Large covariance estimation for spatial functional data with an application to twin studies

Benjamin B. Risk

[a] Statistical and Applied Mathematical Sciences Institute, Research Triangle Park, NC 27709, USA

[b] Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Joint work with Prof. Hongtu Zhu

BIRS NDA Conference, February 2, 2016

Motivation

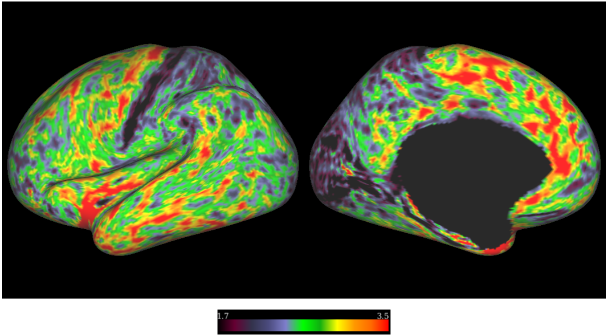


Figure : Cortical thickness (mm) for left hemisphere from a single subject (101006) from the Human Connectome Project.

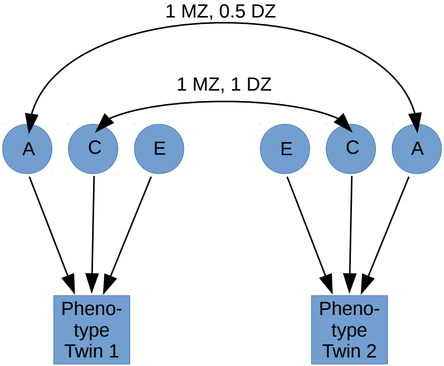
Goals of this talk:

1. Determine “nature vs. nurture” for brain traits.
2. Incorporate spatial information to predict latent effects.
3. Address computational issues from large covariance matrices.

ACE Model

Fisher's model for polygenic effects on a phenotype:
Additive, **C**ommon, and unique **E**nvironmental components

Figure : Path diagram for the SEM. MZ: monozygotic. DZ: dizygotic.



FSEM (Luo et al)

[Luo et al., 2016]: Functional structural equation model for $v \in [0, 1]$:

$$\begin{aligned} y_{ij}(v) &= X_{ij}'\beta(v) + R_{ij}(v), \\ R_{ij}(v) &= \left[\{1 - \mathbb{1}_{DZ}(i)\} + \sqrt{0.5}\mathbb{1}_{DZ}(i)\right] a_i(v) \\ &\quad + \sqrt{0.5}\mathbb{1}_{DZ} a_{ij}(v) + c_i(v) + e_{ij}(v), \end{aligned}$$

with

$$\begin{aligned} a_i(v) &\sim GP(0, \Sigma_a(v, v)), \\ a_{ij}(v) &\sim GP(0, \Sigma_a(v, v)) \\ c_i(v) &\sim GP(0, \Sigma_c(v, v)) \\ e_{ij}(v) &\sim N(0, \sigma_e^2(v)). \end{aligned}$$

FSEM (Luo et al): Three-step estimation

- Estimators:
  1. Univariate analysis at every location using MLE to estimate ACE at every vertex.
  2. MWLE with bandwidth determined using 5-fold CV.
  3. Estimate covariance function with compact support in $\mathbb{R}^1$ using local constant regression with residuals from step 1.

