Inference for Transposable Data: Modeling the Effects of Row and Column Correlations

Genevera I. Allen* & Robert Tibshirani**

August 5, 2010

*Department of Pediatrics-Neurology, Baylor College of Medicine &
Department of Statistics, Rice University

**Departments of Health Research and Policy & Statistics, Stanford University

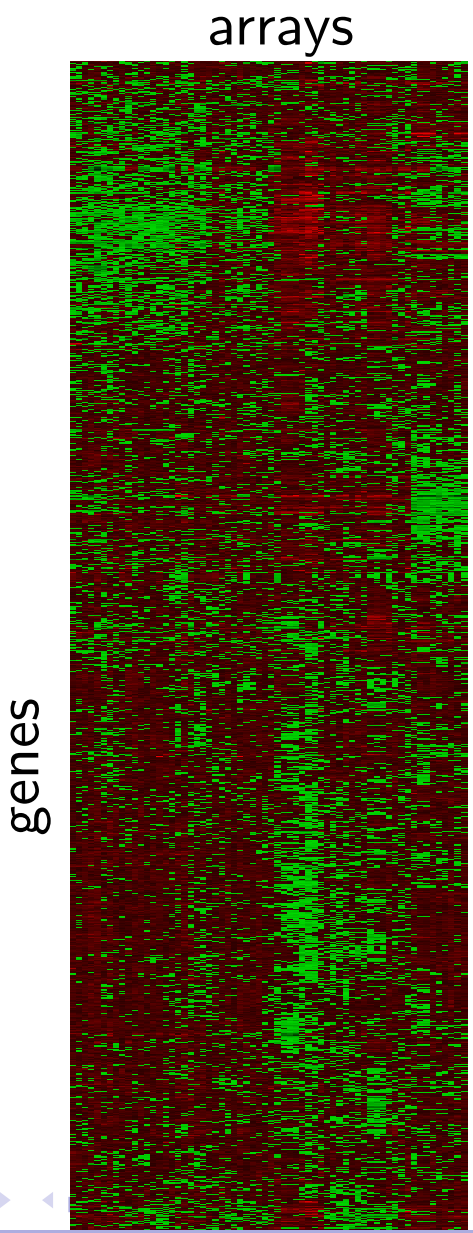
Large-Scale Inference and Genetic Data

Testing the significance of . . .

- Genes in microarrays.
- Isoforms in next-generation sequencing data.
- Biomarkers in protein arrays.

All of these can be arranged in the form of a matrix.

- Question: Is genetic data *transposable*?
  - Rows and/or Columns are features of interest.



In this Talk . . .

1. Introduction: Are our statistical assumptions for large-scale inference correct?

2. What happens when our assumptions are incorrect?
   - Array correlations:

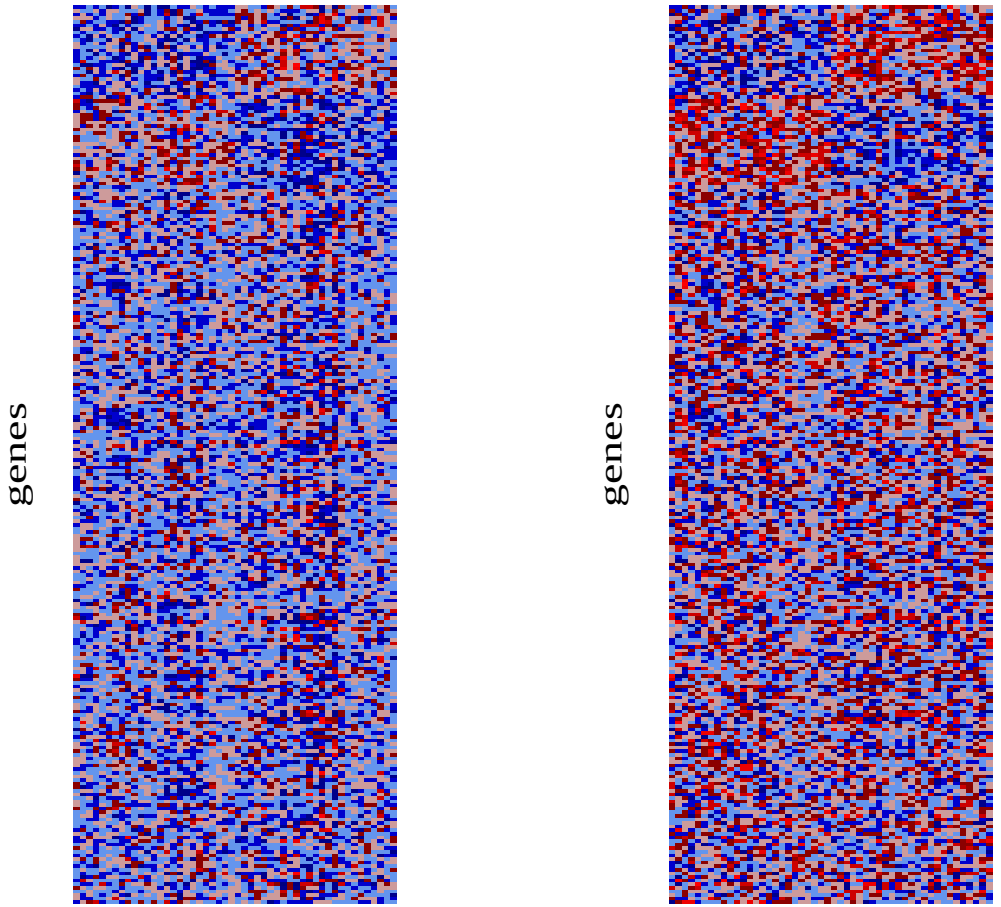
     How does this affect the behavior of our test statistics?
   - Gene and Array correlations:

     How does this affect multiple testing procedures?

3. How do we fix these problems?
   1. Directly model gene and array correlations with *Transposable Regularized Covariance Models*.
   2. De-Correlate or *sphere* the data.

