Treelet Covariance Smoothers

Estimation of Genetic Parameters

B. Draves[1]

[1]Department of Mathematics
Lafayette College

Advisor: T. Gaugler

Moravian College, 2017

Overview

- Principle Component Analysis

Overview

- Principle Component Analysis
- Motivation in Statistical Genetics

Overview

- Principle Component Analysis
- Motivation in Statistical Genetics
- Existing Methods

Overview

- Principle Component Analysis
- Motivation in Statistical Genetics
- Existing Methods
- New Methods & Results

Principle Component Analysis (PCA)

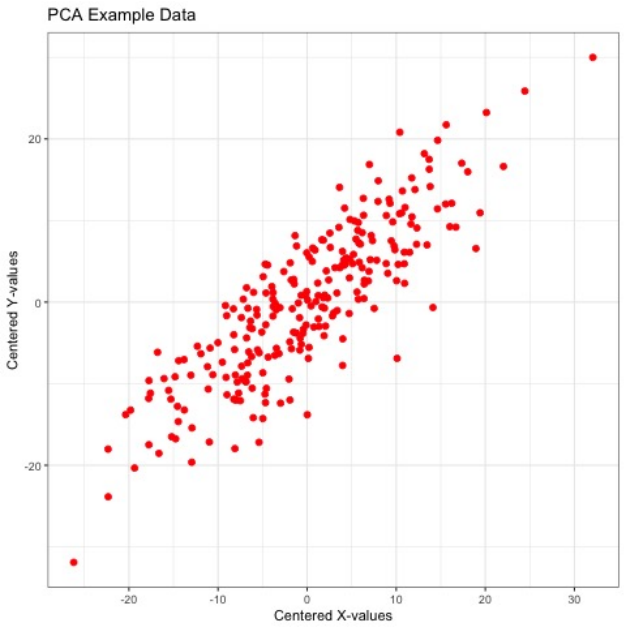
- PCA looks to reduce the number of input variables $X_1, X_2, \ldots, X_k$ (k large) to $k_0 < k$ variables

Principle Component Analysis (PCA)

- PCA looks to reduce the number of input variables $X_1, X_2, \ldots, X_k$ (k large) to $k_0 < k$ variables
- Reduce to fewer inputs while preserving as much variability in the data as possible

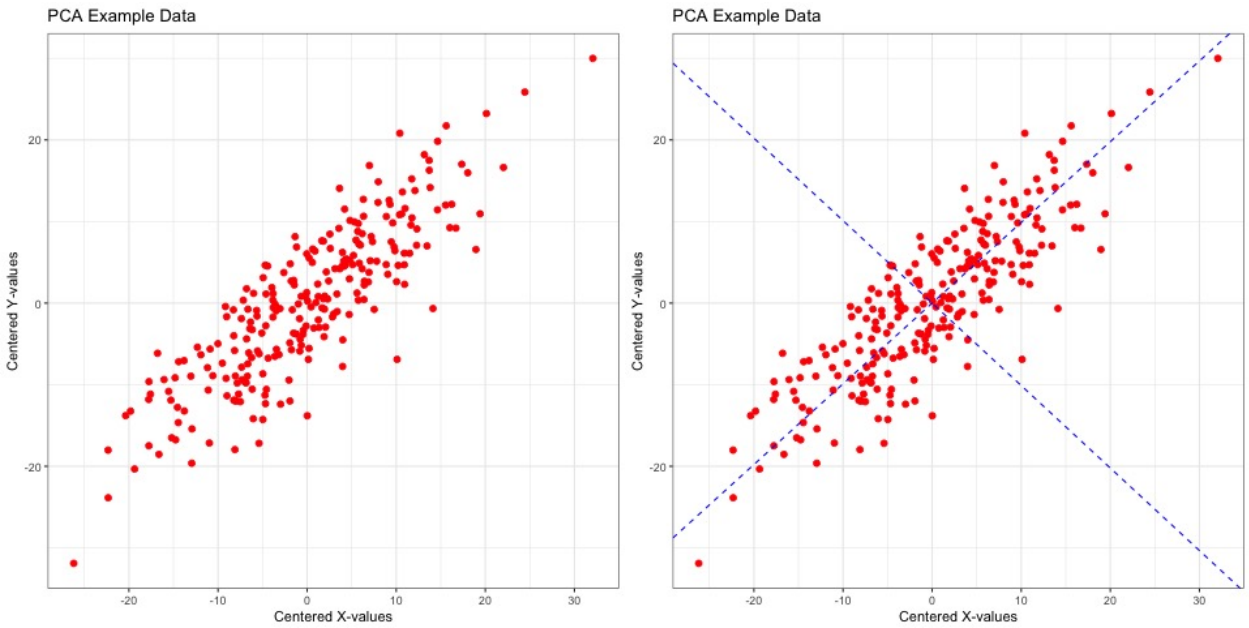
Principle Component Analysis (PCA)