FSEM (Luo et al): Local constant regression for covariance estimation

$$\widehat{U}_{i,j,v_0,v_0'} = \begin{cases} \widehat{R}_{i,j,v_0}\widehat{R}_{i,j,v_0'} & \text{if } v_0 \neq v_0' \\ 0 & \text{if } v_0 = v_0' \end{cases}$$

and

$$\widehat{U}_{i,v_0,v_0'} = \left(\widehat{R}_{i,1,v_0}\widehat{R}_{i,2,v_0'} + \widehat{R}_{i,1,v_0'}\widehat{R}_{i,2,v_0}\right)/2.$$

$$\mathcal{J}_n(v,v') =$$

$$\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{\{v_0,v_0'\in\mathcal{V}_0(v')\}}\left\{\widehat{U}_{ij}(v_0,v_0') - \mathbf{\Sigma}_a(v,v') - \mathbf{\Sigma}_c(v,v')\right\}^2 K_h(v_0,v)K_h(v_0',v')$$

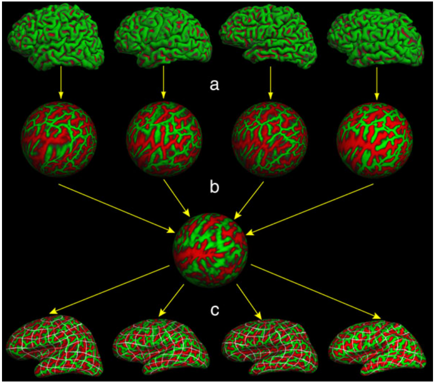
$$+\frac{1}{n_1}\sum_{i=1}^{n_1}\sum_{\{v_0,v_0'\in\mathcal{V}_0(v')\}}\left\{\widehat{U}_i(v_0,v_0') - \mathbf{\Sigma}_a(v,v') - \mathbf{\Sigma}_c(v,v')\right\}^2 K_h(v_0,v)K_h(v_0',v')$$

$$+\frac{1}{n_2}\sum_{i=1}^{n_2}\sum_{\{v_0,v_0'\in\mathcal{V}_0(v')\}}\left\{\widehat{U}_i(v_0,v_0') - 0.5\mathbf{\Sigma}_a(v,v') - \mathbf{\Sigma}_c(v,v')\right\}^2 K_h(v_0,v)K_h(v_0',v').$$

ACE for large spatial data

- Estimators:
  1. Vertex-wise univariate analysis using MLE to estimate ACE at every vertex.
  2. Smooth MLE using biweight kernel with bandwidth determined using GCV.
  3. In our data application, use geodesic distance on Freesurfer 32k spherical template.
  4. Estimate covariance functions at observed locations using a “sandwich” formulation of local constant regression with residuals from step 1. GCV.
  5. Also developing an approach with random projections that scales to massive matrices.

Simplifying the spatial structure



Source: https://surfer.nmr.mgh.harvard.edu/fswiki/FreeSurferAnalysisPipelineOverview

- Our spatial methods will use the geodesic distance on the Freesurfer 32k spherical template.
- Calculations are fast using great-circle formula.

Linear combinations of sample covariances

- “Sample” covariances

$$\text{All: } \mathbf{S}_0 = \frac{1}{N}\left(\mathbf{R}'\mathbf{R}\right)$$

$$\text{MZs: } \mathbf{S}_1 = \frac{1}{2n_1}\left(\mathbf{R}'_{11}\mathbf{R}_{12} + \mathbf{R}'_{12}\mathbf{R}_{11}\right)$$

$$\text{DZs: } \mathbf{S}_2 = \frac{1}{2n_2}\left(\mathbf{R}'_{21}\mathbf{R}_{22} + \mathbf{R}'_{22}\mathbf{R}_{21}\right).$$

  Define *simple estimators*

$$\mathbf{S}_A = \mathbf{S}_0 + \mathbf{S}_1 - 2\mathbf{S}_2 + \operatorname{diag}\mathbf{S}_1 - \operatorname{diag}\mathbf{S}_0$$

$$\mathbf{S}_C = 2\mathbf{S}_2 - 0.5\mathbf{S}_0 - 0.5\mathbf{S}_1 + 0.5\operatorname{diag}\mathbf{S}_0 - 0.5\operatorname{diag}\mathbf{S}_1.$$

- Create PSD estimates by calculating EVD and truncating eigenvalues. Low rank.