Preview: De-Correlating Microarray Data

Data $\xrightarrow{\text{Sphering}}$ De-Correlated Data



1. Re-orders the ranking of the genes.
2. Allows one to reject more truly significant genes.
3. Obtain better estimates of the False Discovery Rate.

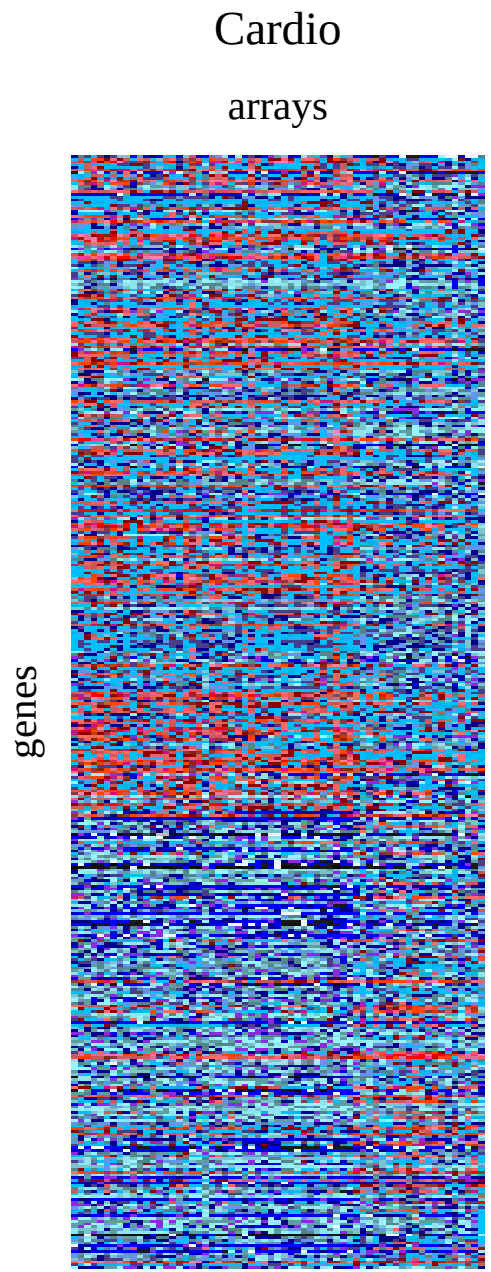
Two-Class Microarray

Goal: Find differentially expressed genes.

Example: "Cardio" data

- Study of cardiovascular disease.
- 20,426 genes and 63 arrays.
- 44 controls and 19 diseased subjects.

(Efron, B., 2009)



Statistical Assumptions

Method

1. For each gene:

   Calculate the two-sample t-test.

Assumptions

Statistical Assumptions

Method

1. For each gene:

   Calculate the two-sample t-test.

Assumptions

1. Independent Arrays.

Statistical Assumptions

Method

1. For each gene:

   Calculate the two-sample t-test.

2. Correct for multiple testing:
   - FDR (False Discovery Rate).
   - Examples: Step-up method (Benjamini & Hochberg, 1995), Permutation methods (SAM, Storey, 2002).

Assumptions

1. Independent Arrays.

Statistical Assumptions

Method

1. For each gene:

   Calculate the two-sample t-test.

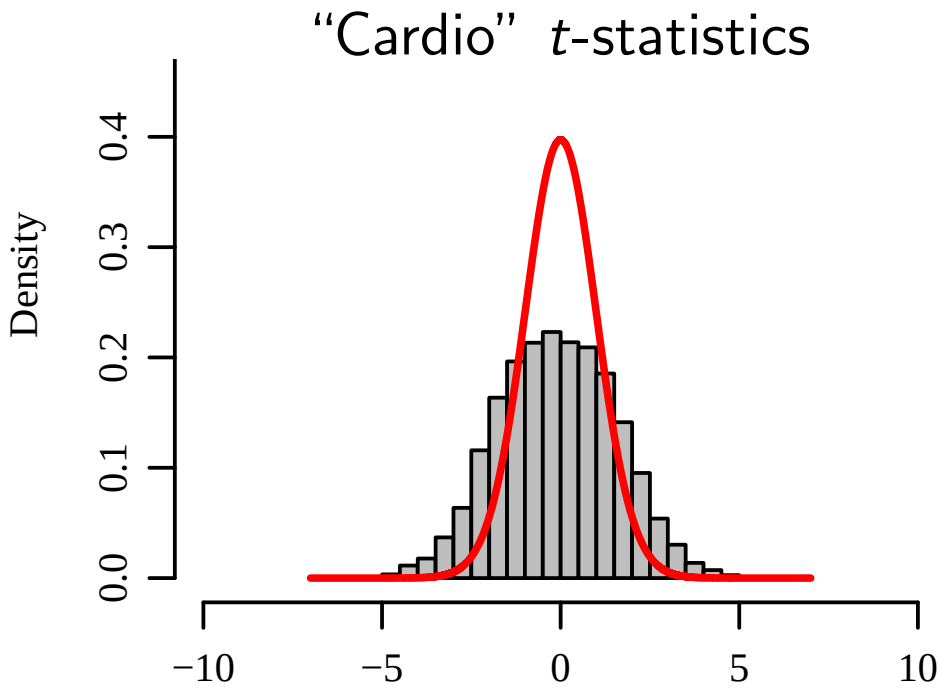
2. Correct for multiple testing:
   - FDR (False Discovery Rate).
   - Examples: Step-up method (Benjamini & Hochberg, 1995), Permutation methods (SAM, Storey, 2002).

Assumptions

1. Independent Arrays.

2. Limited Gene Dependence (positive regression dependence, weak dependence, local dependence).

Are these Realistic?