- PCA looks to reduce the number of input variables $X_1, X_2, \ldots, X_k$ (k large) to $k_0 < k$ variables
- Reduce to fewer inputs while preserving as much variability in the data as possible



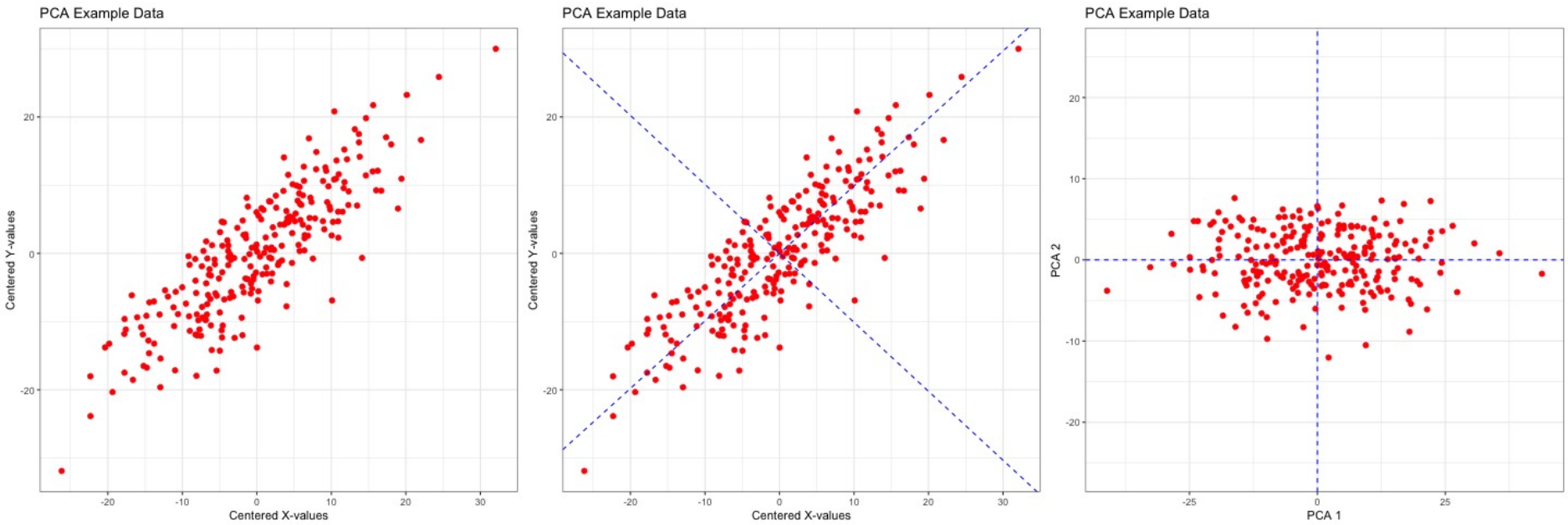
Principle Component Analysis (PCA)

- PCA looks to reduce the number of input variables $X_1, X_2, \ldots, X_k$ (k large) to $k_0 < k$ variables
- Reduce to fewer inputs while preserving as much variability in the data as possible



Principle Component Analysis (PCA)

- PCA looks to reduce the number of input variables $X_1, X_2, \ldots, X_k$ (k large) to $k_0 < k$ variables
- Reduce to fewer inputs while preserving as much variability in the data as possible



Genetics Motivation

- Individual's genetic material can be described by a panel of genotyped SNPs

Genetics Motivation

- Individual's genetic material can be described by a panel of genotyped SNPs
- Using this genetic information, an estimate of the relationship matrix, A, can be calculated