Sandwich formulation of local constant regression

- [Xiao et al., 2013] use sandwich formulation of covariance estimation using bivariate P-splines, $\mathbf{K}\mathbf{S}_A\mathbf{K}'$.
- Facilitates use of GCV, $(\mathbf{K} \otimes \mathbf{K})\ \mathrm{vec}\mathbf{S}_A$
- For twin studies, we have multiple covariance functions to estimate.
- We propose the sandwich formulation of local constant regression estimators.
- Define $\mathbf{K}$ such that $\mathbf{K}_{k,l} = K_h(v_k, v_l)/\sum_{l=1}^{V} K_h(v_k, v_l)$. Then

$$\widehat{\Sigma}_A = \mathbf{K}\mathbf{S}_A^{+}\mathbf{K}' \tag{1}$$

$$\widehat{\Sigma}_C = \mathbf{K}\mathbf{S}_C^{+}\mathbf{K}'. \tag{2}$$

- Smooth eigenvectors only: $(\mathbf{K}\mathbf{\Psi}_A{}^{+})\mathbf{\Lambda}_A{}^{+}(\mathbf{\Psi}_A{}^{+'}\mathbf{K}')$.

eBLUPs for DZ twin pair

- $\boldsymbol{a}_i = [[a_i(1), \ldots, a_i(V)]' \otimes \mathbf{1}_2] \in \mathbb{R}^{2V}$,
  $\boldsymbol{a}_i^* = [a_{i1}(1), a_{i2}(1) \ldots, a_{i1}(V), a_{i2}(V)]' \in \mathbb{R}^{2V}$
- Matrix formulation for DZ pair:

$$\mathbf{Y}_i = (\mathbf{I}_V \otimes \mathbf{X}_i)\,\beta + \sqrt{0.5}(\mathbf{I}_V \otimes \mathbf{I}_2)\boldsymbol{a}_i^* + \sqrt{0.5}(\mathbf{I}_V \otimes \mathbf{J}_2)\boldsymbol{a}_i + (\mathbf{I}_V \otimes \mathbf{J}_2)\boldsymbol{c}_i + \boldsymbol{e}_i$$

  with $\mathbf{Y}_i \in \mathbb{R}^{2V}$, $\boldsymbol{e}_i$ unique environmental variance, $\mathbf{X}_i \in \mathbb{R}^{2\times p}$ design matrix for the twin pair, $\beta \in \mathbb{R}^{Vp}$ fixed effects,
- Derive the BLUPs

$$\hat{\mathbf{a}}_i^* = (0.5\boldsymbol{\Sigma}_a \otimes \mathbf{I}_2)\left\{0.5\boldsymbol{\Sigma}_a \otimes \mathbf{I}_2 + (0.5\boldsymbol{\Sigma}_a + \boldsymbol{\Sigma}_c) \otimes \mathbf{J}_2 + \boldsymbol{\Sigma}_e \otimes \mathbf{I}_2\right\}^{-1}\left\{\mathbf{Y}_i - (\mathbf{I}_V \otimes \mathbf{X}_i)\beta\right\}$$

  and similarly derive predictors for $\hat{\boldsymbol{a}}_i$ and $\hat{\boldsymbol{c}}_i$

Simulation design

- For 50 MZ, 50 DZ, 100 singles, simulate GP at 1002 locations

$$c_i(v) = \sum_{\ell=1}^{4} \xi_{i\ell} f_\ell(v, v'),$$

$$\xi_{i1} \overset{iid}{\sim} \mathcal{N}(0, 2000),$$

$$\xi_{i2} \overset{iid}{\sim} \mathcal{N}(0, 1367),$$

$$\xi_{i3} \overset{iid}{\sim} \mathcal{N}(0, 733),$$

$$\xi_{i4} \overset{iid}{\sim} \mathcal{N}(0, 100),$$

  where $f_\ell(\cdot, \cdot)$ is an orthogonal basis generating local and long-range dependence.
- $a_i(v)$ and $a_{ij}(v)$ modified to have a region with zero variance.
- $\Sigma_e = 2\mathrm{diag}(\Sigma_a + \Sigma_c)$