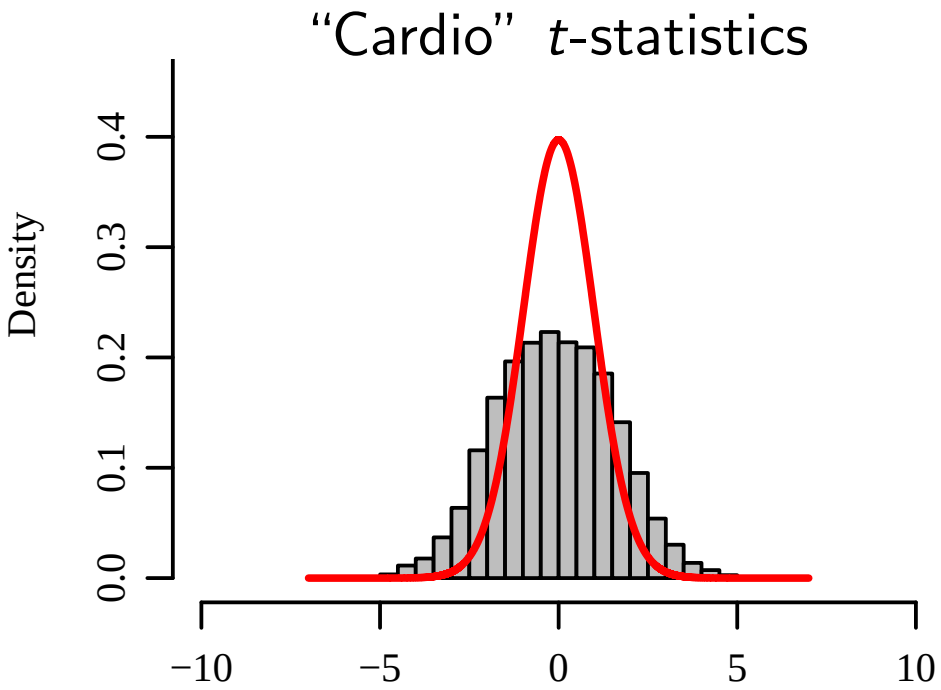


Could this be due to correlations among genes?

Over-dispersion:

- Red: Theoretical Null Distribution: $t_{(61)}$.

Are these Realistic?



Could this be due to correlations among the arrays?

- Measurement process:
  - Instrument drift, batch-effects, time of samples in storage, ...
- Correlated samples:
  - Latent variables such as age, gender or family history ...

Over-dispersion:

- Red: Theoretical Null Distribution: $t_{(61)}$.

In this Talk . . .

1. Introduction: Are our statistical assumptions for large-scale inference correct?

2. What happens when our assumptions are incorrect?
   - Array correlations:

     How does this affect the behavior of our test statistics?
   - Gene and Array correlations:

     How does this affect multiple testing procedures?

3. How do we fix these problems?
   1. Directly model gene and array correlations with *Transposable Regularized Covariance Models*.
   2. De-Correlate or *sphere* the data.

Microarray Matrix Model

$$\mathbf{X}_{m\times n} = \mathbf{M} + \mathbf{S} + \mathbf{N}.$$
$$\text{Data} = \text{Mean} + \text{Signal} + \text{Noise}.$$

where $\mathbf{M}_{m\times n} = \nu \mathbf{1}_{(n)}^T + \mathbf{1}_{(m)}\mu^T$ (mean matrix),

$\mathbf{S}_{m\times n}$ is problem specific (signal matrix),

$\mathbf{N}_{m\times n} \sim N_{m,n}(\mathbf{0}, \mathbf{0}, \mathbf{\Sigma}, \mathbf{\Delta})$ (noise matrix).

- Two-class microarray: $\mathbf{S} = \left[\psi_1 \mathbf{1}_{(n_1)}^T \;\; \psi_2 \mathbf{1}_{(n_2)}^T\right]$, where $\psi_1, \psi_2 \in \Re^m$ are the class signals.
- $\mathbf{X} - \mathbf{S} \sim N_{m,n}(\nu, \mu, \mathbf{\Sigma}, \mathbf{\Delta})$. (*mean-restricted matrix-variate normal*)

Review: Mean-Restricted Matrix-Variate Normal

Matrix extension of the multivariate normal:

$$\mathbf{X}_{m \times n} \sim N_{m,n}(\nu, \mu, \mathbf{\Sigma}, \mathbf{\Delta})$$

- Row means: $\nu \in \Re^m$.
- Column means: $\mu \in \Re^n$.
- Row covariance: $\mathbf{\Sigma} \in \Re^{m \times m}$
- Column covariance: $\mathbf{\Delta} \in \Re^{n \times n}$.

Review: Mean-Restricted Matrix-Variate Normal

Matrix extension of the multivariate normal:

$$\mathbf{X}_{m\times n} \sim N_{m,n}(\nu, \mu, \mathbf{\Sigma}, \mathbf{\Delta})$$

- Row means: $\nu \in \Re^m$.
- Column means: $\mu \in \Re^n$.
- Row covariance: $\mathbf{\Sigma} \in \Re^{m\times m}$
- Column covariance: $\mathbf{\Delta} \in \Re^{n\times n}$.

$$\mathrm{vec}(\mathbf{X}) \sim N\left(\mathrm{vec}(\mathbf{M}), \mathbf{\Omega}\right)$$

Review: Mean-Restricted Matrix-Variate Normal

Matrix extension of the multivariate normal:

$$\mathbf{X}_{m\times n} \sim N_{m,n}(\nu, \mu, \mathbf{\Sigma}, \mathbf{\Delta})$$

- Row means: $\nu \in \Re^m$.
- Column means: $\mu \in \Re^n$.
- Row covariance: $\mathbf{\Sigma} \in \Re^{m\times m}$
- Column covariance: $\mathbf{\Delta} \in \Re^{n\times n}$.

$$\mathrm{vec}(\mathbf{X}) \sim N\left(\mathrm{vec}(\mathbf{M}), \mathbf{\Omega}\right)$$

- $\mathbf{M} = \nu \mathbf{1}_{(n)}^T + \mathbf{1}_{(m)} \mu^T$.

