Multi-modal Association Testing, with Applications to Imaging Genetics

Dustin Pluta, Tong Shen, Hernando Ombao, Zhaoxia Yu
University of California, Irvine

JSM 2017

Overview of Talk

1. Motivation from Imaging Genetics
2. The Mantel Test
3. Score Tests for Fixed Effects, Random Effects, and Ridge Regression
4. Connecting the Mantel and Score Tests
5. Simulation Study

Motivating Application

- **Imaging genetics** studies include genetic data (SNPs) and neuroimaging data (fMRI, EEG, DTI).
- We are interested in **testing for association** of **genetic similarity** with similarity of particular **neurological phenotypes**.
- This can be difficult since
  - The data is **high-dimensional**: 500K SNPs, 500K voxels at 0.5 Hz for fMRI data,
  - **Small effect sizes** distributed across many genetic locations,
  - The data is **noisy** with many possible **confounders**,
  - Results are sensitive to numerous **pre-processing** choices,

Motivating Application

The Data

- 209 subjects,
- 500K SNP values for each subject,
- fMRI readings from 375 parcellated regions for resting state, decision-making task, and working memory task,
- Behavioral data for decision-making and working memory tasks.

Cups Task

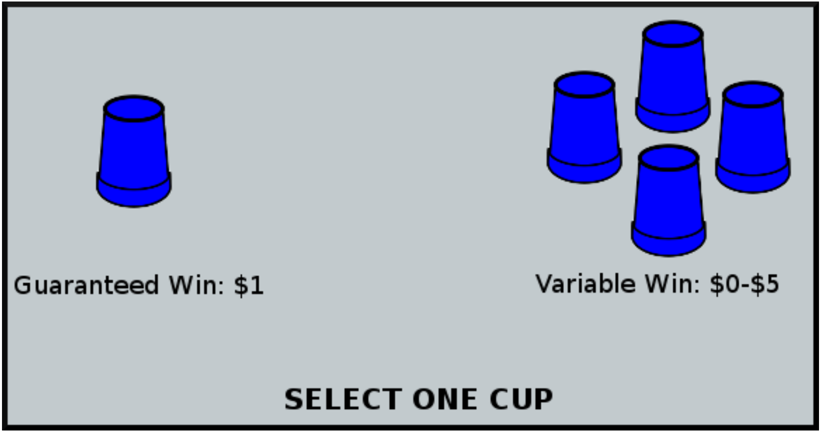


Figure 1: Example of Cups Task trial.

fMRI

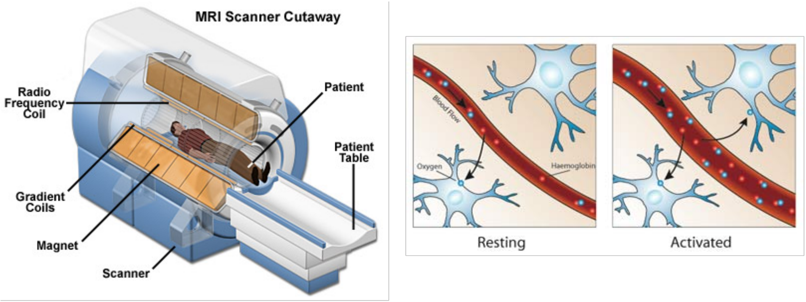


Figure 2: fMRI scanner diagram, and illustration of BOLD response.

Genetics

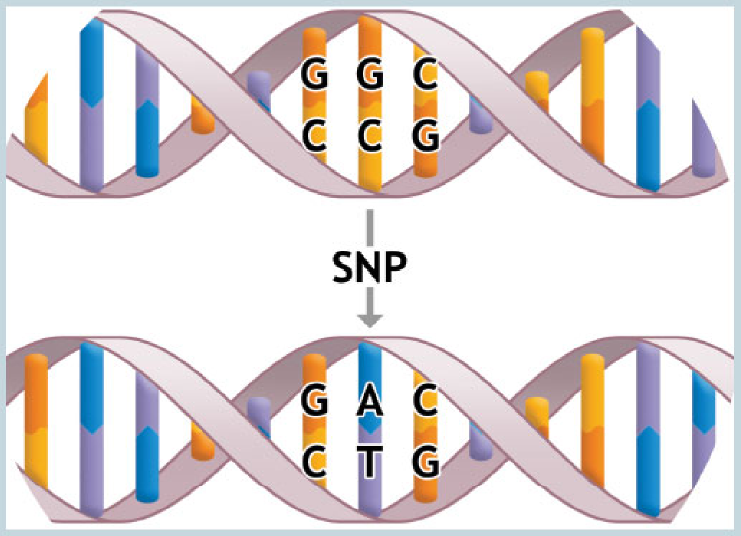


Figure 3: Illustration of a SNP.