Example simulation: four seeds and $\widehat{\Sigma}_a$

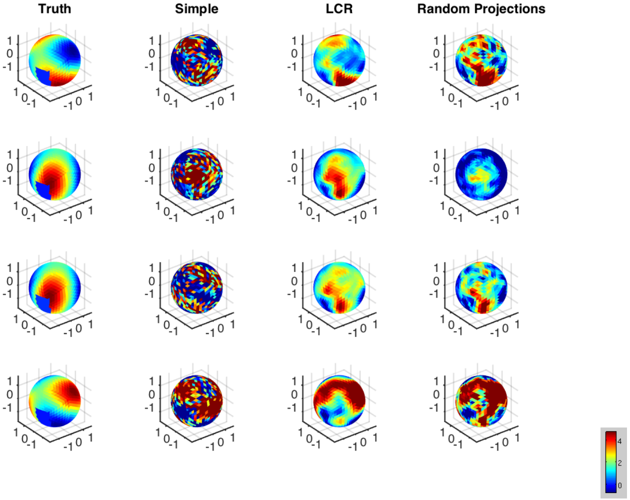
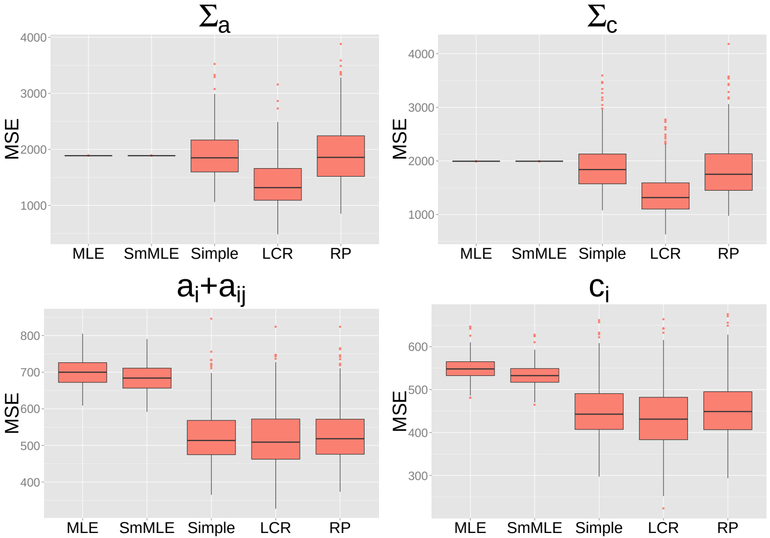


Figure : Estimated covariance for four randomly chosen seeds from the simulation associated with median MLE error.

Simulation Results



Simulation example: Predictions for two subjects

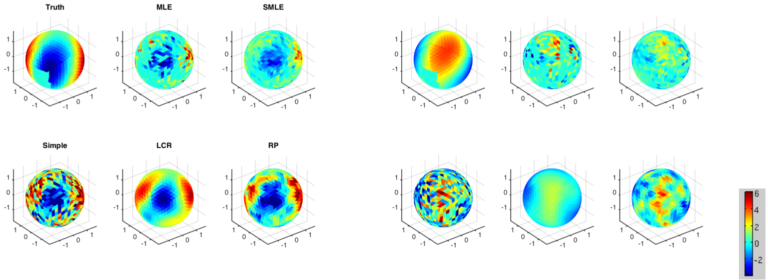


Figure : Predicted $\boldsymbol{a}_i$ for an MZ (left) and $\boldsymbol{a}_i + \boldsymbol{a}_{ij}$ for a DZ (right).