$$\mathbf{M}_{m\times n} =$$

$$\begin{pmatrix} \nu_1 + \mu_1 & \nu_1 + \mu_2 & \cdots & \nu_1 + \mu_n \\ \nu_2 + \mu_1 & \nu_2 + \mu_2 & & \\ \vdots & & \ddots & \vdots \\ \nu_m + \mu_1 & & \cdots & \nu_m + \mu_n \end{pmatrix}$$

Review: Mean-Restricted Matrix-Variate Normal

Matrix extension of the multivariate normal:

$$\mathbf{X}_{m\times n} \sim N_{m,n}(\nu, \mu, \mathbf{\Sigma}, \mathbf{\Delta})$$

- Row means: $\nu \in \Re^m$.
- Column means: $\mu \in \Re^n$.
- Row covariance: $\mathbf{\Sigma} \in \Re^{m\times m}$
- Column covariance: $\mathbf{\Delta} \in \Re^{n\times n}$.

$$\mathbf{\Omega}_{mn\times mn} = \begin{pmatrix} \Delta_{11}\mathbf{\Sigma} & \Delta_{12}\mathbf{\Sigma} & \dots & \Delta_{1n}\mathbf{\Sigma} \\ \Delta_{21}\mathbf{\Sigma} & \Delta_{22}\mathbf{\Sigma} & & \\ \vdots & & \ddots & \vdots \\ \Delta_{n1}\mathbf{\Sigma} & & \dots & \Delta_{nn}\mathbf{\Sigma} \end{pmatrix}$$

$$\mathrm{vec}(\mathbf{X}) \sim N\left(\mathrm{vec}(\mathbf{M}), \mathbf{\Omega}\right)$$

- $\mathbf{M} = \nu \mathbf{1}_{(n)}^T + \mathbf{1}_{(m)}\mu^T$.
- $\mathbf{\Omega} = \mathbf{\Delta} \otimes \mathbf{\Sigma}$.

(Gupta & Nagar, 1999; G. I. Allen & R. Tibshirani, 2010)

Test Statistic Null Distributions

Question: How do test statistics behave when arrays are correlated?

Two-sample Z-test:

- Independent arrays:

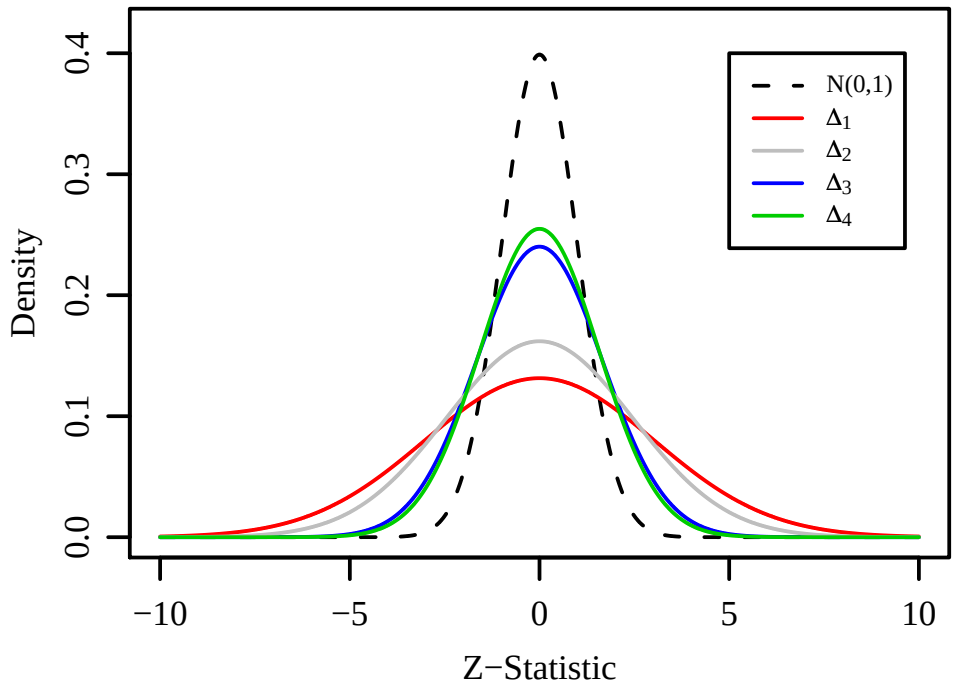
$$Z \sim N(0, 1).$$

- Theorem: Under matrix-variate normal,

$$Z \sim N(0, \eta / c_n),$$

where $c_n = \frac{1}{n_1} + \frac{1}{n_2}$,

η is a function of $\mathbf{\Delta}$.

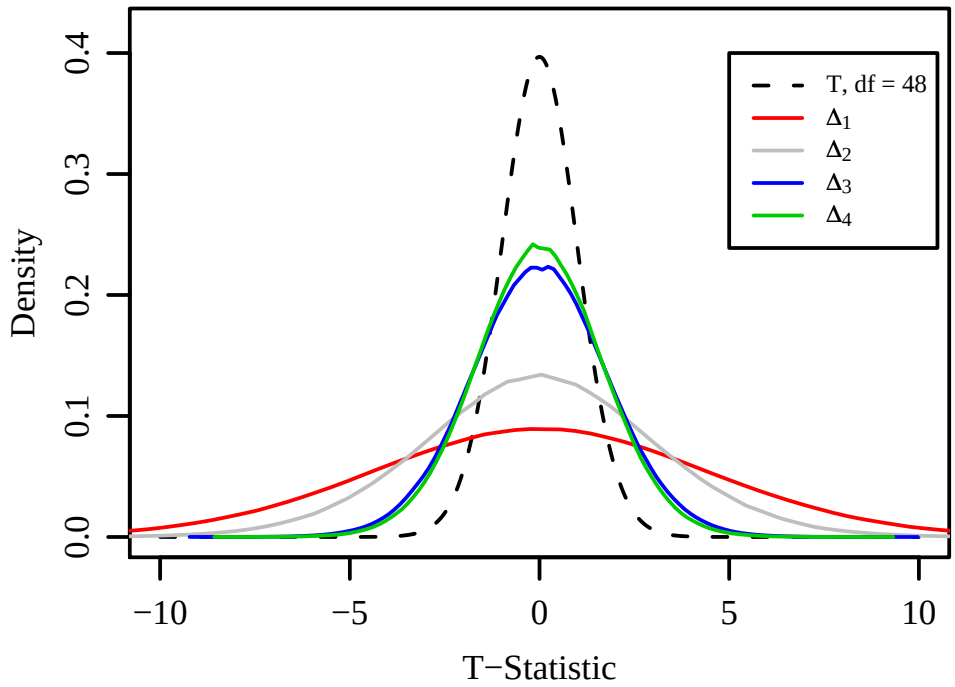


Test Statistic Null Distributions

Question: How do test statistics behave when arrays are correlated?