Multi-modal Association Testing

The Inference Goal

Given a sample of N subjects containing two data modalities X and Y, is distance in X significantly associated with distance in Y?

Multi-modal Association Testing

Application Context

- For our application, X is SNP data and Y is a measure functional connectivity from fMRI.
- We wish to know: is **genetic similarity** significantly correlated with **similarity of functional connectivity**?
- This is closely related to the concept of *heritability* of phenotypes commonly considered in genetics studies.

Mantel Test

- Assume X is an $N \times P$ column-centered matrix and Y is an $N \times 1$ centered vector.
- Given metrics $d_X : \mathbb{R}^P \times \mathbb{R}^P \to \mathbb{R}$ and $d_Y : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, we can form two $N \times N$ **distance (or dissimilarity)** matrices K and H, where

$$K_{ij} = d_X(X_i, X_j)$$
$$H_{ij} = d_Y(Y_i, Y_j).$$

- The **correlation** of these distance matrices is

$$\rho = \frac{\langle K, H \rangle}{\|K\|_2 \|H\|_2},$$

Mantel Test

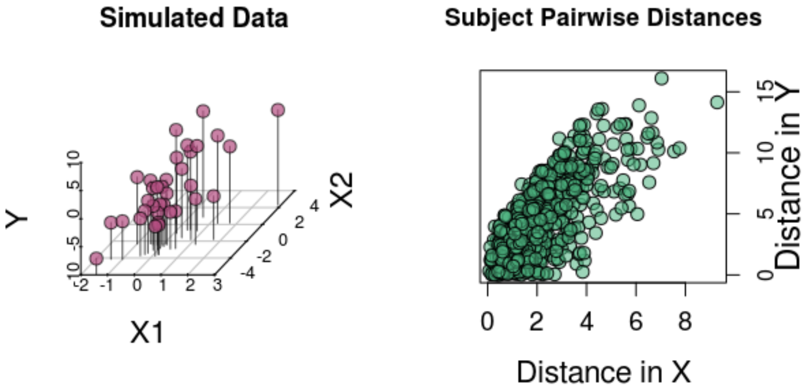


Figure 4: Simulated multi-modal data.

Mantel Test

Statistical Question

How should we test the significance of the correlation?

Mantel Test

Statistical Question

How should we test the significance of the correlation?

One Approach

The common approach, originally suggested by Mantel (1967), is to permute rows and columns of one of the matrices to generate the reference distribution.

Mantel Test

Classical Mantel Test Statistic (Mantel 1967)

Since $\|K\|$ and $\|H\|$ are constant under permutations, we can take the test statistic to be

$$Z = \frac{1}{2}\langle K, H\rangle = \sum_{i=1}^{N}\sum_{j>i} K_{ij}H_{ij}.$$

- **Note:** Since the diagonals for both H and K are 0, they do not affect the calculation of Z, so a total of $\binom{N}{2}$ pairwise distances are used.

Mantel Test

Mantel with Similarity Matrices

- The **Mantel test** is most often applied using **distance** or **dissimilarity matrices**.
- Some applications have used **similarity matrices**, but still used the same test statistic given above.
- However, with K and H as **similarity matrices**, we can instead use a **modified Mantel statistic**

$$Z^* = \langle K, H \rangle = \sum_{i=1}^{N} \sum_{j=1}^{N} K_{ij} H_{ij} = \mathrm{tr}(KH),$$

which uses $\binom{N+1}{2}$ inner products.

