Bayesian Modular & Multiscale Regression

Michele Peruzzi
with David Dunson

ISBA World Meeting, Edinburgh, 29 June 2018

Università Bocconi and Duke University

Topics

- Multiscale points of view
- BM&Ms regression model
- Gender classification using task fMRI data

Introduction

Introduction

Data can be measured at different resolutions:

- **Coarse scales/Low resolutions** are simpler, more manageable, more interpretable
- **Fine scales/High resolutions** are high-dimensional, unlabeled, but possibly more informative?

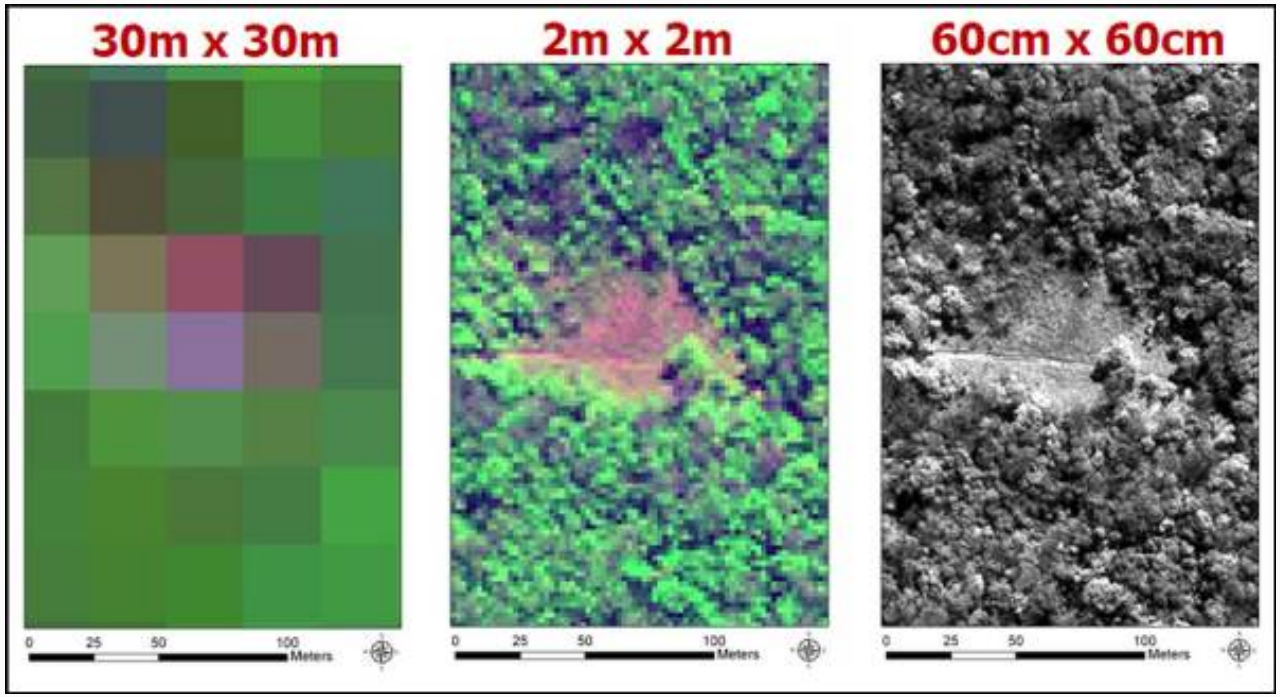
Examples

- EEG, fMRI, DTI, motion sensors, imaging
- Item categorizations, aggregate levels, time-series frequency

Dilemma:

- wish to use low-res but afraid not enough
- wish to use high-res but afraid of false discoveries

Example


30m x 30m
2m x 2m
60cm x 60cm
0 25 50 100 Meters
0 25 50 100 Meters
0 25 50 100 Meters

Introduction

Multiple *scales* (or *resolutions*) coexist:

- second, minute, hour, day, month, year
- individual, household, neighborhood, city, region, country
- voxel, region, “macroregion”, lobe, hemisphere

$\rightarrow$ choosing 1 scale before analysis hides uncertainty, reproducibility? validity?

$\rightarrow$ Haar wavelets’ shrinkage: simple multiscale interpretation of (processed, single-scale) data. Flexibility? Fully Bayesian inference in high-dimensions? Preprocessing?

$\rightarrow$ Sensor readouts tend to be spatially or temporally correlated, especially at higher resolutions

Introduction /2

Two *multiscale points of view*:

1. multiple scales determined a priori for the same data
   - examples
     - hierarchical/nested cortical parcellations
     - nested aggregate product categories
     - time series: second-minute-hour-day-month-year

   GOAL: use unlabeled fine scales as refinements on labeled coarse scales
2. no meaningful coarse scales available, but still interested in *multiscale interpretation* of results

   GOAL: flexible multiscale modeling

BM&Ms: Bayesian Modular & Multiscale regression

BM&Ms

Problem setup

- data available at K resolutions $X_1, \ldots, X_K$
  where $X_{i,j}$ is a $p_j \times 1$ vector of data at resolution j for subj i
  - $p_1 < \cdots < p_K$
  - we assume $X_j = X_{j+1} L_j$, i.e. low resolutions are coarsening of high resolutions
- y_i is a scalar

  GOAL: contribution of scale j to regression function,
  quantify uncertainty, improve prediction

  Overall model

$$\begin{aligned} y &= X_K(\mathcal{L}_1\theta_1 + \cdots + \mathcal{L}_{K-1}\theta_{K-1} + \theta_K) + \varepsilon \\ &= X_1\theta_1 + \cdots + X_K\theta_K + \varepsilon \\ &= X_K\beta + \varepsilon \end{aligned}$$

- $\theta_1, \ldots, \theta_K$ not identifiable by construction

BM&Ms /1

$$y = X_1\theta_1 + \cdots + X_K\theta_K + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2 I_n)$$

$\theta_1, \ldots, \theta_K$ not identifiable $\rightarrow$ use **modularization**:

- break dependence, "cut"
- split the problem into smaller subproblems

for multiscale linear regression:

1. **Module 1**: $y = X_1\theta_1 + \varepsilon$, prior $\theta_1 \sim p(\theta_1)$
   sample $\tilde{\theta}_1 \sim \pi_1(\theta_1|y, X_1)$, get $e_1 = y - X_1\tilde{\theta}_1$
2. **Module 2**: $y - X_1\tilde{\theta}_1 = e_1 = X_2\theta_2 + \varepsilon$, prior $\theta_2 \sim p(\theta_2)$
   sample $\tilde{\theta}_2 \sim \pi_2(\theta_2|e_1, X_2)$, get $e_2 = e_1 - X_2\tilde{\theta}_2$
   ...
K. sample $\tilde{\theta}_K \sim \pi_K(\theta_K|e_{K-1}, X_K)$

Modular posterior = collection of modules' posteriors

$$p_M(\theta_1, \ldots, \theta_K|y, X_{1:K}) = \pi_1(\theta_1|y, X_1) \cdots \pi_K(\theta_K|e_{K-1}, X_K)$$

BM&Ms /2

Modular posterior from 2 linear regression modules:
$p_M(\theta_1, \theta_2 | y, X_{1:2}, \sigma_1^2, \sigma_2^2) = \pi_1(\theta_1 | y, X_1, \sigma_1^2)\pi_2(\theta_2 | e_1, X_2, \sigma_2^2)$ is Normal with mean $\mu_{1:2}$ and covariance matrix $\Sigma_{1:2}$:

$$\mu_{1:2} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} \mu_{\beta_1} \\ \mu_{\beta_2} - Q_1 \mu_{\beta_1} \end{bmatrix} \qquad \Sigma_{1:2} = \begin{bmatrix} \sigma_1^2 \Sigma_1 & -\sigma_1^2 \Sigma_1 Q_1' \\ -\sigma_1^2 Q_1 \Sigma_1 & \sigma_2^2 \Sigma_2 + \sigma_1^2 Q_1 \Sigma_1 Q_1' \end{bmatrix}$$

with $Q_1 = \Sigma_2 X_2' X_1$, and μ_{β_j}, Σ_j $j \in \{1, 2\}$ are the posterior means and variances we would obtain from single resolution models of the form

$$y = X_j \beta_j + \varepsilon_j,$$

and Normal priors for β_j

BM&Ms /3

In general

- we may use $K > 2$ modules
- modular posterior is not fully Bayesian
- corresponds to data-dependent prior
- θ_j prioritized over θ_{j+1}
- if the posterior from each component module is easy to sample from (e.g. conjugacy), then it is easy to sample from the modular posterior.

In large samples, BM&Ms are equivalent to step-wise least squares on residuals: The large sample modular posterior mean is

$$\bar{\mu}_{1:2} \approx \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 - L_1 \hat{\beta}_1 \end{bmatrix}$$

where $\hat{\beta}_j = (X_j' X_j)^{-1} X_j' y$ and L_1 stretches $\hat{\beta}_1$ to the dimension of $\hat{\beta}_2$.

BM&Ms in action!

Gender classification

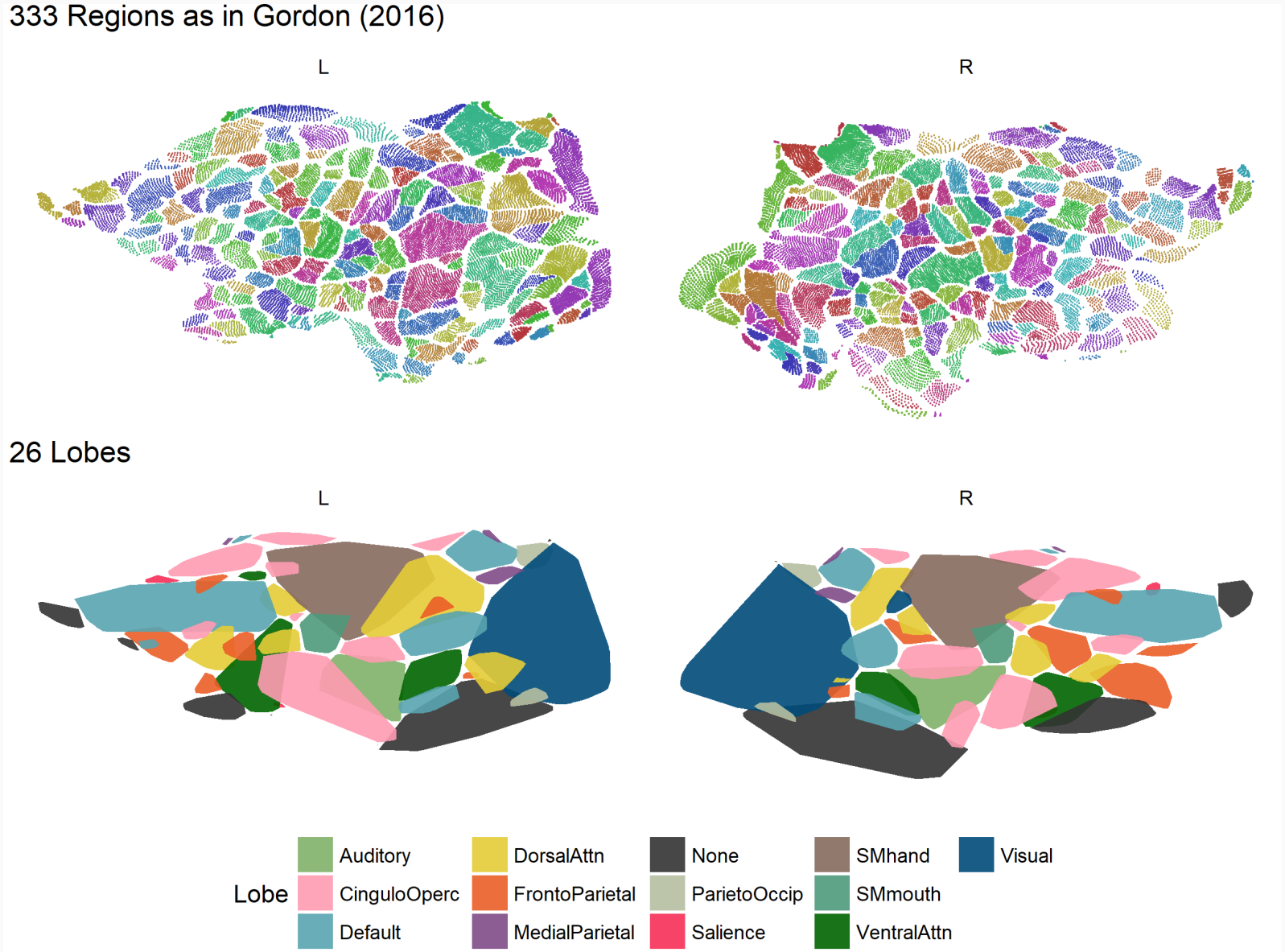
We consider the task of gender classification using tfMRI data (Human Connectome Project data)

- observe subjects while performing a task
- record brain activation
- use X_i = brain activation to predict y_i = gender

We use the Gordon (2016) parcellation with 2 corresponding modules:

- 26 labeled lobes, each associated to some function/interpretation; 333 regions, each nested into a lobe
- $n = 100$ subjects
- use spike-and-slab priors

