# Environmental and Genetic Effects in High Dimensional Imaging Data

Benjamin B. Risk[a]
Joint work with Hongtu Zhu[b]

[a] SAMSI and the University of North Carolina, Chapel Hill and Emory University
[b] MD Anderson Cancer Center

samsi
NSF▾Duke▾NCSU▾UNC

Statistical Methods in Imaging Conference
University of Pittsburgh
2 June 2017

## Importance of heritability studies

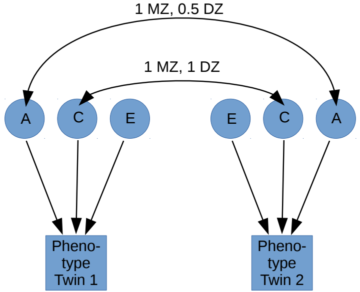
- Are brain traits environmentally or genetically determined?
- Create images of overall genetic effects.
- Classic twin designs [Fisher, 1919] important for future molecular genetic studies [Van Dongen et al., 2012].
- Heritability of brain and psychological disorders: 0.93 for bipolar [Kieseppä et al., 2004], 0.82 schizophrenia [Kendler, 2001], 0.74 Alzheimer's [Gatz et al., 1997], 0.25-0.76 multiple sclerosis [Hawkes and Macgregor, 2009], 0.33 major depression [Kendler, 2001].
- Here we study healthy young adults to understand heritability of brain traits.

## ACE Model

Fisher's model for polygenic effects on a phenotype:
**A**dditive genetic, **C**ommon environmental, and unique **E**nvironmental

- No dominant effects (non-additive), gene-gene interaction (epistasis), or gene-environment interaction.
- No assortative mating.

Figure: Path diagram for the SEM. MZ: monozygotic. DZ: dizygotic. Heritability defined as $\sigma_a^2/(\sigma_a^2 + \sigma_c^2 + \sigma_e^2)$.



# Surfaces: White and Pial

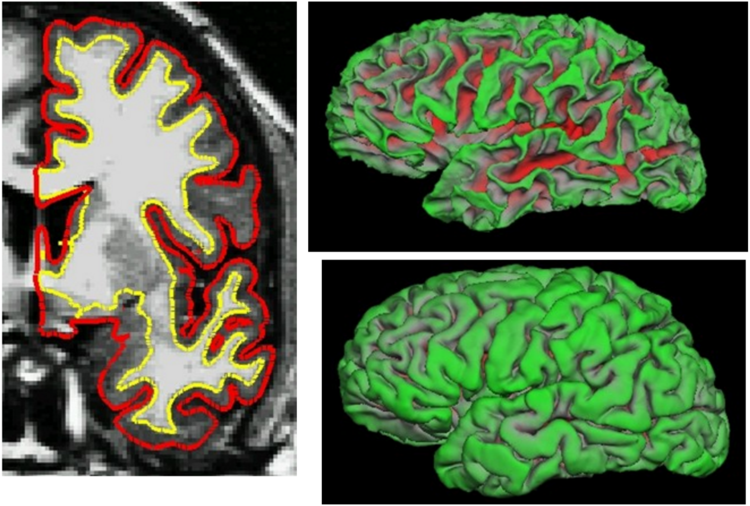
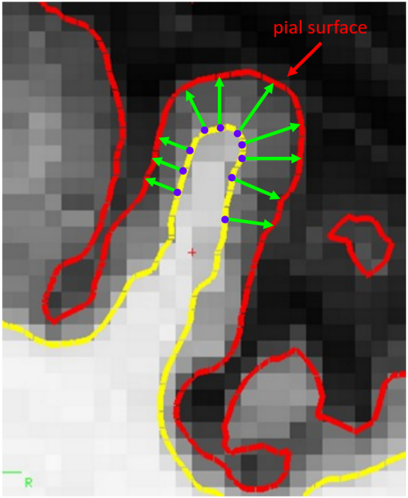


Figure: Source: Introduction to Freesurfer. In HCP preprocessed data, Freesurfer is used to delineate cortical thickness for 0.7 mm voxels.

# Cortical Thickness




- Distance between white and pial surfaces along normal vector.
- 1-5mm

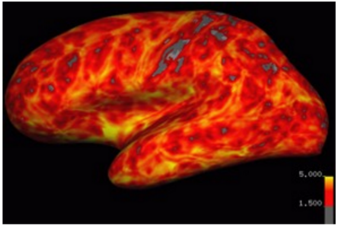



Figure: Source: Introduction to Freesurfer.

# SURFACE REGISTRATION: VERTICES

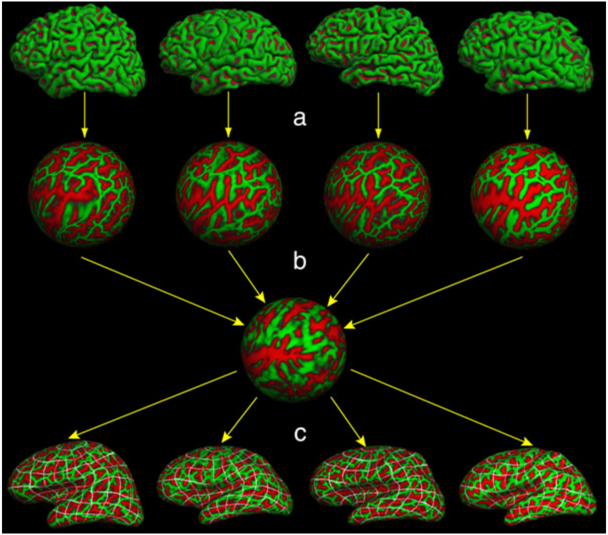



Figure: Source: Introduction to Freesurfer.

