamiltonian Monte Carlo (HMC)
- Reimannian Manifold HMC (RMHMC)

Theory guarantees approx. error can be made arbitrarily small

Should assess sensitivity to priors

Deterministic

- Variational Bayes (VB)
- Integrated Nested Laplacian Approximation (INLA)

No such theory

Should conduct thorough sim. studies to assess approx. error

Should assess sensitivity to priors

HMC

$$H(q, p) = U(q) + K(p)$$

HMC

$$H(q,p) = U(q) + K(p)$$

- $H(q,p)$—The (separable) Hamiltonian

HMC

$$H(q,p) = U(q) + K(p)$$

- $H(q,p)$—The (separable) Hamiltonian
- $U(q)$—Potential energy function

HMC

$$H(q,p) = U(q) + K(p)$$

- $H(q,p)$—The (separable) Hamiltonian
- $U(q)$—Potential energy function
- $K(p)$—Kinetic energy function

HMC

$$H(q,p) = U(q) + K(p)$$

- $H(q,p)$—The (separable) Hamiltonian
- $U(q)$—Potential energy function
- $K(p)$—Kinetic energy function
- q—Position vector of the particles (parameter vector)

HMC

$$H(q, p) = U(q) + K(p)$$

- $H(q, p)$—The (separable) Hamiltonian
- $U(q)$—Potential energy function
- $K(p)$—Kinetic energy function
- q—Position vector of the particles (parameter vector)
- p—Momentum vector of particles (latent vector)

HMC

$$H(q,p) = U(q) + K(p)$$

- $H(q,p)$—The (separable) Hamiltonian
- $U(q)$—Potential energy function
- $K(p)$—Kinetic energy function
- q—Position vector of the particles (parameter vector)
- p—Momentum vector of particles (latent vector)

Partial derivatives of the Hamiltonian determine how q and p change over time

HMC

Hamiltonian (partial differential) equations:

$$\begin{aligned}
\frac{dq_i}{dt} &= \frac{\partial H(q,p)}{\partial p_i} &&= \frac{\partial K(p)}{\partial p_i} \\
\frac{dp_i}{dt} &= -\frac{\partial H(q,p)}{\partial q_i} &&= -\frac{\partial U(q)}{\partial q_i}
\end{aligned}$$

HMC

Hamiltonian (partial differential) equations:

$$\frac{dq_i}{dt} = \frac{\partial H(q,p)}{\partial p_i} = \frac{\partial K(p)}{\partial p_i}$$
$$\frac{dp_i}{dt} = -\frac{\partial H(q,p)}{\partial q_i} = -\frac{\partial U(q)}{\partial q_i}$$

For HMC:

- $U(q)$—minus the log posterior density
- $K(p) = \frac{1}{2}p'M^{-1}p$
  - M is a SPD matrix, typically a scalar multiple of the identity matrix
- IF analytic solution to Hamilton equations, we have a deterministic solution to our Bayesian problem
- Typically need to solve equations numerically

HMC

Hamiltonian (partial differential) equations:

$$\frac{dq_i}{dt} = \frac{\partial H(q,p)}{\partial p_i} = \frac{\partial K(p)}{\partial p_i}$$
$$\frac{dp_i}{dt} = -\frac{\partial H(q,p)}{\partial q_i} = -\frac{\partial U(q)}{\partial q_i}$$

For HMC:

- Integrals approx. by iterating with the *Leapfrog Method*
- Solution will be biased (due to approx. error) unless
  - Metropolis update performed (either accept or reject current state)
  - Acceptance rates typically high (so almost deterministic solution)
- Mixing typically much faster than Metropolis-Hastings
  - Don't have to draw as many samples
  - **`http://mc-stan.org`**

RMHMC

For RMHMC:

- $K(q,p) = \frac{1}{2} p' M^{-1}(q) p$
- Don't need to guess $M(q)$
  - Automatically adjusts to geometry of parameter manifold
- $M(q)$ is expected Fisher info. matrix + negative Hessian of log-prior
- For RMHMC, need the inverse of $M(q)$ (no longer diagonal)
  - In most imaging problems the dim. of M(q) is too large to invert

VB

- Approximates solution to $\pi(\theta \mid y)$ with a density $q(\theta)$

VB

- Approximates solution to $\pi(\theta \mid y)$ with a density $q(\theta)$
- Restrict q to a manageable class of densities
  - $q(\theta) = \prod_{i=1}^{p} q_i(\theta_i)$ (mean-field approximation)
  - q is a member of a parametric family

VB

- Approximates solution to $\pi(\theta \mid y)$ with a density $q(\theta)$
- Restrict q to a manageable class of densities
  - $q(\theta) = \prod_{i=1}^{p} q_i(\theta_i)$ (mean-field approximation)
  - q is a member of a parametric family
- Minimize the K-L distance between $q(\theta)$ and $\pi(\theta \mid y)$:

$$\int q(\theta) \ln \left[\frac{q(\theta)}{\pi(\theta \mid y)} \right] d\theta$$