Mantel Test

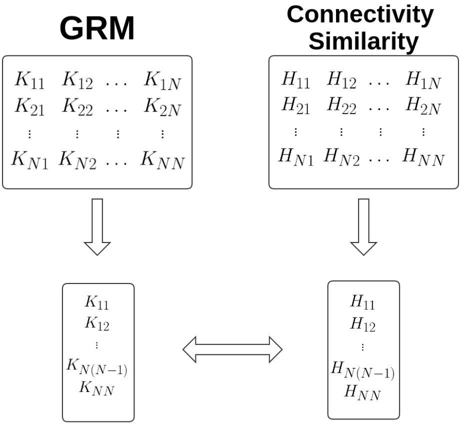
Classical Mantel and Modified Mantel

Classical Mantel $$Z = \sum_{i=1}^{N} \sum_{j>i} K_{ij} H_{ij}$$

Modified Mantel $$Z^* = \sum_{i=1}^{N} \sum_{j=1}^{N} K_{ij} H_{ij} = \mathrm{tr}(KH)$$

- If K and H are **distance matrices**, then $Z^* = 2Z$.
- If K and H instead contain the **inner products**, then Z^* will not be equivalent to Z whenever the diagonals of K and H are not constant.

Mantel Test



$$Z^* = \langle K, H \rangle \equiv \sum_i \sum_j K_{ij} H_{ij}$$

Figure 5: Diagram of the Mantel Test.

Mantel Test

Inner Product Similarity

$$K_{ij} = \langle X_i, X_j \rangle, \quad H_{ij} = \langle Y_i, Y_j \rangle$$

Similarity with General Inner Products

Allowing for **general inner products** defined by some positive semi-definite matrix $\mathcal{W}$, we can write

$$K = X\mathcal{W}X^T, \quad H = YY^T$$

Kernel Mantel Test Statistic

$$Z^* = \mathrm{tr}(X\mathcal{W}X^T YY^T) = Y^T X\mathcal{W}X^T Y = \|X^T Y\|^2_{\mathcal{W}}.$$

Mantel Test

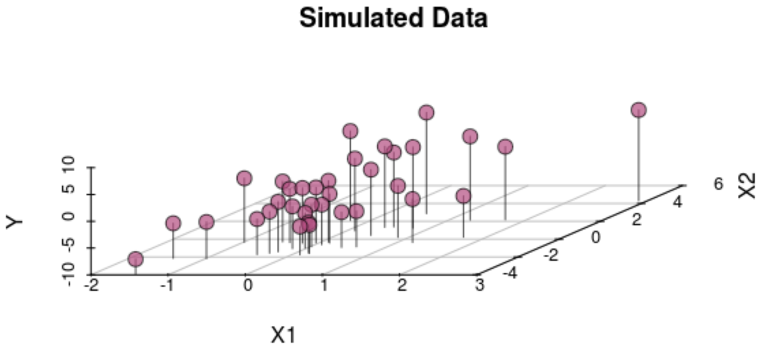


Figure 6: Simulated multi-modal data.

Mantel Test

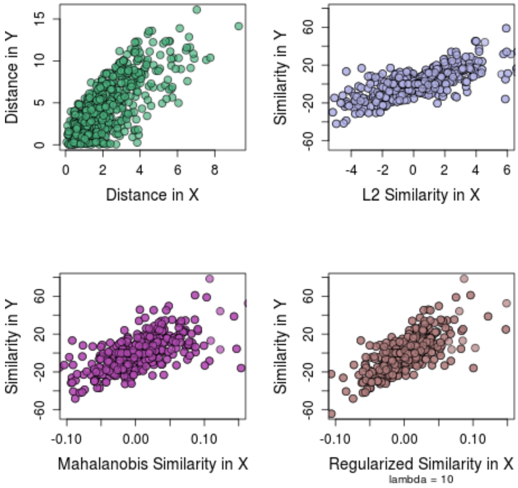


Figure 7: Comparison of L_2 distance and similarity measures.

The Score Test

Score Test Statistic

$$S = \mathcal{U}(\beta^0)^T \mathcal{I}^{-1}(\beta^0)\mathcal{U}(\beta^0).$$

Some nice properties of the **Score Test** are

- It uses only the **null hypothesis parameter values**,
- Can be useful when dealing with **boundaries of the parameter space**,
- It's the **most powerful test for small effect sizes**,
- It's **asymptotically equivalent** to the Wald and Likelihood Ratio tests.

