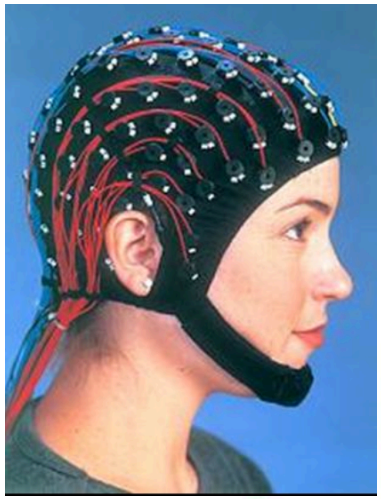
Posterior Mean Screening for Scalar-on-Image Regression

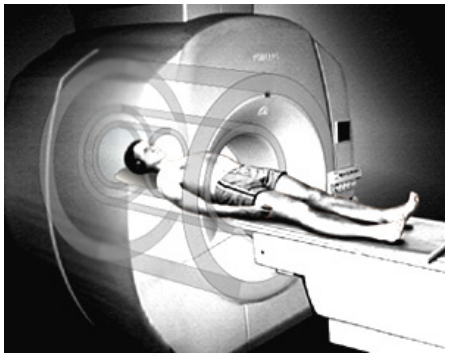
Jian Kang

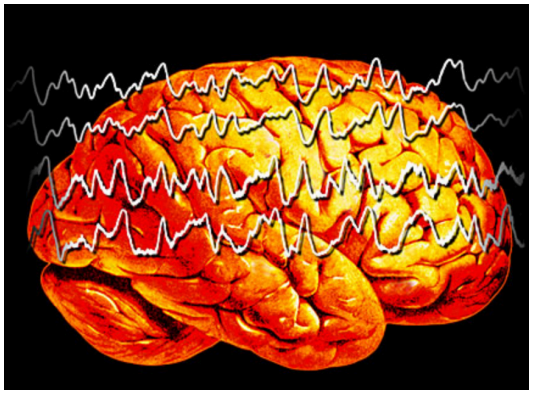
Department of Biostatistics
University of Michigan, Ann Arbor

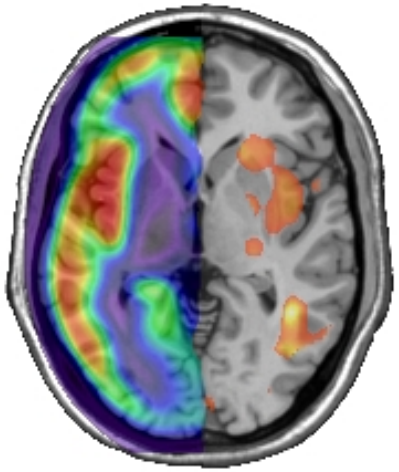
Feb 01, 2016

Neuroimaging



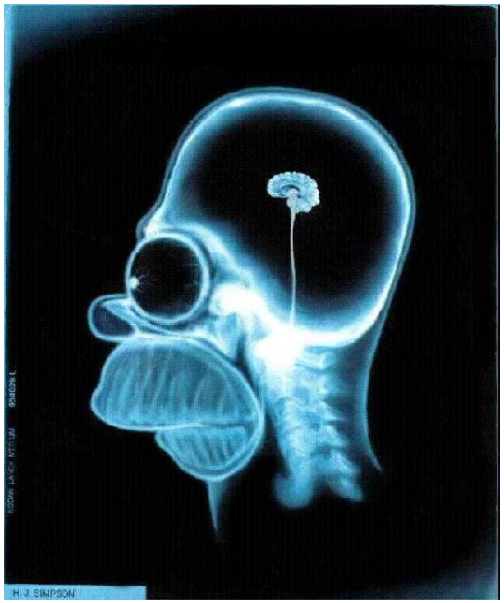






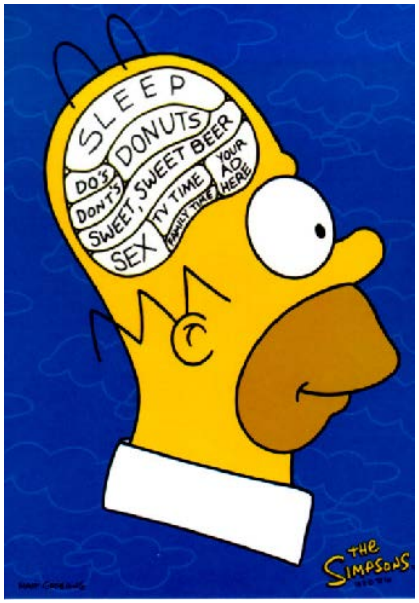
Neuroimaging

Brain anatomy



Structural neuroimaging
shows contrast
between different tissues
MRI, DTI

Brain functions



Functional neuroimaging
indirectly measure
neural activity
fMRI, PET

A Big Data Science

The amount of neuroimaging data per study reported from published articles in *NeuroImage*.