Two sample T-test:

- Independent Arrays: $T \sim t_{(n-2)}$.
- Correlated Arrays (matrix-variate normal): No closed form distribution.
- Variances estimated by Monte Carlo.



Study: Multiple Testing and Dependence

Simulation Study:

- Data from matrix-variate normal model.
- Used two-sample t-statistics.
- Applied various FDR-controlling procedures.

Conclusions:

1. Good News: FDR controlled under many types of gene dependence.
2. Bad News: FDR NOT controlled under gene AND array dependence.

In this Talk . . .

1. Introduction: Are our statistical assumptions for large-scale inference correct?

2. What happens when our assumptions are incorrect?
   - Array correlations:

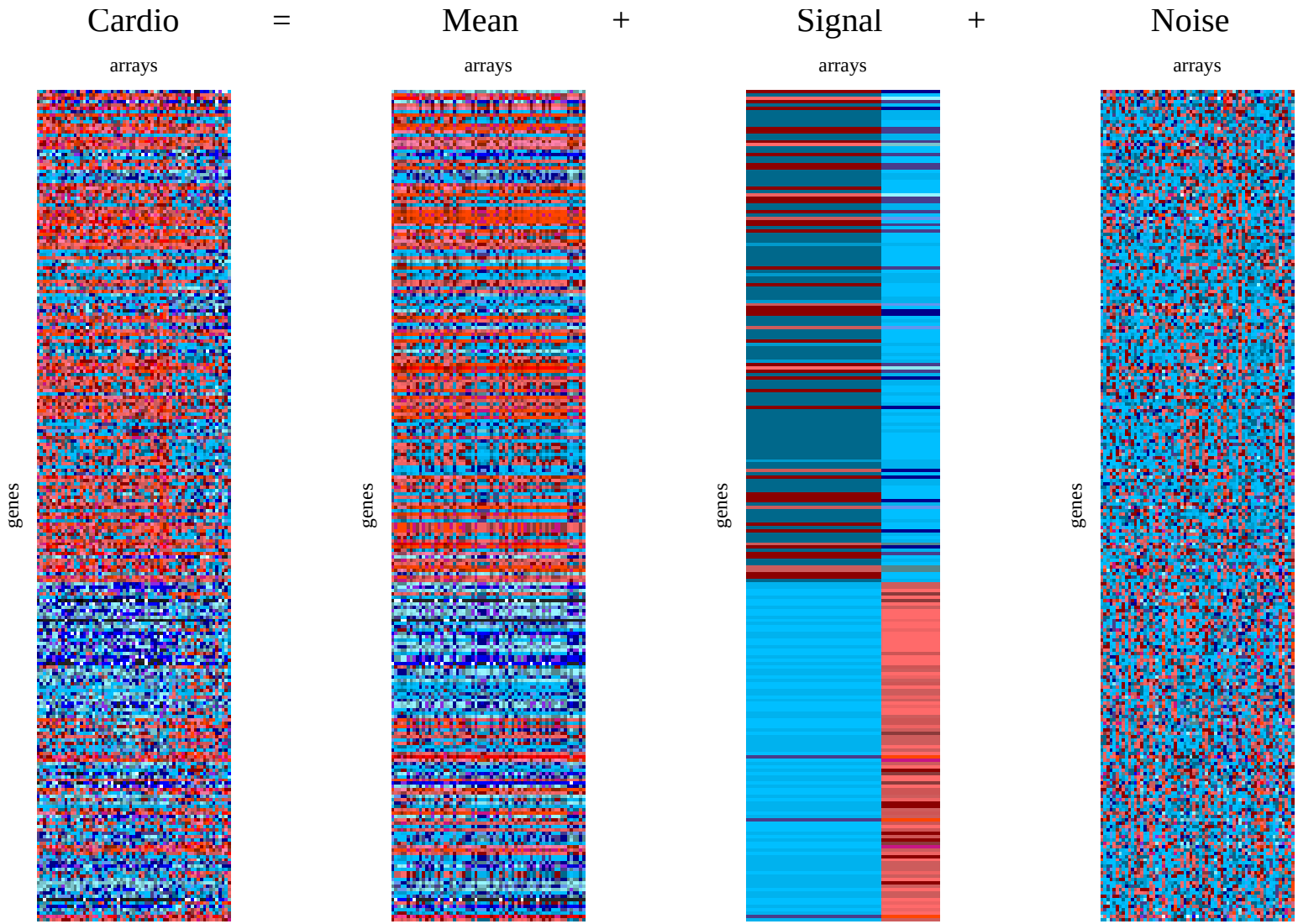
     How does this affect the behavior of our test statistics?
   - Gene and Array correlations:

     How does this affect multiple testing procedures?

3. How do we fix these problems?
   1. Directly model gene and array correlations with *Transposable Regularized Covariance Models*.
   2. De-Correlate or *sphere* the data.

De-Correlating the Data

Step 1: Decompose data into Mean + Signal + Noise.