  - Iterate until some convergence criteria is met

VB

- Approximates solution to $\pi(\theta \mid y)$ with a density $q(\theta)$
- Restrict q to a manageable class of densities
  - $q(\theta) = \prod_{i=1}^{p} q_i(\theta_i)$ (mean-field approximation)
  - q is a member of a parametric family
- Minimize the K-L distance between $q(\theta)$ and $\pi(\theta \mid y)$:

$$\int q(\theta) \ln \left[\frac{q(\theta)}{\pi(\theta \mid y)} \right] d\theta$$

  - Iterate until some convergence criteria is met
- KEY: find a good variational density q that is much easier to deal with than $\pi(\theta \mid y)$

VB

- Approximates solution to $\pi(\theta \mid y)$ with a density $q(\theta)$
- Restrict q to a manageable class of densities
  - $q(\theta) = \prod_{i=1}^{p} q_i(\theta_i)$ (mean-field approximation)
  - q is a member of a parametric family
- Minimize the K-L distance between $q(\theta)$ and $\pi(\theta \mid y)$:

$$\int q(\theta) \ln \left[\frac{q(\theta)}{\pi(\theta \mid y)} \right] d\theta$$

  - Iterate until some convergence criteria is met
- KEY: find a good variational density q that is much easier to deal with than $\pi(\theta \mid y)$
- Typically much faster than MCMC

VB

- Approximates solution to $\pi(\theta \mid y)$ with a density $q(\theta)$
- Restrict q to a manageable class of densities
  - $q(\theta) = \prod_{i=1}^{p} q_i(\theta_i)$ (mean-field approximation)
  - q is a member of a parametric family
- Minimize the K-L distance between $q(\theta)$ and $\pi(\theta \mid y)$:

$$\int q(\theta) \ln \left[\frac{q(\theta)}{\pi(\theta \mid y)} \right] d\theta$$

  - Iterate until some convergence criteria is met
- KEY: find a good variational density q that is much easier to deal with than $\pi(\theta \mid y)$
- Typically much faster than MCMC
  - However, posterior variances underestimated—sometimes severely

INLA

Consider the posterior of a latent Gaussian model: $\pi(x, \theta \mid y)$
Posterior marginals are

$$\pi(x_i \mid y) = \int \pi(x_i \mid \theta, y)\pi(\theta \mid y)d\theta$$

$$\pi(\theta_j \mid y) = \int \pi(\theta \mid y)d\theta_{-j}$$

INLA

Consider the posterior of a latent Gaussian model: $\pi(x, \theta \mid y)$
Posterior marginals are

$$\begin{aligned} \pi(x_i \mid y) &= \int \pi(x_i \mid \theta, y)\pi(\theta \mid y)d\theta \\ \pi(\theta_j \mid y) &= \int \pi(\theta \mid y)d\theta_{-j} \end{aligned}$$

Construct nested approximations:

$$\begin{aligned} \tilde{\pi}(x_i \mid y) &= \int \tilde{\pi}(x_i \mid \theta, y)\tilde{\pi}(\theta \mid y)d\theta \\ \tilde{\pi}(\theta_j \mid y) &= \int \tilde{\pi}(\theta \mid y)d\theta_{-j} \end{aligned}$$

INLA

Now

$$\pi(\theta \mid y) \approx \tilde{\pi}(\theta \mid y) \propto \left. \frac{\pi(x, \theta, y)}{\tilde{\pi}_G(x \mid \theta, y)} \right|_{x = x^*(\theta)}$$

This is a Laplace approx.

INLA

Now

$$\pi(\theta \mid y) \approx \tilde{\pi}(\theta \mid y) \propto \left. \frac{\pi(x, \theta, y)}{\tilde{\pi}_G(x \mid \theta, y)} \right|_{x = x^*(\theta)}$$

This is a Laplace approx.

Also

$$\pi(x_i \mid \theta, y)$$

approximated by another Laplacian approximation.

INLA

Now

$$\pi(\theta \mid y) \approx \tilde{\pi}(\theta \mid y) \propto \left. \frac{\pi(x, \theta, y)}{\tilde{\pi}_G(x \mid \theta, y)} \right|_{x = x^*(\theta)}$$

This is a Laplace approx.

Also

$$\pi(x_i \mid \theta, y)$$

approximated by another Laplacian approximation.

- Numerical integration used to approximation the full marginals

INLA

Now

$$\pi(\theta \mid y) \approx \tilde{\pi}(\theta \mid y) \propto \left. \frac{\pi(x, \theta, y)}{\tilde{\pi}_G(x \mid \theta, y)} \right|_{x = x^*(\theta)}$$

This is a Laplace approx.

Also

$$\pi(x_i \mid \theta, y)$$

approximated by another Laplacian approximation.

- Numerical integration used to approximation the full marginals
- Very fast and accurate

INLA

Now

$$\pi(\theta \mid y) \approx \tilde{\pi}(\theta \mid y) \propto \left. \frac{\pi(x, \theta, y)}{\tilde{\pi}_G(x \mid \theta, y)} \right|_{x = x^*(\theta)}$$

This is a Laplace approx.

Also

$$\pi(x_i \mid \theta, y)$$

approximated by another Laplacian approximation.

- Numerical integration used to approximation the full marginals
- Very fast and accurate
- R package available that solves many problems
  - **http://www.r-inla.org**

Which One?

- It all depends

Which One?

- It all depends
- INLA or STAN—little time setting up model and fast

Which One?