# Data Example

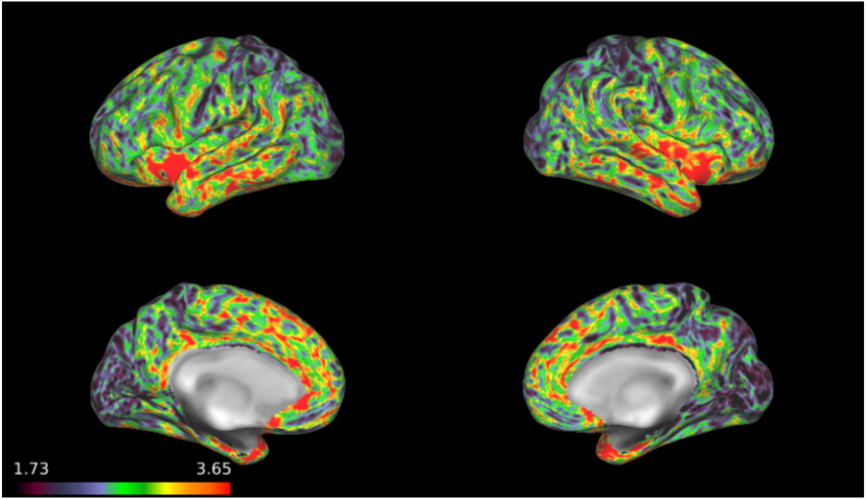
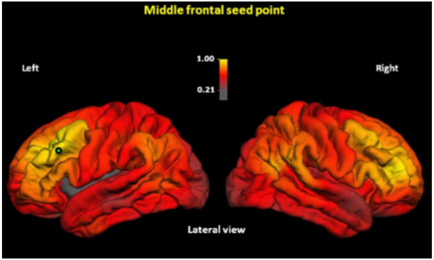



Figure: Cortical thickness (mm) in the left hemisphere from subject 101006 from the Human Connectome Project.

## Importance of cortical thickness

- Cortical thickness is important to intelligence [Karama et al., 2011].
- Cortical thinning is associated with dementia [Dickerson et al., 2009].
- Cortical network: correlations between cortical thickness.
- Abnormalities in cortical networks have been associated with psychiatric disorders such as depression [Wang et al., 2016].
- Develop an atlas of genetic patterning in cortical thickness.

# Structural networks in the literature



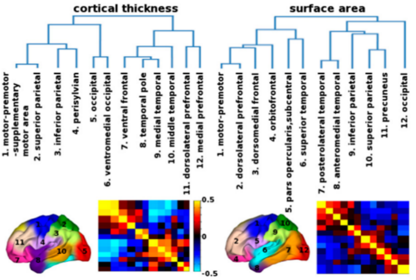




Figure: Genetic correlations from a seed in the middle frontal gyrus (Figure 3 in [Rimol et al., 2010]) and structural networks [Chen et al., 2013].

## Shortcomings of current methods

- Previous approaches: massive bivariate approach [Rimol et al., 2010] with large bias from smoothing.
- Recent improvements for small data: [Luo et al., 2017] FSEM for linear space, 93 locations, 300 subjects.
- [Luo et al., 2017] estimate a symmetric function rather than PSD.
- Problematic when $V >> N$, ultra-high dimensional setting.

## FSEM

[Luo et al., 2017]: Functional structural equation model (ACE) for $v \in [0, 1]$:

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

with

$$
\begin{aligned}
a_i(v) &\sim (0, \boldsymbol{\Sigma}_a(v,v)), \\
a_{ij}(v) &\sim (0, \boldsymbol{\Sigma}_a(v,v)) \\
c_i(v) &\sim (0, \boldsymbol{\Sigma}_c(v,v)) \\
e_{ij,G}(v) &\sim (0, \boldsymbol{\Sigma}_{e,G}(v,v)) \\
e_{ij,M}(v) &\sim (0, \sigma^2_{e,M}(v)).
\end{aligned}
$$

Measurement-error corrected heritability:

$$
h^2(v) = \boldsymbol{\Sigma}_a(v,v)/(\boldsymbol{\Sigma}_a(v,v) + \boldsymbol{\Sigma}_c(v,v) + \boldsymbol{\Sigma}_{e,G}(v,v))
$$

## Symmetric function from [Luo et al., 2017]

$$\hat{U}_{ijv_0v_0'} = R_{ijv_0}R_{ijv_0'}$$
$$\hat{U}_{iv_0v_0'} = 0.5\left(R_{i1v_0}R_{i2v_0'} + R_{i1v_0'}R_{i2v_0}\right)$$

Class of symmetric functions, optimization independent for each pair $\{v, v'\}$:

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

$$\text{(self)}\quad \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{v_0\neq v_0'}\left\{\hat{U}_{ijv_0v_0'} - \boldsymbol{\Sigma}_a(v,v') - \boldsymbol{\Sigma}_c(v,v') - \boldsymbol{\Sigma}_{e,G}(v,v')\right\}^2 k_h(v_0-v)k_h(v_0'-v')$$

$$\text{(MZ)}\quad + \frac{1}{n_1}\sum_{i=1}^{n_1}\sum_{v_0\neq v_0'}\left\{\hat{U}_{iv_0v_0'} - \boldsymbol{\Sigma}_a(v,v') - \boldsymbol{\Sigma}_c(v,v')\right\}^2 k_h(v_0-v)k_h(v_0'-v')$$

$$\text{(DZ)}\quad + \frac{1}{n_2}\sum_{i=n_1+1}^{n_1+n_2}\sum_{v_0\neq v_0'}\left\{\hat{U}_{iv_0v_0'} - 0.5\boldsymbol{\Sigma}_a(v,v') - \boldsymbol{\Sigma}_c(v,v')\right\}^2 k_h(v_0-v)k_h(v_0'-v').$$

## CONTRIBUTIONS

- **COV**ariance **F**unction **E**stimation in the **F**unctional structural **E**quation model for spatial domains in high dimensional setting: 60,000 locations with 1094 subjects.

