Bayesian Structure Learning for Functional Neuroimaging

Mijung Park; Oluwasanmi Koyejo, Joydeep Ghosh, Russell A.Poldrack, Jonathan W.Pillow

The University of Texas at Austin, Electrical and Computer Engineering, Psychology and Neurobiology

October 10, 2013

Presented by Zhengming Xing

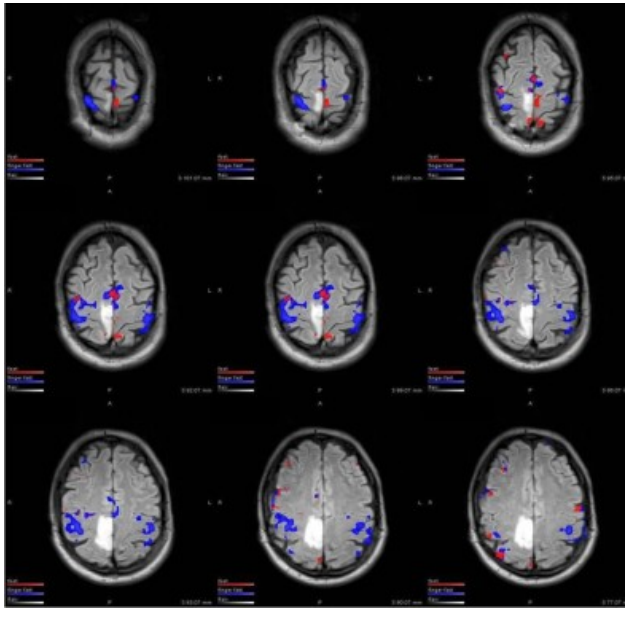
Outline

- Introduction
- Generative model
- Prior covariance design
- BSL for regression
- BSL for classification
- Experiment results

Introduction

fMRI data:

- measurements of blood oxygenation, which is sensitive to the amount of neuronal activity.
- collected while the participant is presented with a stimulus or cognitive task.
- 3-D data, high dimensional, noisy, highly correlated.



Introduction

Objective: Brain reading. Identify the brain region that exhibit a systematic response to the stimulus.
Main idea: Exploit both spatial sparsity and smoothness of the high-demensional and large correlation Brain images.

- Sparsity results from the fact that only small brain regions are activated during a particular task.
- Smoothness results from the fact that the brain regions activated extend across many voxels.

Generative model

Let $\mathbf{x} \in \mathbb{R}^D$ be a feature vector representing the whole brain voxel activation levels and y represents the stimulus.
Denotes $\mathbf{X} = [\mathbf{x}_1^T | ... | \mathbf{x}_N^T] \in \mathbb{R}^{N \times D}$. The linear model and its prior are defined as following.

$$p(\mathbf{y}|\mathbf{x}, \zeta) = \mathcal{N}(f(\mathbf{x}), \sigma^2), f(x) = \omega^T \mathbf{x}, p(\omega|\theta) = \mathcal{N}(0, \mathbf{C})$$

The objective is to parametrize this covariance matrix $\mathbf{C}$ to capture prior smoothness and sparsity assumptions.

Prior covariance structure design

Smoothness

- The weight vector ω is considered smooth if the signal power of $\mathbf{w} = DFT(\omega)$ is concentrated near zero.
- Let $\mathbf{e}_l \in \mathbb{R}^3$ represent the index location in the frequency domain corresponding to the DFT of a three dimensional spatial signal. (i.e. $\mathbf{e}_l = \mathbf{0}$ corresponds to zero frequency). Assume $\mathbf{w} \sim \mathcal{N}(0, \mathbf{G})$ and $\mathbf{G} \in \mathbb{R}^{D \times D}$ is diagonal with entries
$$\mathbf{G}_{l,l} = exp(-\tfrac{1}{2}\mathbf{e}_l^T \mathbf{\Psi} \mathbf{e}_l - \rho)$$
- Let $\mathbf{B} \in \mathbb{R}^{D \times D}$ be the matrix representation of the 3-dimensional discrete Fourier transform. $\mathbf{w} = \mathbf{B}\boldsymbol{\omega} = DFT(\omega)$ and $\boldsymbol{\omega} = \mathbf{B}^T\mathbf{w} = IDFT(\mathbf{w})$.
- Integrate out the prior $\mathbf{w}$, we will have
$$\boldsymbol{\omega} \sim \mathcal{N}(0, \mathbf{B}^T\mathbf{G}\mathbf{B})$$