Genetics Motivation

- Individual's genetic material can be described by a panel of genotyped SNPs
- Using this genetic information, an estimate of the relationship matrix, A, can be calculated
- Genetically inferred relationship matrices are typically very noisy

Genetics Motivation

- Individual's genetic material can be described by a panel of genotyped SNPs
- Using this genetic information, an estimate of the relationship matrix, A, can be calculated
- Genetically inferred relationship matrices are typically very noisy
- Idea: use PCA to set the data on top of a more "natural" basis

Treelet Covariance Smoothing

- Goal: preserve local structure of data

Treelet Covariance Smoothing

- Goal: preserve local structure of data
- Iteratively change basis via PCA to preserve the variability between two most closely related individuals

Treelet Covariance Smoothing

- Goal: preserve local structure of data
- Iteratively change basis via PCA to preserve the variability between two most closely related individuals
- Repeat this process until all individuals are processed

Treelet Covariance Smoothing

- Goal: preserve local structure of data
- Iteratively change basis via PCA to preserve the variability between two most closely related individuals
- Repeat this process until all individuals are processed
- Once all individuals are processed, enforce sparsity to improve the estimator

TCS Properties

P1

P2

P3

P4

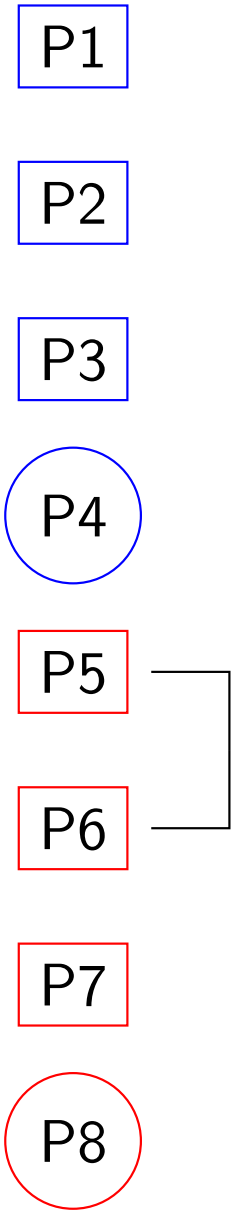
P5

P6

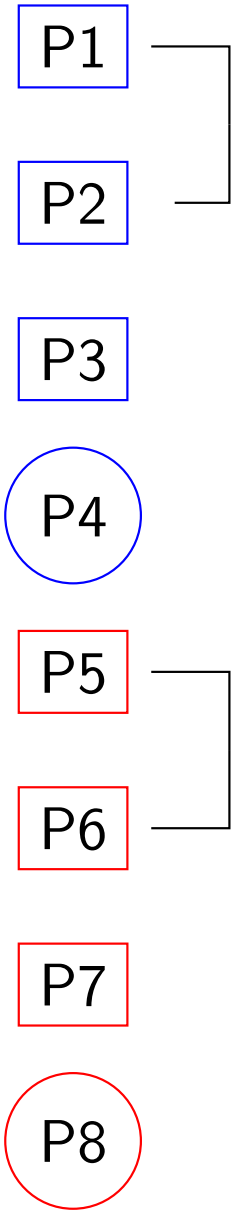
P7

P8

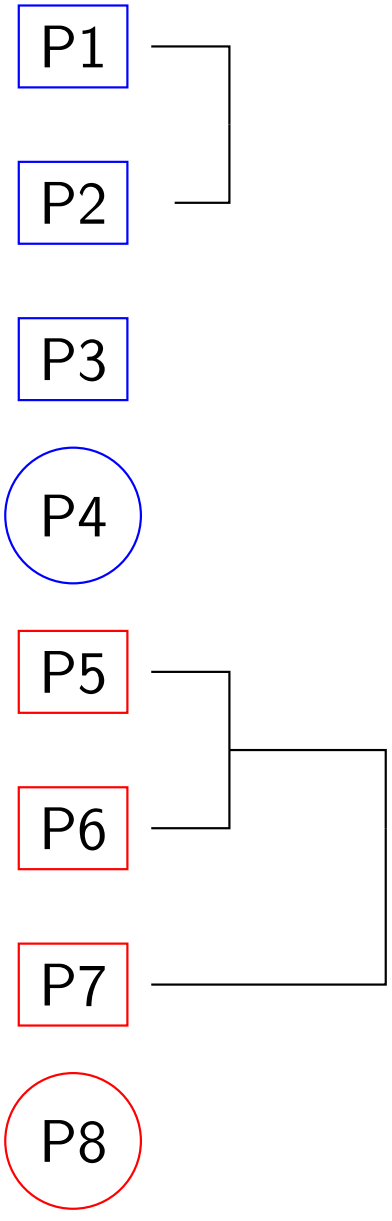
TCS Properties



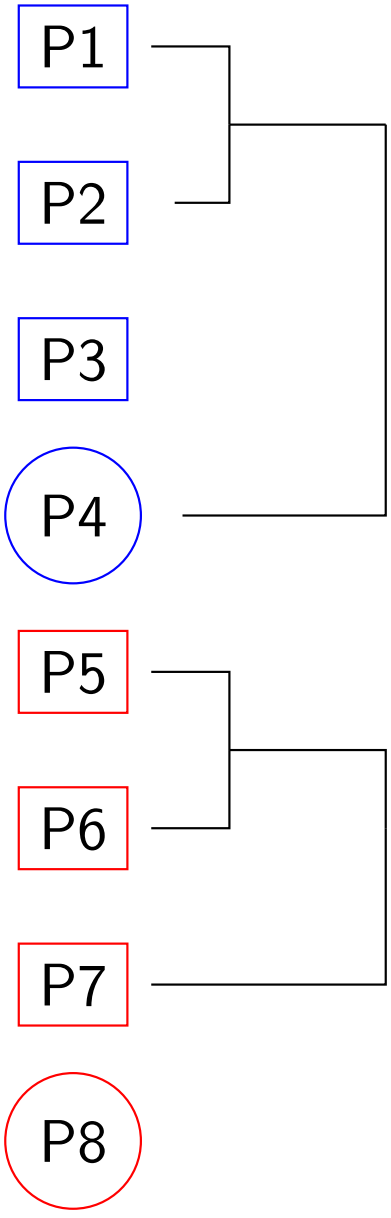
TCS Properties



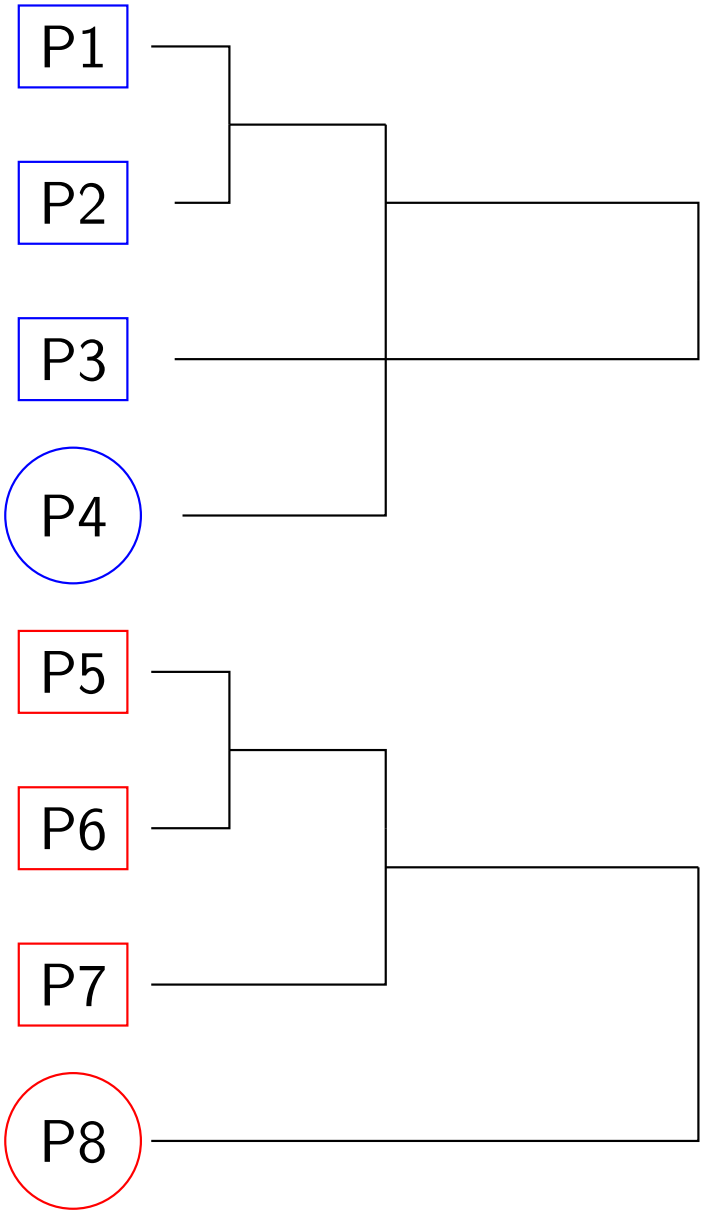
TCS Properties



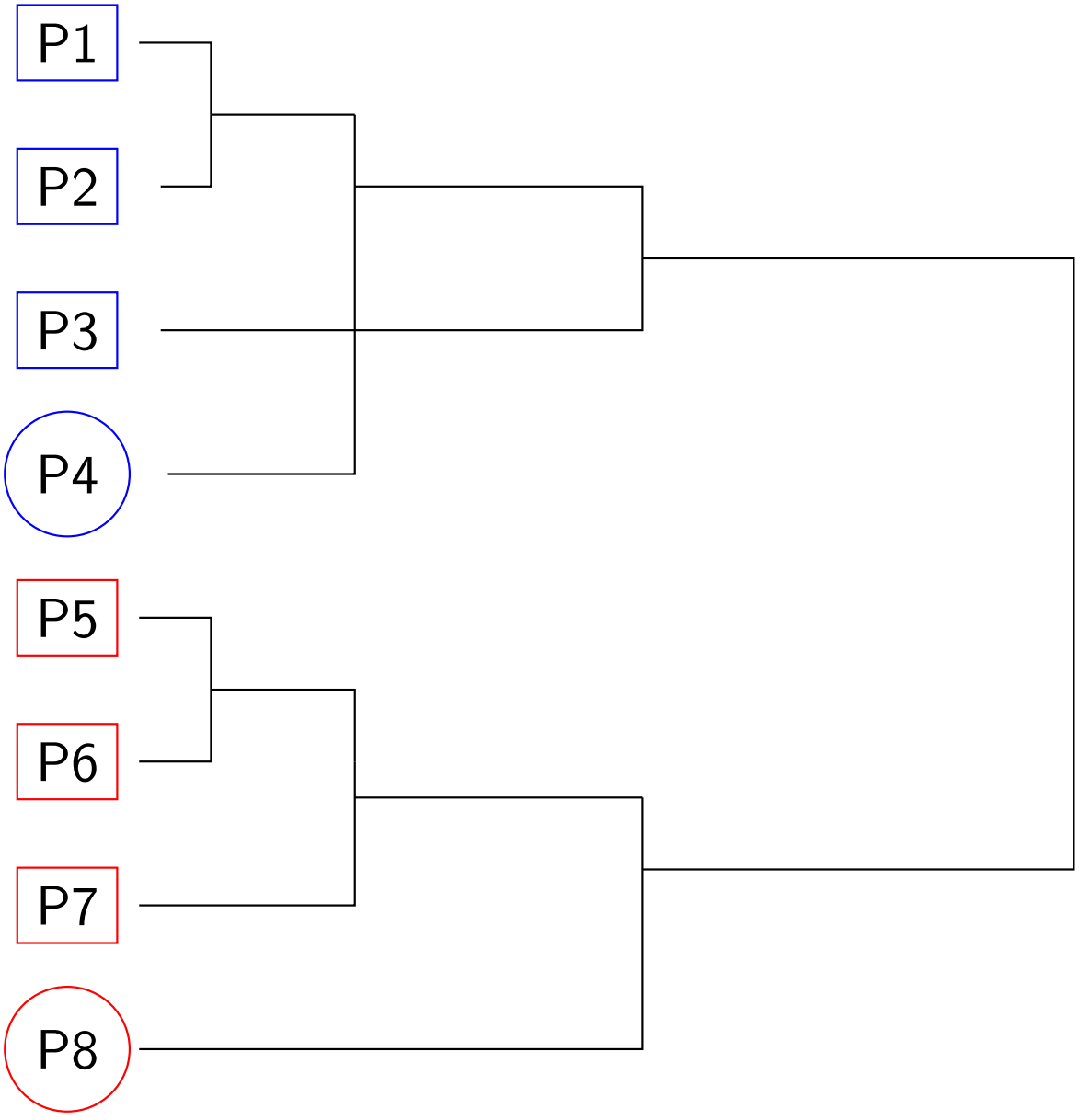
TCS Properties



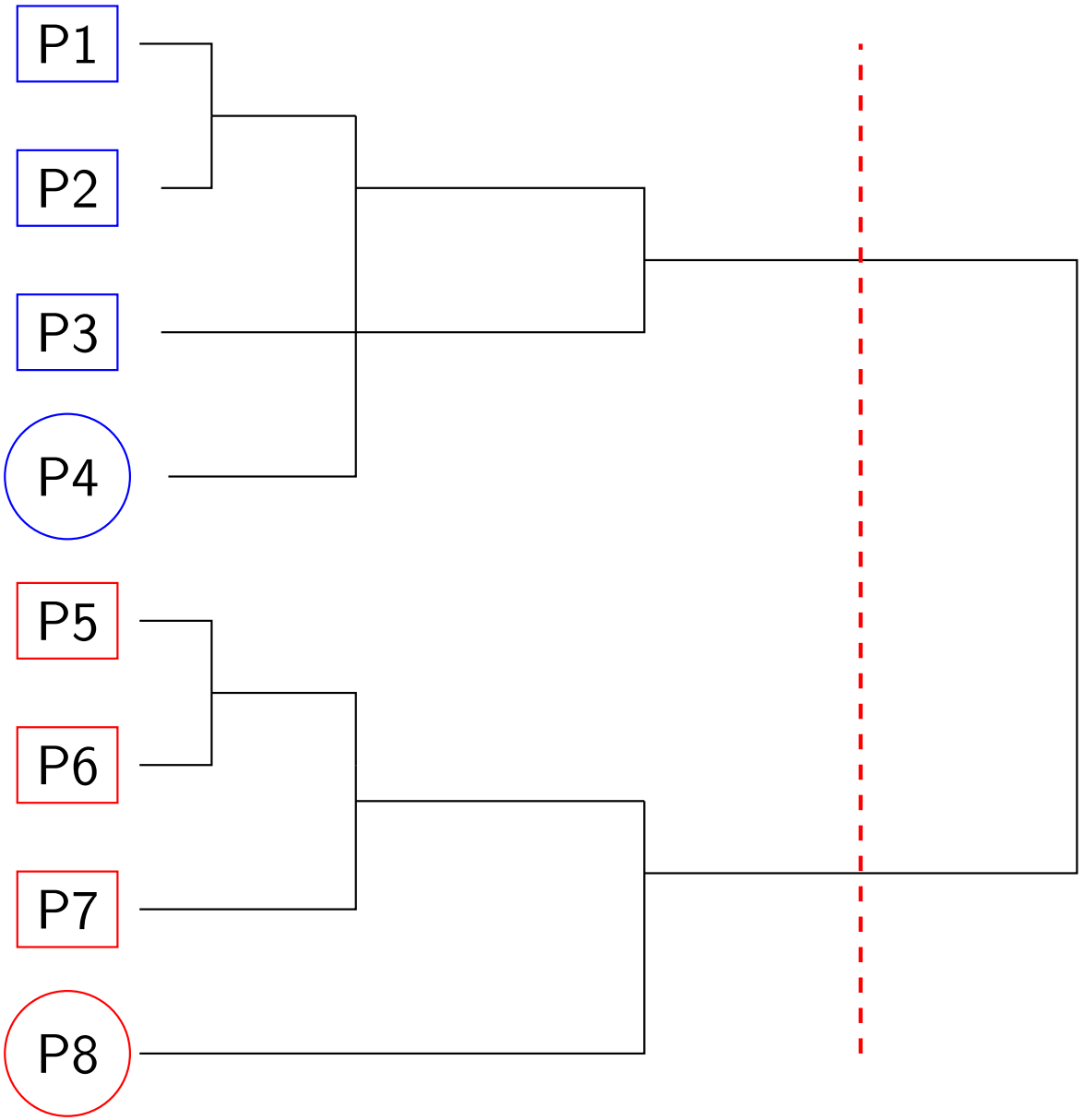
TCS Properties



TCS Properties



TCS Properties



Expanding TCS

- Why do we require merging into one cluster?