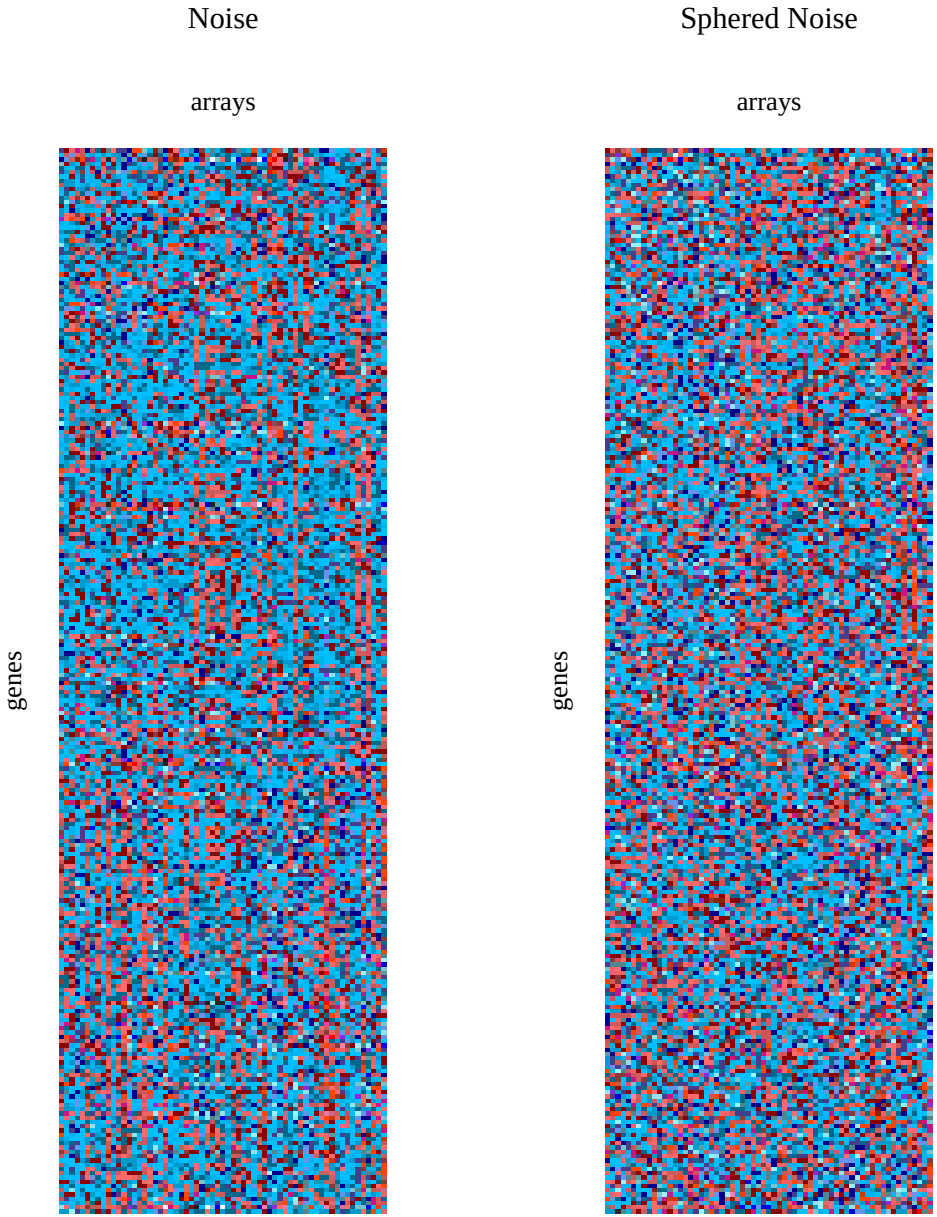


De-Correlating the Data

Step 2: Estimate the Gene and Array Covariances of the Noise. *Sphere* the Noise.

- $\tilde{\mathbf{N}} = \hat{\mathbf{\Sigma}}^{-1/2} \hat{\mathbf{N}} \hat{\mathbf{\Delta}}^{-1/2}$.
- $\hat{\mathbf{\Sigma}}$ & $\hat{\mathbf{\Delta}}$ estimated via *Transposable Regularized Covariance Models*.

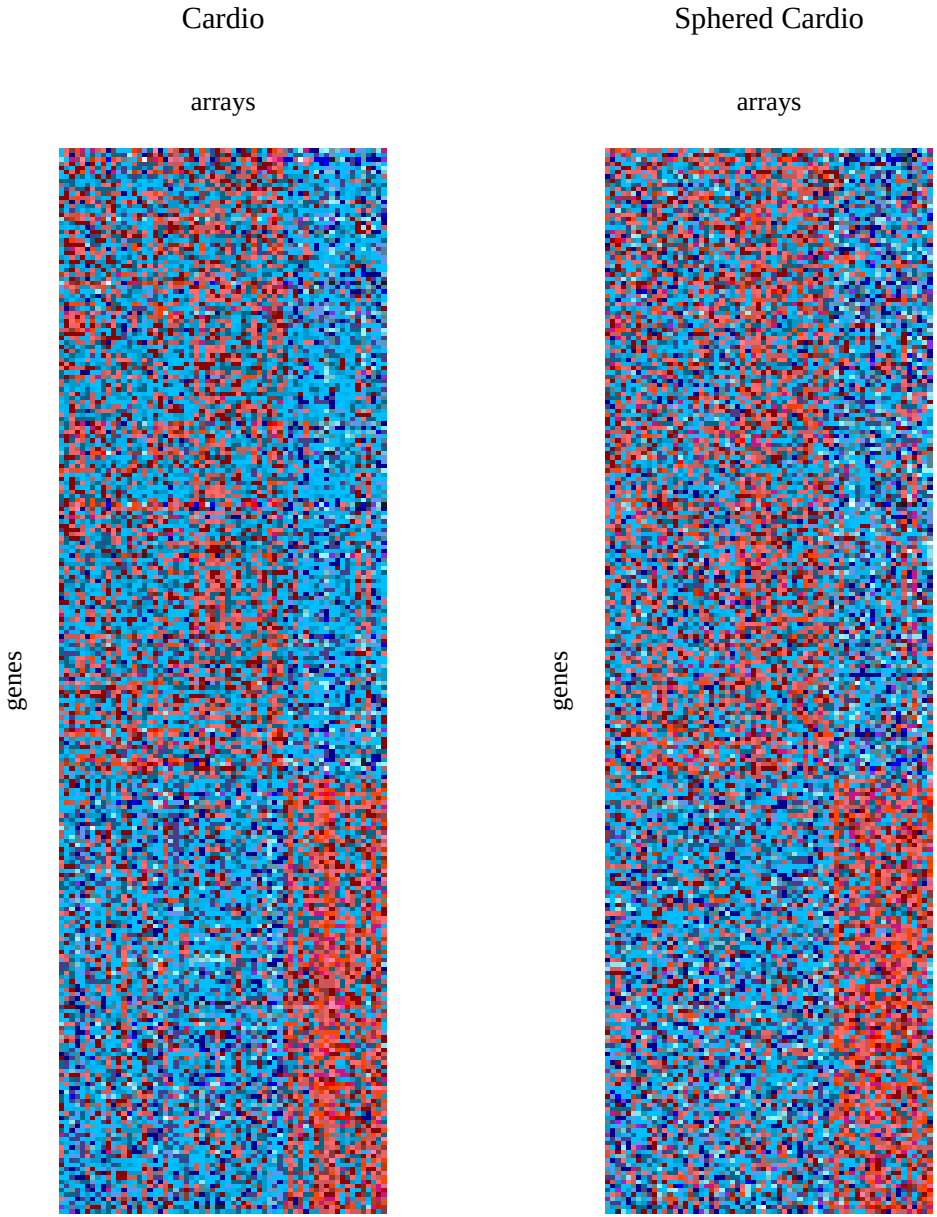
(G. I. Allen & R. Tibshirani, 2010)