- It all depends
- INLA or STAN—little time setting up model and fast
- If special problem—must program (VB, INLA and HMC)

Which One?

- It all depends
- INLA or STAN—little time setting up model and fast
- If special problem—must program (VB, INLA and HMC)
  - Programming will be time consuming

Which One?

- It all depends
- INLA or STAN—little time setting up model and fast
- If special problem—must program (VB, INLA and HMC)
  - Programming will be time consuming
  - INLA and VB should run faster than HMC

Which One?

- It all depends
- INLA or STAN—little time setting up model and fast
- If special problem—must program (VB, INLA and HMC)
  - Programming will be time consuming
  - INLA and VB should run faster than HMC
  - INLA and VB:
    - No theory to guarantee arbitrarily small approximation error
    - Since both rely on approximations, I recommend simulation studies to assess statistical properties of estimators—time consuming and should compare to MCMC or HMC solution

Which One?

- It all depends
- INLA or STAN—little time setting up model and fast
- If special problem—must program (VB, INLA and HMC)
  - Programming will be time consuming
  - INLA and VB should run faster than HMC
  - INLA and VB:
    - No theory to guarantee arbitrarily small approximation error
    - Since both rely on approximations, I recommend simulation studies to assess statistical properties of estimators—time consuming and should compare to MCMC or HMC solution
  - HMC:
    - Need analytic derivatives
    - Need to tune numerical integration step
    - Theory ensures approximation error can be made arbitrarily small

Simulation Study: HMC, VB or INLA?

- Interested in LGCPs on the brain (3D problem)

Simulation Study: HMC, VB or INLA?

- Interested in LGCPs on the brain (3D problem)
- Will require coding
  - Regardless of whether I choose INLA, VB or HMC

Simulation Study: HMC, VB or INLA?

- Interested in LGCPs on the brain (3D problem)
- Will require coding
  - Regardless of whether I choose INLA, VB or HMC
- Conducting simulation study on small 64×64 grid (note: 2D)
  - INLA can fit LGCP models on 2D grids

Simulation Study: HMC, VB or INLA?

- Interested in LGCPs on the brain (3D problem)
- Will require coding
  - Regardless of whether I choose INLA, VB or HMC
- Conducting simulation study on small 64×64 grid (note: 2D)
  - INLA can fit LGCP models on 2D grids
- Generated 2D intensity function (LGCP)

Simulation Study: HMC, VB or INLA?

- Interested in LGCPs on the brain (3D problem)
- Will require coding
  - Regardless of whether I choose INLA, VB or HMC
- Conducting simulation study on small 64×64 grid (note: 2D)
  - INLA can fit LGCP models on 2D grids
- Generated 2D intensity function (LGCP)
- Simulated 1000 point patterns based on this intensity

Simulation Study: HMC, VB or INLA?

- Interested in LGCPs on the brain (3D problem)
- Will require coding
  - Regardless of whether I choose INLA, VB or HMC
- Conducting simulation study on small 64×64 grid (note: 2D)
  - INLA can fit LGCP models on 2D grids
- Generated 2D intensity function (LGCP)
- Simulated 1000 point patterns based on this intensity
- Assessed stat. properties of estimators from INLA, VB and HMC