Expanding TCS

- Why do we require merging into one cluster?
- Idea: Stop merging variables to utilize familiar blocks

Expanding TCS

- Why do we require merging into one cluster?
- Idea: Stop merging variables to utilize familiar blocks
- Smooth by projecting data onto sum variables
  - Treelet Covariance Blocking (TCB)

Expanding TCS

- Why do we require merging into one cluster?
- Idea: Stop merging variables to utilize familiar blocks
- Smooth by projecting data onto sum variables
  - Treelet Covariance Blocking (TCB)
- Further enforce sparsity by thresholding estimates
  - Treelet Covariance Blocked Smoothing (TCBS)

Treelet Covariance Smoothers

- At each level of the tree there are basis vectors $\{v_1^{(\ell)}, \ldots, v_N^{(\ell)}\}$

Treelet Covariance Smoothers

- At each level of the tree there are basis vectors $\{v_1^{(\ell)}, \ldots, v_N^{(\ell)}\}$
- We can then write our estimate of A, $\hat{\Sigma}$, by

$$\widehat{\Sigma} = \sum_{i \in \widehat{S}_\ell} \widehat{\gamma}_{i,i}^{(\ell)} \widehat{\mathbf{v}}_i^{(\ell)} \left(\widehat{\mathbf{v}}_i^{(\ell)}\right)^t + \sum_{i,j \in \widehat{S}_\ell, i \neq j} \widehat{\gamma}_{i,j}^{(\ell)} \widehat{\mathbf{v}}_i^{(\ell)} \left(\widehat{\mathbf{v}}_j^{(\ell)}\right)^t$$

Treelet Covariance Smoothers

- At each level of the tree there are basis vectors $\{v_1^{(\ell)}, \ldots, v_N^{(\ell)}\}$
- We can then write our estimate of A, $\hat{\Sigma}$, by

$$\widehat{\Sigma} = \sum_{i \in \widehat{S}_\ell} \widehat{\gamma}_{i,i}^{(\ell)} \widehat{\mathbf{v}}_i^{(\ell)} \left(\widehat{\mathbf{v}}_i^{(\ell)}\right)^t + \sum_{i,j \in \widehat{S}_\ell, i \neq j} \widehat{\gamma}_{i,j}^{(\ell)} \widehat{\mathbf{v}}_i^{(\ell)} \left(\widehat{\mathbf{v}}_j^{(\ell)}\right)^t$$

- Here $\hat{\gamma}_{i,j}^{(\ell)}$ are the variance - covariance estimates of transformed relationship variables