Preliminary HCP Results: Covariance from a seed

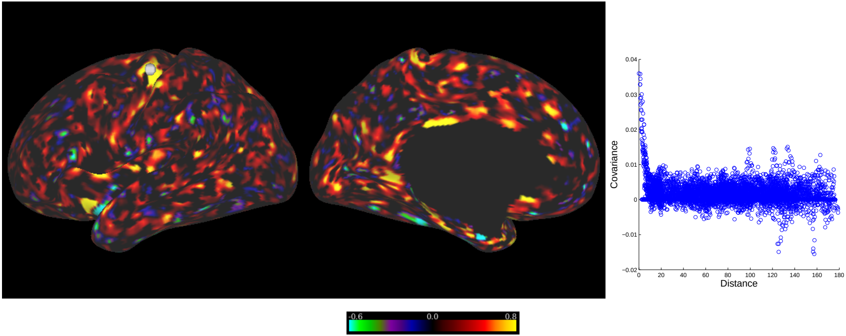


Figure : 100*Covariance of genetic effects in cortical thickness (left hemisphere) from seed 5062 using the LCR method (left) and covariance versus distance (right).

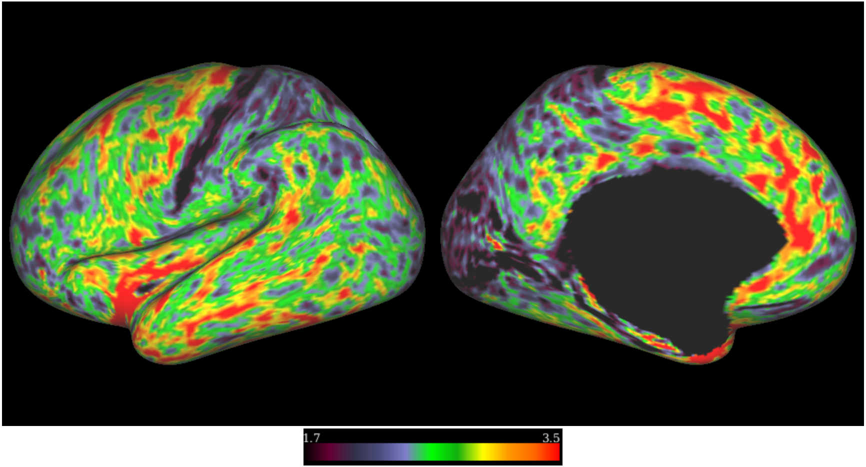


Figure : Cortical thickness (left hemisphere) for subject 101006.

Preliminary results

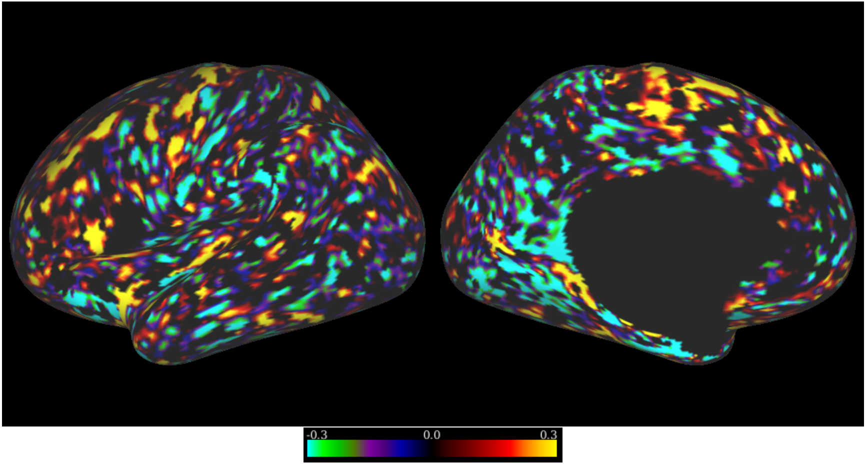


Figure : Additive genetic effect for subject 100106 estimated from covariance function.