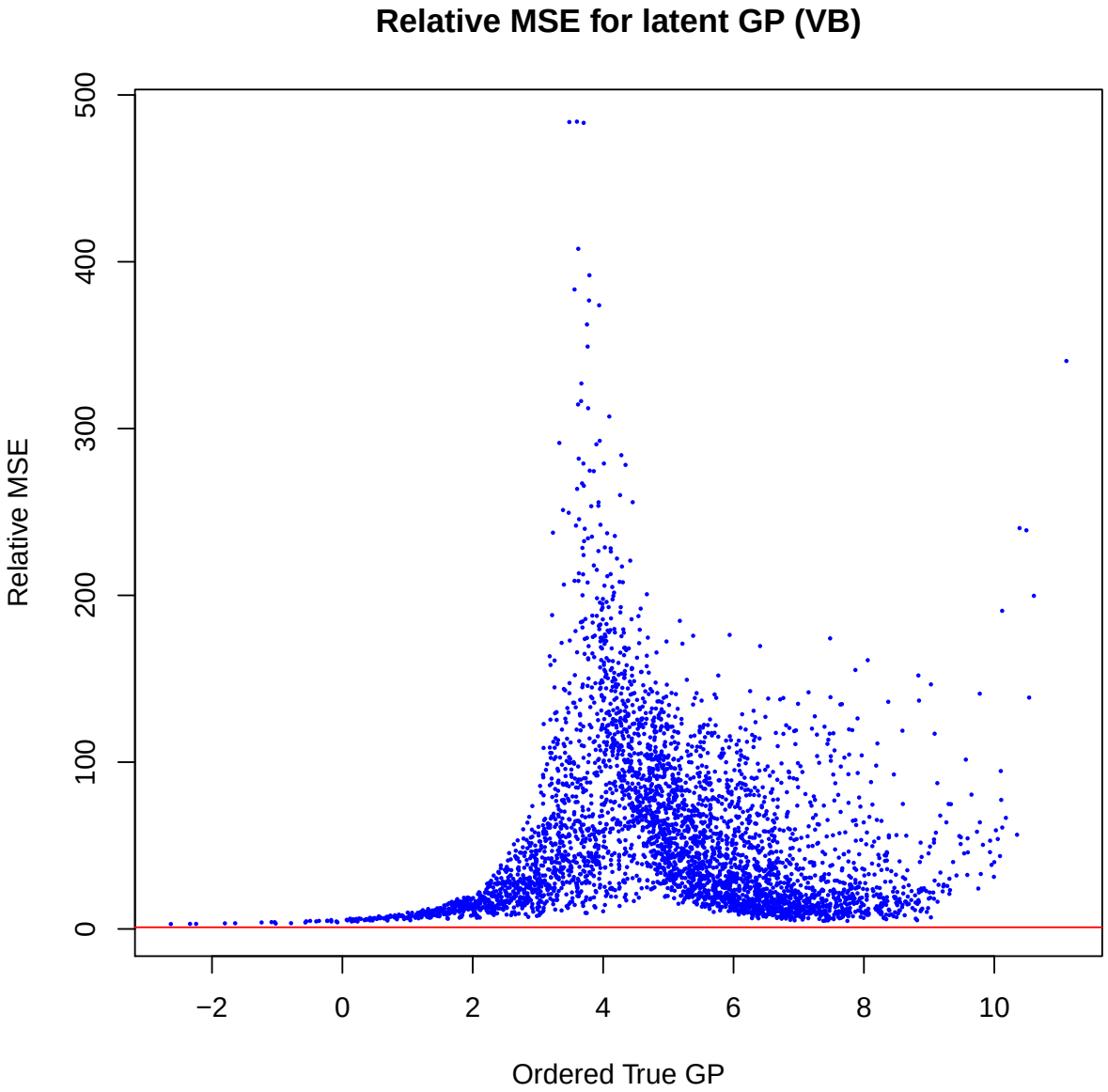
Simulation Study—Results

True Simulation Values: $\mu = 5 \quad \sigma^{-2} = 0.286 \quad E(N) = 792.1$

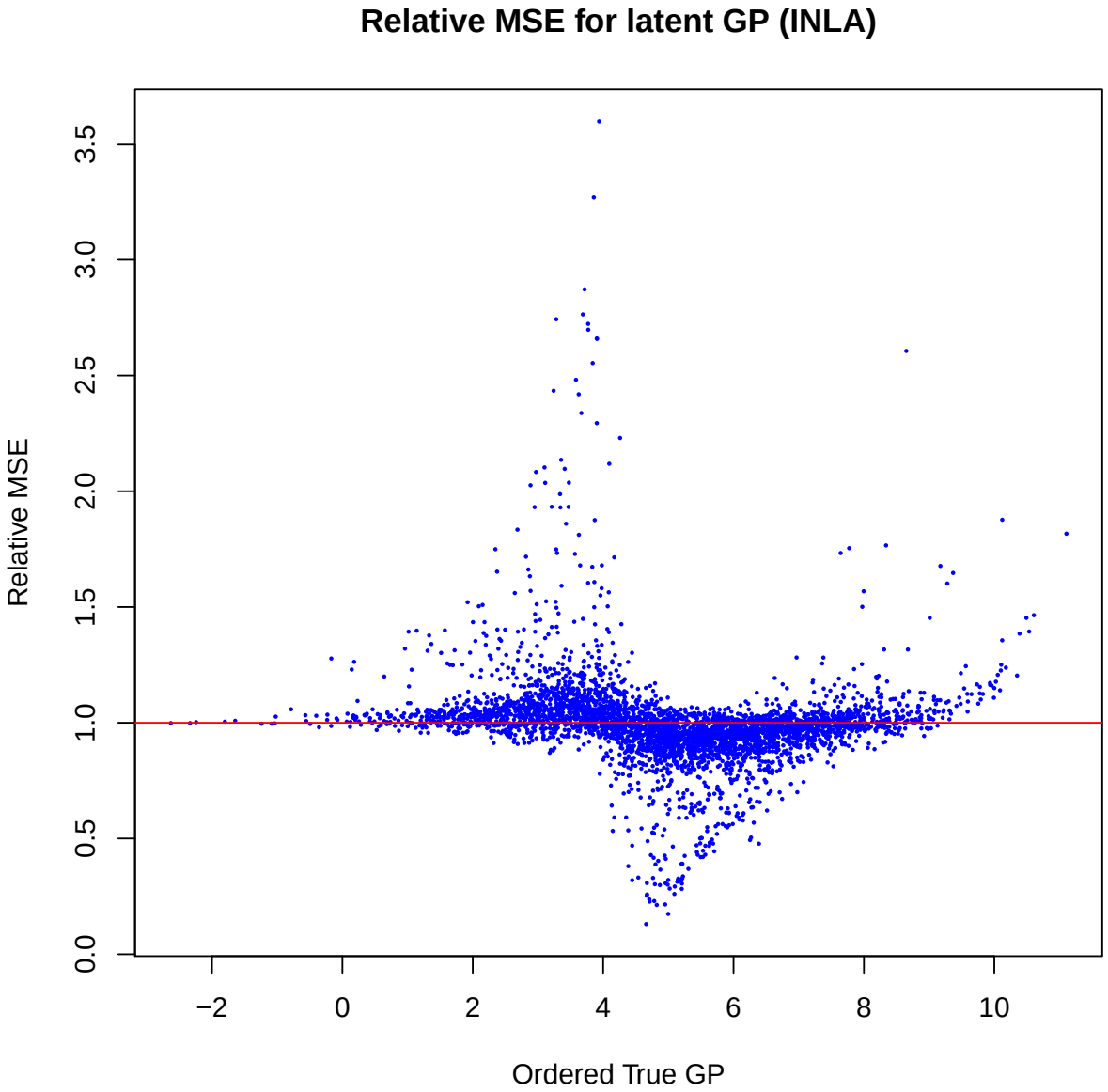
	HMC		INLA		VB	
Parm	Bias	MSE	Bias(rel)	MSE(rel)	Bias(rel)	MSE(rel)
μ	0.071	0.014	0.261(3.69)	0.077(5.39)	1.23(17.07)	1.52(106.73)
σ^{-2}	0.015	$6e^{-4}$	0.061(4.15)	0.004(7.01)	0.281(19.20)	0.079(131)
$E(N)$	0.877	791	-179(-204)	32666(41.28)	-11.36(-13)	958.5(1.21)