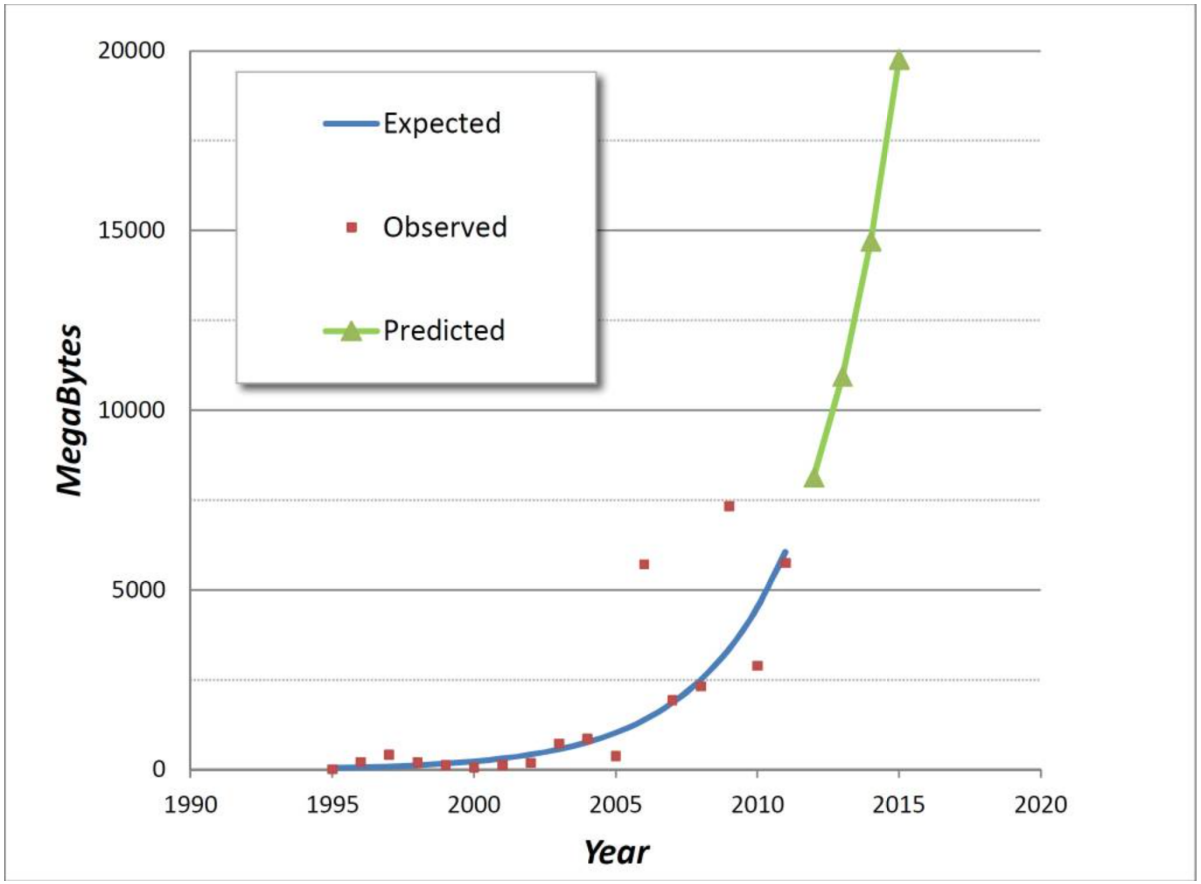
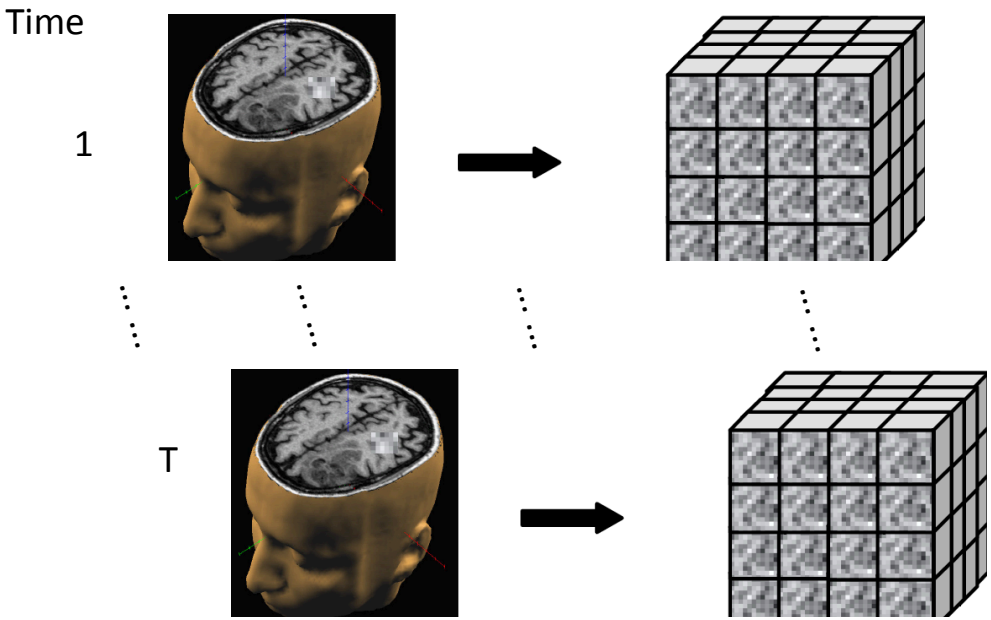