De-Correlating the Data

Step 3: De-Correlated Data.

- $\tilde{\mathbf{X}} = \hat{\mathbf{S}} + \tilde{\mathbf{N}}$.
- Approximately independent genes AND arrays.
- T-statistics distributed approximately $t_{(n-2)}$.

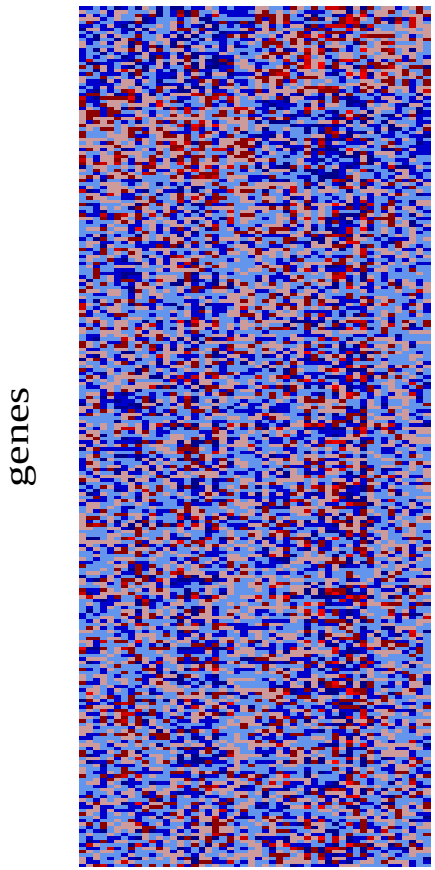


Cardio Results: Data Images

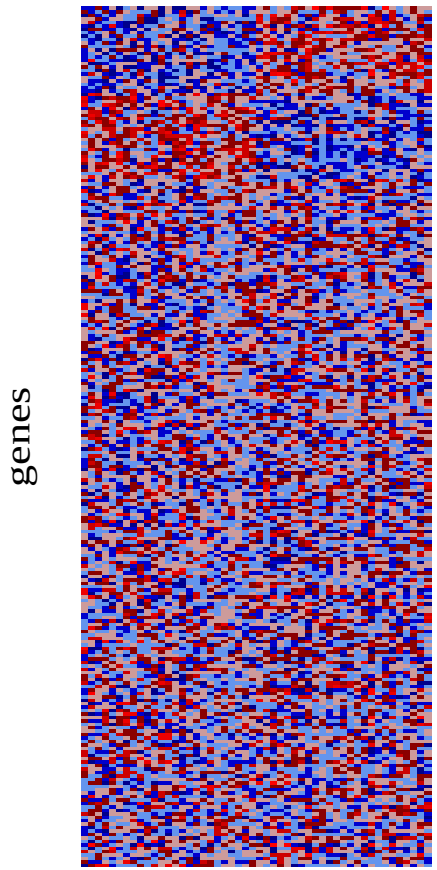
"Cardio"-Inspired Simulation:

- 250 genes, 50 differentially expressed.
- Gene & Array correlations: randomly selected Cardio genes & arrays.