$E(N)$ is the expected number of points over the region.
It is the integrated intensity function.

Simulation Study—Results



Simulation Study—Results



Parallelization

- Are Bayesian methods amenable to parallelization?

Parallelization

- Are Bayesian methods amenable to parallelization?
  - Stochastic algorithms serial in nature

Parallelization

- Are Bayesian methods amenable to parallelization?
  - Stochastic algorithms serial in nature
  - Likelihood, however, typically conditionally independent

Parallelization

- Are Bayesian methods amenable to parallelization?
  - Stochastic algorithms serial in nature
  - Likelihood, however, typically conditionally independent
  - Can parallelize likelihood, which is usually the most computationally expensive component

Parallelization

- Are Bayesian methods amenable to parallelization?
  - Stochastic algorithms serial in nature
  - Likelihood, however, typically conditionally independent
  - Can parallelize likelihood, which is usually the most computationally expensive component
  - Particle filtering—easy to parallelize. Each particle independent.

Parallelization

- Are Bayesian methods amenable to parallelization?
  - Stochastic algorithms serial in nature
  - Likelihood, however, typically conditionally independent
  - Can parallelize likelihood, which is usually the most computationally expensive component
  - Particle filtering—easy to parallelize. Each particle independent.
- Clusters/multi-CPU machines
  - Best for task parallelization

Parallelization

- Are Bayesian methods amenable to parallelization?
  - Stochastic algorithms serial in nature
  - Likelihood, however, typically conditionally independent
  - Can parallelize likelihood, which is usually the most computationally expensive component
  - Particle filtering—easy to parallelize. Each particle independent.
- Clusters/multi-CPU machines
  - Best for task parallelization
- GPUs
  - Best for data parallelization
  - Extremely good at “embarrassingly parallel” operations

GPU Example

Ge, T., Müller-Lenke, N., Bendfeldt, K., Nichols, T. E., Johnson, T. D. (2014) Analysis of Multiple Sclerosis lesions via spatially varying coefficients. *AOAS* **8** 1095–1118.

- A study of Multiple Sclerosis MRI data

GPU Example

Ge, T., Müller-Lenke, N., Bendfeldt, K., Nichols, T. E., Johnson, T. D. (2014) Analysis of Multiple Sclerosis lesions via spatially varying coefficients. *AOAS* **8** 1095–1118.

- A study of Multiple Sclerosis MRI data
- Want to correlate clinical symptoms with lesion location

GPU Example

Ge, T., Müller-Lenke, N., Bendfeldt, K., Nichols, T. E., Johnson, T. D. (2014) Analysis of Multiple Sclerosis lesions via spatially varying coefficients. *AOAS* **8** 1095–1118.

- A study of Multiple Sclerosis MRI data
- Want to correlate clinical symptoms with lesion location
- Lesions segmented by Neuroradiologists

GPU Example

Ge, T., Müller-Lenke, N., Bendfeldt, K., Nichols, T. E., Johnson, T. D. (2014) Analysis of Multiple Sclerosis lesions via spatially varying coefficients. *AOAS* **8** 1095–1118.

- A study of Multiple Sclerosis MRI data
- Want to correlate clinical symptoms with lesion location
- Lesions segmented by Neuroradiologists
- Work with binary images as outcomes
  - Spatial generalized linear model (probit or logit link)