Figure 1 from Van Horn and Toga, 2014

Data Characteristics

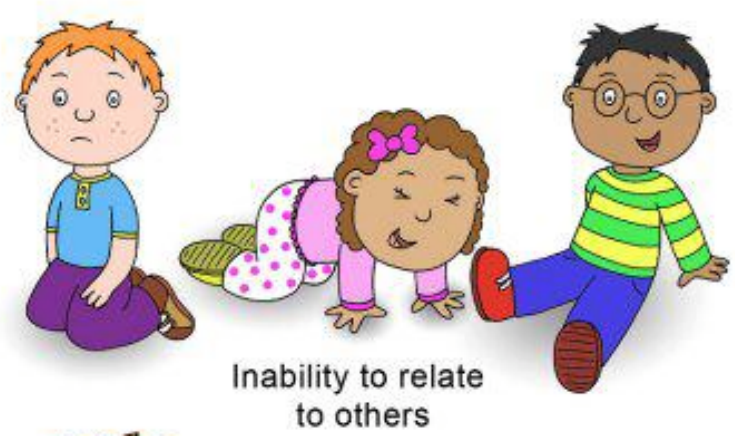
- 3D Data (Spatial Data)



- Massive dat sets: up to 300,000 voxels in a standard brain template
- Important features: contiguous regions, sharp edges and jumps
- Complex spatial correlations (neighbors, long-range between ROIs)
- Temporal correlations

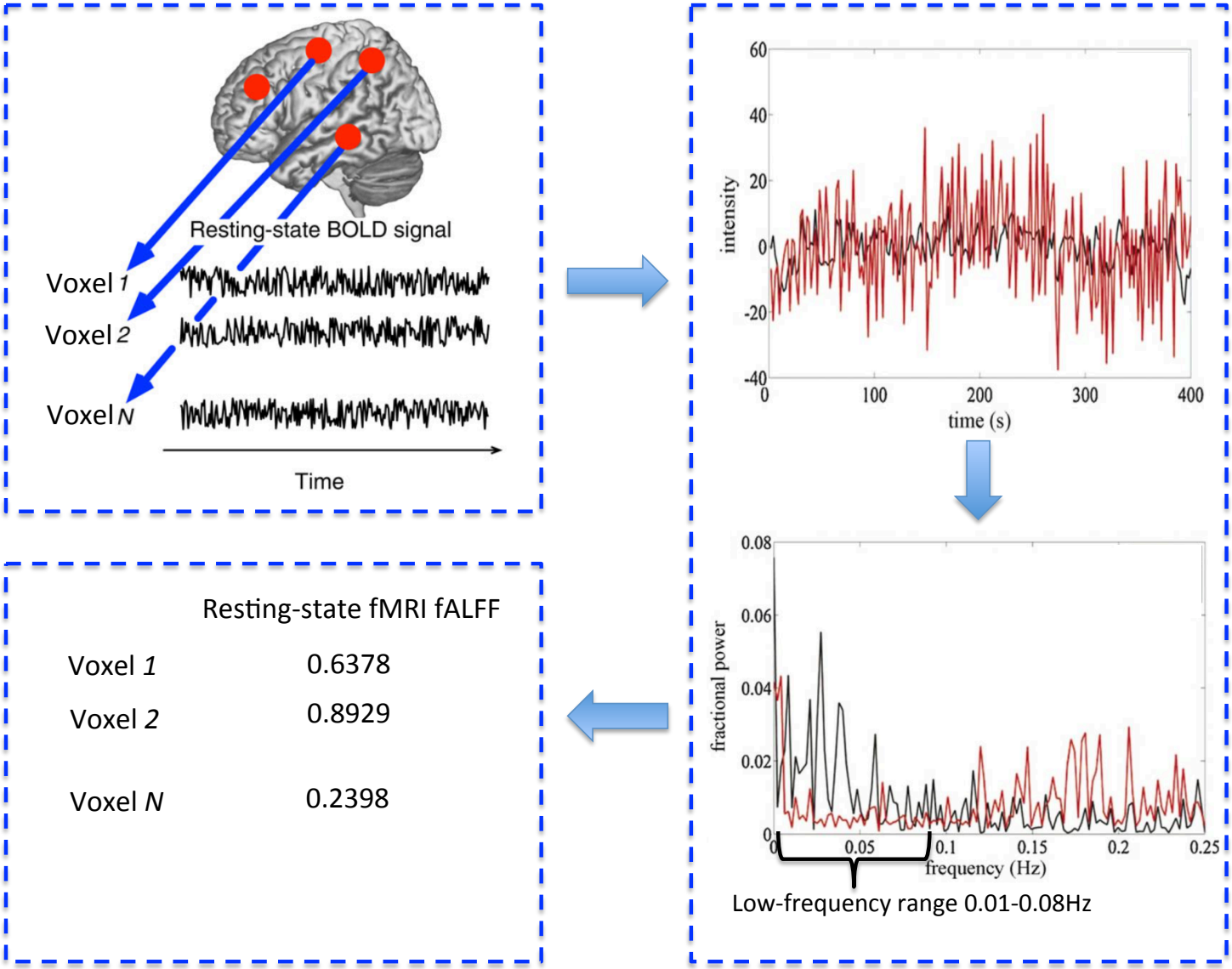
An Example

- **Autism Brain Imaging Data Exchange (ABIDE)**
- **ASD: Autism spectrum disorder**



- **Resting-state fMRI**
  - sidesteps the challenge of designing tasks
  - aggregates data sets from multiple imaging sites
  - Voxel-wise fALFF (fractional amplitude of low-frequency fluctuations) to characterize the local brain activity.