## Contributions

- **COV**ariance **F**unction **E**stimation in the **F**unctional structural **E**quation model for spatial domains in high dimensional setting: 60,000 locations with 1094 subjects.
- COVFEFE

## CONTRIBUTIONS

- **COV**ariance **F**unction **E**stimation in the **F**unctional structural **E**quation model for spatial domains in high dimensional setting: 60,000 locations with 1094 subjects.
- PSD-ACE

## CONTRIBUTIONS

- **COV**ariance **F**unction **E**stimation in the **F**unctional structural **E**quation model for spatial domains in high dimensional setting: 60,000 locations with 1094 subjects.
- PSD-ACE
- Automate smoothing via geodesic kernels.

## CONTRIBUTIONS

- **COV**ariance **F**unction **E**stimation in the **F**unctional structural **E**quation model for spatial domains in high dimensional setting: 60,000 locations with 1094 subjects.
- PSD-ACE
- Automate smoothing via geodesic kernels.
- Improve estimates of heritability.

## CONTRIBUTIONS

- **COV**ariance **F**unction **E**stimation in the **F**unctional structural **E**quation model for spatial domains in high dimensional setting: 60,000 locations with 1094 subjects.
- PSD-ACE
- Automate smoothing via geodesic kernels.
- Improve estimates of heritability.
- Create an atlas of the genetic and environmental networks.

## New estimators with positive semi-definite constraints

- PSD estimators: truncating to positive eigenvalues should decrease MISE, also allows prediction.
- However, we found that this leads to huge bias!
- Derive alternative estimators. Let $\boldsymbol{\Sigma}_a(v, v') = (\mathbf{z}_v^a)^T \mathbf{z}_{v'}^a$, and $\boldsymbol{\Sigma}_c(v, v') = (\mathbf{z}_v^c)^T \mathbf{z}_{v'}^c$, and $\boldsymbol{\Sigma}_{e,G}(v, v') = (\mathbf{z}_v^{e,G})^T \mathbf{z}_{v'}^{e,G}$.
- Here, $\mathbf{z}_v^a \in \mathbb{R}^{n_1+n_2}$; let $\hat{U}^*_{ijv_0v_0'} = \hat{U}_{ijv_0v_0'} - \{\mathbb{1}_{v_0=v_0'}\}\hat{\sigma}^2_{e,M}(v_0)$.

$$\mathcal{J}^{PSD} = \underset{\mathbf{Z}_a \in \mathbb{R}^{V\times d_a}, \mathbf{Z}_c \in \mathbb{R}^{V\times d_c}}{\operatorname{argmin}}$$

$$\text{(self)}\ \frac{1}{N}\sum_{i,j}\sum_{v,v'}\sum_{v_0,v_0'}\left\{\hat{U}^*_{ijv_0v_0'} - (\mathbf{z}_v^a)^T\mathbf{z}_{v'}^a - (\mathbf{z}_v^c)^T\mathbf{z}_{v'}^c - (\mathbf{z}_v^{e,G})^T\mathbf{z}_{v'}^{e,G}\right\}^2 k_h(v_0, v)k_h(v_0', v')$$

$$\text{(MZ)}\ + \frac{1}{n_1}\sum_{i=1}^{n_1}\sum_{v,v'}\sum_{v_0,v_0'}\left\{\hat{U}_{iv_0v_0'} - (\mathbf{z}_v^a)^T\mathbf{z}_{v'}^a - (\mathbf{z}_v^c)^T\mathbf{z}_{v'}^c\right\}^2 k_h(v_0, v)k_h(v_0', v')$$

$$\text{(DZ)}\ + \frac{1}{n_2}\sum_{i=n_1+1}^{n_1+n_2}\sum_{v,v'}\sum_{v_0,v_0'}\left\{\hat{U}_{iv_0v_0'} - 0.5(\mathbf{z}_v^a)^T\mathbf{z}_{v'}^a - (\mathbf{z}_v^c)^T\mathbf{z}_{v'}^c\right\}^2 k_h(v_0, v)k_h(v_0', v')$$

- Information: (1) psd; (2) smoothness.
- Decomposition $\boldsymbol{\Sigma}_a = \mathbf{Z}_a\mathbf{Z}_a^T$ not unique, but in practice convergence is not an issue.

- Costly objective function, $O(V^4)$, greater than $1.3 \times 10^{19}$!
- We can not evaluate the objective function.
- Remarkably, we can optimize it.
- We derived a gradient-descent algorithm.
- Parameter space dramatically reduced because rank is $n_1 + n_2$ (229) for $\widehat{\boldsymbol{\Sigma}}_a$, $\widehat{\boldsymbol{\Sigma}}_c$, and $N - n_1$ (943) for $\widehat{\boldsymbol{\Sigma}}_{e,G}$.
- Initialize with truncated symmetric, “Sandwich” estimator.
- Updates are $O(V^2 N)$.

**Proposition**

$\widehat{\boldsymbol{\Sigma}}_a$ *is psd in the sense that for any* $\boldsymbol{v} \in \mathcal{M}^p$, $\mathbf{x} \in \mathbb{R}^p$, $\mathbf{x}^T \widehat{\boldsymbol{\Sigma}}_a(\boldsymbol{v}, \boldsymbol{v})\mathbf{x} \geq 0$.