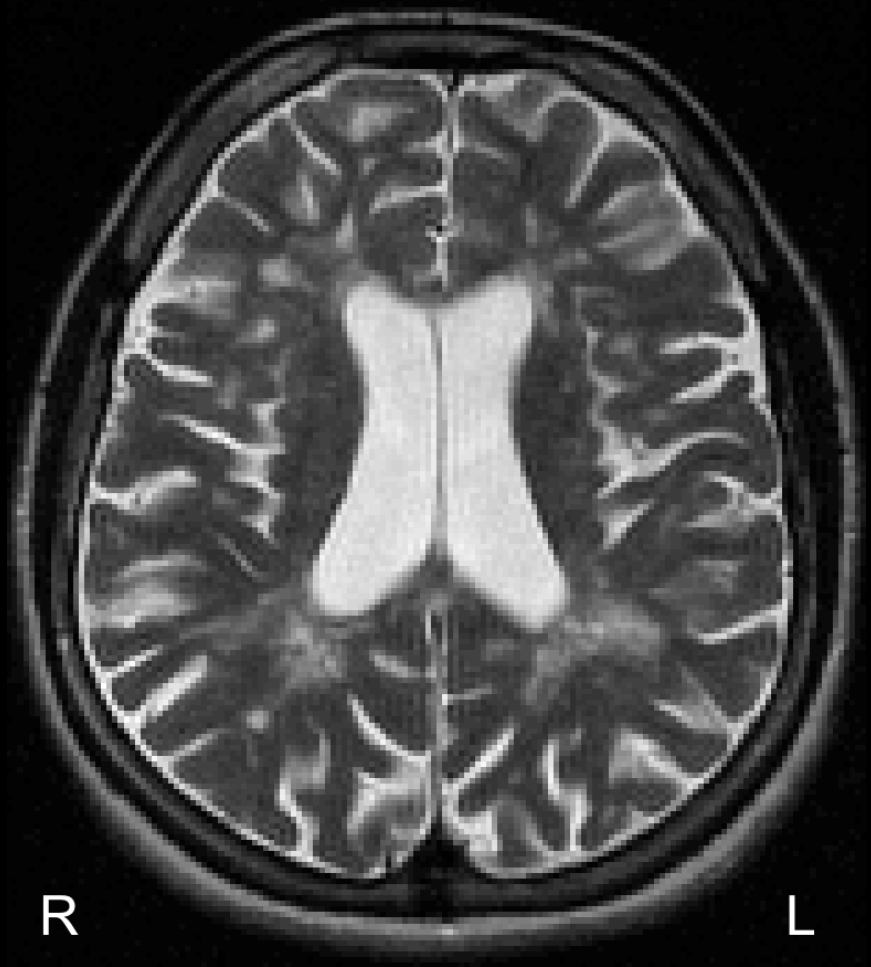
GPU Example

Ge, T., Müller-Lenke, N., Bendfeldt, K., Nichols, T. E., Johnson, T. D. (2014) Analysis of Multiple Sclerosis lesions via spatially varying coefficients. *AOAS* **8** 1095–1118.

- A study of Multiple Sclerosis MRI data
- Want to correlate clinical symptoms with lesion location
- Lesions segmented by Neuroradiologists
- Work with binary images as outcomes
  - Spatial generalized linear model (probit or logit link)
- Clinical symptoms + nuisance covariates
  - Parameters are spatially varying over the brain and are spatially correlated
  - GMRF used to model the spatial correlation

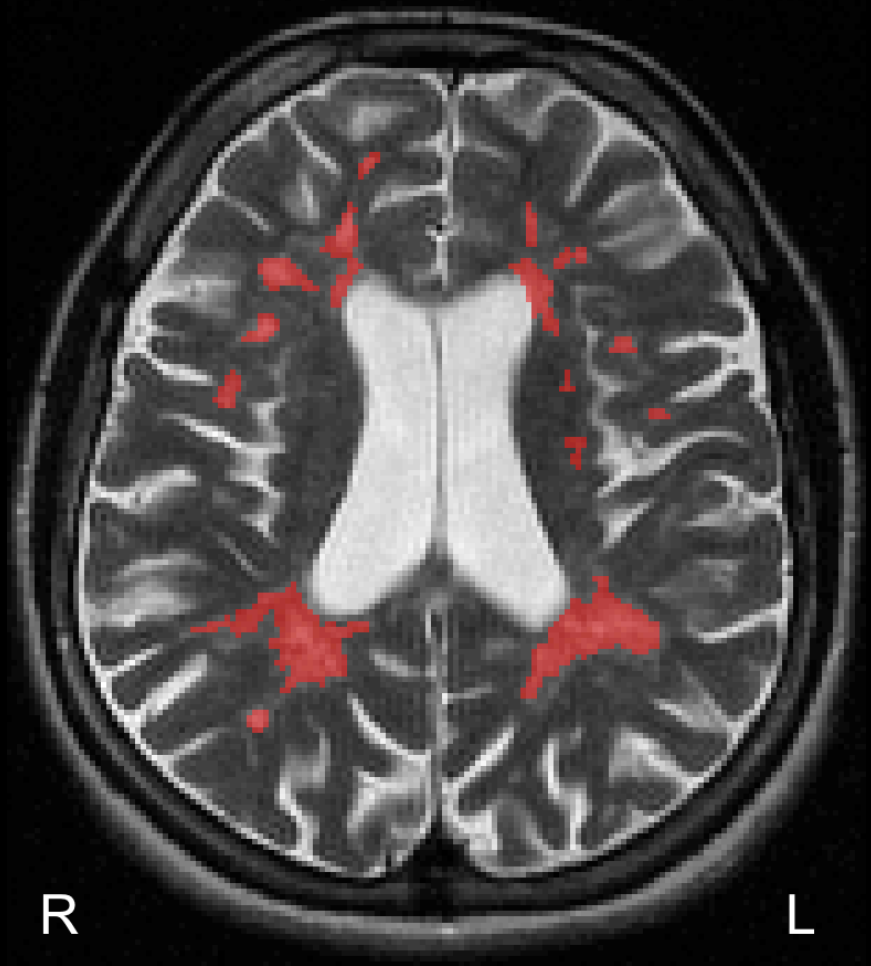
GPU Example—MS high-resolution imaging

Data are T_2 hyperintense lesions



GPU Example—MS high-resolution imaging

Data are T_2 hyperintense lesions



GPU Example

Covariates:

- 15 subject specific covariates
  - 7 FSS, PASAT score, age, gender, disease duration
  - 4 MS subtypes (dummy coded into 4 variables)

GPU Example

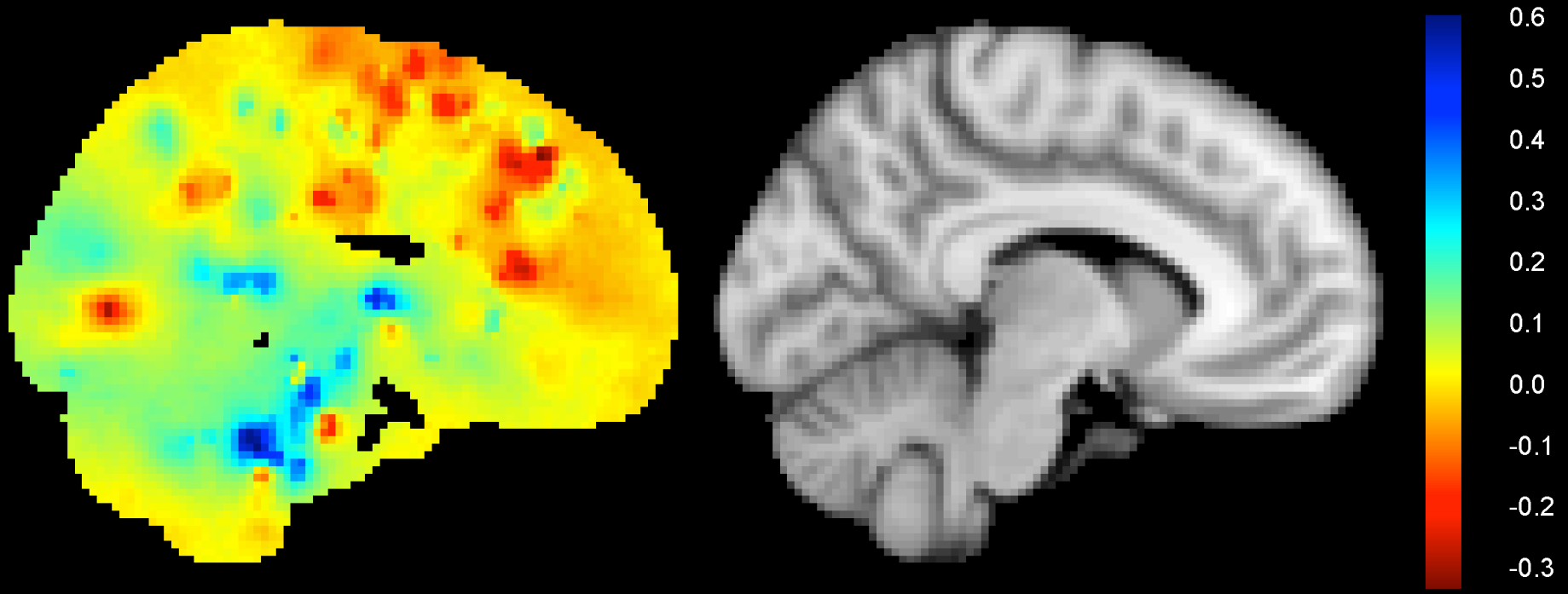
Covariates:

- 15 subject specific covariates
  - 7 FSS, PASAT score, age, gender, disease duration
  - 4 MS subtypes (dummy coded into 4 variables)

Problem Size:

- $\approx$ 66 million observations (275K voxels $\times$ 239 subjects)
- $\approx$ 41 million spatially varying coefficients (275K voxels $\times$ 15 covariates)