Discussion

- Incorporating spatial information improves prediction of random effects
- Locally weighted covariance estimators can capture short-range and long-range correlations.
- Future directions: explore better ways for PSD to minimize distance between symmetric function on the sphere and positive semi-definite functions.
- Develop bounds on approximation error from random projections to control balance between accuracy and speed.

Acknowledgments

Data were provided (in part) by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. This material was based upon work partially supported by the NSF grant DMS-1127914 to the Statistical and Applied Mathematical Science Institute.

References I

Luo, S., Song, R., Styner, M., Gilmore, J. H., and Zhu, H. (2016).
FSEM: Functional structural equation model for twin functional data.
In review.

Xiao, L., Li, Y., and Ruppert, D. (2013).
Fast bivariate p-splines: the sandwich smoother.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 75(3):577–599.

Generalized approach for massive matrices

- Idea of random projections: $\mathbf{Q}$ is random semi-orthogonal with dimensions $V \times R$ for some $R << V$
- Since Rank$\mathbf{S}_A$ is small, we let $\mathbf{Q} \in \mathcal{O}^{V \times R}$ for some $R > N + 20$. Then

$$\widehat{\Sigma}_A \approx \mathbf{K}\mathbf{Q}\mathbf{Q}'\widehat{\mathbf{S}}_A\mathbf{Q}\mathbf{Q}'\mathbf{K}'.$$

Approach for massive matrices, cont.

- Define $\mathbf{T} = \mathbf{Q}'\mathbf{S}_A\mathbf{Q}$ and let $\tilde{\mathbf{T}}$ be the PSD matrix closest to $\mathbf{T}$.
- It will turn out that we can calculate $\tilde{\mathbf{T}}$ without constructing $\mathbf{S}_A$.
- We will define a memory efficient approach to obtain the fPCA decomposition that is equivalent to the following estimator:

$$\widehat{\Sigma}_A^{FAST} = \mathbf{K}\mathbf{Q}\tilde{\mathbf{T}}\mathbf{Q}'\mathbf{K}'$$

which is guarenteed PSD.

Approach for massive matrices, cont.

Define $\tilde{\mathbf{R}} = \mathbf{R}\mathbf{Q}$, which is $N \times d$, and similarly define $\tilde{\mathbf{R}}_{11}$, $\tilde{\mathbf{R}}_{12}$, $\tilde{\mathbf{R}}_{21}$, $\tilde{\mathbf{R}}_{22}$. Then

$$\begin{aligned}
\mathbf{T} &= \mathbf{Q}'\mathbf{S}_A\mathbf{Q} \\
&= \mathbf{Q}'\left\{\mathbf{S}_0 + \mathbf{S}_1 - 2\mathbf{S}_2 + \operatorname{diag}\mathbf{S}_1 - \operatorname{diag}\mathbf{S}_0\right\}\mathbf{Q} \\
&= \mathbf{Q}'\left\{\frac{1}{N}\left(\mathbf{R}'\mathbf{R}\right) + \frac{1}{2n_1}\left(\mathbf{R}'_{11}\mathbf{R}_{12} + \mathbf{R}'_{12}\mathbf{R}_{11}\right) + \frac{1}{n_2}\left(\mathbf{R}'_{21}\mathbf{R}_{22} + \mathbf{R}'_{22}\mathbf{R}_{21}\right)\right. \\
&\quad \left. + \operatorname{diag}\mathbf{S}_1 - \operatorname{diag}\mathbf{S}_0\right\}\mathbf{Q} \\
&= \frac{1}{N}\tilde{\mathbf{R}}'\tilde{\mathbf{R}} + \frac{1}{2n_1}\left(\tilde{\mathbf{R}}'_{11}\tilde{\mathbf{R}}_{12} + \tilde{\mathbf{R}}'_{12}\tilde{\mathbf{R}}_{11}\right) + \frac{1}{n_2}\left(\tilde{\mathbf{R}}'_{21}\tilde{\mathbf{R}}_{22} + \tilde{\mathbf{R}}'_{22}\tilde{\mathbf{R}}_{21}\right) \\
&\quad + \mathbf{Q}'\left(\operatorname{diag}\mathbf{S}_1 - \operatorname{diag}\mathbf{S}_0\right)\mathbf{Q}.
\end{aligned}$$

Approach for massive matrices, cont.