Resting-state fMRI and fALFF



Resting-state fMRI fALFF	
Voxel *1*	0.6378
Voxel *2*	0.8929
Voxel *N*	0.2398

ABIDE Data

- A total of 1,112 individuals (539 ASDs v.s. 573 typical controls) across 17 imaging sites
- For each individual, fALFF were computed over 185,405 voxels in 90 regions of interest (ROIs) in the brain based on the Automated Anatomical Labeling (AAL) system

- Demographical variables were also collected, such as age at scan, sex and intelligence quotient (IQ)

Scalar-on-Image Regression

- Prediction/classification studies: use brain image to
  - Classify a subject's group membership (e.g. disease status)
  - Predict clinical response or behavior (e.g. treatment response)
  - Predict neural activity
- Linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

- $\mathbf{y} = (y_1, \ldots, y_n)$: the outcome variable
- $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ be the image predictor with $\mathbf{x}_j = (x_{1,j}, \ldots, x_{n,j})^{\mathrm{T}}$,
- $x_{i,j}$ is the image intensity measurement at spatial location $\mathbf{s}_j \in \mathcal{B}$ subject i.

Variable Selection in High-Dimensional Feature Space

- Regularization methods (Tibshirani (1996), Fan and Li (2001), Zou and Hastie (2005)),
- Point mass mixture (Mitchell and Beauchamp (1988), George and McCulloch (1993, 1997), West (2003), Clyde and George (2004))
- Continuous shrinkage priors (Park and Casella (2008), Polson and Scott (2010), Carvalho et al. (2010), Bhattacharya et al. (2012))
- Non-local priors (Johnson and Rossell, (2012), Johnson (2013))
- Ising or binary Markov random field priors (Li and Zhang (2010), Stingo et al. (2011) Smith and Fahrmeir (2007), Goldsmith et al. (2012), Li et al (2015))
- Heuristic methods (Berger and Molina (2005), Hans, Dobra and West (2007), Scott and Carvalho (2009), Bottolo and Richardson (2010))

Ultra-high dimensional variable screening

- Straightforward Approach: Let β_j^M be one screening statistic. Given a threshold parameter γ, then the selected indices is given by

$$\mathcal{M} = \{j : |\beta_j^M| > \gamma\}.$$

- Sure Independence Screening (Fan and Lv (2008), Fan and Song (2010), Zhao and Li (2012)),
  - Screening statistics $\beta_j^{\mathrm{SIS}} = \mathbf{X}^{\mathrm{T}}\mathbf{y}$.
  - Theoretical results need strong conditions.
  - Computation order $O(np)$.
  - Performance is not good in many cases.

- High-dimensional Ordinary Least-squares Projection (Wang and Leng, 2015)
  - Screening statistics $\beta_j^{\mathrm{HOLP}} = \mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}\mathbf{y}$.
  - Theoretical results do not need very strong conditions.
  - Computation complexity $O(n^2p) + O(p^2n)$.
  - Performance is improved.

A New Approach: Posterior Mean Screening

- A new feature screening approach:
  - Motivated by scalar-on-image regression, but can be more general
  - Derived from the Bayesian modeling framework, but also has good theoretical results as frequentist methods.
  - Performance can be much better compared to SIS and HOLP for the scalar-on-image regression model
  - Computation complexity is the same as HOLP.
  - The algorithm can be paralleled using the GPU techniques.

A Bayesian Model

- Assign multivariate Gaussian distribution to $\boldsymbol{\beta}$.

$$\mathbf{y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n), \text{ and } \quad \boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0}_p, \tau^2\boldsymbol{\Lambda}),$$

  where τ^2 and σ^2 are variance parameters.

- The correlation matrix $\boldsymbol{\Lambda}$ can be chosen flexible: e.g.
  - Incorporate the spatial smoothness within local regions using correlation kernel in Gaussian processes.
  - Between-region correlation structure.
  - When $\boldsymbol{\Lambda} = \mathbf{I}_p$, it has a close link to ridge regression and HOLP.
  - When $\boldsymbol{\Lambda} = g(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$, it becomes the Zellner's g prior.
- **Key idea**: Using the marginal posterior mean of β_j as the screening statistics.

How to efficiently compute the posterior mean for ultra-high dimensional case?

For each j, coefficient $\boldsymbol{\beta}$ can be split into two parts β_j and $\boldsymbol{\beta}_{-j} = (\beta_k, k \neq j)^{\mathrm{T}}$. The conditional prior distribution of $\boldsymbol{\beta}_{-j}$ given β_j is given by

$$\boldsymbol{\beta}_{-j} \sim \mathrm{N}(\mathbf{0}_{p-1}, \tau^2 \boldsymbol{\Gamma}_{-j}).$$

where $\boldsymbol{\Gamma}_{-j} = \boldsymbol{\Lambda}_{-j,-j} - \boldsymbol{\Lambda}_{-j,j}\boldsymbol{\Lambda}_{-j,j}^{\mathrm{T}}$ with $\boldsymbol{\Lambda}_{-j,-j} = (\lambda_{k,l})_{k\neq j, l\neq j}$ and $\boldsymbol{\Lambda}_{-j,j} = (\lambda_{k,j}, k \neq j)^{\mathrm{T}}$.