Gordon (2016) parcellation

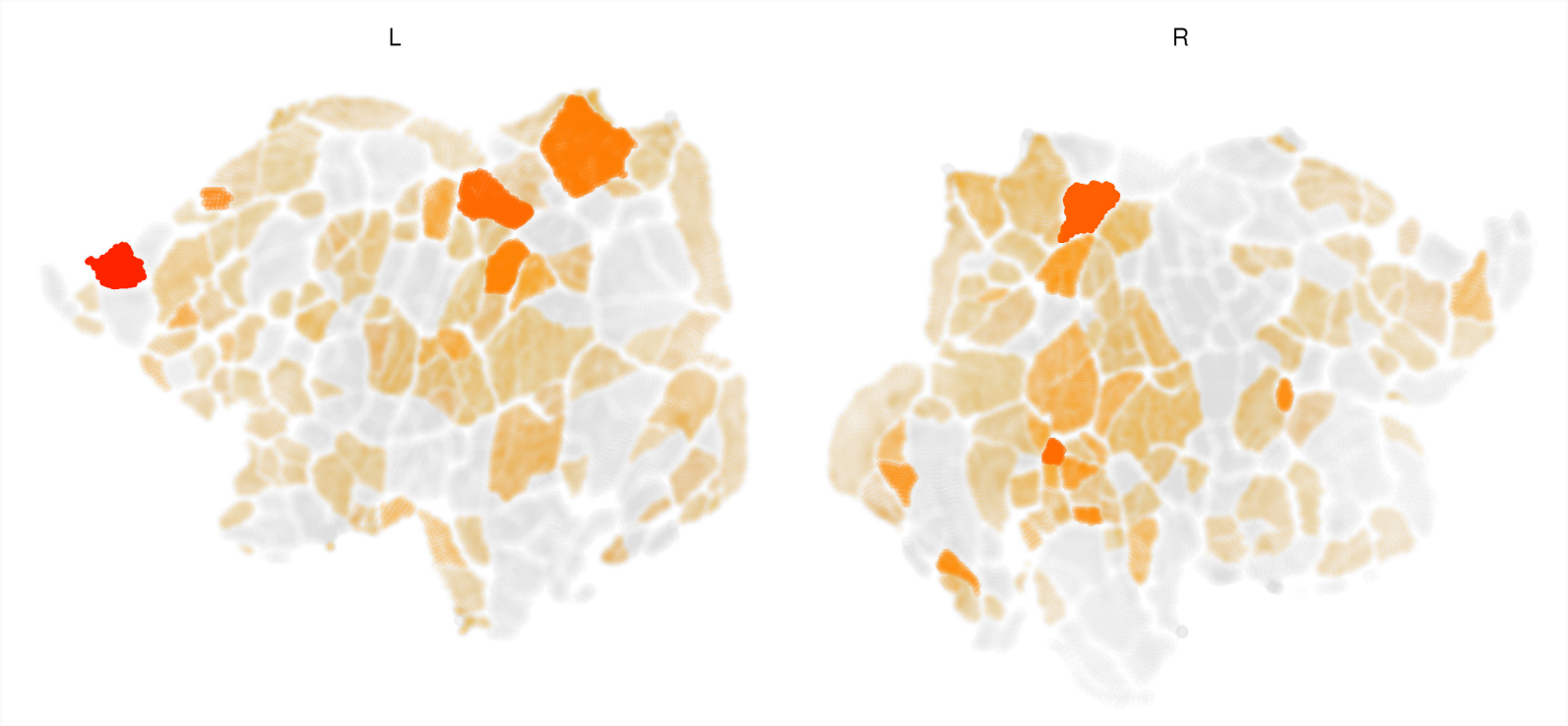


Gender classification

Lasso-selected regions on same data:

Gender classification

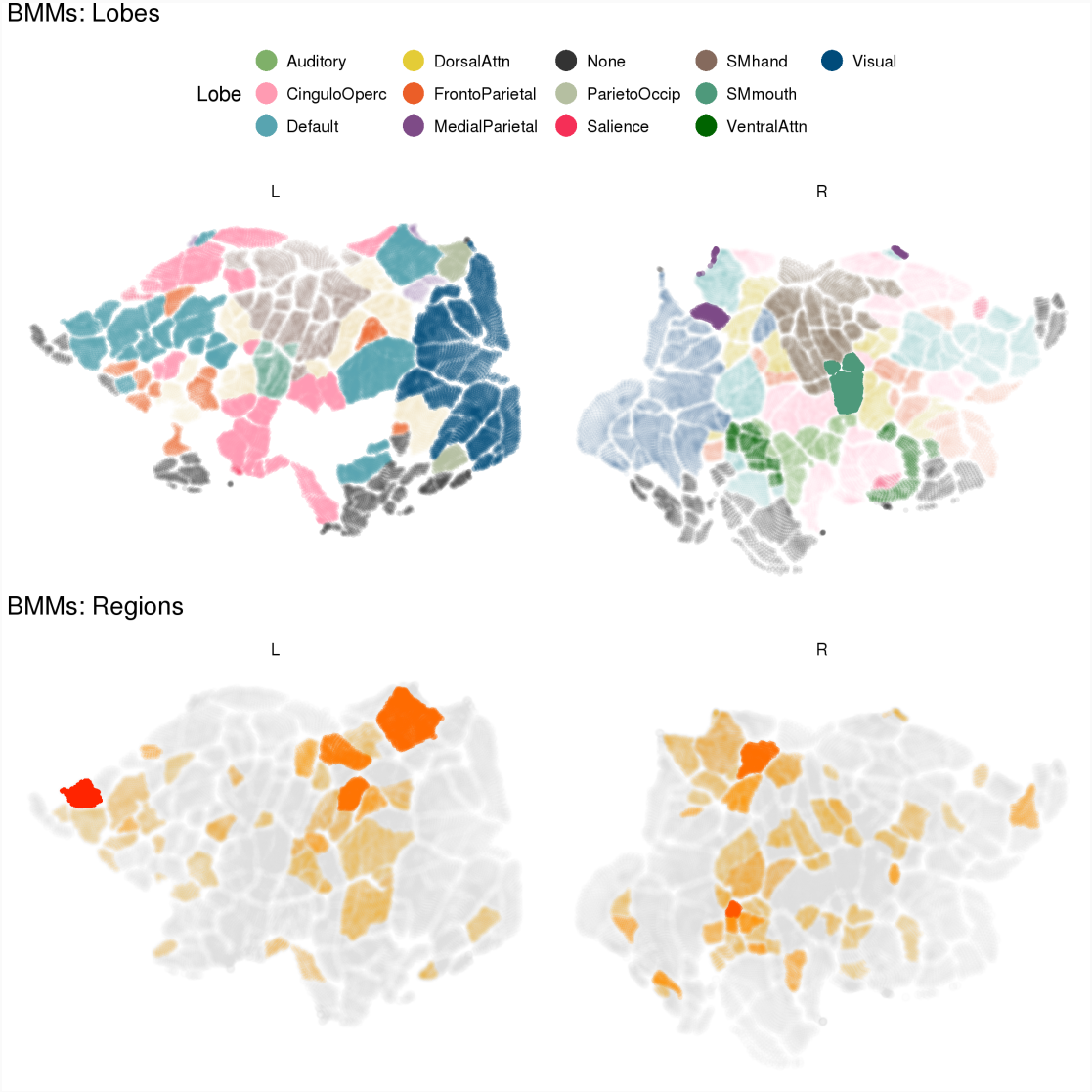
Single-scale spike-and-slab:



(thresholded) posterior selection probability of regions

Gender classification

BM&Ms with spike-slab priors:



(thresholded) posterior selection probability of regions

BM&Ms using Gordon (2016)

Information on lobes reduces need to use many regions

- most lobes are selected, indicating diffuse low-resolution information
- after using lobes, BM&Ms select $\approx$ 50% the number of regions of the Bayesian variable-selection model
- BM&Ms achieve essentially the same performance as Bayesian variable selection

Ok, but...

We have used Gordon (2016) because

- interpretable
- manageable dimension
- nested parcellation

However

- no spatial consideration
- choice of parcellation is *still* ad-hoc
- what about the *actual voxels*?

We now take the second perspective: $\rightarrow$ multiscale interpretation of high-resolution data

Multiscale scalar-on-image regression

We used BM&Ms on data at pre-specified resolutions X_1 (lobes) and X_2 (regions). We now consider the highest-resolution data $\mathbf{X}$:

- for subject i, $\mathbf{X}_i$ is a 341×896 image
- effective dimension of B: $p = 59063$
- single-scale model: $y_i = vec(X_i) \cdot vec(B) + \varepsilon$
- B is the coefficient *matrix* that we want to estimate

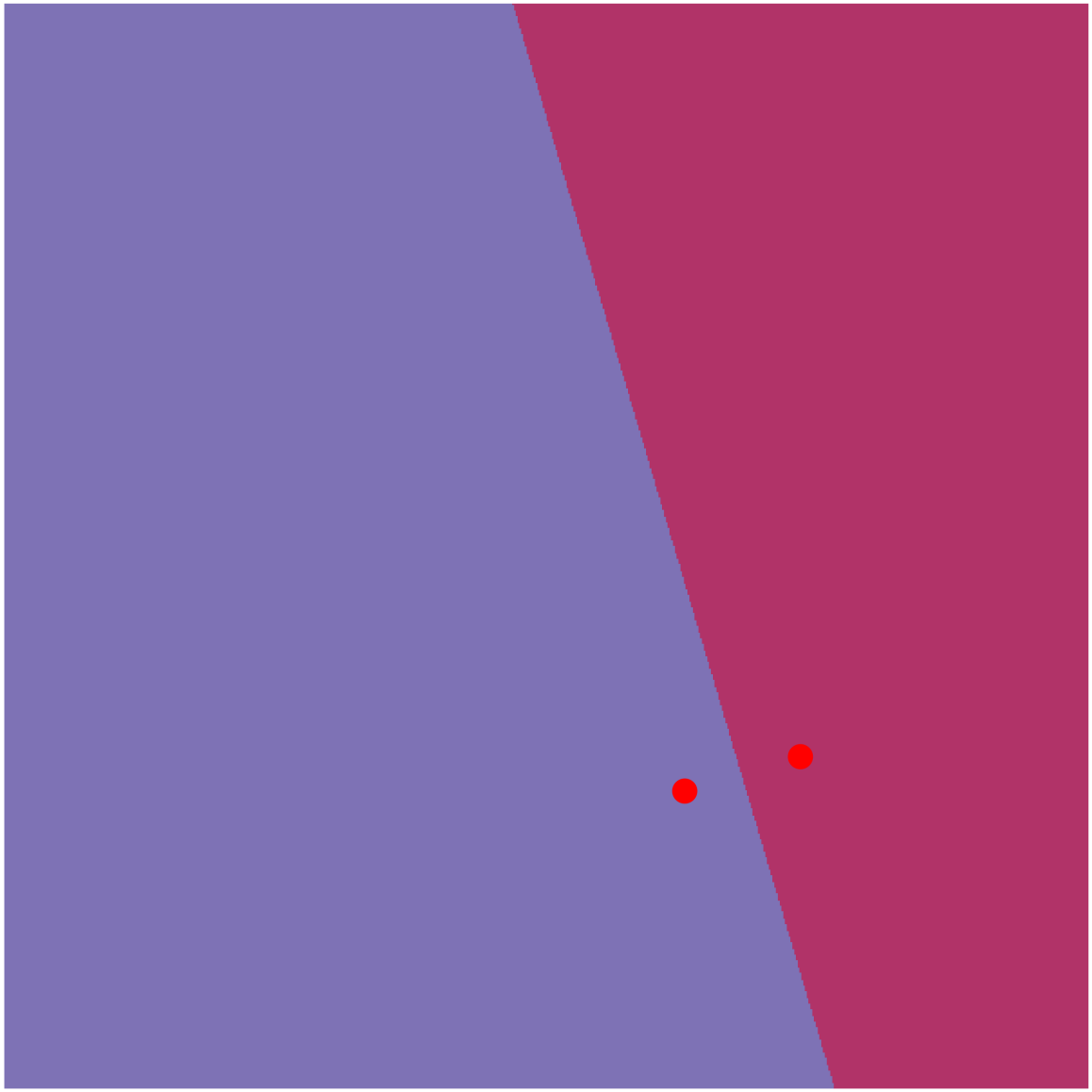
Overall model:

$$y = vec(\mathbf{X}) \cdot (\mathcal{L}_1\theta_1 + \cdots + \mathcal{L}_{K-1}\theta_{K-1} + \theta_K) + \varepsilon$$

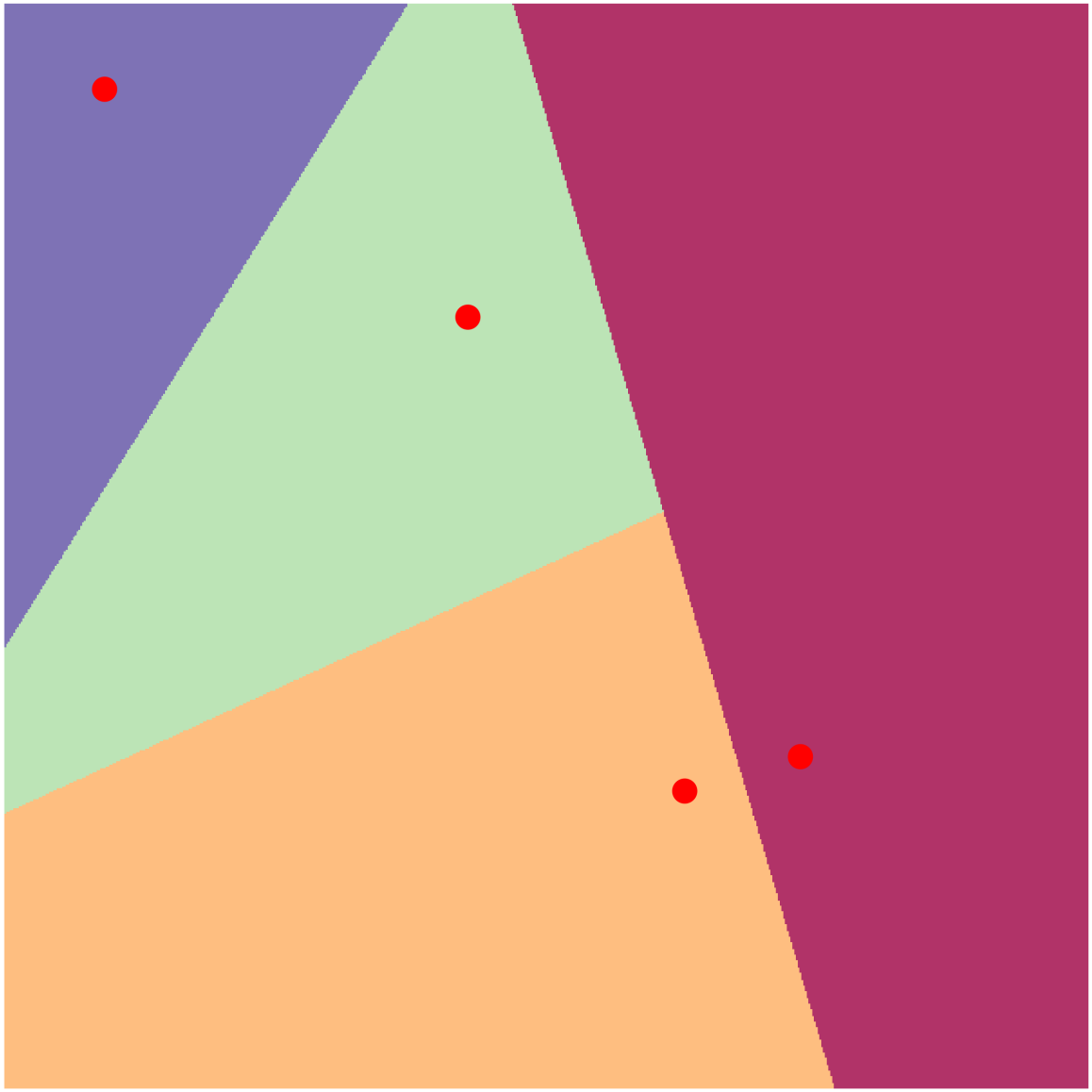
where we now let $\mathcal{L}_j$ as parameters to be estimated.