Prior covariance design

Block spatial sparsity

- Let z_d represent the three dimensional sampling grid. Each cluster is defined by proximity to a central vector $\boldsymbol{\kappa}_c \in \mathbb{R}^3$

  $s(d) = \sum_{c=1}^{C} s_c(d) = \sum_{c=1}^{C} \gamma_c exp(-\frac{1}{2}(\mathbf{z}_d - \boldsymbol{\kappa_c})^T \Omega_c^{-1}(\mathbf{z}_d - \boldsymbol{\kappa}_c))$

To sum up, we have the covariance matrix as following.

$$\mathbf{C} = \mathbf{SB}^T\mathbf{GBS}$$

where $\mathbf{S} = diag(\mathbf{s}^{1/2})$ is the diagonal covariance matrix to impose spatial sparseness. **B** is the discrete Fourier transform matrix. **G** is a diagonal matrix to impose smoothness.

Inference

- Hyperparameter estimation maximize the evidence function.
  $\boldsymbol{\theta} = \{\boldsymbol{\Psi}, \rho, \{\gamma_c, \boldsymbol{\kappa}_c, \boldsymbol{\Omega}_c\}_{c=1}^{C}\}$

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\Theta}) = \int p(\mathbf{y}|\omega, \mathbf{X}, \zeta) p(\omega|\boldsymbol{\theta}) d\omega$$

- Stimulus prediction for held-out images.

$$p(y_*|\mathbf{x}_*, \mathcal{D}, \boldsymbol{\Theta}) = \int p(y_*|\mathbf{x}_*, \omega, \zeta) p(\omega|\boldsymbol{\theta}, \mathcal{D}) d\omega$$

- Point estimate of weight vector ω

$$arg_\omega min[-log(p(\mathbf{y}|\omega, \mathbf{X}, \zeta)) + \tfrac{1}{2}\omega^T \mathbf{C}^{-1} \omega]$$

BSL for regression

Stimuli are modeled as independent Gaussian distributed variables $p(\mathbf{y}|\mathbf{x}, \zeta) = \mathcal{N}(f(\mathbf{x}), \sigma^2)$.

- Hyperparameter estimation. Maximize the evidence function.

  $p(\mathbf{y}|\mathbf{X}, \mathbf{\Theta}) = \mathcal{N}(0, \mathbf{K}_y)$, where $\mathbf{K}_y = \mathbf{K} + \sigma^2\mathbf{I}$ and $\mathbf{K} = \mathbf{X}\mathbf{C}\mathbf{X}^T$

- Stimulus prediction for held-out images.

$$p(f_*|\mathbf{x}_*, \mathcal{D}, \mathbf{\Theta}) = \mathcal{N}(\mu, \Sigma)$$
$$\mu = \mathbf{x}_*^T\mathbf{C}\mathbf{X}^T(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{y},$$
$$\Sigma = \mathbf{x}_*^T\mathbf{C}\mathbf{x}_* + \mathbf{x}_*^T\mathbf{C}\mathbf{X}^T(K + \sigma^2\mathbf{I})^{-1}\mathbf{X}\mathbf{C}\mathbf{x}_*$$

- Point estimate of weight vector $\boldsymbol{\omega}$

  $p(\mathbf{w}|\mathcal{D}, \boldsymbol{\theta}) = \mathcal{N}(\sigma^2\Sigma\mathbf{X}^T\mathbf{y}, \Sigma)$, and $\Sigma = (\mathbf{C}^{-1} + \sigma^2\mathbf{X}^T\mathbf{X})^{-1}$

BSL for Classification

Let J be the total number of classes and y_n^j be an indicator variable with $y_n^j = 1$ if the nth image is from class j, $y_n^j = 0$ otherwise. Linear function response for each class is denoted as f_n^j and multinomial logistic link function is employed,

$$p(y_n^j | \{f_n^j\}_{j=1}^J) = \frac{exp(f_n^j)}{\sum_{l=1}^J exp(f_n^l)}$$

Denotes $\mathbf{f}^j = [f_1^j, ..., f_N^j]^T$ and with $\mathbf{f}^j \sim \mathcal{N}(0, \mathbf{K}^j)$, $\mathbf{K}^j = \mathbf{X}\mathbf{C}^j\mathbf{X}^T$

