Smooth Scalar-on-Image Regression via Spatial Bayesian Variable Selection

Jeff Goldsmith, Lei Huang and Ciprian M. Crainiceanu

STATMOS Journal Club Presentation
Presenter: Zhou (Joe) Lan
Introduction

Motivation

- A neuroimaging study relating differences in intracranial white matter microstructure to cognitive disability in multiple sclerosis (MS) patients
- Our goal in this article is to investigate the relationship between the scalar outcomes (you may ask What is it?) and the images (predictor) using regression model.
- scalar outcomes: measures cognitive function (cognitive score, e.g)

Methods Overlook

- scalar-on-image regression models when images are registered multidimensional manifolds.
- a fast and scalable Bayes inferential procedure to estimate the image coefficient
- The central idea is the combination of an Ising prior distribution, which controls a latent binary indicator map, and an intrinsic Gaussian Markov random field, which controls the smoothness of the nonzero coefficients
- The model is fit using a single-site Gibbs sampler, which allows fitting within minutes for hundreds of subjects with predictor images containing thousands of locations
Introduction

Example

Estimated coefficient image from the scalar-on-image regression analysis.

- On the left, nonzero coefficients are shown in blue and the corpus callosum (It connects the left and right cerebral hemispheres and facilitates interhemispheric communication) is shown in transparent red.

- On the right, the coefficient image is overlayed on a subjects scan for anatomical reference.
Methods

Scalar-on-Image Regression Model

Observations

- Each Subject $i = 1, 2, ..., I$
- Observed Data $\{y_i, \omega_i, X_i\}$
  - $y_i$: scalar outcome
  - $\omega_i$: a vector of scalar covariates
  - $X_i$: image predictor measured over a lattice

Model

- $y_i = \omega_i^T \alpha + X_i \cdot \beta + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2_\epsilon)$, $i.i.d$
- Parameters to be estimated
  - $\alpha$: fixed effects vector
  - $\beta$: a collection of regression coefficients defined on the same lattice as the image predictor
    - sparse and organised into spatially contiguous regions
    - smooth in nonzero regions
  - $\gamma$ a latent binary predictor indicating image locations

Jeff Goldsmith, Lei Huang and Ciprian M. Crăiăniţă
Smooth Scalar-on-Image Regression via Spatial Bayesian Variable Selection
Methods
Scalar-on-Image Regression Model

Example
Example of Observation $X_i$
Methods
Parameter Estimation Using Single-Site Gibbs Sampler

Ising Prior

- **γ Definition:**
  \[ \beta_l \begin{cases} 
  0, & \text{if } \gamma_l = 0 \\
  \neq 0, & \text{if } \gamma_l \neq 0 
  \end{cases} \]  
  (1)

- **Expression:**
  \[ p(\gamma) = \phi(a, b) \exp \left( a \cdot \gamma + \sum_l \{ \sum_{l' \in \delta_l} b_{li} I(\gamma_l = \gamma_{l'}) \} \right) \]
  - **a**: Sparsity
  - **b**: Interaction between neighboring points
  - Simplify \((a, b)\) to \((a, b)\) for smoothness assumption
  - How to define \(a\) and \(b\)? (Discuss later)
Methods
Scalar-on-Image Regression Model

Example
Ising Prior/Model on Grid

Example of Ising Prior with different Smooth Parameter
Methods
Parameter Estimation Using Single-Site Gibbs Sampler

Gaussian Markov Random Field (MRF) Prior

- General Expression: $\beta_l | \gamma_l = 1, \beta_{-l}, \gamma_{-l} \sim N(\bar{\beta}_{\delta_l}, \sigma^2_{\beta}/d_l)$
  
  $$\bar{\beta}_{\delta_l} = \frac{\sum_{l'} \beta_{l'} \gamma_{l'}}{d_l}$$; $d_l$: number of neighbouring elements

- Specification expression:

  $$\beta_l | \gamma_l = 1, \beta_{-l}, \gamma_{-l}, \alpha \propto p(y | \beta, \gamma_l = 1, \alpha)p(\beta_l | \gamma_l = 1, \beta_{-l})$$
  
  $$\sim N(\mu_l, \sigma^2_l)$$

- $\sigma^2_l = \left(\frac{1}{\sigma^2_{\epsilon}} X_{.l}^T X_{.l} + \frac{d_l}{\sigma^2_{\beta}}\right)^{-1}$

- $\mu_l = \sigma^2_l \left(\frac{1}{\sigma^2_{\epsilon}} (y - \omega \cdot \alpha - X_{.(-l)} \cdot \beta_{-l})^T X_{.l} + \frac{d_l}{\sigma^2_{\beta}} \bar{\beta}_{\delta_l}\right)$

- $\gamma_l; \beta_l = \beta^* | y, \beta_{-l}, \gamma_{-l} \sim Bernoulli\left(\frac{1}{1+g_l}\right)$

- $g_l = \frac{p(\gamma_l=0, \beta_l=0) | y, \gamma_{-l}, \beta_{-l})}{p(\gamma_l=1, \beta_l=\beta^*) | y, \gamma_{-l}, \beta_{-l})}$
Methods
Parameter Estimation Using Single-Site Gibbs Sampler

**Full Hierarchical Model**

\[
y_i \sim N(\omega_i^T \alpha + X_i \cdot \beta, \sigma^2)
\]
\[
\beta_l \sim N(\bar{\beta}_\delta, \sigma^2_\beta/d_l) \text{(Markov Random Fields)}
\]
\[
\gamma_l \sim Ising[a, b] \text{(Gibbs Fields)}
\] (4)

**Markov Random Fields v.s Gibbs Fields**

- Markov Random Fields: Only specify the **conditional independence**
- Gibbs fields: Have an **implicit probability function for each clique**
Methods
Parameter Estimation Using Single-Site Gibbs Sampler

Markov Random Fields

Figure 3: Markov Random Field example.