- $\mathcal{L}_j$ correspond to "groupings of adjacent locations"
- $vec(X) \cdot \mathcal{L}_j$ is low-dimensional
- we let $\mathcal{L}_j$ correspond to a Voronoi tessellation mask

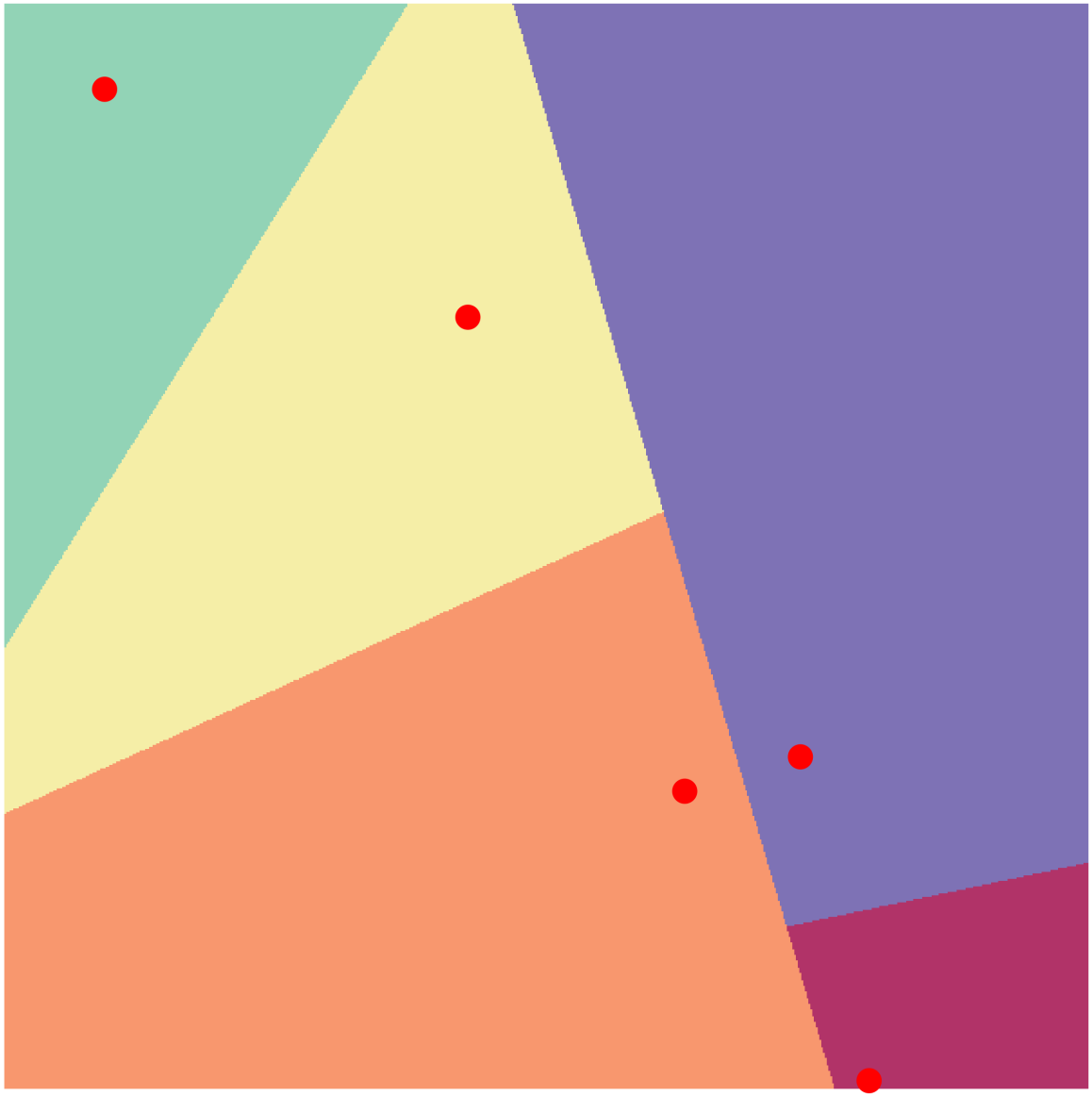
Multiscale Voronoi tessellation



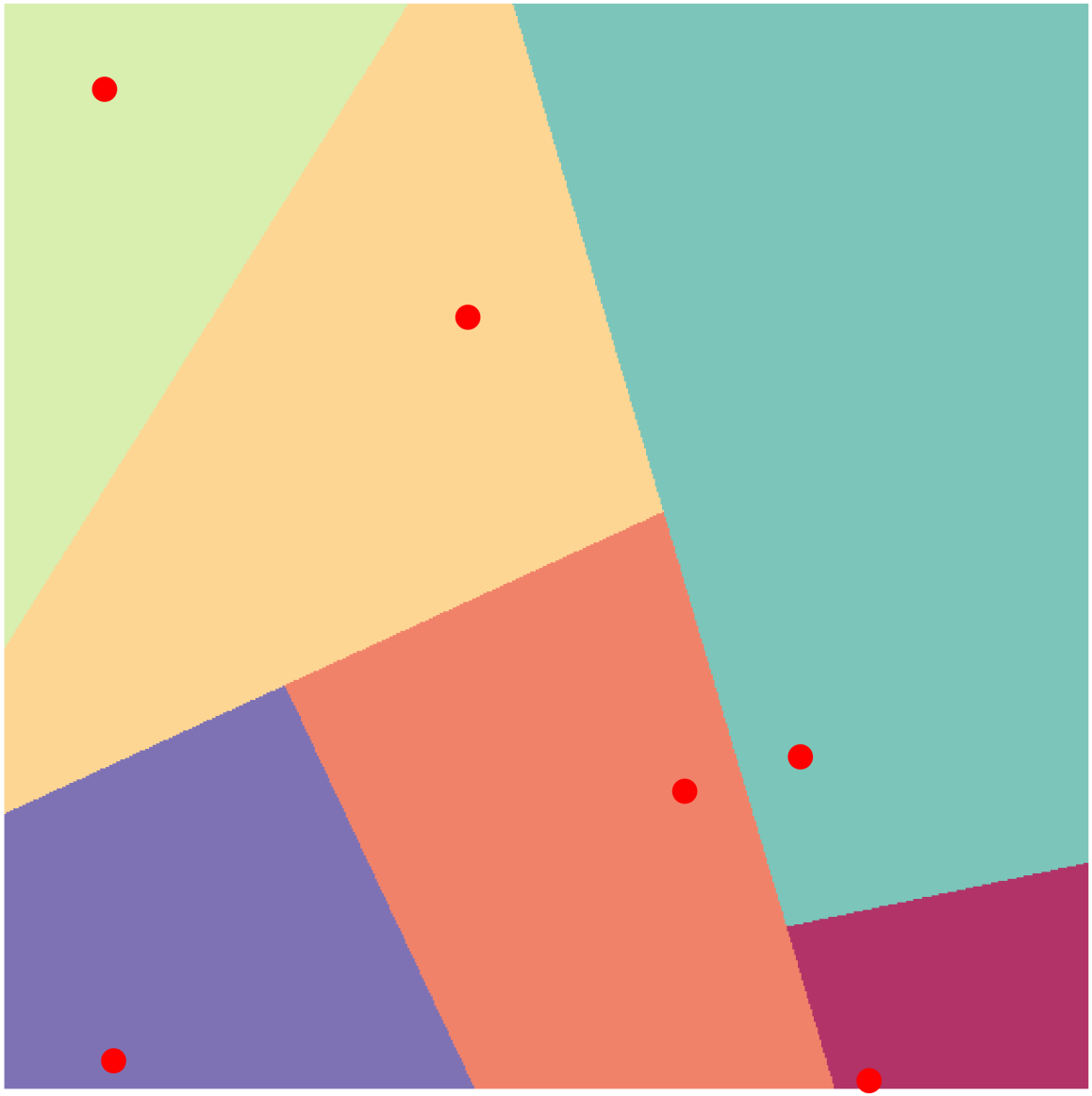
Multiscale Voronoi tessellation



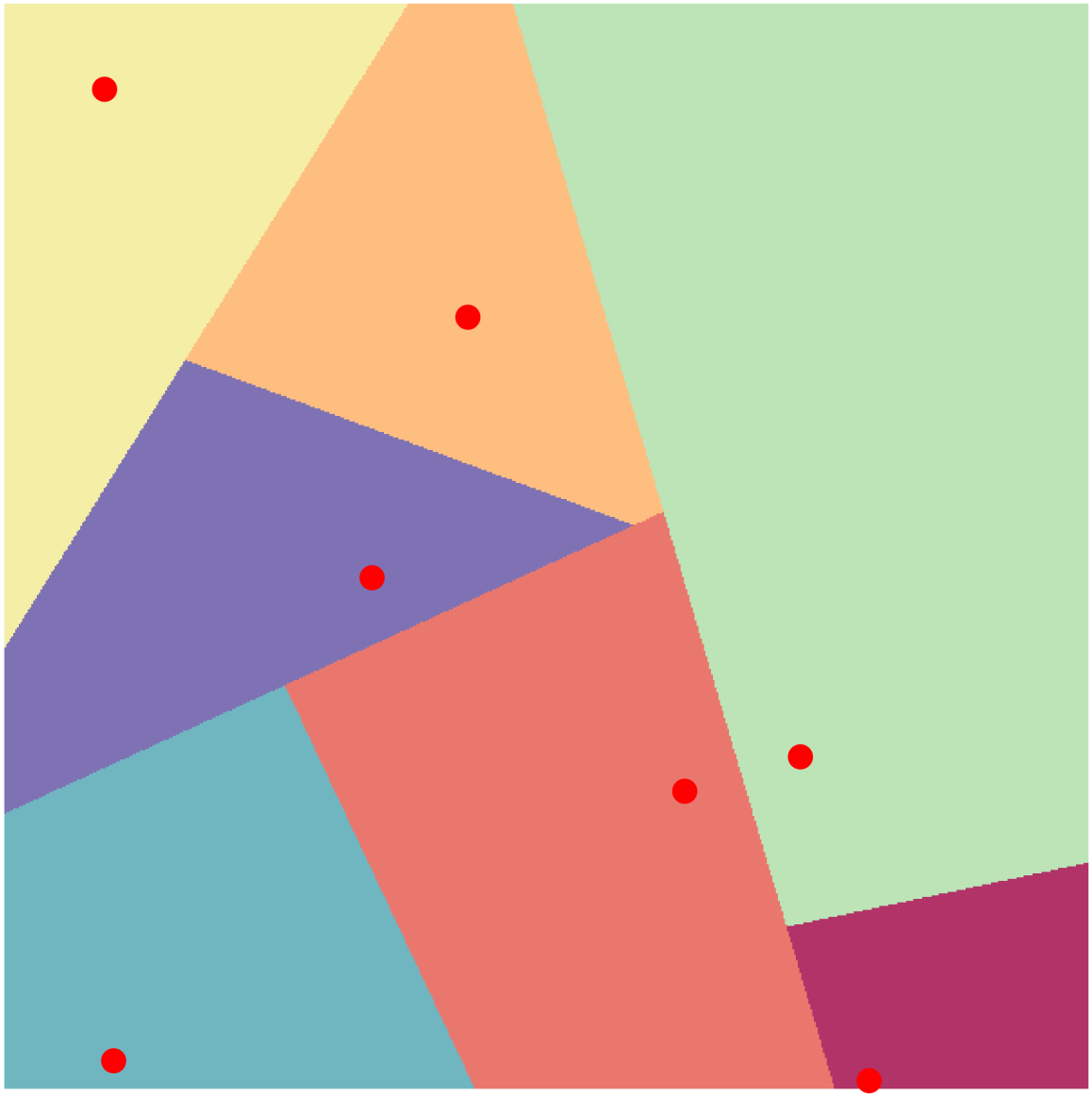
Multiscale Voronoi tessellation



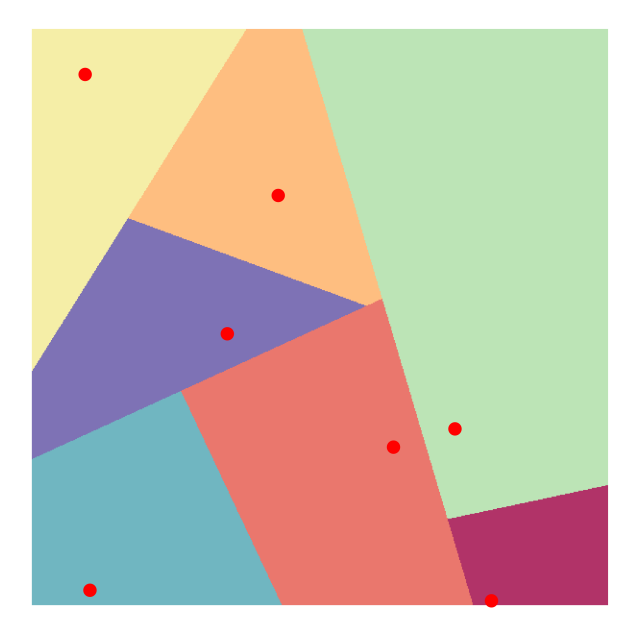
Multiscale Voronoi tessellation



Multiscale Voronoi tessellation



Multiscale scalar-on-image regression



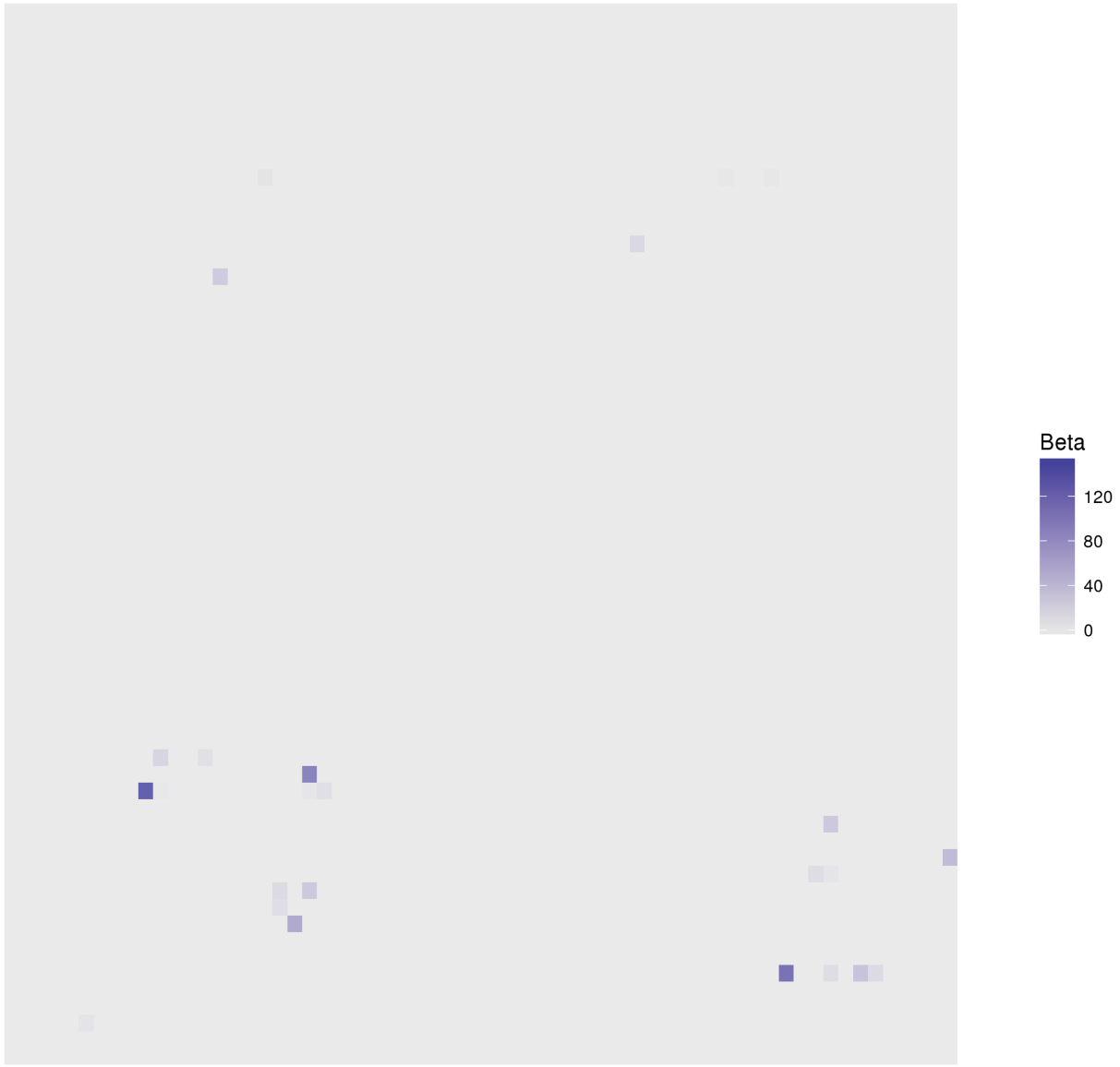
- “nested” Voronoi can be described by centers (red dots)
- their location and number can be explored via MCMC
- each Voronoi region is associated to a single regression coefficient
- interpretation in 1D: centers = splits and the coefficient vector would be modeled as a step function
- we fix the number of BM&Ms scales at $K = 3$