Three Classes of Linear Models

Fixed Effects

$$Y \sim N(X\beta, \sigma_\varepsilon^2 I_N)$$

Ridge Regression

$$Y \sim N(X\beta, \sigma_\varepsilon^2 I_N), \qquad \|\beta\|_2^2 < c(\lambda)$$

Random Effects

$$Y \sim N(0, \sigma_b^2 K + \sigma_\varepsilon^2 I_N), \qquad K = XX^T$$

The Score Test: Fixed Effects Model

For the fixed effects model, the log-likelihood, score vector and Fisher Information are

$$\ell(\beta|\sigma_\varepsilon^2) \propto (Y - X\beta)^T(Y - X\beta) + c$$
$$\mathcal{U}(\beta|\sigma_\varepsilon^2) \propto X^T(Y - X\beta).$$
$$\mathcal{I}(\beta|\sigma_\varepsilon^2) \propto X^T X.$$

The resulting global score test for $H_0 : \beta = 0$ is

Fixed Effects Score Test

$$S_F = \frac{1}{\sigma_\varepsilon^2} Y^T X(X^T X)^{-1} X^T Y \mathrel{\dot\sim} \chi_r^2.$$

The Score Tests

Score Tests for Three Classes of Linear Models

Model	Equiv. Score Stat.†	Equiv. Norm
Fixed	$S_F = Y^T X(X^T X)^{-1} X^T Y$	$\|X^T Y\|_{\mathcal{M}}^2$
Ridge	$S_\lambda = Y^T X(X^T X + \lambda I)^{-1} X^T Y$	$\|X^T Y\|_{\mathcal{M}_\lambda}^2$
Random	$S_R = Y^T X X^T Y$	$\|X^T Y\|_2^2$

† These statistics yield equivalent P-values when using the permutation procedure to produce the reference distribution.

Connecting the Mantel and Score Tests

Similarity Mantel Test Statistic

The Similarity Mantel test statistics can be formulated as

$$Z^*_{\mathcal{W}} = \mathrm{tr}(K_{\mathcal{W}} H)$$

The Three Classes of Linear Models and Corresponding Kernels

Model	**Mantel Stat.**	**Kernel $\mathcal{W}$**
Fixed	$Z^*_F = \mathrm{tr}(YY^TX(X^TX)^{-1}X^T)$	$(X^TX)^{-1}$
Ridge	$Z^*_\lambda = \mathrm{tr}(YY^TX(X^TX+\lambda I)^{-1}X^T)$	$(X^TX+\lambda I)^{-1}$
Random	$Z^*_R = \mathrm{tr}(YY^TXX^T)$	I_P

Comparing Fixed, Random, and Penalized Models

Question

How does the statistical performance of the three classes of models compare for different values of N, P, and effect size?

- Intuitively, we may expect
  - **Fixed effects** is best when $N >> P$,
  - **Random effects** is best when $P >> N$,
  - **Penalized models** somewhere in between.

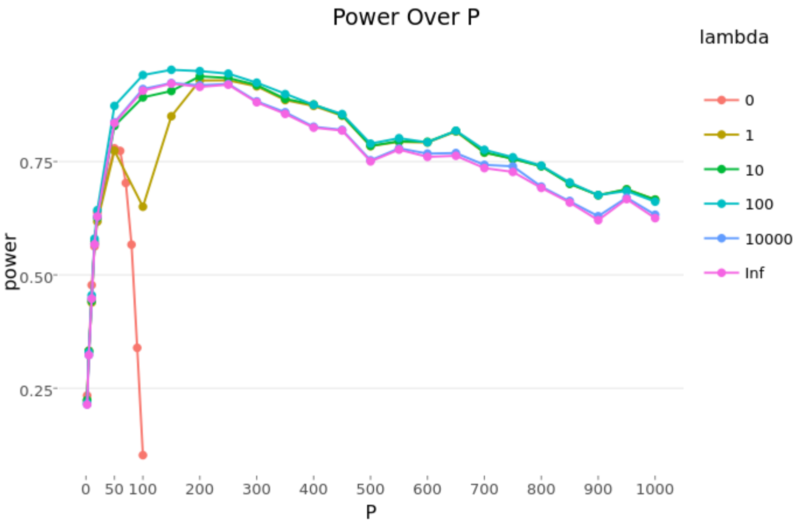
The Score Test

- We see that if we have **orthogonal design** with $X^T X \propto I_P$, then the fixed effects, random effects, and penalized models will give identical results.
- The discrepancy of testing with S_F vs S_R will increase as $X^T X$ deviates further from the identity.
- For penalized models, $S_0 = S_F$. To examine S_λ as $\lambda \to \infty$, apply the Woodbury formula