## Gradient descent for covariance estimation

Input: The $N \times V$ data matrix $\mathbf{Y}$ and design matrix $\mathbf{X}$; tolerance $\epsilon$, 0.0001; step size $\lambda$, 0.001.

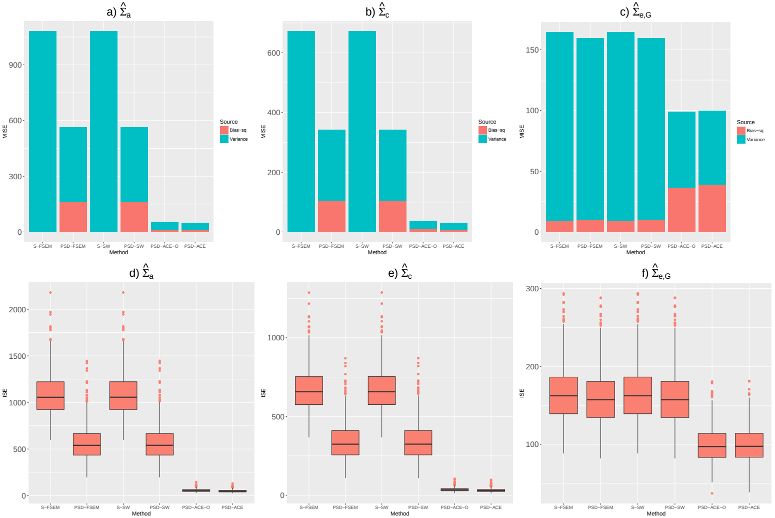
Result: $\widehat{\boldsymbol{\Sigma}}_a^{PSD-ACE}$, $\widehat{\boldsymbol{\Sigma}}_c^{PSD-ACE}$, and $\widehat{\boldsymbol{\Sigma}}_{e,G}^{PSD-ACE}$.

1. Estimate measurement error, $\hat{\boldsymbol{\sigma}}_{e,M}$, and residuals, $\widehat{\mathbf{R}}$, using SMLE with input $\mathbf{Y}$ and $\mathbf{X}$.
2. Calculate $\widehat{\boldsymbol{\Sigma}}_a^{S-SW}$, $\widehat{\boldsymbol{\Sigma}}_c^{S-SW}$, and $\widehat{\boldsymbol{\Sigma}}_{e,G}^{S-SW}$ in which the bandwidths are chosen using GCV. These bandwidths will be used in subsequent estimators.
3. Choose the rank $d_a$ based on the scree plot for $\widehat{\boldsymbol{\Sigma}}_a^{S-SW}$. Use the selected eigenvalue-eigenvector pairs to generate an initial value $\mathbf{Z}_a^{(0)}$. Repeat this process for $\mathbf{Z}_c^{(0)}$ and $\mathbf{Z}_{e,G}^{(0)}$.
4. Calculate $\nabla_a^{(0)}$, $\nabla_c^{(0)}$, and $\nabla_{e,G}^{(0)}$ using the initial values and calculate $\alpha = \sqrt{||\nabla_a^{(0)}||_F^2 + ||\nabla_c^{(0)}||_F^2 + ||\nabla_{e,G}^{(0)}||_F^2}$.
5. While $\sqrt{||\nabla_a^{(n)}||_F^2 + ||\nabla_c^{(n)}||_F^2 + ||\nabla_{e,G}^{(0)}||_F^2} > \epsilon\alpha$, increment $n$ and calculate $\mathbf{Z}_a^{(n)} = \mathbf{Z}_a^{(n-1)} - \lambda\nabla_a^{(n-1)}$, and similarly for $\mathbf{Z}_c^{(n)}$ and $\mathbf{Z}_{e,G}^{(n)}$.
6. Calculate the covariance functions.

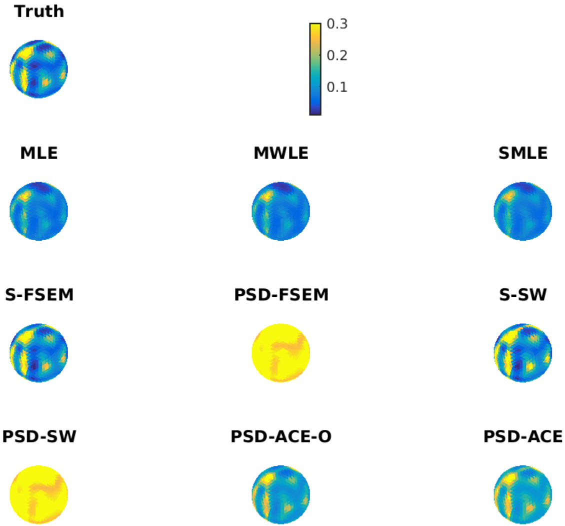
## Simulations: covariance functions

- Construct $\boldsymbol{\Sigma}_a$, $\boldsymbol{\Sigma}_c$, and $\boldsymbol{\Sigma}_{e,G}$ from sixth order even spherical harmonics (28 functions).
- For 100 MZ, 100 DZ, 200 singletons, simulate GP at 1002 locations
- Scaled to match empirical estimates from HCP analysis, resulting in heritability ranging from 0.016 to 0.498 with mean equal to 0.126.
- $\hat{\sigma}^2_{e,M}$ spatially varying, average 0.03.