Simulated data

True B matrix in $y_i = vec(\mathbf{X}_i) \cdot vec(B) + \varepsilon$

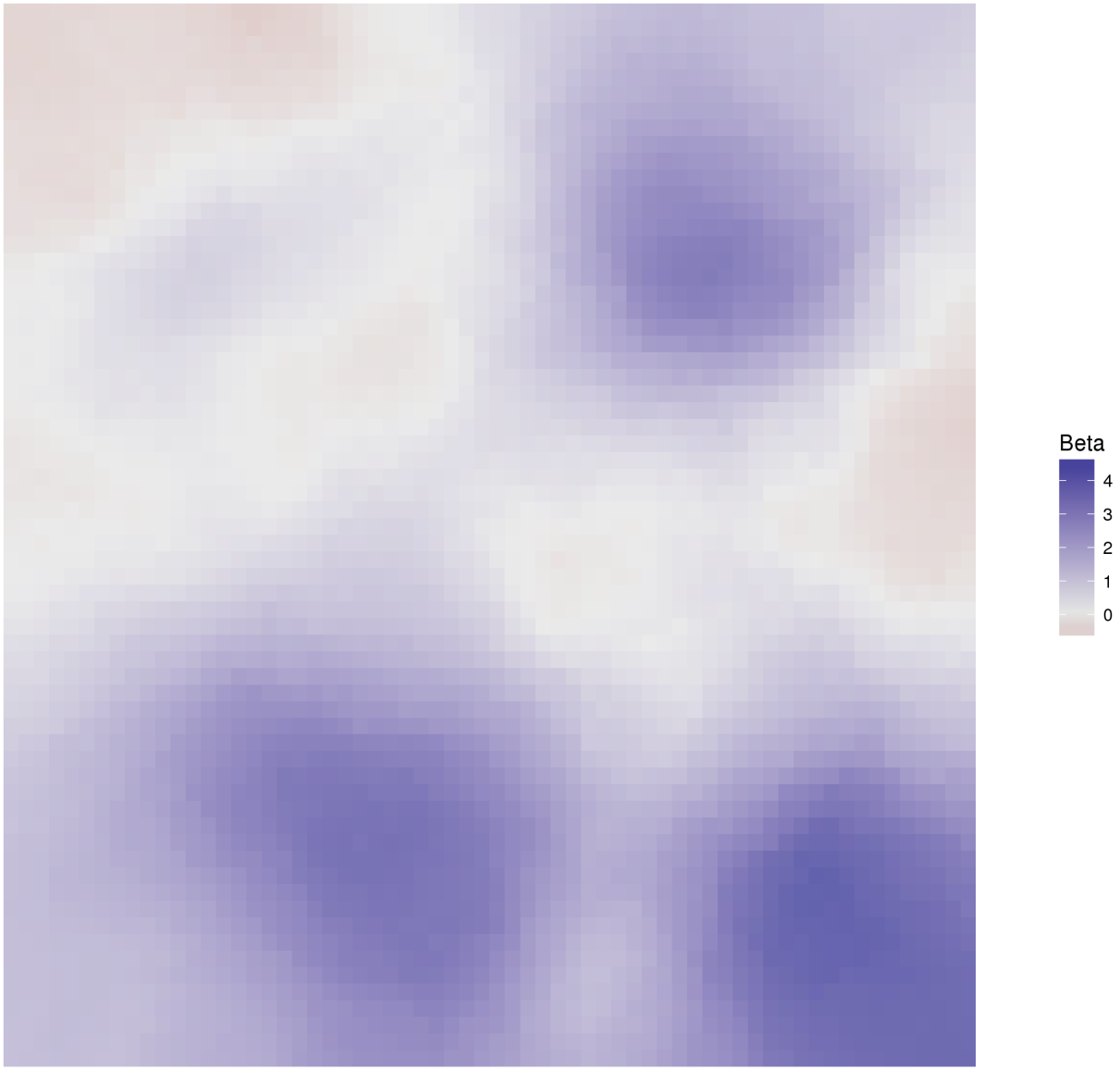
Simulated data

Estimated *B* matrix – Lasso



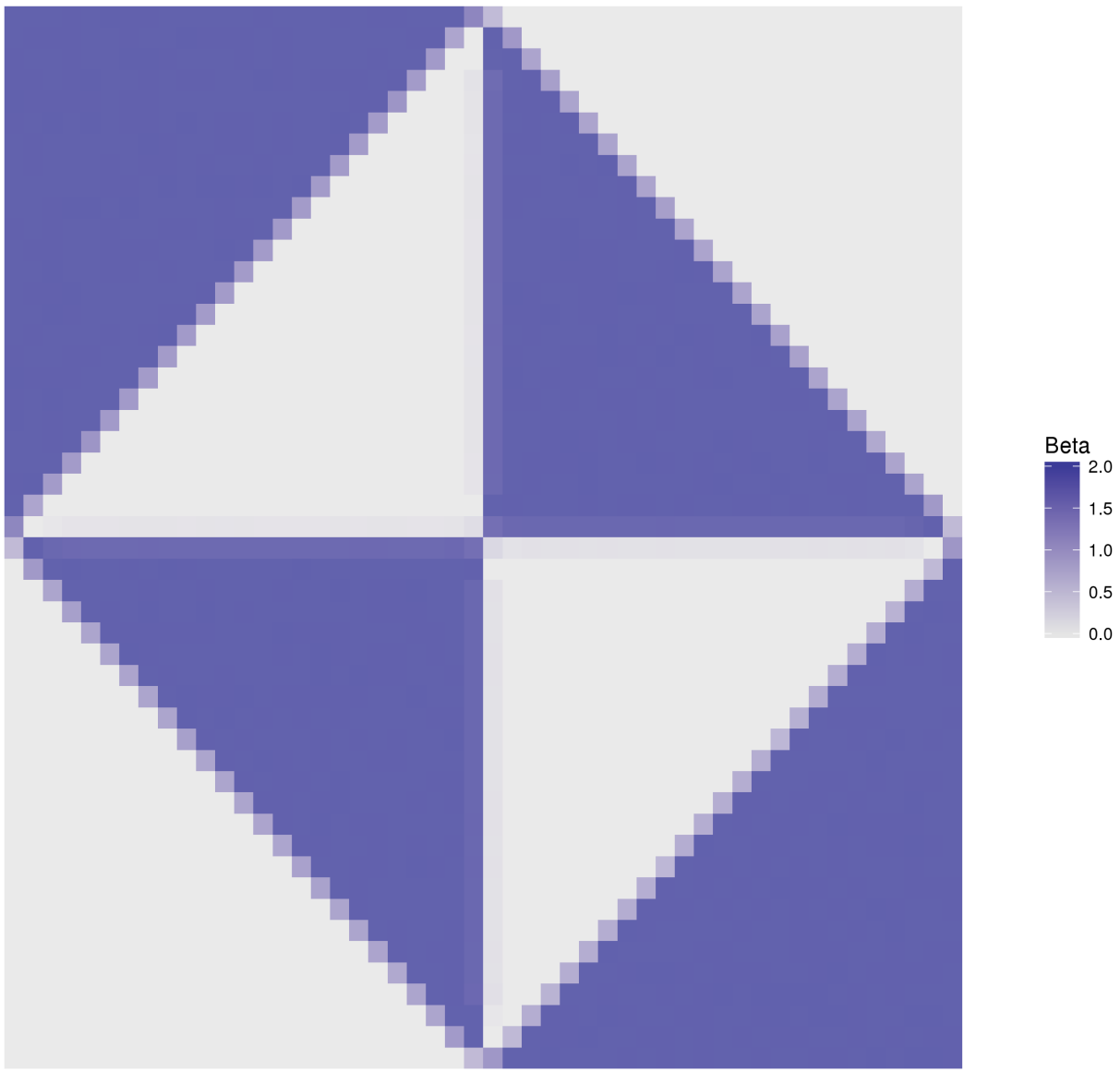
Simulated data

Estimated *B* matrix – BMMs



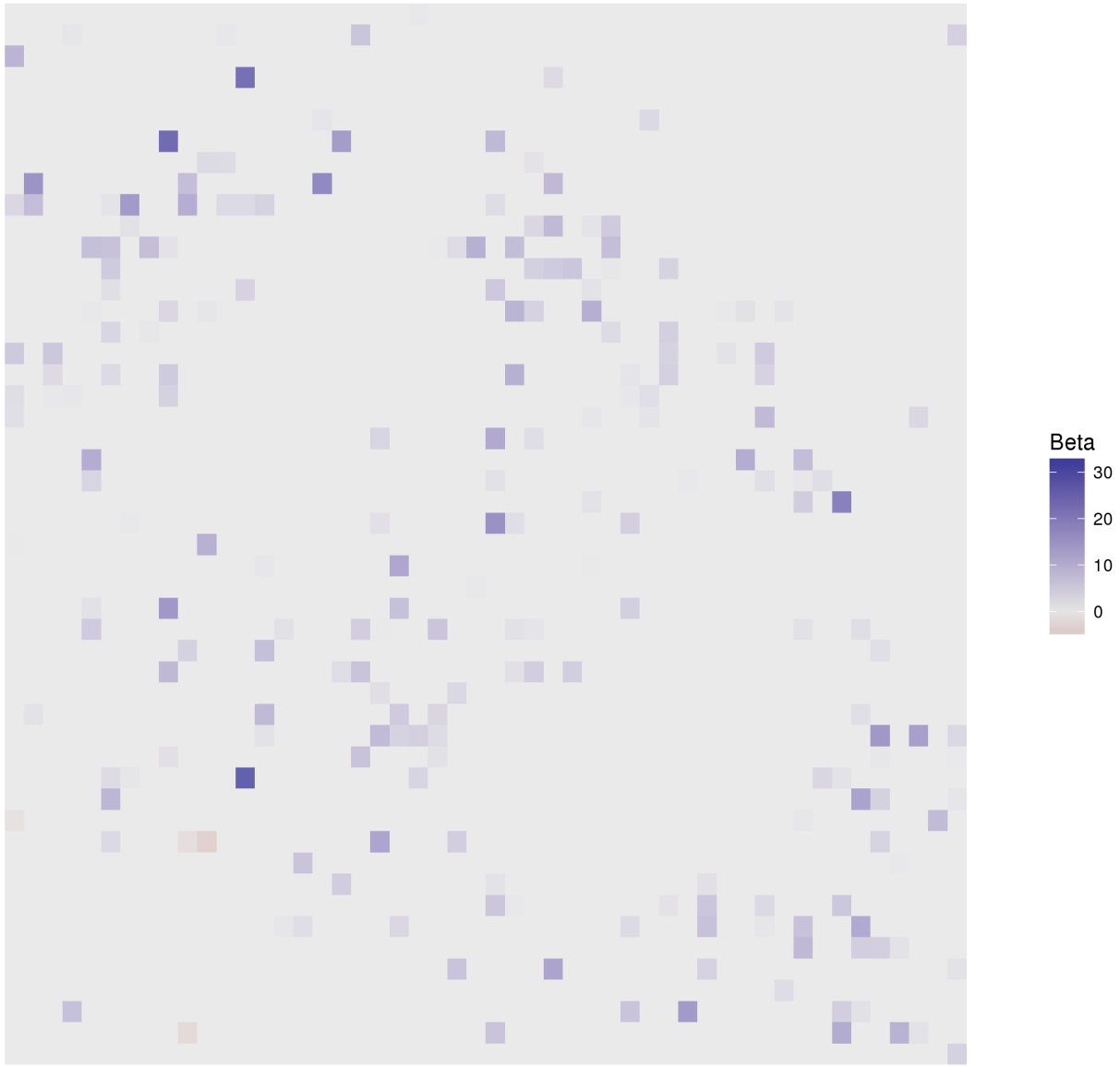
Simulated data