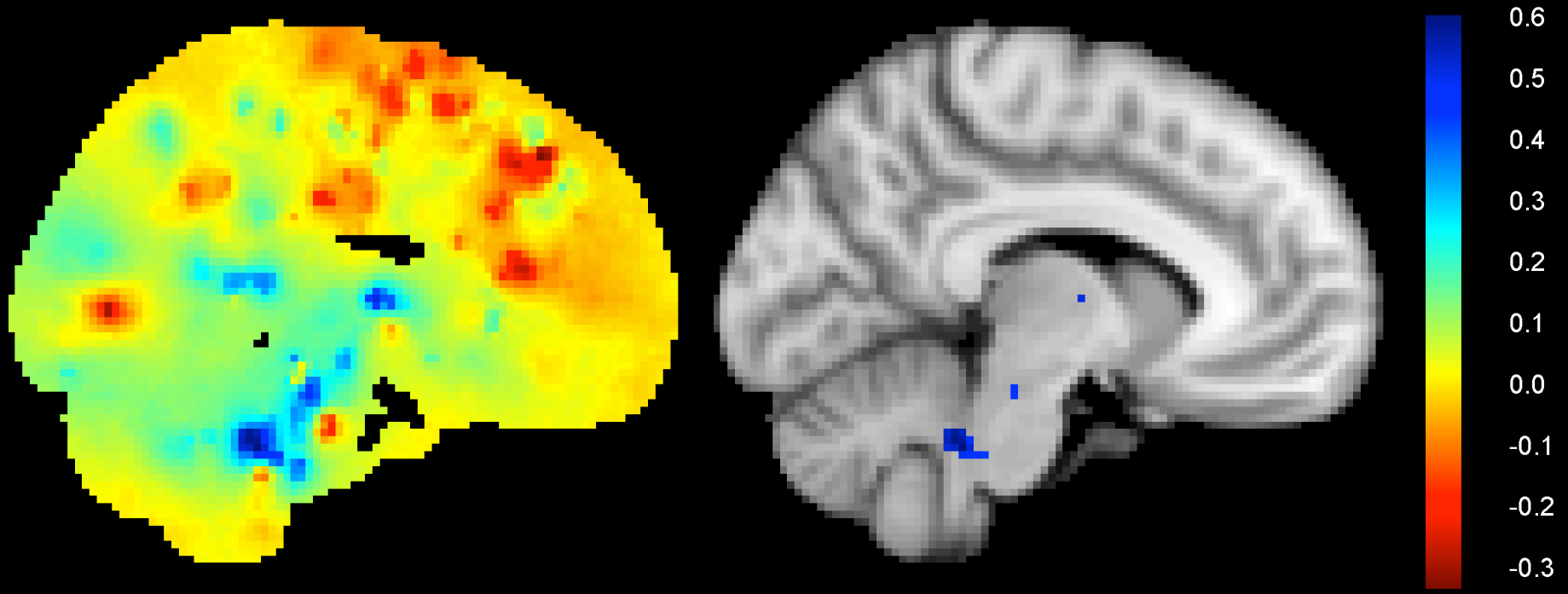
Spatially Varying Coefficients: Cerebellar Func. System Score



6 mm left of midline

β

Spatially Varying Coefficients: Cerebellar Func. System Score



6 mm left of midline

β

GPU Example

Timing:

- 10K iterations after 20K of burning
  - CPU: (Serial code). 38.67 sec/iteration (3.3 GHz processor, Linux)
  - GPU: (Parallel code). 0.21 sec/iteration (NVIDIA K20c, 2496 threads)

- Speed up: approximately 184 times faster.
  - 13.4 days (CPU) vs. 1hr 45min (GPU)

Conclusion

- Recent work in simulation algorithms is drastically reducing the computational expense/workload

Conclusion

- Recent work in simulation algorithms is drastically reducing the computational expense/workload
- Combined with the ability to easily incorporate
  - prior information
  - spatial and temporal correlation

Conclusion

- Recent work in simulation algorithms is drastically reducing the computational expense/workload
- Combined with the ability to easily incorporate
  - prior information
  - spatial and temporal correlation

The future of Bayesian Analysis in Neuroimaging appears Bright

References

INLA

1. Rue, H., Martino, S., Chopin, N. (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (w/discussion). *JRSSB* **71**(2) 319–392.
2. Illian, J. B., Sørbye, S. H., Rue, H. (2010) A toolbox for fitting complex spatial point process models using integrated nested laplace approximations (INLA). *Technical Report* Trondheim University, Trondheim, Norway.
3. Lindgren, F., Rue, H., Lindström, J. (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach (w/discussion). *JRSSB* **73**(4) 423–498.

References

Variational Bayes

1. Jaakkola, T., Jordan, M. I. (2000) Bayesian parameter estimation via variational methods. *Statistics and Computing* **10** 25–37.
2. Ormerod, J. T., Wand, M. P. (2010) Explaining Variational Approximations. *The American Statistician* **64**(2) 140–153.

HMC

1. Neal, R. M. (2011) MCMC using Hamiltonian dynamics. *in* Handbook of Markov chain Monte Carlo. Eds: Broosk, S., Gelman, A., Jones, G. L., Meng, X-L., 113–162. CRC Press.
2. Girolami, M., Calderhead, B. (2011) Riemann manifold Langevin and Hamiltonian Monte Carlo methods (w/discussion). *JRSSB* **73**(2) 123–214.

References

1. Zhang, L., Guindani, M., Vannucci, M. (2014) Bayesian models for fMRI data analysis. To appear in *WIRES: Computational Statistics*
2. Aston, J. D., Johansen, A. M. (2014) Bayesian inference on the brain: Bayesian solutions to selected problems in Neuroimaging. To appear in *Proceedings of the IWBCTA 2013*, Varanasi, India.

1. Liu, Z., Berrocal, V. J., Bartsch, A. J., Johnson, T. D. (2014) Pre-Surgical fMRI data analysis using a spatially adaptive conditionally autoregressive model. Submitted to *Bayesian Analysis*.
2. Xu, L., Johnson, T. D., Nichols, T. E., Nee, D. (2009) Modeling inter-subject variability in fMRI activation location: a Bayesian hierarchical spatial model. *Biometrics* **65** 1041–1051.
3. Ge, T., Müller-Lenke, N., Bendfeldt, K., Nichols, T. E., Johnson, T. D. (2014) Analysis of Multiple Sclerosis lesions via spatially varying coefficients. *AOAS* **8** 1095–1118.