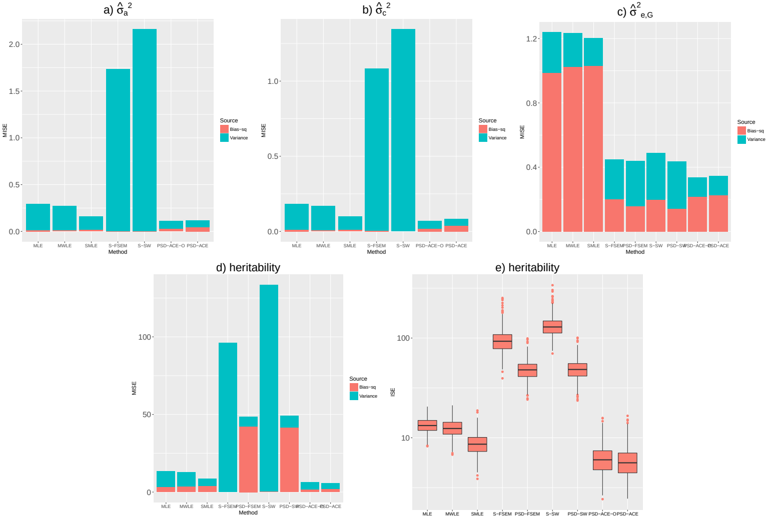
# Simulation Results



# VISUALIZING BIAS IN HERITABILITY



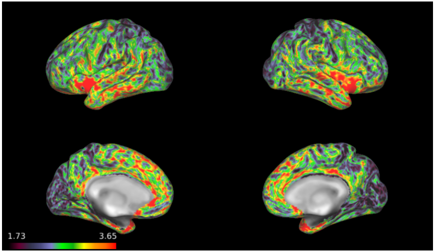
# Variances and heritability



## Conclusions from simulation studies

- PSD-ACE results in much lower MISE for heritabilities, variances, and covariances
- It does introduce some bias.
- Howevever, heritability is *less* biased than the MLE and MWLE.

## HCP Analysis




- HCP: map all structural and functional connections in the healthy brain.
- Preprocessed data from HCP [Glasser et al., 2013]: cortical thickness estimated using FreeSurfer.
- 1094 subjects, 595 females; 151 MZ pairs, 78 DZ pairs.
- No direct smoothing.
- Age: $28.8 \pm 3.7$.
- Assessed covariates: gender, age, handedness, height, weight, BMI, ICV.
- Kept gender, age and ICV.

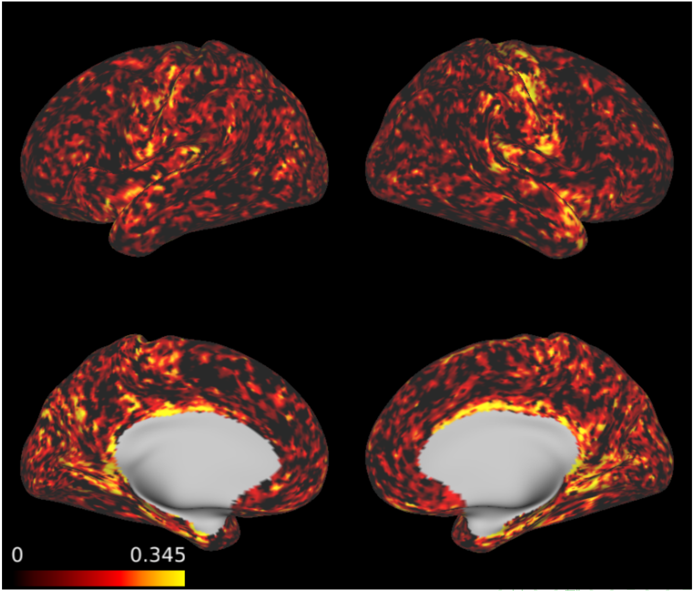
## Distances on the cortical manifold

- We use biweight kernels with geodesic distance in the group template 32k_fs_LR.
- Distance between hemispheres is infinite.
- Only affects the *local* smoothing – long-distance correlations learned from data.
- GCV selected bandwidths are very small: bw=1.3, average weights are 0.878, 0.044, 0.044, 0.015, 0.015, 0.002, and 0.002.

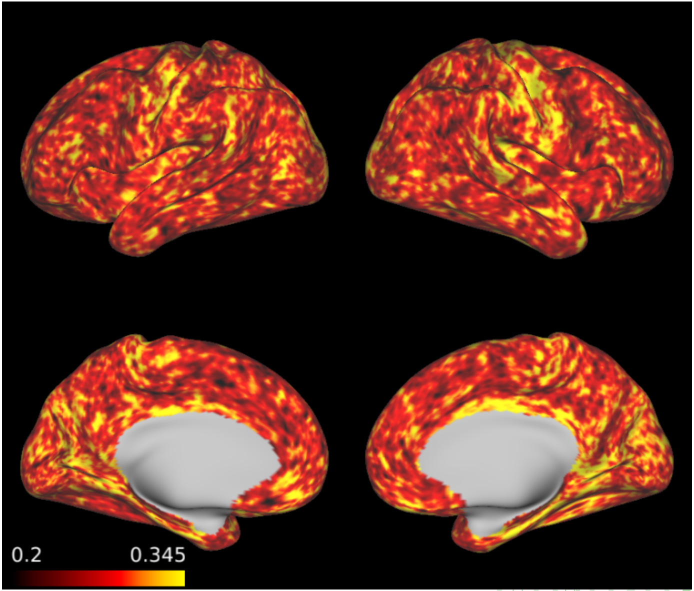
## Strategy for very big data

- Calculations involve four $V \times V$ dense matrices, plus other memory usage.
- We have estimated the model for the HCP data using HPC with 3 TB of RAM.
- Also have code with a divide and conquer approach: allows the use of high resolution atlases on personal computers.
- Divide and conquer may work for smoother data, e.g., fMRI?