True B matrix in $y_i = vec(\mathbf{X}_i) \cdot vec(B) + \varepsilon$



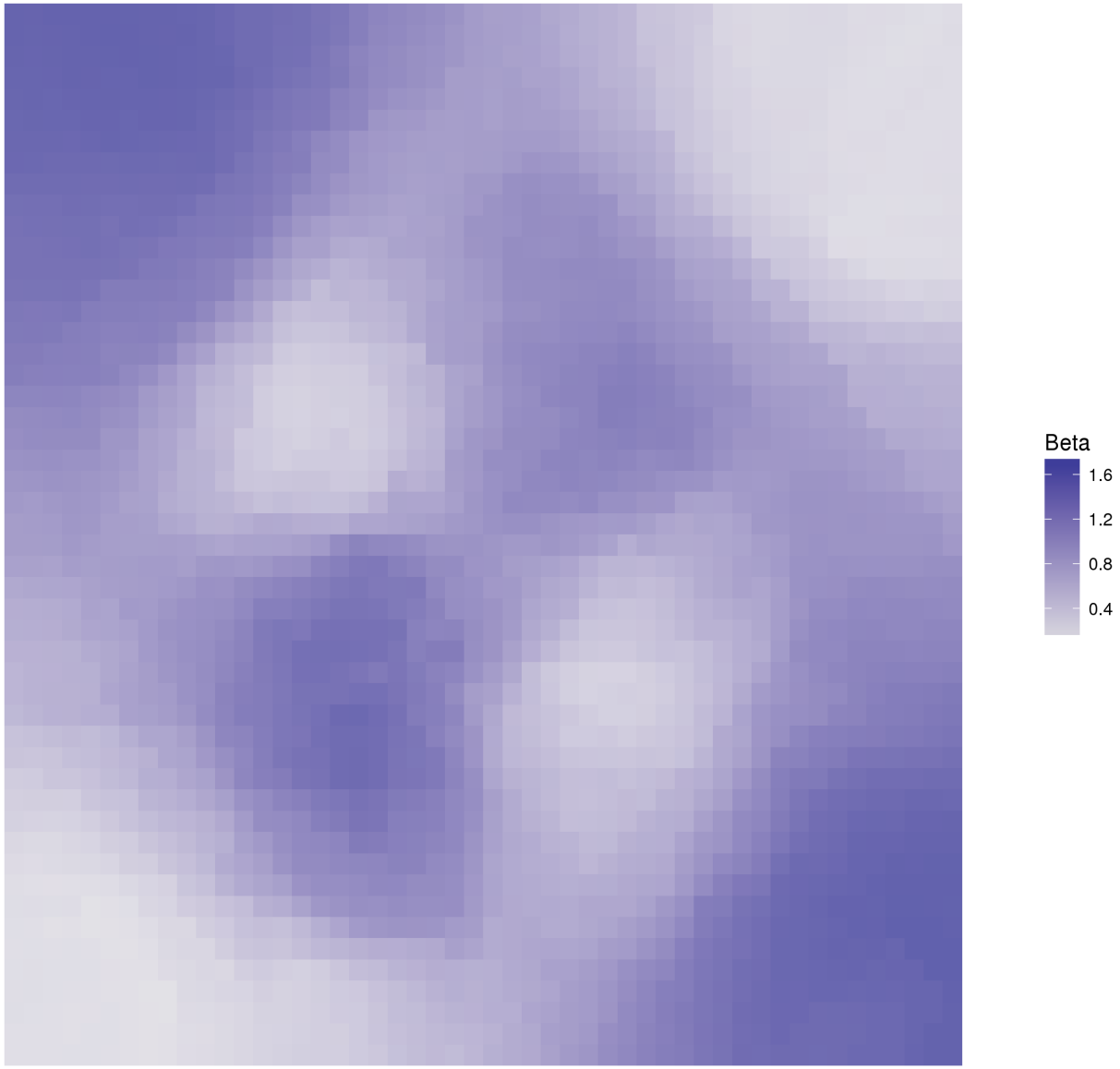
Simulated data

Estimated *B* matrix – Lasso



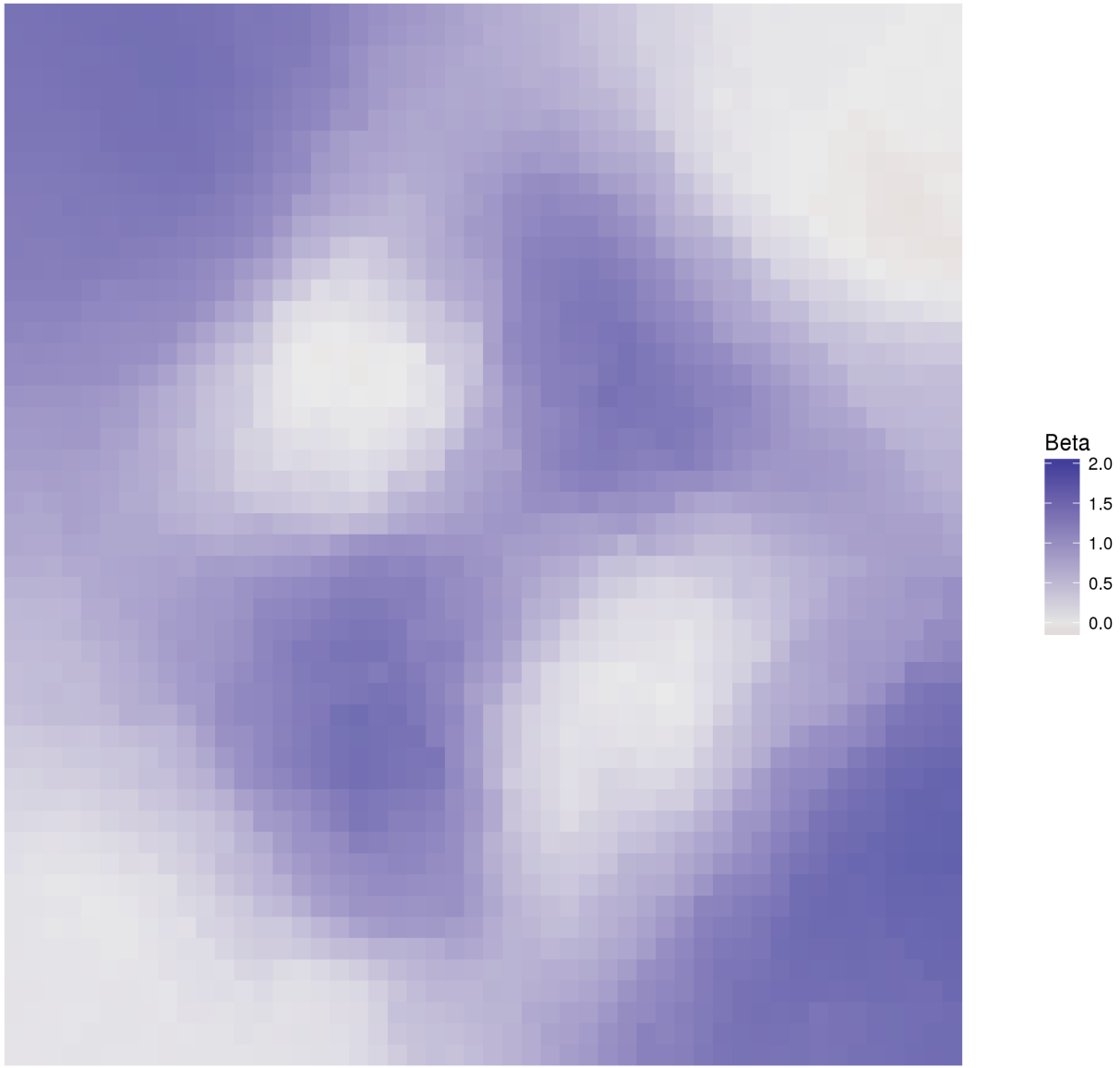
Simulated data

Estimated *B* matrix – BMMs (low resolution)



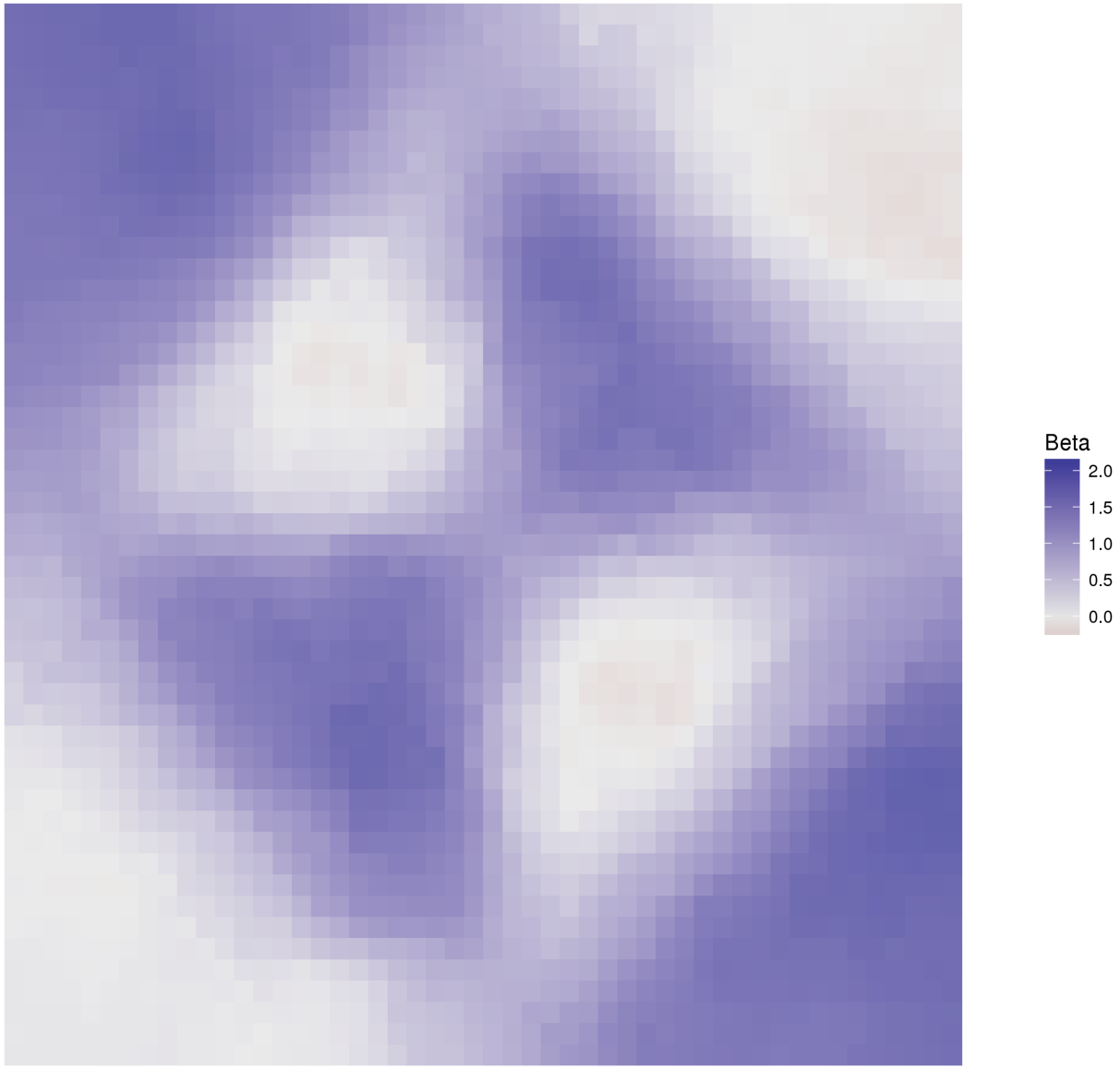
Simulated data

Estimated *B* matrix – BMMs (mid resolution)