- Hyperparameter estimation. Laplace approximate of the posterior function, the evidence function is given by

$$p(\mathbf{y}|\theta) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\theta)}{p(\mathbf{f}|\mathbf{y},\theta)} \approx \frac{exp(L(\mathbf{f}))\mathcal{N}(\mathbf{f}|0,\mathbf{K})}{c\mathcal{N}(\mathbf{f}_{map},\mathbf{\Lambda}^{-1})}$$

BSL for Classification

- Use the approximated posterior distribution, obtain the predictive distribution.

$$p(f_*^j|\mathbf{x}_*, \mathcal{D}, \Theta) = \mathcal{N}(\mu, \Sigma)$$
$$\mu = \mathbf{x}_*^T \mathbf{C}^j \mathbf{X}^T (\mathbf{K}^j)^{-1} \mathbf{f}_{map}^j$$
$$\Sigma = diag(\mathbf{k}(\mathbf{x}_*, \mathbf{x}_*)) - \mathbf{Q}_*^T (\mathbf{K} + \mathbf{H}^{-1})^{-1} \mathbf{Q}_*$$

where $\mathbf{k}(\mathbf{x}_*, \mathbf{x}_*)$ is the vector of covariances with the j^{th} element given by $\mathbf{x}_*^T \mathbf{C} \mathbf{x}_*$ and $\mathbf{Q}_*$ is the $JN \times J$ matrix

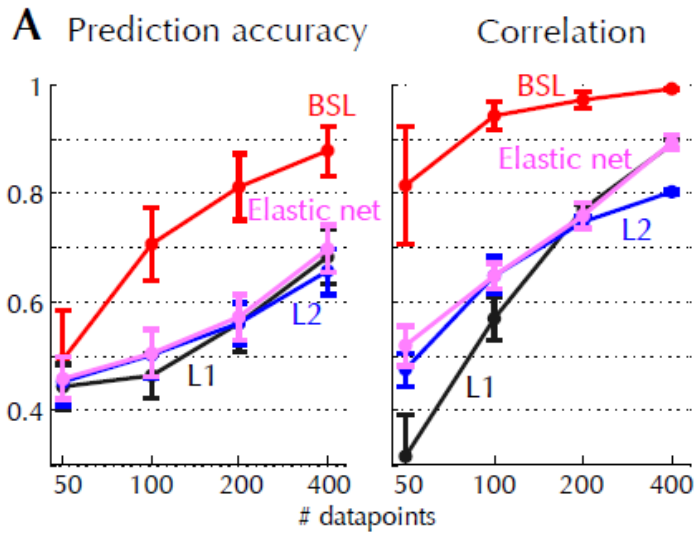
$$\begin{pmatrix} \mathbf{X}^T \mathbf{C}^1 \mathbf{x}_*^T & 0 & \dots & 0 \\ 0 & \mathbf{X}^T \mathbf{C}^2 \mathbf{x}_*^T & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \mathbf{X}^T \mathbf{C}^J \mathbf{x}_*^T \end{pmatrix}$$

- Point estimate of weight vector ω

$$arg_\omega min[\mathbf{y}^T \mathbf{f} + \sum_{n=1}^N log(\sum_{l=1}^J exp(f_n^l)) + \omega^T \mathbf{C}^{-1} \omega]$$