# Heritability MLE estimates: $\sigma_a^2(v)/(\sigma_a^2(v) + \sigma_c^2(v) + \sigma_e^2(v))$



# Heritability PSD-ACE estimates: $\sigma_a^2(v)/(\sigma_a^2(v)+\sigma_c^2(v)+\sigma_{e,G}^2(v))$



# HERITABILITY ESTIMATES: RECENT LITERATURE

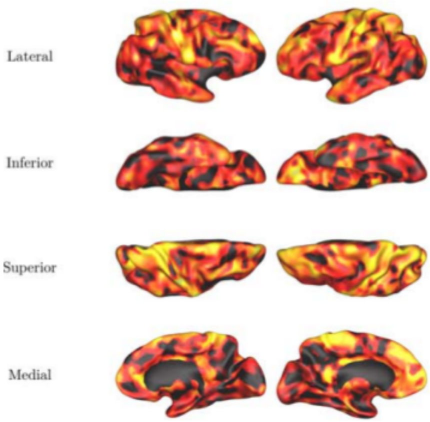



Figure: Heritability estimates from [Shen et al., 2016].

# Comparison with literature

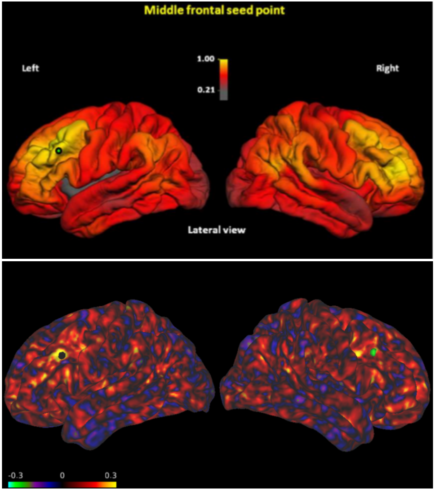
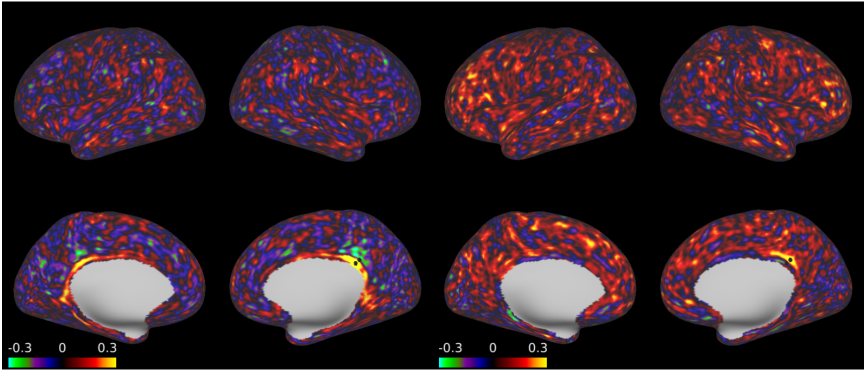
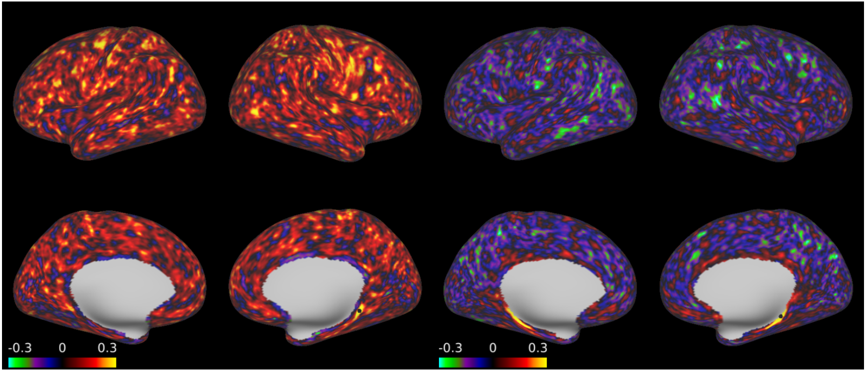



Figure: Genetic correlations from a seed in the middle frontal gyrus (top: Figure 3 in [Rimol et al., 2010]; bottom: PSD-ACE).

# Structural connectivity in genetic component, medial locations



# Hub versus isolated in genetic component



# DISCUSSION

- Average heritability was 0.28 in PSD-ACE (max 0.43) versus 0.09 (max 0.55) in MLE. Mostly due to PSD constraints.
- Automated smoothing using GCV – chooses small bandwidth.
- We present a data-principled approach to determine the bias-variance trade-off.
- We developed the first atlas of genetic patterning in cortical thickness networks.
- Future work: predict the genetic effects in individuals, which can then be related to the genetic components of behavior.

## ACKNOWLEDGMENTS

Thank you!

- Data were provided (in part) by the Human Connectome Project, WU-Minn Consortium funded NIH.
- BBR was partly supported by NSF grant DMS-1127914 to SAMSI and the National Institute of Mental Health of the National Institutes of Health under award number T32MH106440.

# References I


Chen, C.-H., Fiecas, M., Gutierrez, E., Panizzon, M. S., Eyler, L. T., Vuoksimaa, E., Thompson, W. K., Fennema-Notestine, C., Hagler, D. J., Jernigan, T. L., et al. (2013).
Genetic topography of brain morphology.
*Proceedings of the National Academy of Sciences*, 110(42):17089–17094.

Dickerson, B. C., Bakkour, A., Salat, D. H., Feczko, E., Pacheco, J., Greve, D. N., Grodstein, F., Wright, C. I., Blacker, D., Rosas, H. D., et al. (2009).
The cortical signature of Alzheimer's disease: regionally specific cortical thinning relates to symptom severity in very mild to mild ad dementia and is detectable in asymptomatic amyloid-positive individuals.
*Cerebral cortex*, 19(3):497–510.