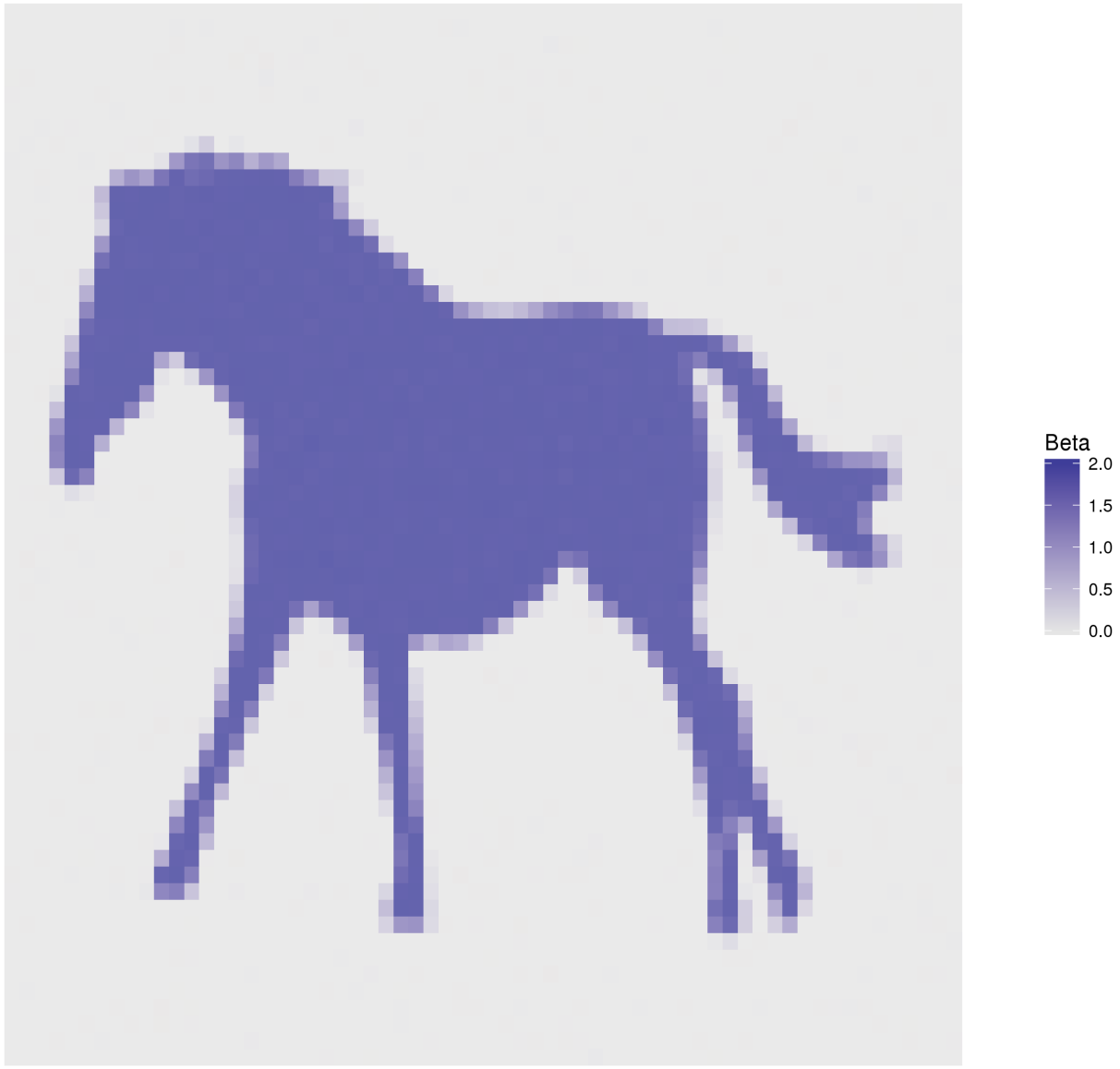
Simulated data

Estimated *B* matrix – BMMs (high resolution)



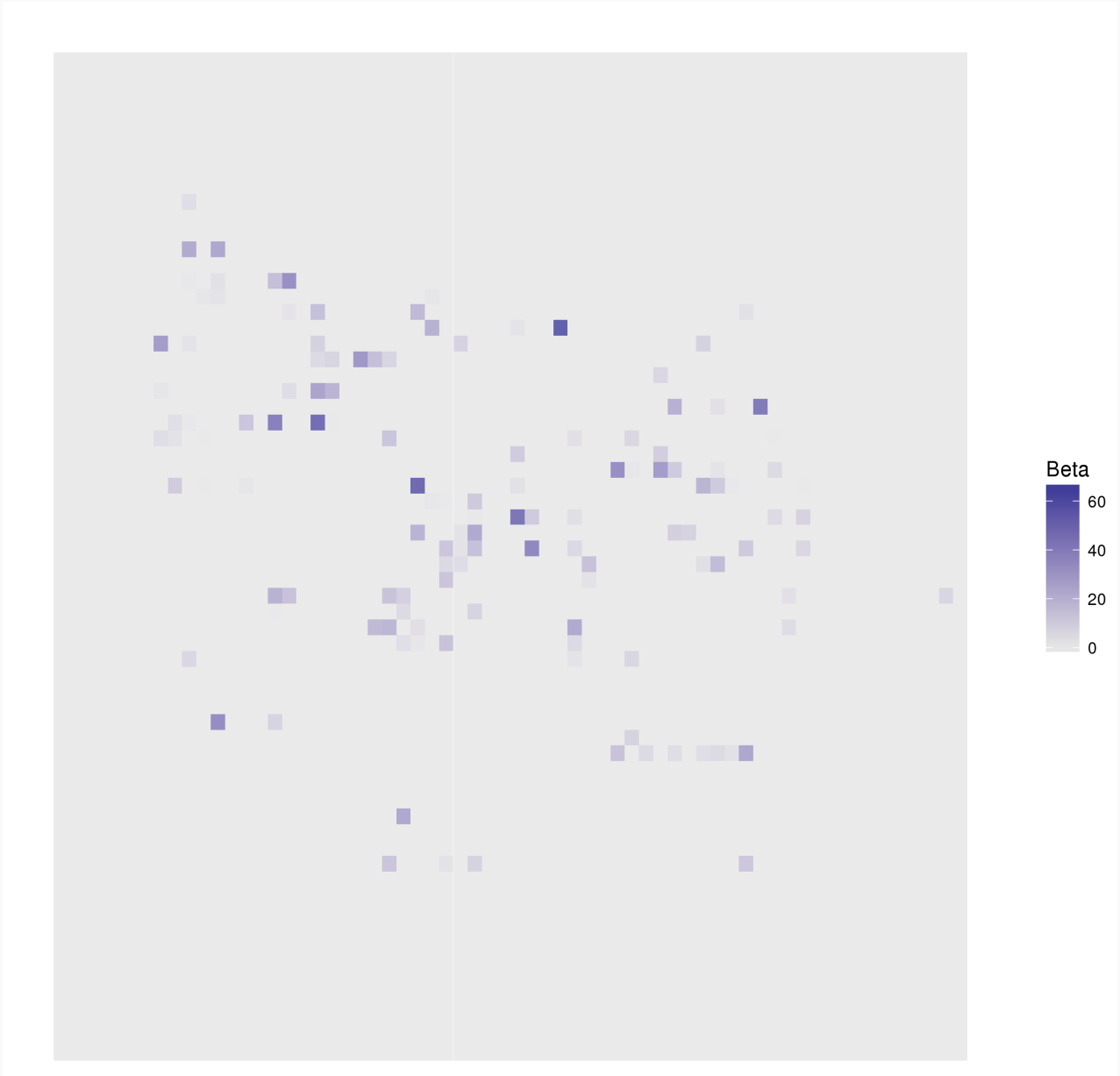
Simulated data

True B matrix in $y_i = vec(\mathbf{X}_i) \cdot vec(B) + \varepsilon$



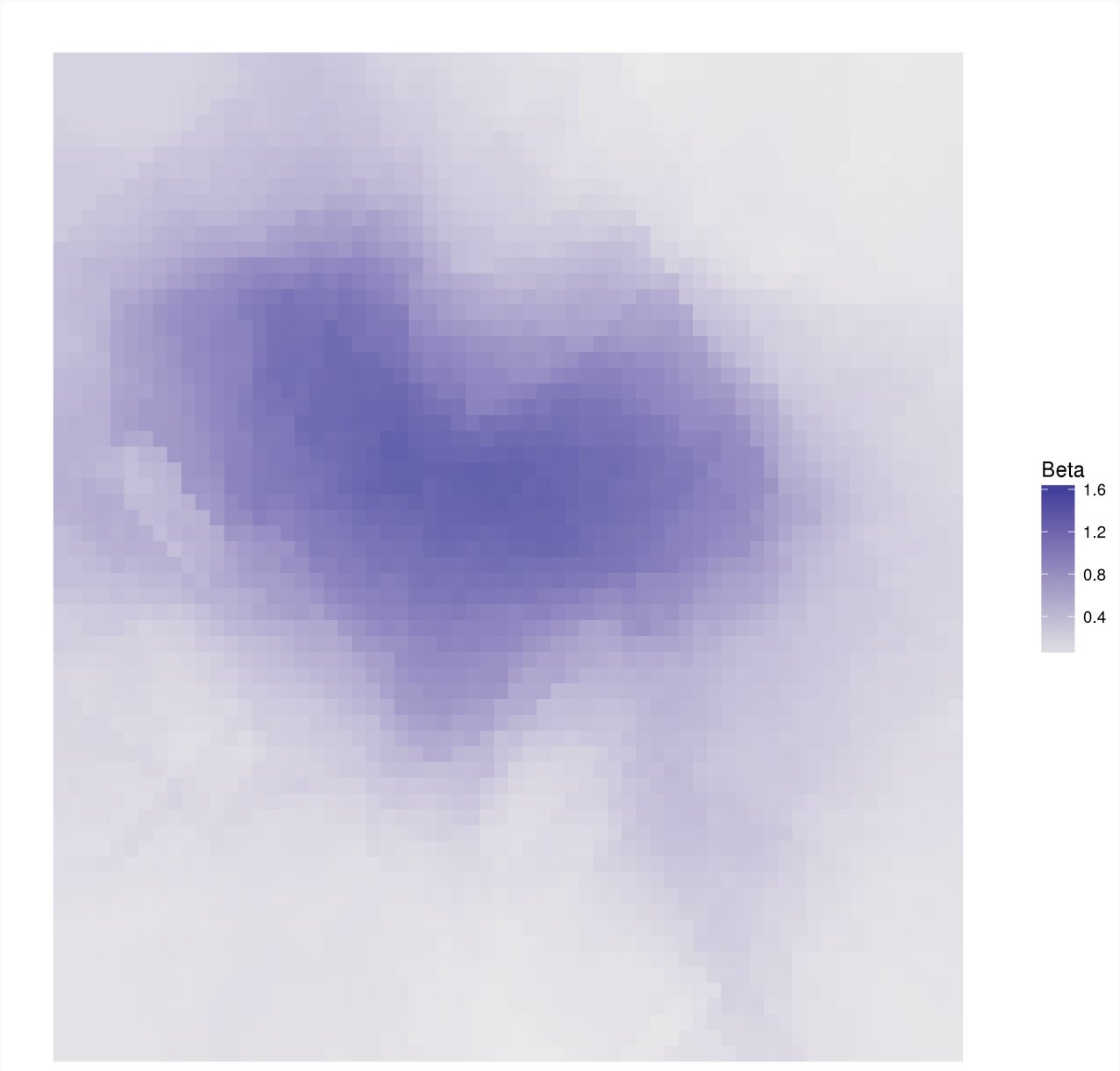
Simulated data

Estimated *B* matrix – Lasso



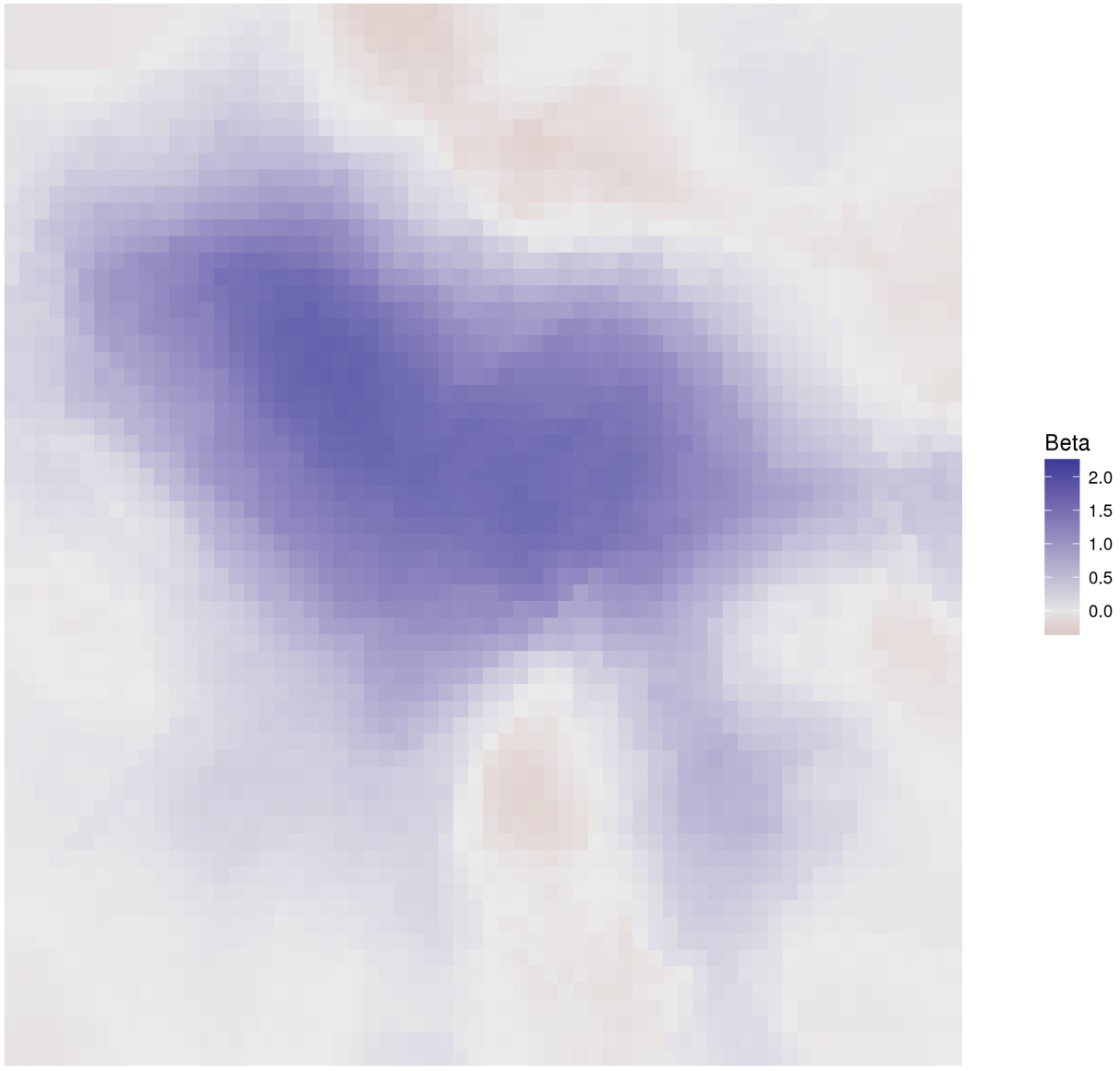
Simulated data

Estimated *B* matrix – BMMs (low resolution)