Model Equivalence

Recall the joint model

$$\mathbf{y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n), \text{ and } \quad \boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0}_p, \tau^2\boldsymbol{\Lambda}), \tag{1}$$

The marginal posterior distribution of β_j given $\mathbf{y}$ in model (1) is equivalent to the posterior distribution of β_j given $\mathbf{y}$ from model (2),

$$\mathbf{y} \sim \mathrm{N}\left[\mathbf{x}_j\beta_j, \tau^2\boldsymbol{\Omega}_{-j}^{-1}\right], \text{ and } \qquad \beta_j \sim \mathrm{N}[0, \tau^2], \tag{2}$$

where $\boldsymbol{\Omega}_{-j} = [\mathbf{X}_{-j}\boldsymbol{\Gamma}_{-j}\mathbf{X}_{-j}^{\mathrm{T}} + \theta^2\mathbf{I}_n]^{-1}$ with $\theta^2 = \sigma^2/\tau^2$. It is given by

$$[\beta_j \mid \mathbf{y}, \mathbf{X}, \sigma^2, \tau^2\ ] \sim \mathrm{N}(\nu_j, \kappa_j^2),$$

where

$$\beta_j^{\mathrm{PMS}} = \nu_j = \frac{\mathbf{x}_j^{\mathrm{T}}\boldsymbol{\Omega}_{-j}\mathbf{y}}{\mathbf{x}_j^{\mathrm{T}}\boldsymbol{\Omega}_{-j}\mathbf{x}_j + 1}, \qquad \kappa_j^2 = \frac{\tau^2}{\mathbf{x}_j^{\mathrm{T}}\boldsymbol{\Omega}_{-j}\mathbf{x}_j + 1}$$

Useful Identities

To simplify the computation, we introduce $\mathbf{X}\mathbf{\Lambda} = [\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_p]$. Define $\mathbf{\Omega} = (\mathbf{X}\mathbf{\Lambda}\mathbf{X}^{\mathrm{T}} + \theta^2\mathbf{I}_n)^{-1}$. Then $\mathbf{\Omega}_{-j} = (\mathbf{\Omega}^{-1} - \widetilde{\mathbf{x}}_j\widetilde{\mathbf{x}}_j^{\mathrm{T}})^{-1}$. Furthermore,

$$\mathbf{\Omega}_{-j} = \mathbf{\Omega} + \frac{\mathbf{\Omega}\widetilde{\mathbf{x}}_j\widetilde{\mathbf{x}}_j^{\mathrm{T}}\mathbf{\Omega}}{1 - \widetilde{\mathbf{x}}_j^{\mathrm{T}}\mathbf{\Omega}\widetilde{\mathbf{x}}_j},$$

Then

$$\mathbf{x}_j^{\mathrm{T}}\mathbf{\Omega}_{-j}\mathbf{x}_j = \mathbf{x}_j^{\mathrm{T}}\mathbf{\Omega}\mathbf{x}_j + \frac{\widetilde{\mathbf{x}}_j^{\mathrm{T}}\mathbf{\Omega}\mathbf{x}_j}{1 - \widetilde{\mathbf{x}}_j^{\mathrm{T}}\mathbf{\Omega}\widetilde{\mathbf{x}}_j}(\widetilde{\mathbf{x}}_j^{\mathrm{T}}\mathbf{\Omega}\mathbf{x}_j),$$

$$\mathbf{x}_j^{\mathrm{T}}\mathbf{\Omega}_{-j}\mathbf{y} = \mathbf{x}_j^{\mathrm{T}}\mathbf{\Omega}\mathbf{y} + \frac{\widetilde{\mathbf{x}}_j^{\mathrm{T}}\mathbf{\Omega}\mathbf{x}_j}{1 - \widetilde{\mathbf{x}}_j^{\mathrm{T}}\mathbf{\Omega}\widetilde{\mathbf{x}}_j}(\widetilde{\mathbf{x}}_j^{\mathrm{T}}\mathbf{\Omega}\mathbf{y}).$$