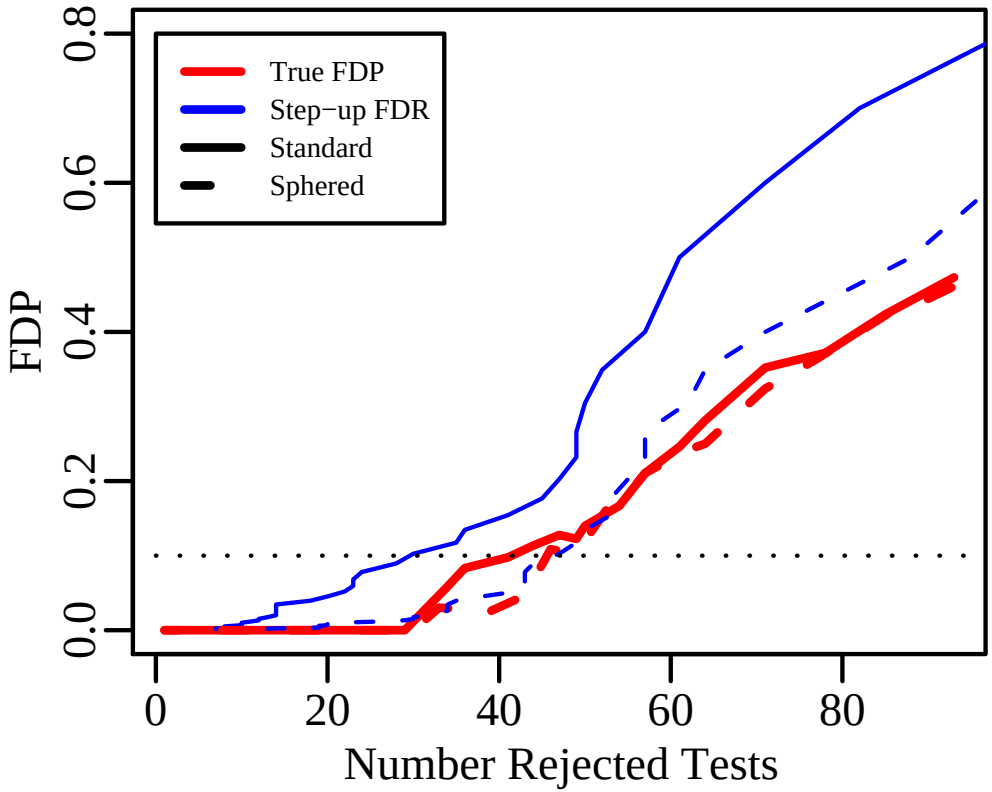
Signal + Cardio Noise



Signal + Sphered Noise



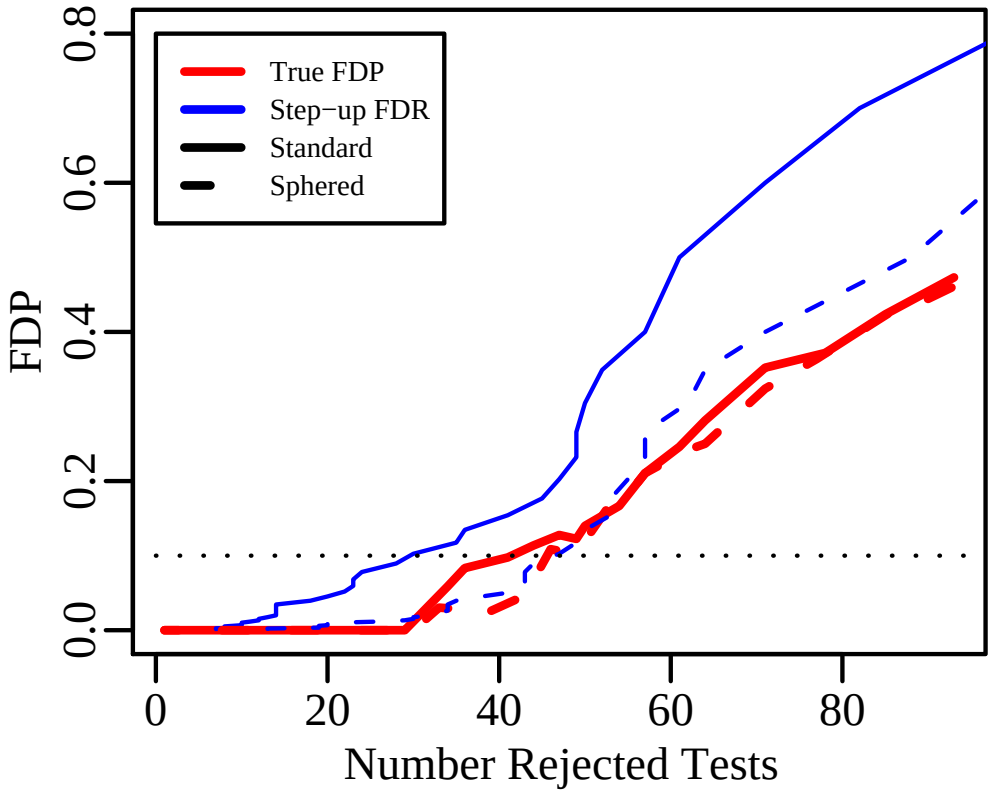
Cardio Results: FDR Curves



Benefits of Sphering:

1. Increased statistical power. (Gene rank is re-ordered.)
   - Standard Method: 38 genes rejected.
   - Sphering: 43 genes rejected

Cardio Results: FDR Curves



Benefits of Sphering:

1. Increased statistical power. (Gene rank is re-ordered.)
   - Standard Method: 38 genes rejected.
   - Sphering: 43 genes rejected
2. Correct estimation of FDR.
   - Standard Method: 30 genes rejected.
   - Sphering: 43 genes rejected

Results: Other Models

	Standard		Sphered	
	FDP	$\widehat{\text{FDR}}$	FDP	$\widehat{\text{FDR}}$
Latent Variable Model*	0.189	0.383	0.167	0.166
Random Effects Model	0.52	0.0229	0.154	0.207
Gene Correlations	0.169	0.19	0.141	0.185
Gene & Array Correlations	0.111	0.426	0.105	0.124

True FDP and FDR estimated by the step-up method for 55/250 rejected tests averaged over 10 simulations.

*(J. Leek & J. Storey, 2008)

Conclusions & Future Work

Conclusions:

1. Gene and especially Array correlations pose a major problem for large-scale inference.
2. *Sphering* the data can correct these problems.

Conclusions & Future Work

Conclusions:

1. Gene and especially Array correlations pose a major problem for large-scale inference.
2. *Sphering* the data can correct these problems.

Future Work:

- Extensions to categorical data.
  - Application: Next-generation sequencing data.
- Approximations for high-dimensional data.
  - Application: Functional MRIs.

Acknowledgments & References

Acknowledgments:

SF Bay Area Chapter of the American Statistical Association Student Travel Award

References:

G. I. Allen & R. Tibshirani, Transposable regularized covariance models with an application to missing data imputation, (To Appear) *Annals of Applied Statistics*, 2010.

B. Efron, Are a set of microarrays independent of each other?, *Annals of Applied Statistics*, **13**: 3 (922-942), 2009.

A. K. Gupta & D. K. Nagar, *Matrix variate distributions*, Chapman & Hall, CRC Press, 1999.