Fisher, R. A. (1919).
XV. The correlation between relatives on the supposition of Mendelian inheritance.
*Transactions of the royal society of Edinburgh*, 52(02):399–433.

Gatz, M., Pedersen, N. L., Berg, S., Johansson, B., Johansson, K., Mortimer, J. A., Posner, S. F., Viitanen, M., Winblad, B., and Ahlbom, A. (1997).
Heritability for alzheimer's disease: the study of dementia in swedish twins.
*The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 52(2):M117–M125.

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., et al. (2013).
The minimal preprocessing pipelines for the Human Connectome Project.
*NeuroImage*, 80:105–124.

Hawkes, C. and Macgregor, A. (2009).
Twin studies and the heritability of ms: a conclusion.
*Multiple Sclerosis*, 15(6):661–667.

Karama, S., Colom, R., Johnson, W., Deary, I. J., Haier, R., Waber, D. P., Lepage, C., Ganjavi, H., Jung, R., Evans, A. C., et al. (2011).
Cortical thickness correlates of specific cognitive performance accounted for by the general factor of intelligence in healthy children aged 6 to 18.
*Neuroimage*, 55(4):1443–1453.

# References II

Kendler, K. S. (2001).
Twin studies of psychiatric illness: an update.
*Archives of general psychiatry*, 58(11):1005–1014.

Kieseppä, T., Partonen, T., Haukka, J., Kaprio, J., and Lönnqvist, J. (2004).
High concordance of bipolar i disorder in a nationwide sample of twins.
*American Journal of Psychiatry*.

Luo, S., Song, R., Styner, M., Gilmore, J. H., and Zhu, H. (2017).
FSEM: Functional structural equation model for twin functional data.
*In review*.

Rimol, L. M., Panizzon, M. S., Fennema-Notestine, C., Eyler, L. T., Fischl, B., Franz, C. E., Hagler, D. J., Lyons, M. J., Neale, M. C., Pacheco, J., et al. (2010).
Cortical thickness is influenced by regionally specific genetic factors.
*Biological psychiatry*, 67(5):493–499.

Shen, K.-K., Dore, V., Rose, S., Fripp, J., McMahon, K. L., de Zubicaray, G. I., Martin, N. G., Thompson, P. M., Wright, M. J., and Salvado, O. (2016).
Heritability and genetic correlation between the cerebral cortex and associated white matter connections.
*Human brain mapping*, 37(6):2331–2347.

Van Dongen, J., Slagboom, P. E., Draisma, H. H., Martin, N. G., and Boomsma, D. I. (2012).
The continuing value of twin studies in the omics era.
*Nature Reviews Genetics*, 13(9):640–653.

Wang, T., Wang, K., Qu, H., Zhou, J., Li, Q., Deng, Z., Du, X., Lv, F., Ren, G., Guo, J., et al. (2016).
Disorganized cortical thickness covariance network in major depressive disorder implicated by aberrant hubs in large-scale networks.
*Scientific reports*, 6.

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.

## Infinite dimensional problem

$$
\begin{aligned}
U_{ijv_0v_0'} &= R_{ijv_0}R_{ijv_0'} \\
U_{iv_0v_0'} &= 0.5\left(R_{i1v_0}R_{i2v_0'} + R_{i1v_0'}R_{i2v_0}\right)
\end{aligned}
$$

$$
\begin{aligned}
&\underset{\boldsymbol{\Sigma}_a,\boldsymbol{\Sigma}_c,\boldsymbol{\Sigma}_{e,G}:\mathcal{M}\mapsto\mathcal{F}^+}{\operatorname{argmin}} \\
&\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\int_{v,v'\in\mathcal{M}}\sum_{v_0\neq v_0'}\left\{U_{i,j,v_0,v_0'} - \boldsymbol{\Sigma}_a(v,v') - \boldsymbol{\Sigma}_c(v,v') - \boldsymbol{\Sigma}_{e,G}(v,v')\right\}^2 k_h(v_0,v;v_0',v')d\mathcal{M} \\
&+\frac{1}{n_1}\sum_{i=1}^{n_1}\int_{v,v'\in\mathcal{M}}\sum_{v_0,v_0'}\left\{U_{i,v_0,v_0'} - \boldsymbol{\Sigma}_a(v,v') - \boldsymbol{\Sigma}_c(v,v')\right\}^2 k_h(v_0,v;v_0',v')d\mathcal{M} \\
&+\frac{1}{n_2}\sum_{i=1}^{n_2}\int_{v,v'\in\mathcal{M}}\sum_{v_0,v_0'}\left\{U_{i,v_0,v_0'} - 0.5\boldsymbol{\Sigma}_a(v,v') - \boldsymbol{\Sigma}_c(v,v')\right\}^2 k_h(v_0,v;v_0',v')d\mathcal{M}.
\end{aligned}
$$

## LINEAR COMBINATIONS OF SAMPLE COVARIANCES

- "Sample" covariances

$$\text{All: } \mathbf{S}_0 = \frac{1}{N}\left(\mathbf{R}^T\mathbf{R}\right)$$
$$\text{MZs: } \mathbf{S}_1 = \frac{1}{2n_1}\left(\mathbf{R}_{11}^T\mathbf{R}_{12} + \mathbf{R}_{12}^T\mathbf{R}_{11}\right)$$
$$\text{DZs: } \mathbf{S}_2 = \frac{1}{2n_2}\left(\mathbf{R}_{21}^T\mathbf{R}_{22} + \mathbf{R}_{22}^T\mathbf{R}_{21}\right).$$