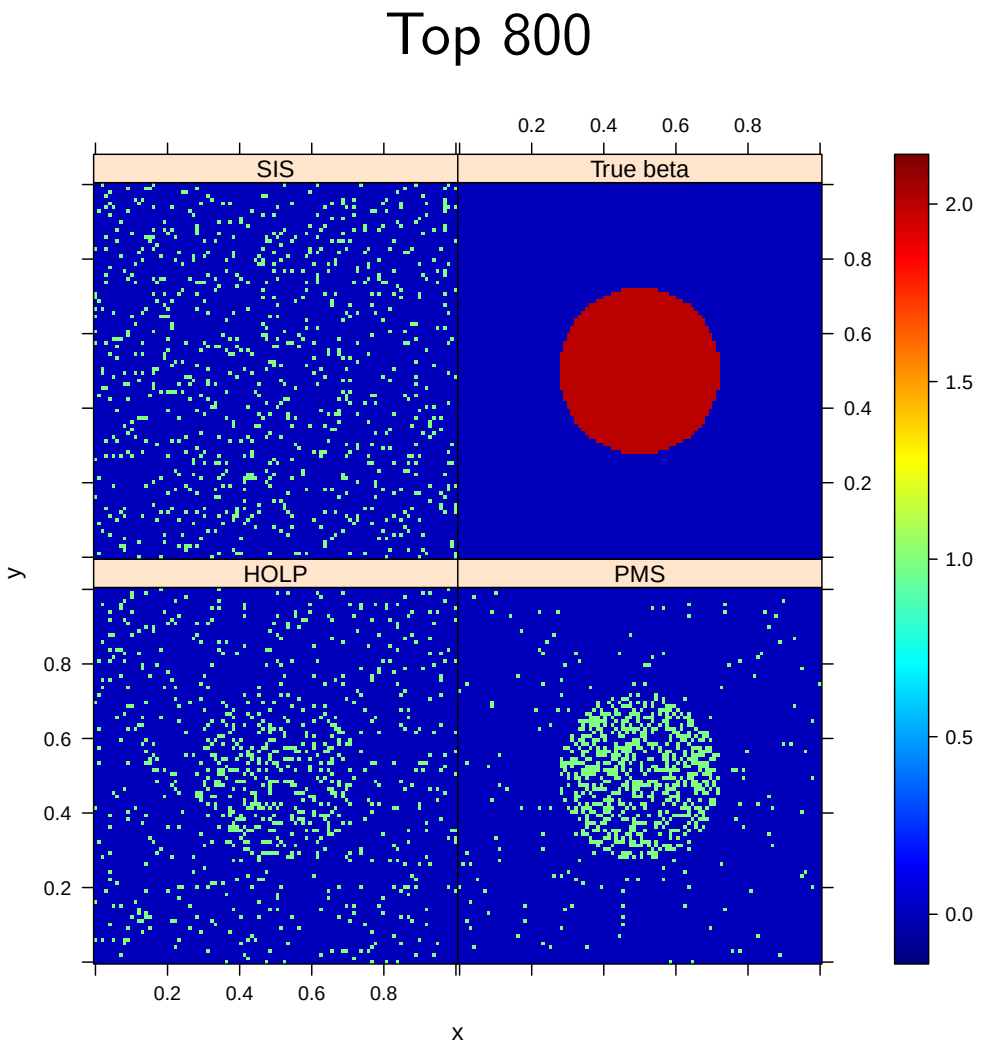
Alogrithm

- Input: $\mathbf{y}$, $\mathbf{X}$, $\mathbf{\Lambda}$, θ^2,
  1. Compute $\widetilde{\mathbf{X}} = \mathbf{X}\mathbf{\Lambda}$. Note that $\widetilde{\mathbf{X}} = [\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_p]$.
  2. Compute $\mathbf{\Omega} = (\widetilde{\mathbf{X}}\mathbf{X}^{\mathrm{T}} + \theta^2 \mathbf{I}_n)^{-1}$.
  3. For $j = 1, \ldots, p$,
     1. Compute $[\mathbf{x}_j^*, \widetilde{\mathbf{x}}_j^*, \mathbf{y}^*] = \mathbf{\Omega}[\mathbf{x}_j, \widetilde{\mathbf{x}}_j, \mathbf{y}]$
     2. Compute $(a_j, b_j, c_j) = \widetilde{\mathbf{x}}_j^{\mathrm{T}}[\mathbf{x}_j^*, \widetilde{\mathbf{x}}_j^*, \mathbf{y}^*]$
     3. Compute $(d_j, e_j) = \mathbf{x}_j^{\mathrm{T}}[\mathbf{x}_j^*, \mathbf{y}^*]$
     4. Compute $f_j = a_j/(1 - b_j)$
     5. Compute $\nu_j = (e_j + f_j c_j)/(d_j + f_j a_j + 1)$
- Output: $\{\nu_j, j = 1, \ldots, p\}$.

Hyperprior Specifications

- $\theta^2 = \sigma^2/\tau^2$ can be very small for a less-informative prior. It can be zero as long as $\mathbf{X}\mathbf{\Lambda}\mathbf{X}^{\mathrm{T}}$ is non-singular (similar to HOLP).
- For the scalar on image regression, we can choose $\mathbf{\Lambda} = [\lambda_{j,j'}]$ with $\lambda_{j,j'} \exp\left(-\rho\|\mathbf{s}_j - \mathbf{s}_{j'}\|^2\right)$ within certain region.
- The choice of γ_n can be based on how many variables that are included in model. It can be proportional to the sample size n.