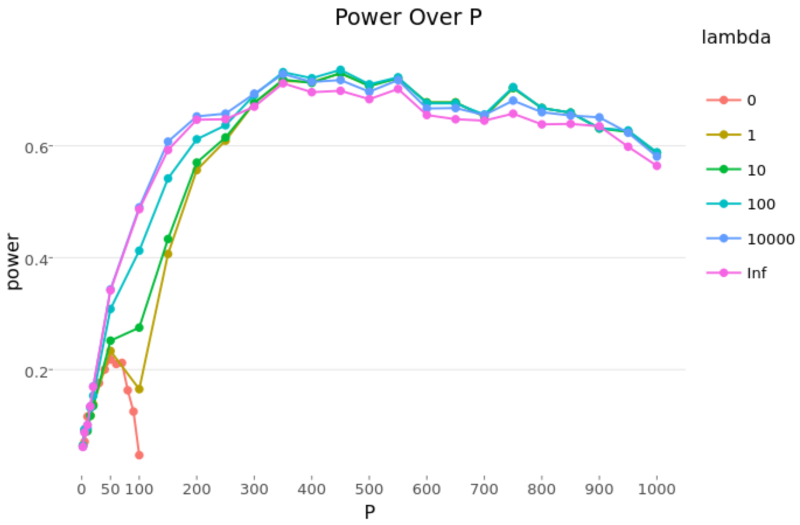
$$(X^T X + \lambda I)^{-1} = \frac{1}{\lambda} I_P - \frac{1}{\lambda^2} X^T (I_N + \frac{1}{\lambda} X X^T)^{-1} X$$

- Thus, for large λ, $S_\lambda \approx S_R$.

Simulation with Identity Covariance for X



Simulation with AR(1) + Diagonal Trend Cov. for X



Multivariate Mantel

- We can easily extend the Mantel test framework to accommodate multivariate responses.
- Suppose Y is an $N \times Q$ response matrix, and define the similarity matrices

$$
\begin{aligned}
K &= X(X^T X)^{-1} X^T \\
H &= Y(Y^T Y)^{-1} Y^T.
\end{aligned}
$$

Multivariate Mantel

- The Mantel test procedure can be performed exactly the same as before with test statistic $Z^* = \text{tr}(KH)$.
- Assuming $\text{rank}(K) = P$ and $\text{rank}(H) = Q$:

$$\text{tr}(KK) = \text{tr}(K) = \text{rank}(K) = P$$
$$\text{tr}(HH) = \text{tr}(H) = \text{rank}(H) = Q$$
$$\rho(K, H) = \frac{1}{\sqrt{PQ}}\text{tr}(KH) = \frac{1}{\sqrt{PQ}}Z^*$$

Summary

- The **Similarity Mantel Test** is **equivalent to the Score test** for a linear model whose form depends on the choice of inner product.
- Consequently, the **Similarity Mantel Test** is most powerful for small effects
- The **Mantel Test** implies an underlying parametric model through the choice of similarity or distance measure.
- The **Ridge Regression Score Test** converges to the **Random Effects Score Test** as $\lambda \to \infty$.
- The Mantel test can be easily extended to **multivariate response data**, and can accommodate multi-modal data of arbitrary type and arbitrary dimension in each mode.

Acknowledgements

- **Gui Xue**, PI, Center for Brain and Learning Sciences, Beijing Normal University
- **Chuansheng Chen**, Dept. of Psychology and Social Behavior, UCI
- **Hernando Ombao**, Dept. of Statistics, KAUST & UCI
- **Zhaoxia Yu**, Dept. of Statistics, UCI
- **Tong Shen**, Dept. of Statistics, UCI (PhD Student)

References

- Ge T, et al. Massively Expedited Genome-Wide Heritability Analysis (MEGHA). PNAS. 2015. 112, 2479-2484.
- Xue G, et al. Functional Dissociations of Risk and Reward Processing in the Medial Prefrontal Cortex. Cerebral Cortex. 2009. 19, 1019-1027.
- Yang J, et al. GCTA: A Tool for Genome-wide Complex Trait Analysis. The American Journal of Human Genetics. (2011) 88, 76-82.
- Tzeng et al. (2009) Biometrics 65, 822.
- Visscher et al. (2014) Statistical power to detect genetic (co)variance of complex traits using SNP data in unrelated samples. PLoS Genetics, 10(4): e1004269.