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 $\mathbf{S}_a^+$ and $\mathbf{S}_c^+$ 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, $\mathbf{KSK}^T$.
- Facilitates use of GCV:

$$GCV(h) = ||(\boldsymbol{k}_h \otimes \boldsymbol{k}_h)\ \mathrm{vec}(\mathbf{S}) - \mathrm{vec}(\mathbf{S})||^2/(1 - \mathrm{tr}(\mathbf{K} \otimes \mathbf{K})/V^2)^2$$

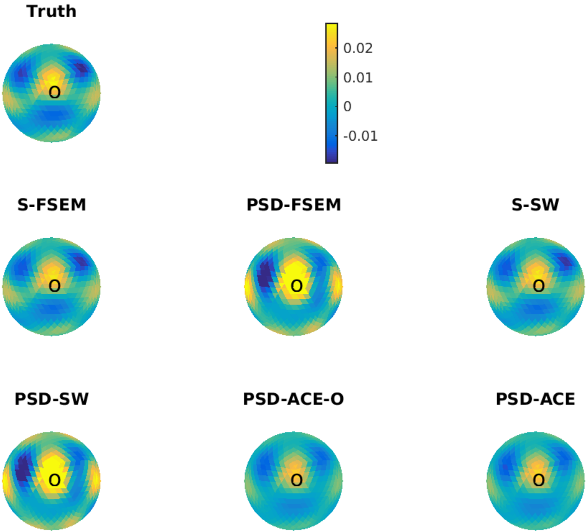
- For twin studies, we have multiple covariance functions to estimate.
- We propose the sandwich formulation of local constant regression estimators.
- Define $\mathbf{K}$ such that $\mathbf{K}_{k,l} = k_h(v_k, v_l)/\sum_{l=1}^{V} k_h(v_k, v_l)$. Then

$$\widehat{\boldsymbol{\Sigma}}_a^{LCR} = \mathbf{K}\mathbf{S}_a^+\mathbf{K}^T \tag{1}$$

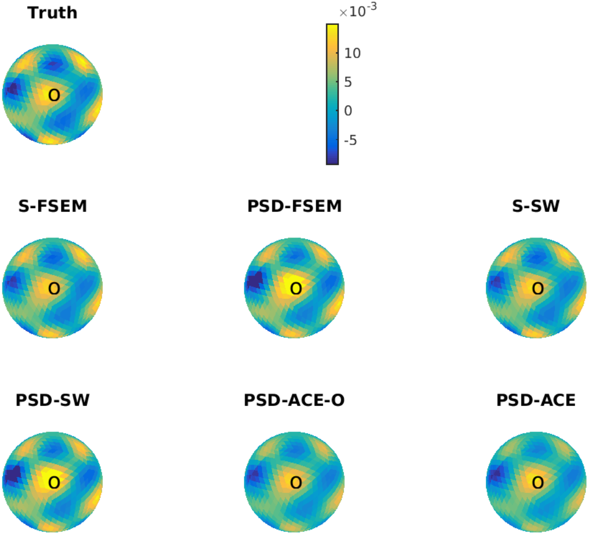
$$\widehat{\boldsymbol{\Sigma}}_c^{LCR} = \mathbf{K}\mathbf{S}_c^+\mathbf{K}^T. \tag{2}$$

- Smooth eigenvectors only: $(\mathbf{K}\boldsymbol{\Psi}_a{}^+)\boldsymbol{\Lambda}_a{}^+(\boldsymbol{\Psi}_a{}^{+T}\mathbf{K}^T)$.

## VISUALIZING BIAS IN GENETIC COVARIANCE

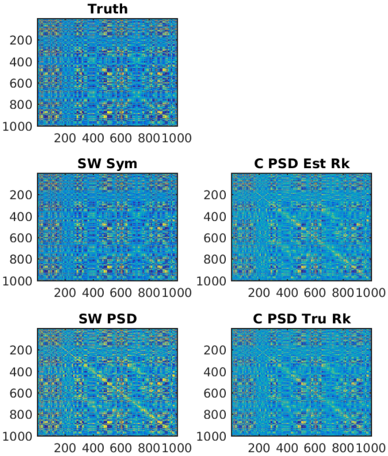


# VISUALIZING BIAS IN GENETIC COVARIANCE

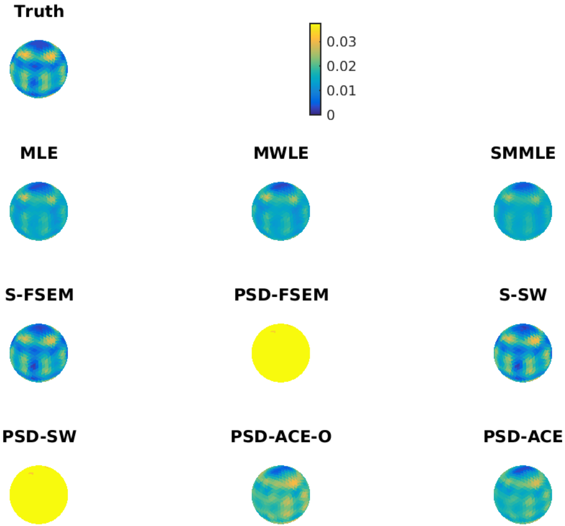


# Average of covariance estimates across simulations.

Figure: Examining bias. Covariance estimates of $\mathbf{\Sigma}_a$ averaged across simulations.



## Visualizing bias in genetic variance



## Geodesic distances

Alternatives to great circle distance:

1. geodesic distance along a group-averaged cortex
   - not advised because folding patterns are averaged resulting in a smoothed surface with distances affected in undesirable ways
2. registered individual surfaces
   - unclear if it would improve or degrade performance
   - creates additional mathematical and computational challenges.