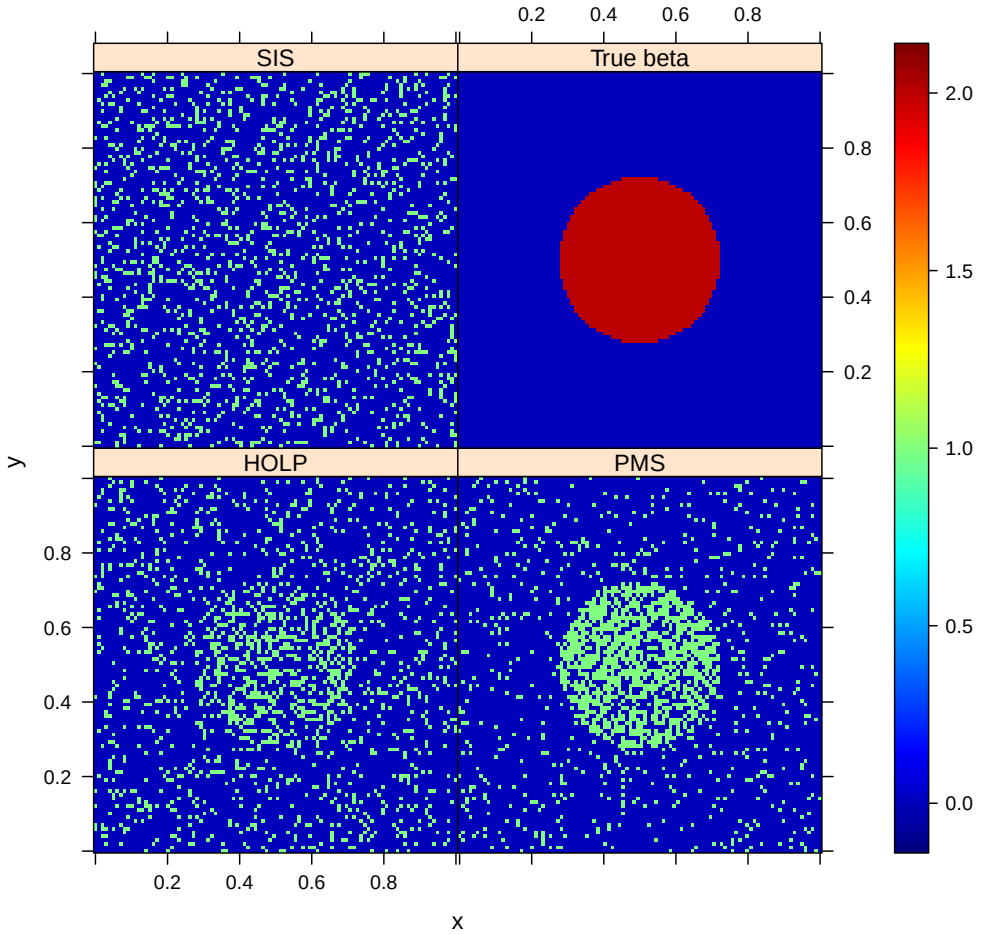
Simulation Results



- True signals (in red): 1,600 (PMS: 649, HOLP: 327, SIS: 126)
- Pixels: 150x150 = 22,500

Simulation Results

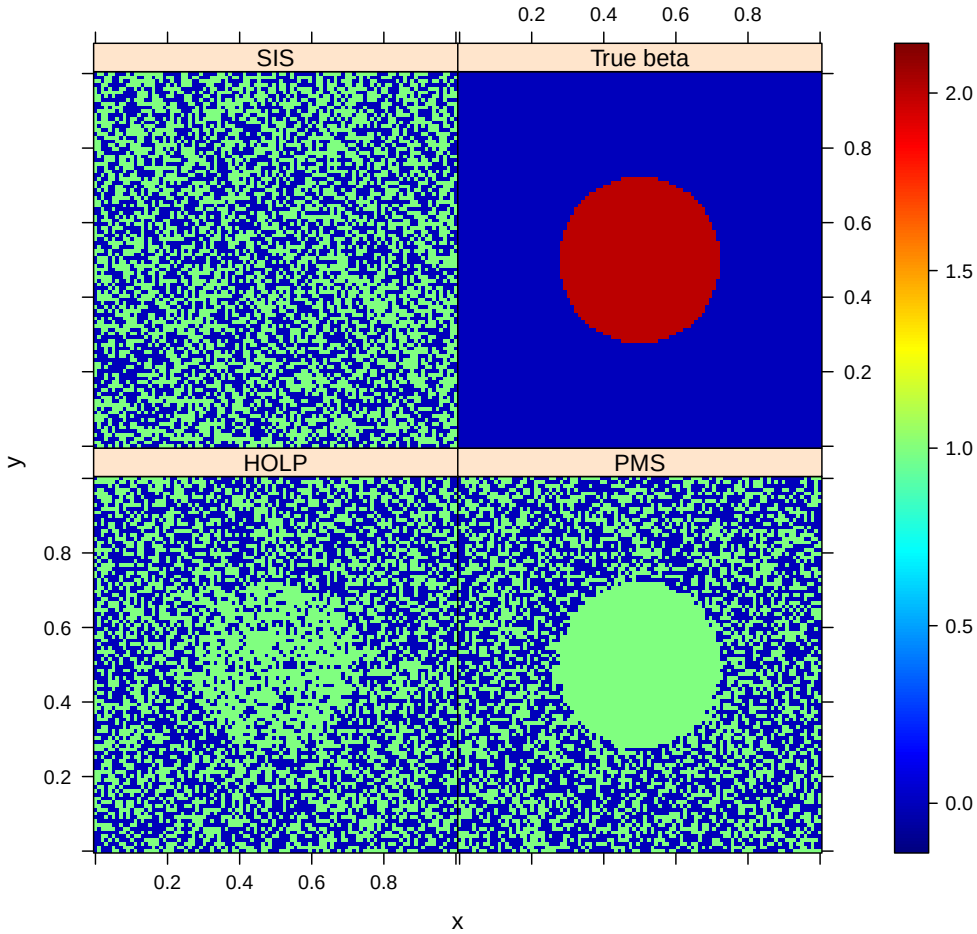
Top 1,600



- True signals (in red): 1,600 (PMS: 948, HOLP: 549, SIS: 254)
- Pixels: 150x150 = 22,500