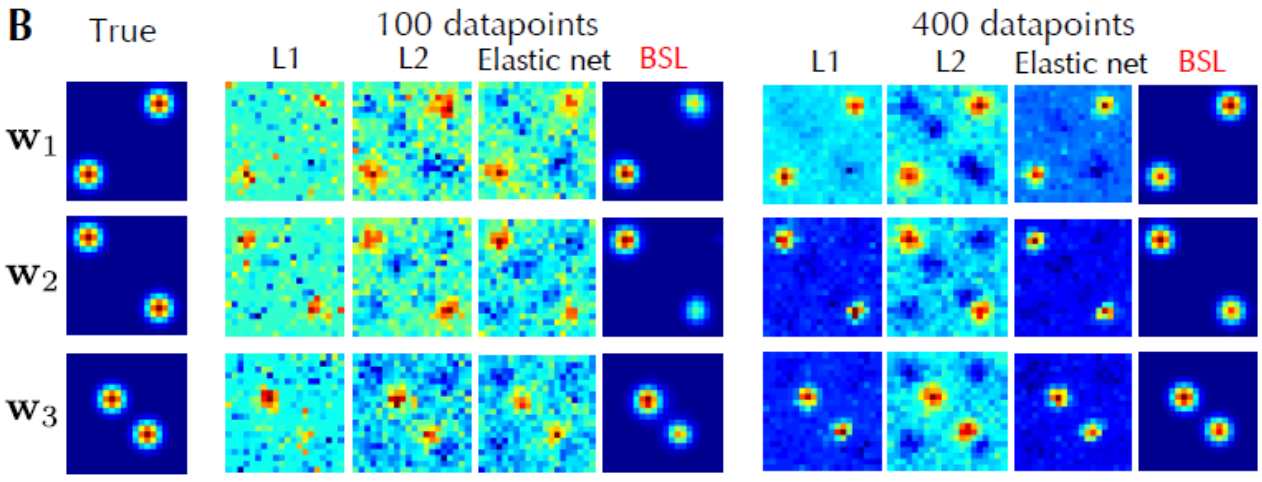
Experiment Results

Test the simulate data in a 3-class classification setting. Generate N random 2-dimensional images where each pixel was generated from $\mathcal{N}(0, 1)$. The dimensionality of each weight vector was 20 by 20, resulting in a D=1200 parameter.



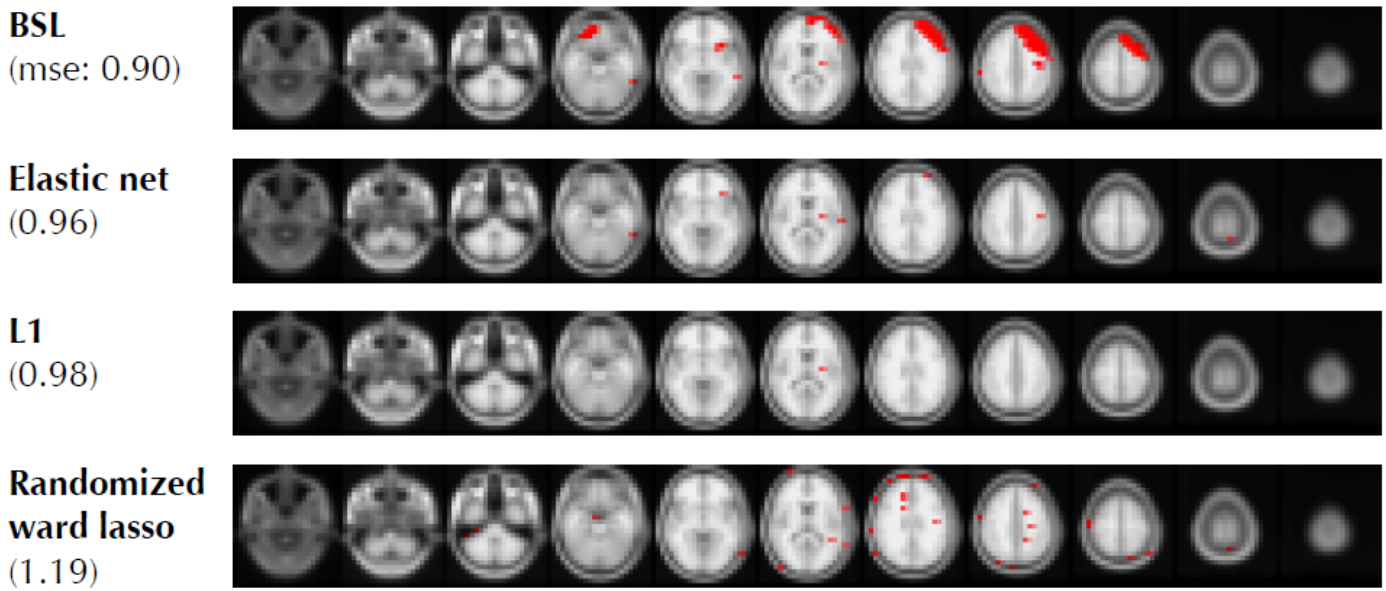
Experiment Results

Test the method in a 3-class classification setting. Generate N random 2-dimensional images where each pixel was generated from $\mathcal{N}(0, 1)$. The dimensionality of each weight vector was 20 by 20, resulting in a D=1200 parameter space.



Experiment Results

fMRI data were collected from 126 participants while the subjects performed a stop-signal task. The fMRI data were down samples to $22 \times 27 \times 22$ voxels and regression with the estimate stop-signal reaction time.



Support (in red) of the estimated weights from each method using real fMRI data.