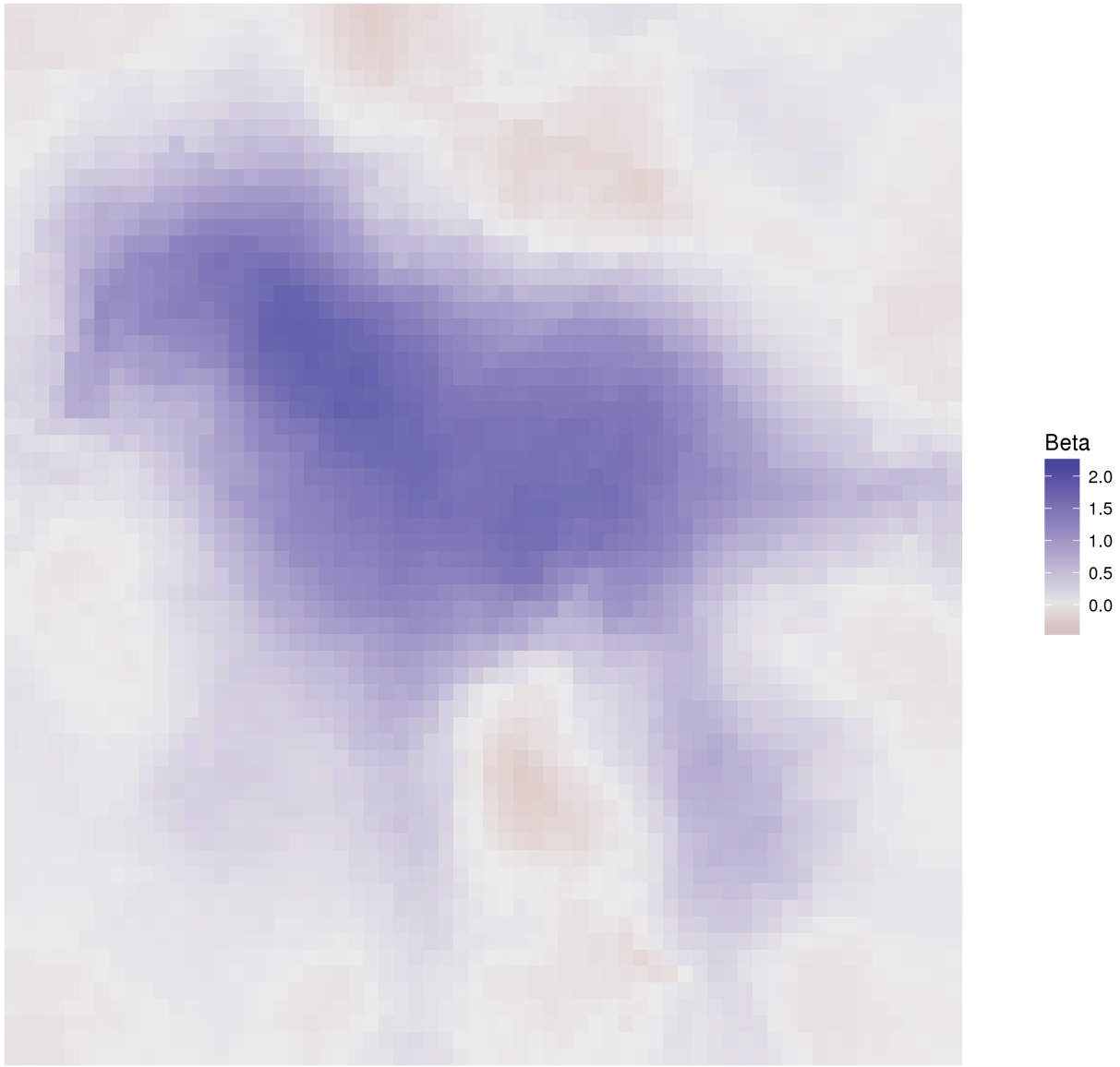
Simulated data

Estimated *B* matrix – BMMs (mid resolution)



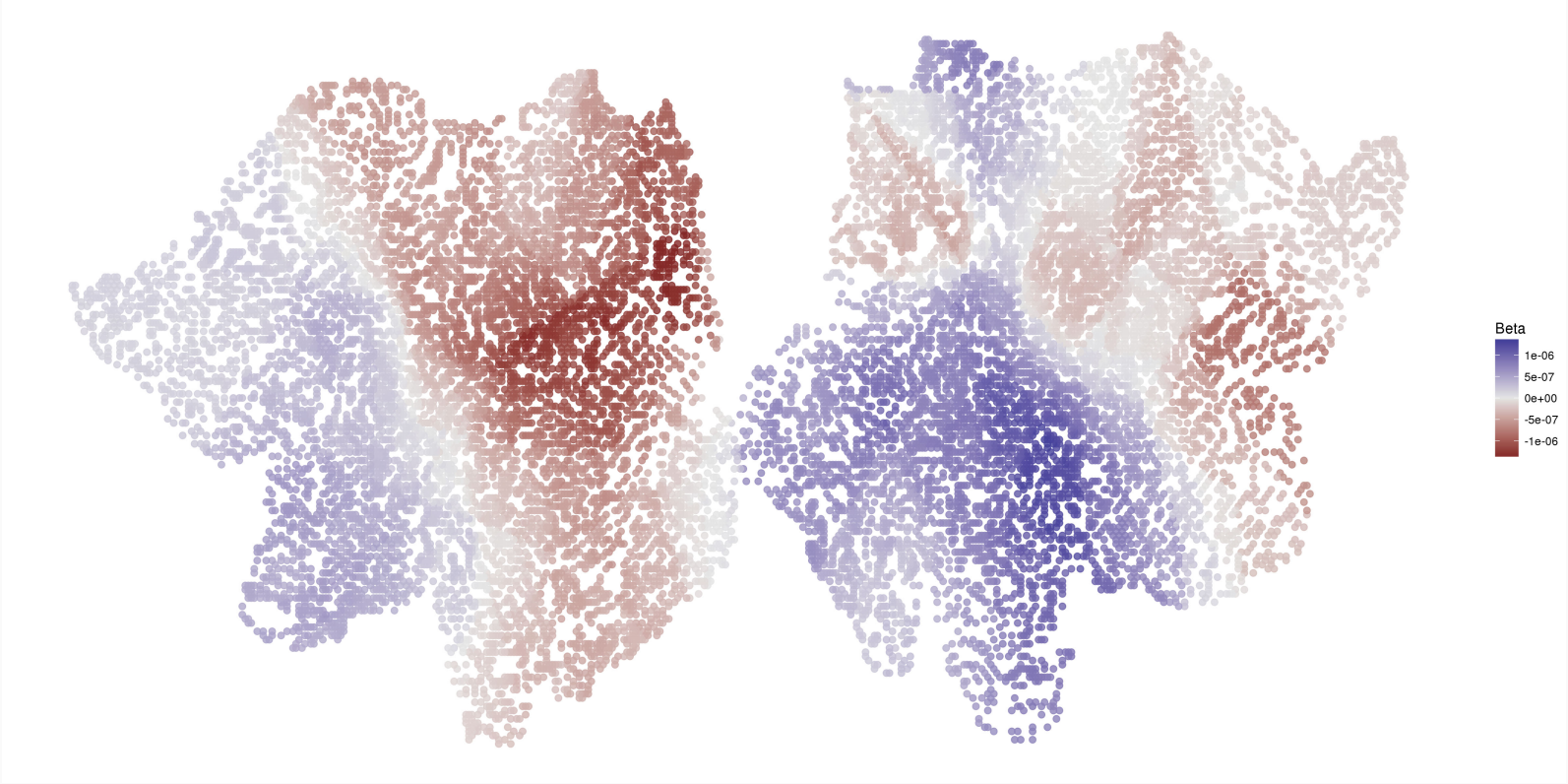
Simulated data

Estimated *B* matrix – BMMs (high resolution)



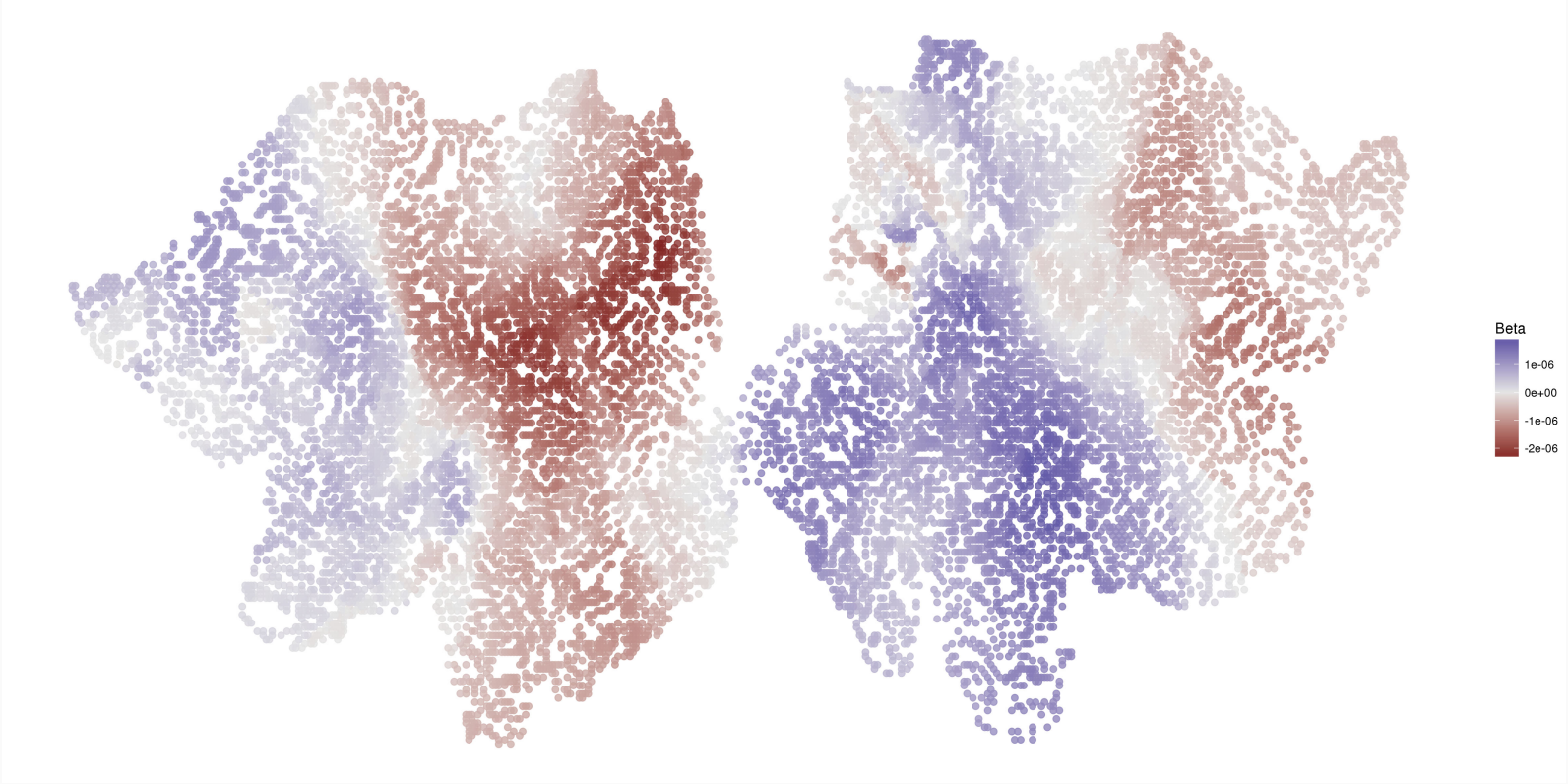
Back to HCP data

Estimated *B* matrix – BMMs (low resolution)



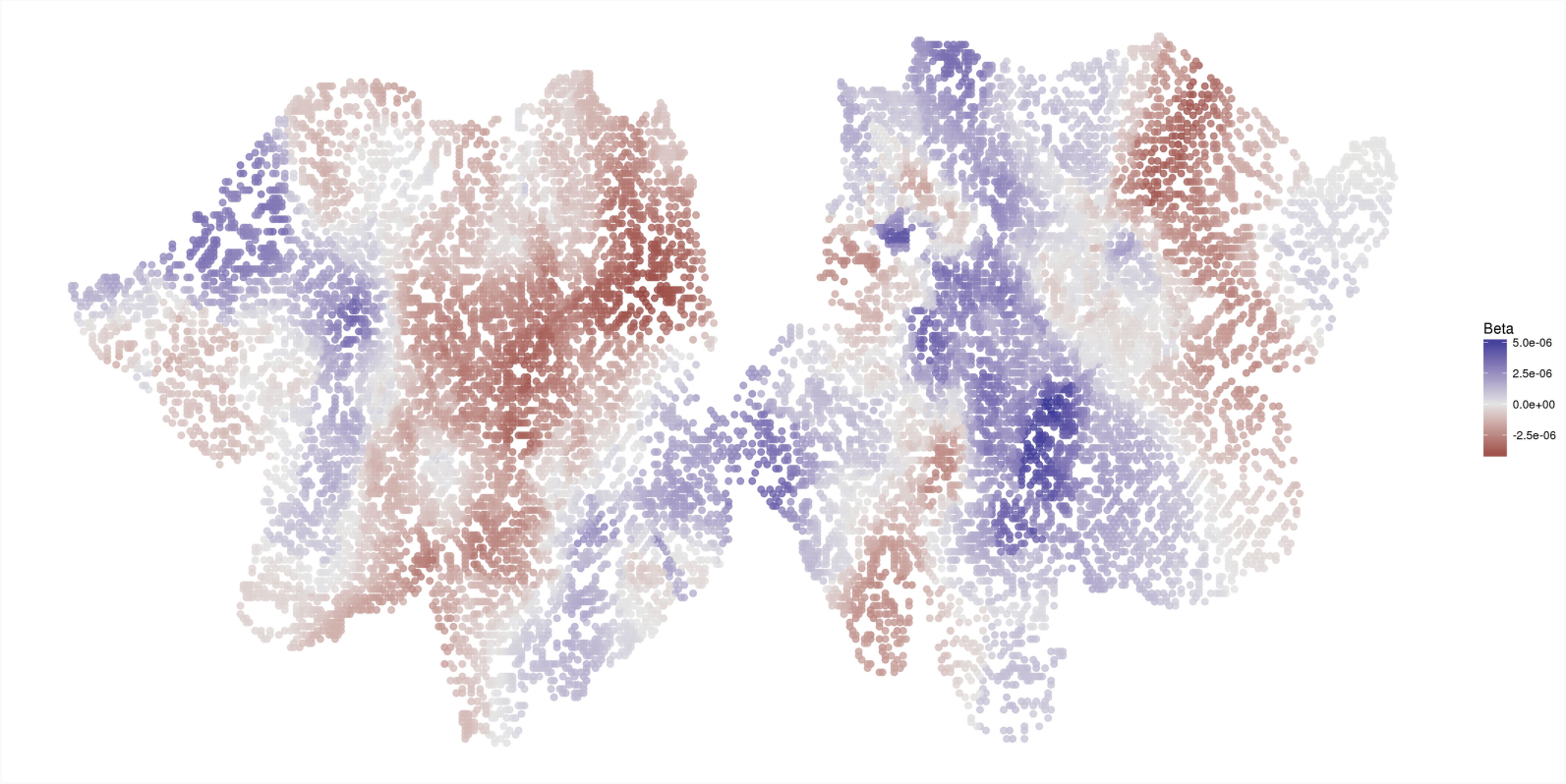
Back to HCP data

Estimated *B* matrix – BMMs (mid resolution)