- We calculate the terms in $\operatorname{diag}\mathbf{S}_1$ and $\operatorname{diag}\mathbf{S}_2$ and use sparse matrix multiplication to calculate $\mathbf{Q}'\operatorname{diag}\mathbf{S}_1\mathbf{Q}$ and $\mathbf{Q}'\operatorname{diag}\mathbf{S}_0\mathbf{Q}$.
- We calculate $\tilde{\mathbf{T}}$ by truncating to the eigenvalue/eigenvector pairs corresponding to positive eigenvalues. Let $\tilde{\mathbf{T}} = \tilde{\mathbf{\Psi}}_A \tilde{\mathbf{\Lambda}}_A \tilde{\mathbf{\Psi}}_A{}'$.
- Define
$$\mathbf{\Psi}_A^{FAST} = \mathbf{KQ}\tilde{\mathbf{\Psi}}_A.$$
- Then our fPCA approximation of the covariance matrix is
$$\widehat{\mathbf{\Sigma}}_A^{FAST} = \mathbf{\Psi}_A^{FAST}\tilde{\mathbf{\Lambda}}_A\mathbf{\Psi}_A^{FAST\prime}.$$

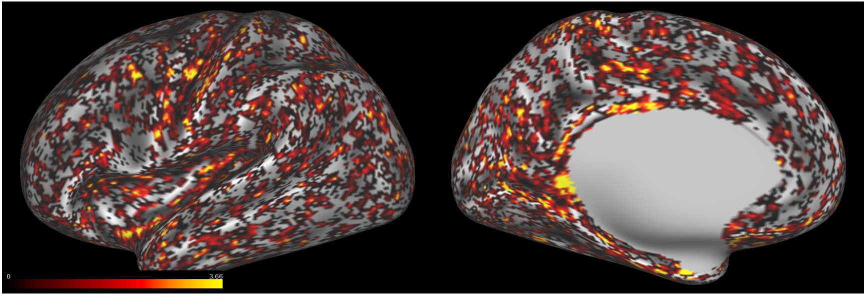


Figure : Point-wise likelihood ratio statistic for model without versus with additive genetic variance.

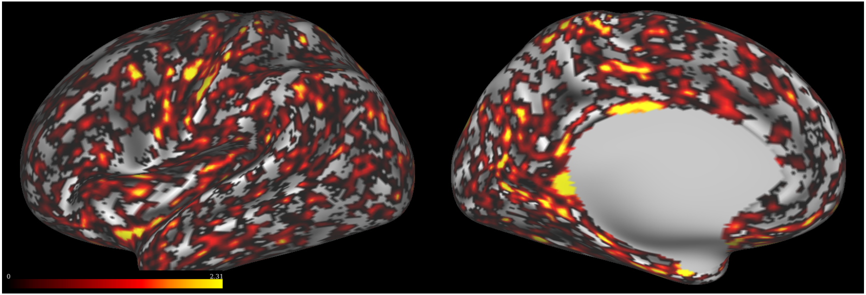


Figure : Point-wise weighted likelihood ratio statistic for model without versus with additive genetic variance.