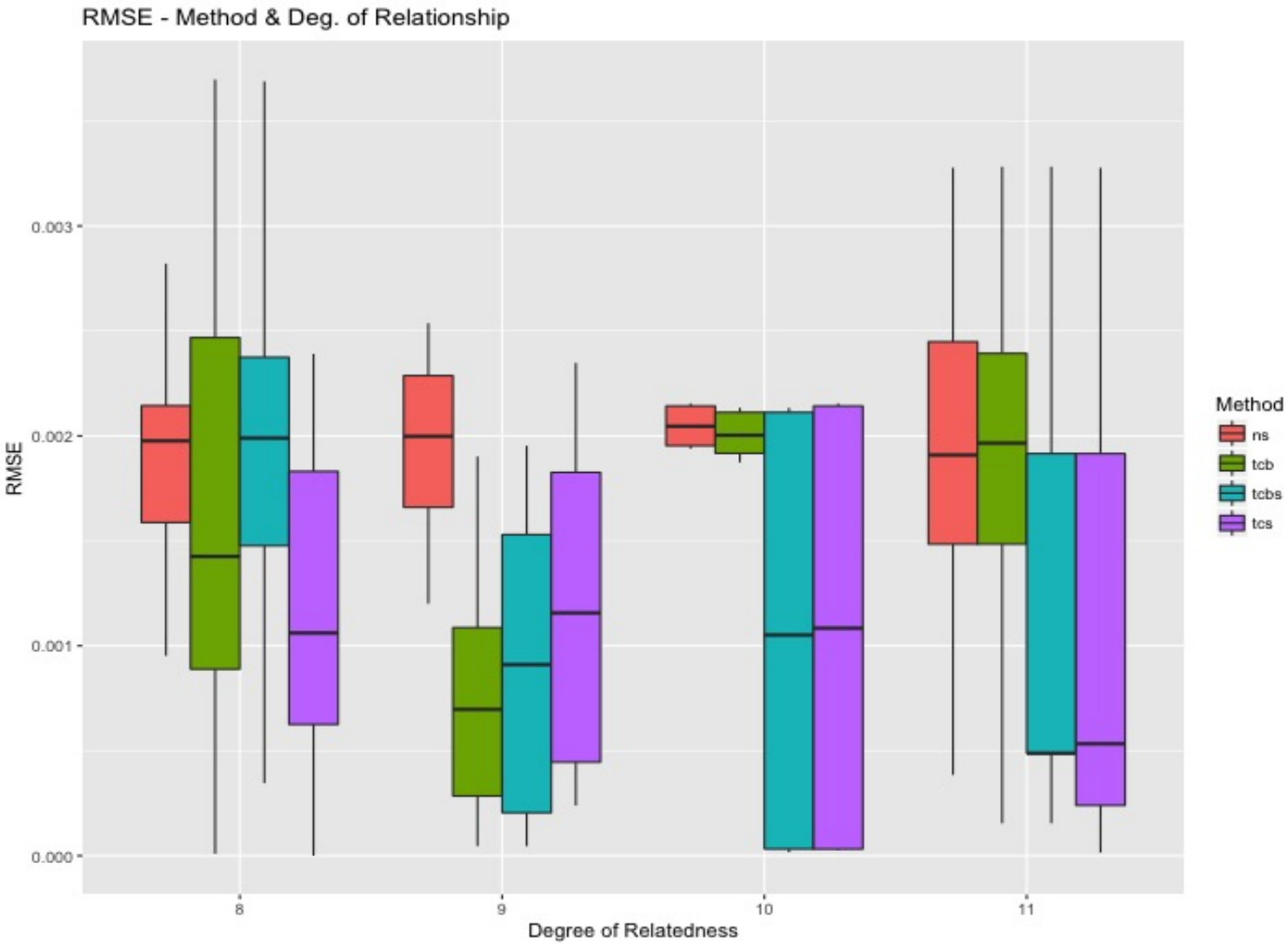
Treelet Covariance Smoothers

- At each level of the tree there are basis vectors $\{v_1^{(\ell)}, \ldots, v_N^{(\ell)}\}$
- We can then write our estimate of A, $\hat{\Sigma}$, by

$$\widehat{\Sigma} = \sum_{i \in \widehat{S}_\ell} \widehat{\gamma}_{i,i}^{(\ell)} \widehat{\mathbf{v}}_i^{(\ell)} \left(\widehat{\mathbf{v}}_i^{(\ell)}\right)^t + \sum_{i,j \in \widehat{S}_\ell, i \neq j} \widehat{\gamma}_{i,j}^{(\ell)} \widehat{\mathbf{v}}_i^{(\ell)} \left(\widehat{\mathbf{v}}_j^{(\ell)}\right)^t$$

- Here $\hat{\gamma_{i,j}}^{(\ell)}$ are the variance - covariance estimates of transformed relationship variables
- We can enforce sparsity by further thresholding these $\hat{\gamma_{i,j}}^{(\ell)}$ values

Simulation Results



Conclusion

- Treelet Covariance Smoothers is a class of methods which improve the estimation of distant relationships

Conclusion

- Treelet Covariance Smoothers is a class of methods which improve the estimation of distant relationships
- Estimating relationships is the centerpiece of successful estimation of other genetic parameters such as heritability

Conclusion

- Treelet Covariance Smoothers is a class of methods which improve the estimation of distant relationships
- Estimating relationships is the centerpiece of successful estimation of other genetic parameters such as heritability
- Understanding of genetic basis of heritability can lead to better treatment

Thanks for listening

Questions? Comments?