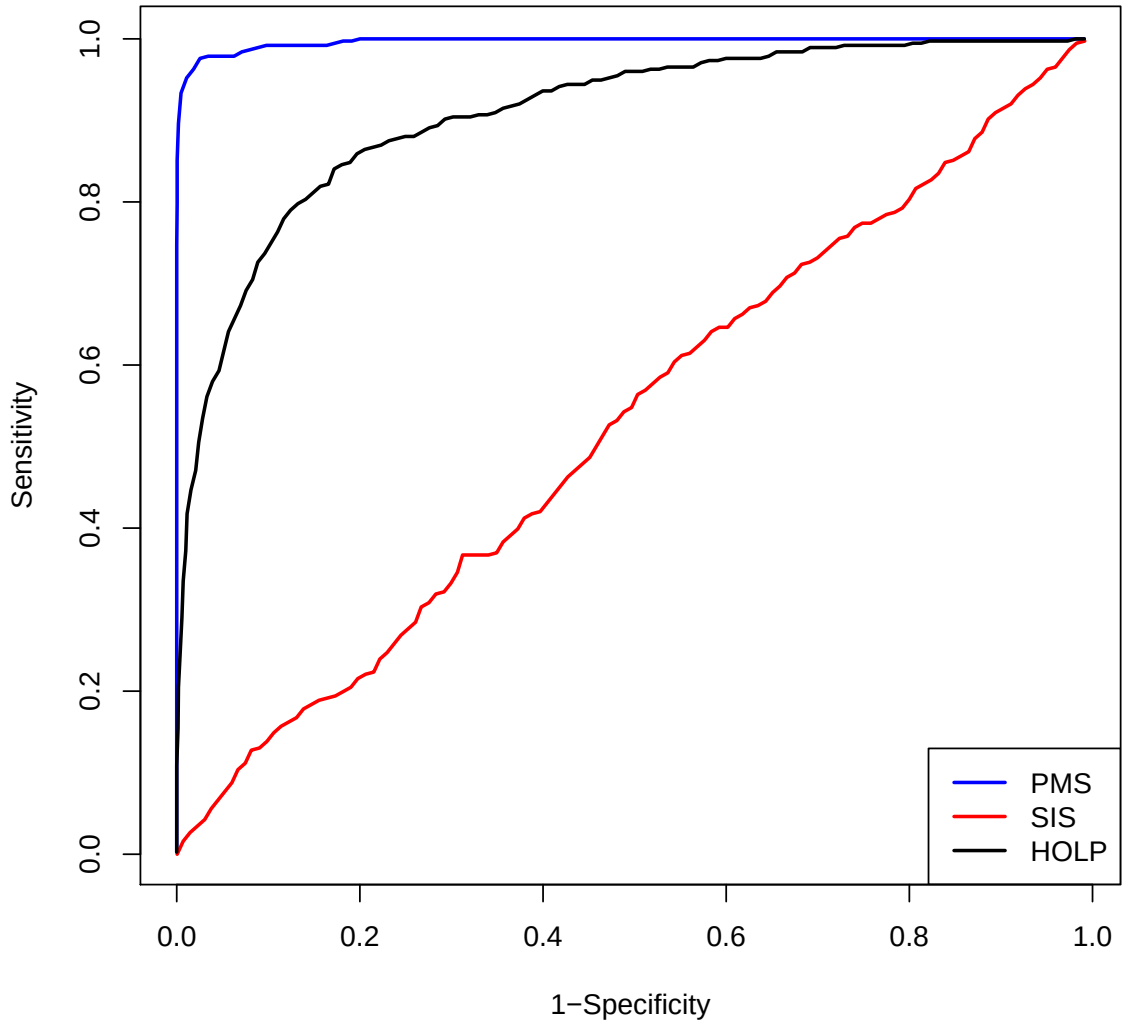
Simulation Results

Top 5,000



- True signals (in red): 1,600 (PMS: 1600, HOLP: 1148, SIS: 746)
- Pixels: 150x150 = 22,500

ROC Curves



Preliminary Analysis of ABIDE Data

- 1112 healthy subjects with fALFF values on 185,405 voxels over 90 regions
- Make prediction on the disease status $y_i \in \{-1, 1\}$.
- Iteratively PMS in remove the half voxels each time. Using cross validation to determine the stopping time
- Selected major regions: PoCG-R and IFGtriang-R, which can achieve 84% accuracy percent prediction accuracy. (I-HOLP: 75% and I-SIS: 65%).
- Computational time: 69 seconds (GPU parallel computing, CUDA 7.5, Macbook Pro, C++, ArrayFire).

Future Work

- Theoretical Justifications
  - Sure Screening
  - Sure Consistency
- More simulation studies for more complex correlation structure
- PMS for generalized linear model (better for data analysis).
- Package for implementing GPU computation in R.