Preliminary results

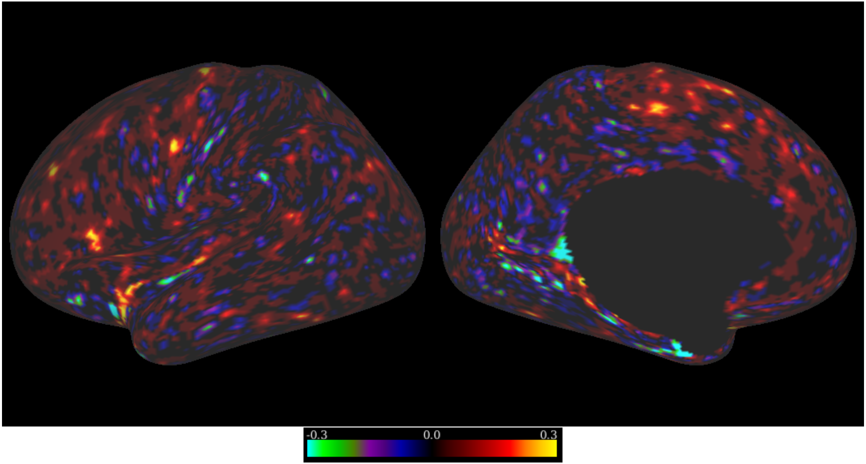
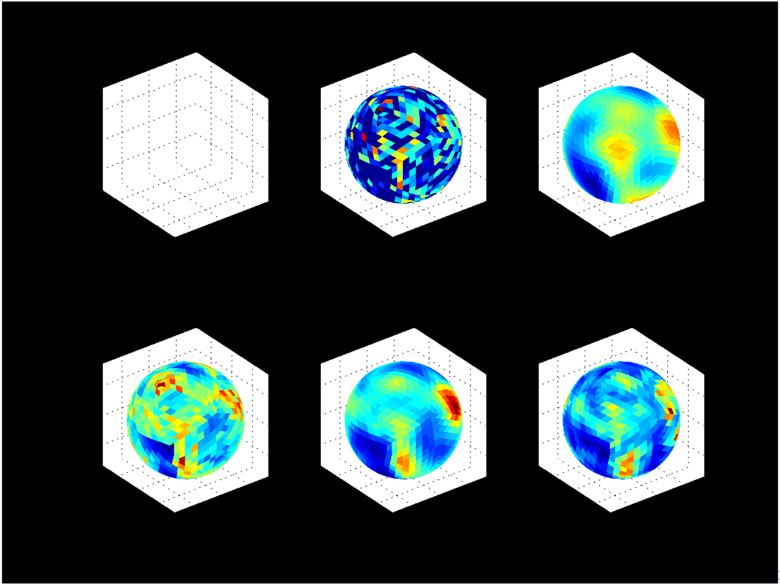
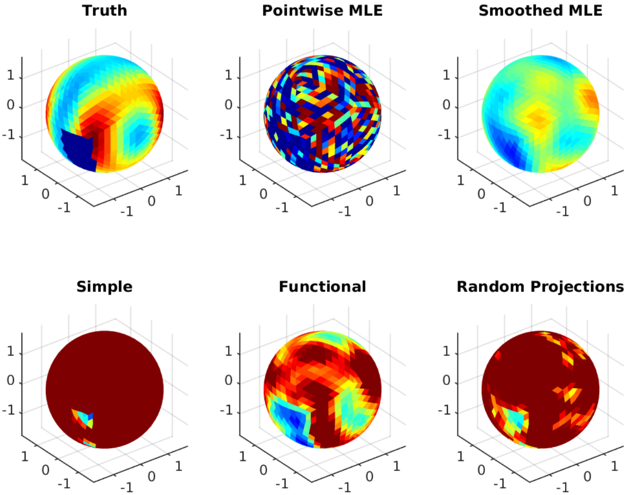


Figure : Additive genetic effect estimated from pointwise SmMLE for subject 100106.

Example from single simulation: $\text{diag}\widehat{\Sigma}_a$



Example from a single simulation: $\mathrm{diag}\widehat{\Sigma}_a$



Simulation design and example: $\widehat{\Sigma}_c$