(a) Node $x_1$ is conditionally independent to node $x_3$ given $x_2$, $x_4$ and $x_7$. There is no unblocked path between $x_1$ and $x_3$.

(b) There is at least one path from $x_1$ to $x_3$ given $x_2$, $x_4$ and $x_6$; $x_1$ and $x_3$ are not (necessarily) conditionally independent.
Hammersley-Clifford Theorem (Proof Omitted)

It proves that a **Markov Random Field** and **Gibbs Field** are equivalent with regard to the same graph. In other words:

- Given any Markov Random Field, all joint probability distributions that satisfy the conditional independencies can be written as clique potentials over the maximal cliques of the corresponding Gibbs Field.
- Given any Gibbs Field, all of its joint probability distributions satisfy the conditional independence relationships specified by the corresponding Markov Random Field.

Theoretical Properties

\[
p(\gamma_{-i} | \gamma_i) \quad \text{and} \quad p(\beta_{-i} | \gamma_i = 1, \beta_{-i}, \gamma_{-i})
\]
satisfy the conditions of the **Hammersley-Clifford Theorem**

\[
\exists \text{ joint prior distributions } p(\gamma) \quad \text{and} \quad p(\beta | \gamma)
\]

\[
\therefore \text{ We have to consider if } p(\beta | \gamma) \text{ is proper.}
\]
Methods
Parameter Estimation Using Single-Site Gibbs Sampler

**Theorem**

If there exists at least one location $l$ for which $\gamma_l = 0$, then $p(\beta | \gamma)$ is proper.

**Proof.**

Recall Brook’s Lemma

$\Rightarrow f(\beta_L) \propto \exp[-\frac{1}{2} \beta_L^T D^{-1} (I - B) \beta_L] \Rightarrow D^{-1} (I - B)$ is Positive Definite

**Brook’s Lemma (Proof Omitted)**

For any $x, y \in \Gamma^S$ with strictly positive probability:

$$\frac{\pi(x)}{\pi(y)} = \prod_{i=1}^{K} \frac{\pi(x_i | x_1, x_2, \ldots, x_i-1, y_{i+1}, y_{i+2}, \ldots, y_K)}{\pi(y_i | x_1, x_2, \ldots, x_i-1, y_{i+1}, y_{i+2}, \ldots, y_K)}$$
Methods
Parameter Estimation Using Single-Site Gibbs Sampler

Hyper-Parameter Specification:

- Parameters to be Specified: $a, b, \sigma_\epsilon^2, \sigma_\beta^2$
- Methods:
  - Tuning Parameters (If we have $b$ other than $b$, it would be horrible)
  - A Double Metropolis-Hastings Sampler (Faming Liang, 2010 “A double Metropolis-Hastings sampler for spatial models with intractable normalizing constants”):
    - Assign a sampler for hyper-parameter sampling simultaneously with posterior sampler
    - **Parameter Tuning is AVOIDED!**

Single-Site Gibbs Sampling

- Li and Zhang (2010): the sampler is computational expensive for inverting $p_i \times p_i$ matrix
- Smith and Fahrmeir (2007): the computational time needed for matrix calculation increase with **square** of nonzero coefficients
Simulation and Application

Example

Two-Dimensional Coef.

\[ \text{MSE}_1: .986; \quad \text{True Pos: .747} \]
\[ \text{MSE}_0: .015; \quad \text{True Neg: .986} \]

Three-Dimensional Coef.

\[ \text{MSE}_1: 1.96; \quad \text{True Pos: .524} \]
\[ \text{MSE}_0: .011; \quad \text{True Neg: .977} \]
### Example

<table>
<thead>
<tr>
<th>( \sigma_x^2 / \sigma_e^2 )</th>
<th>( I = 100 )</th>
<th>( I = 500 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td><strong>GMRF – 2D</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True Pos.</td>
<td>0.680</td>
<td>0.627</td>
</tr>
<tr>
<td>True Neg.</td>
<td>0.990</td>
<td>0.969</td>
</tr>
<tr>
<td><strong>EX. – 2D</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True Pos.</td>
<td>0.535</td>
<td>0.428</td>
</tr>
<tr>
<td>True Neg.</td>
<td>0.955</td>
<td>0.968</td>
</tr>
<tr>
<td><strong>GMRF – 3D</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True Pos.</td>
<td>0.589</td>
<td>0.499</td>
</tr>
<tr>
<td>True Neg.</td>
<td>0.992</td>
<td>0.987</td>
</tr>
<tr>
<td><strong>EX. – 3D</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True Pos.</td>
<td>0.558</td>
<td>0.441</td>
</tr>
<tr>
<td>True Neg.</td>
<td>0.956</td>
<td>0.958</td>
</tr>
</tbody>
</table>
Limitations

- It is possible to overfit a model at the expense of prediction on future data.
- Not an uncommon issue when the number of parameters vastly exceeds the number of subjects.
- Assumes that the coefficient image is sparse, and it is currently unclear what ramifications the sparsity assumption will have when it is inaccurate.

Tuning parameters

Further Work:

- Apply network information for modelling spatially relationship.
- Considering Bayesian sampling methodologies.
- Apply location-specific hyperparameter specification.
A. Author.

*Handbook of Everything.*

S. Someone.

On this and that.

*Journal of This and That, 2(1):50–100, 2000.*