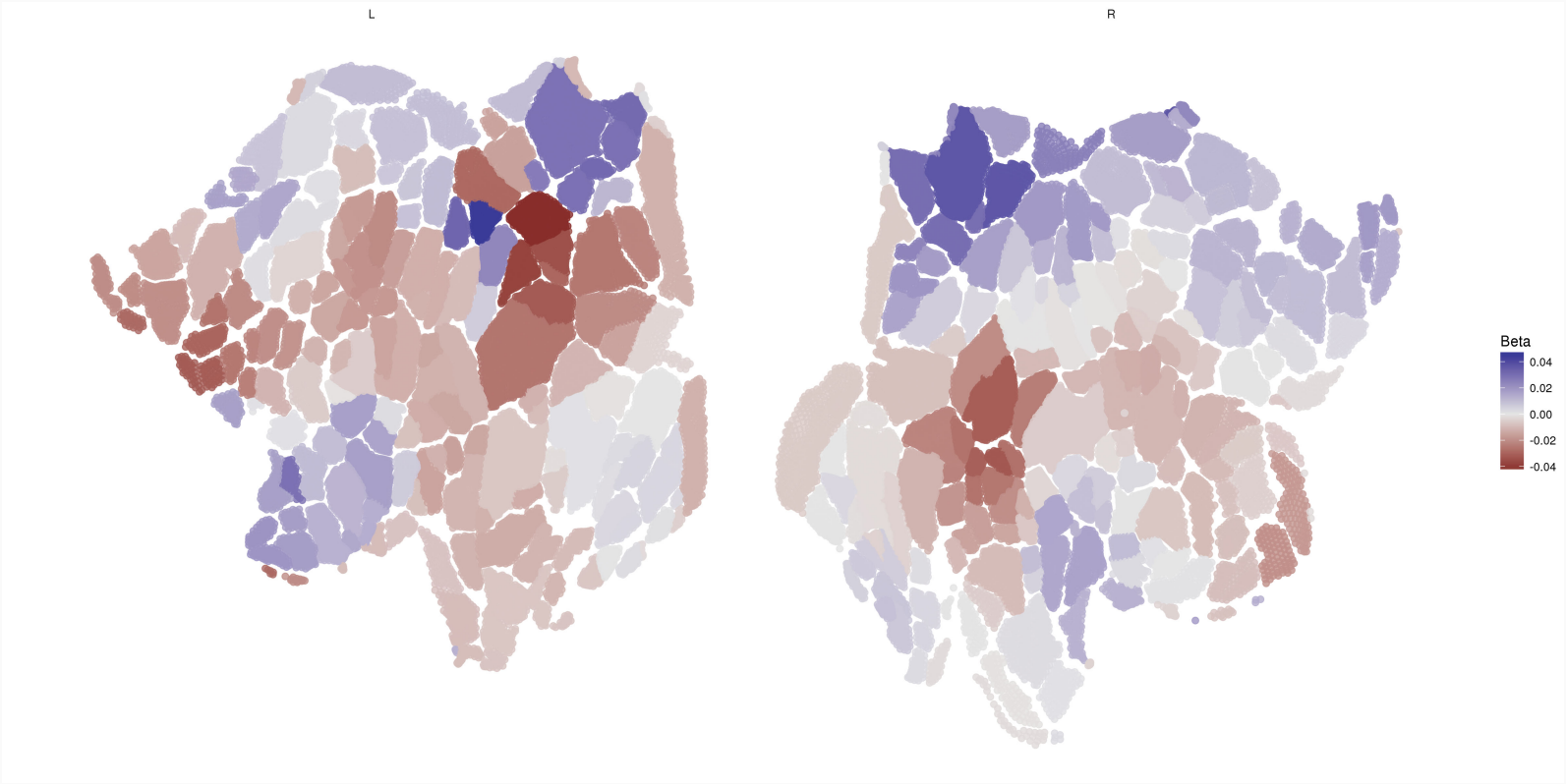
Back to HCP data

Estimated *B* matrix – BMMs (high resolution)



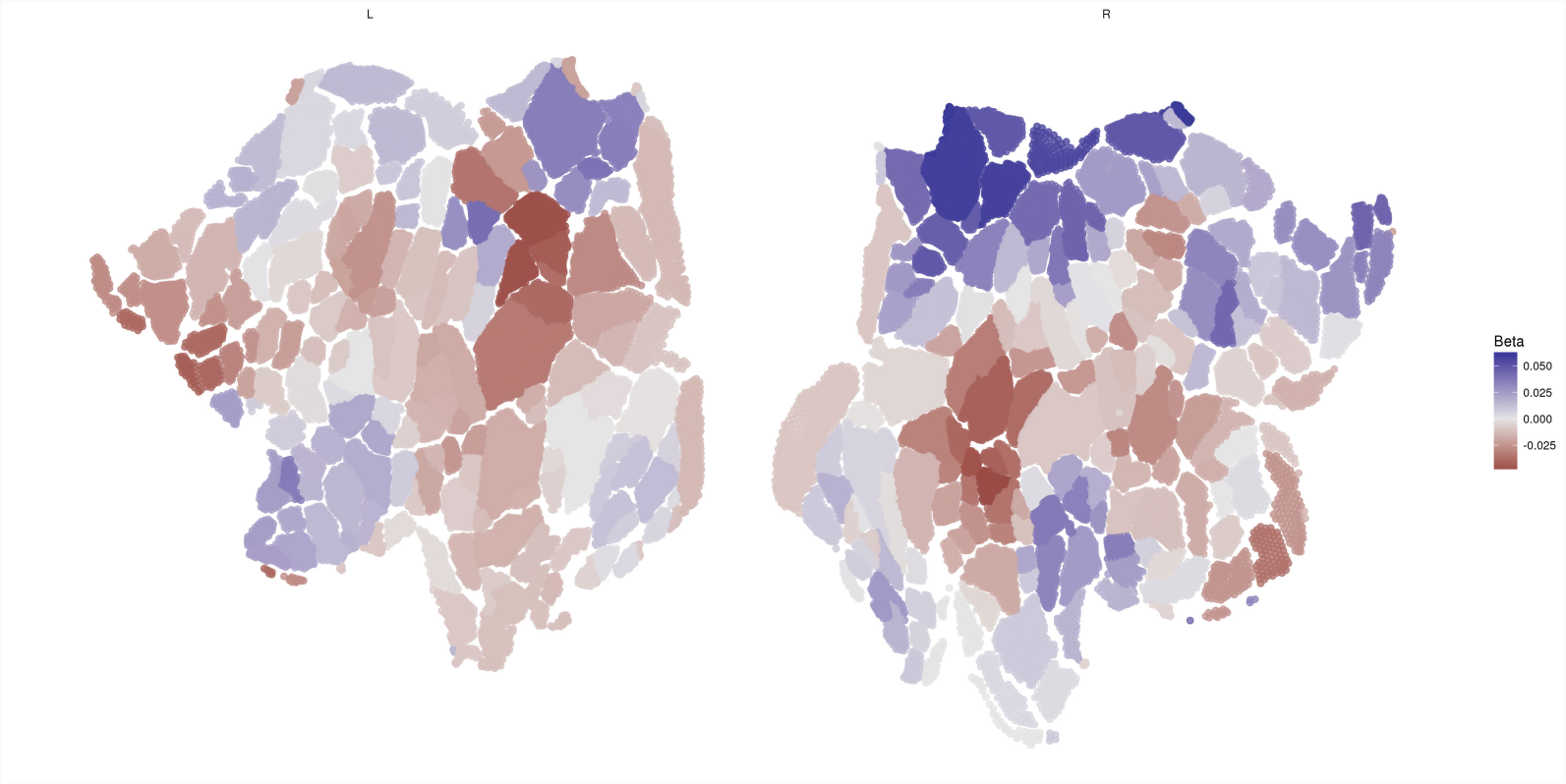
Back to HCP data

Estimated *B* matrix – BMMs on Gordon333 (low resolution)



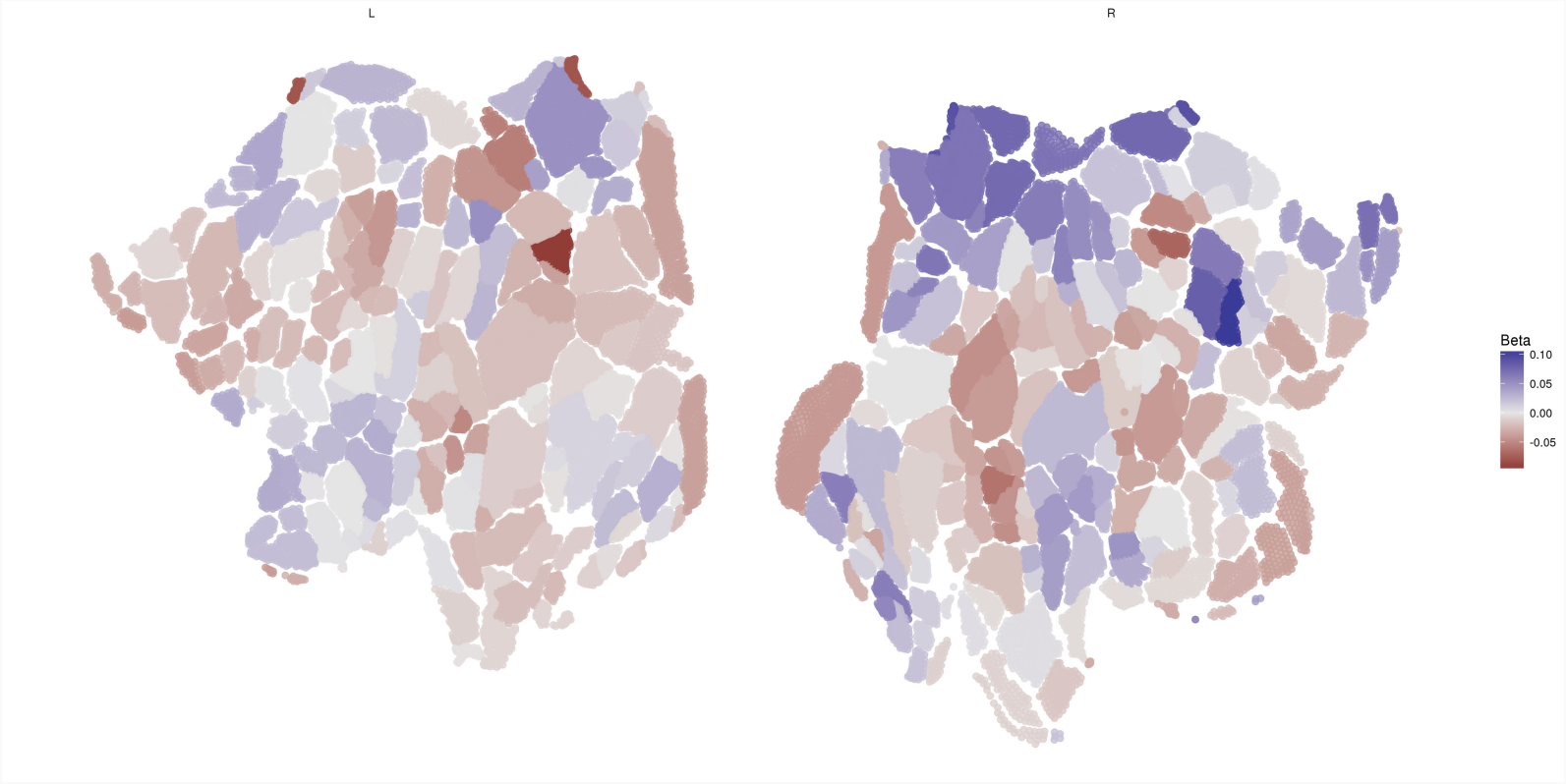
Back to HCP data

Estimated *B* matrix – BMMs on Gordon333 (mid resolution)

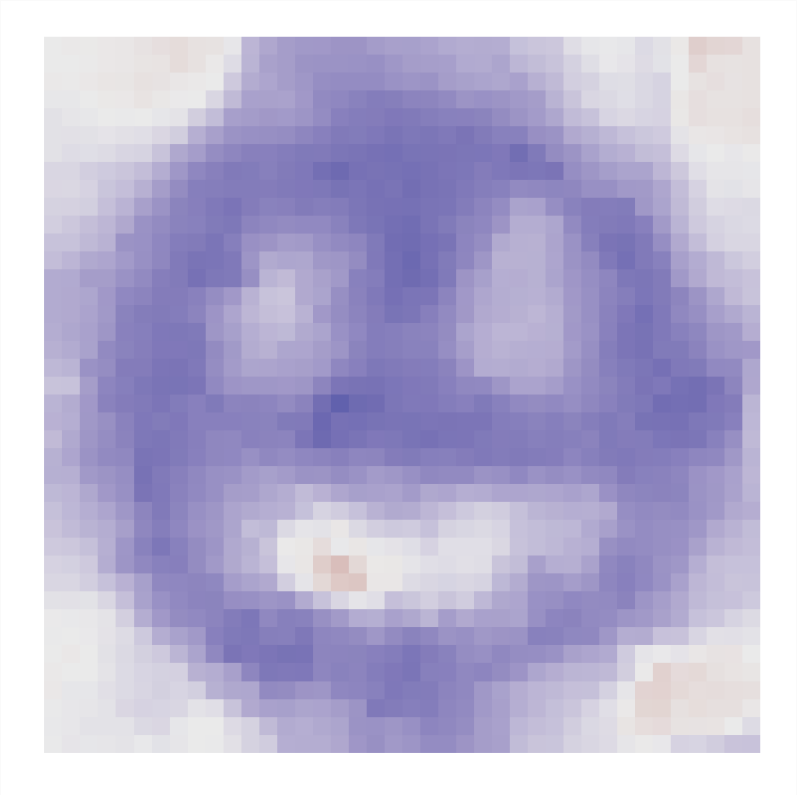


Back to HCP data

Estimated *B* matrix – BMMs on Gordon333 (high resolution)



The end



Thank you!