Population structure and eigenanalysis

Cambridge. Isaac Newton Institute: December 14 2006

Nick Patterson

Population structure, how to detect?

- Why we care
- Main idea: Principal components
- Review of some modern statistics
- How well does it work?
- Can estimate data size needed to see structure
- Relationship with other (Bayesian) models

Given genotype data: is it from homogeneous population?

Need statistic and formal test.

Want to test for additional structure when some found.

Should work when markers are in LD.

Statistic

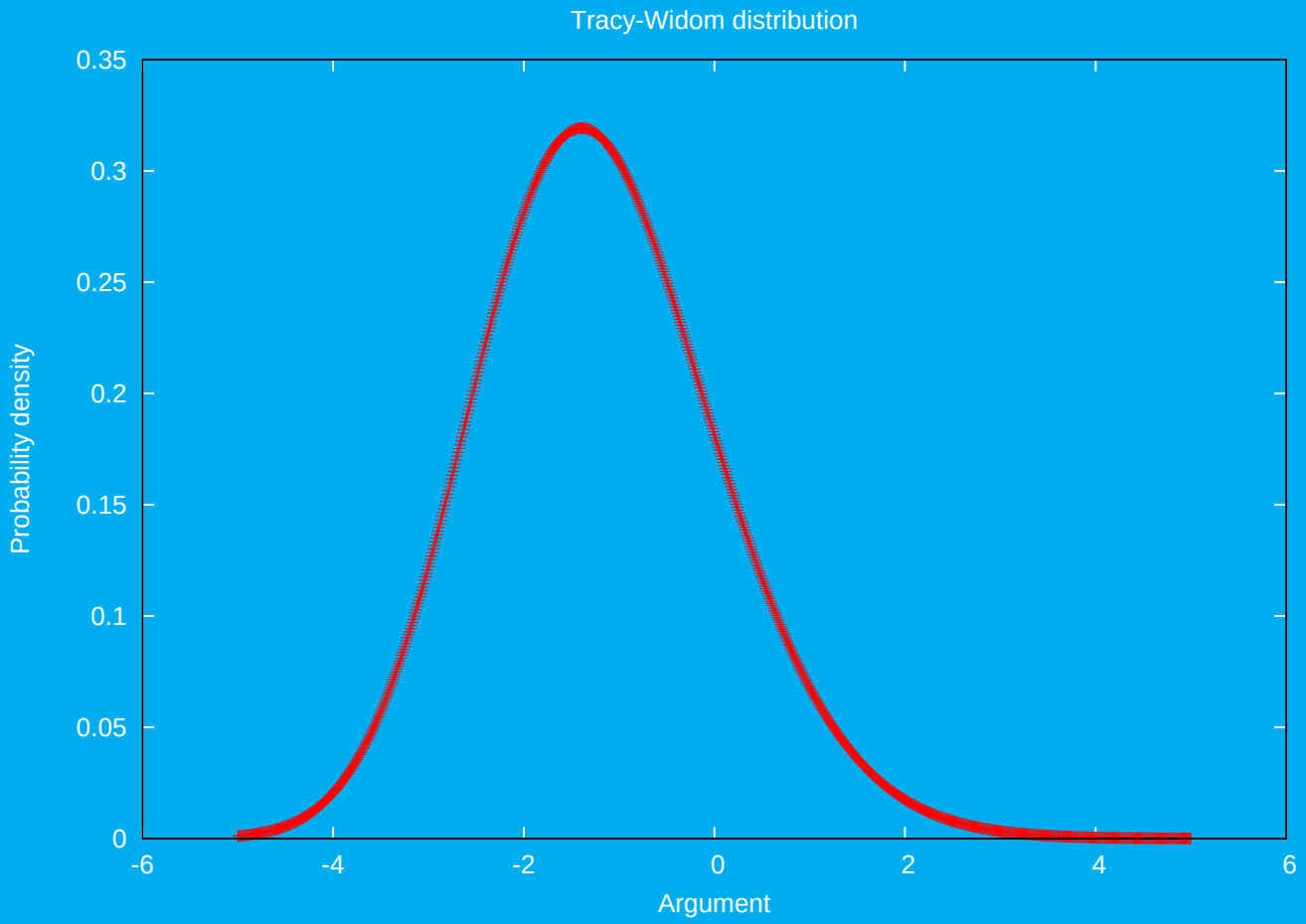
Data: Big rectangular matrix M: $p \times n$
For each column: Set π to be mean value. Subtract π.
Divide column by $\sqrt{(\pi/2)(1-\pi/2)}$
Compute sample covariance matrix X (of columns)
Statistic λ_1: largest eigenvalue

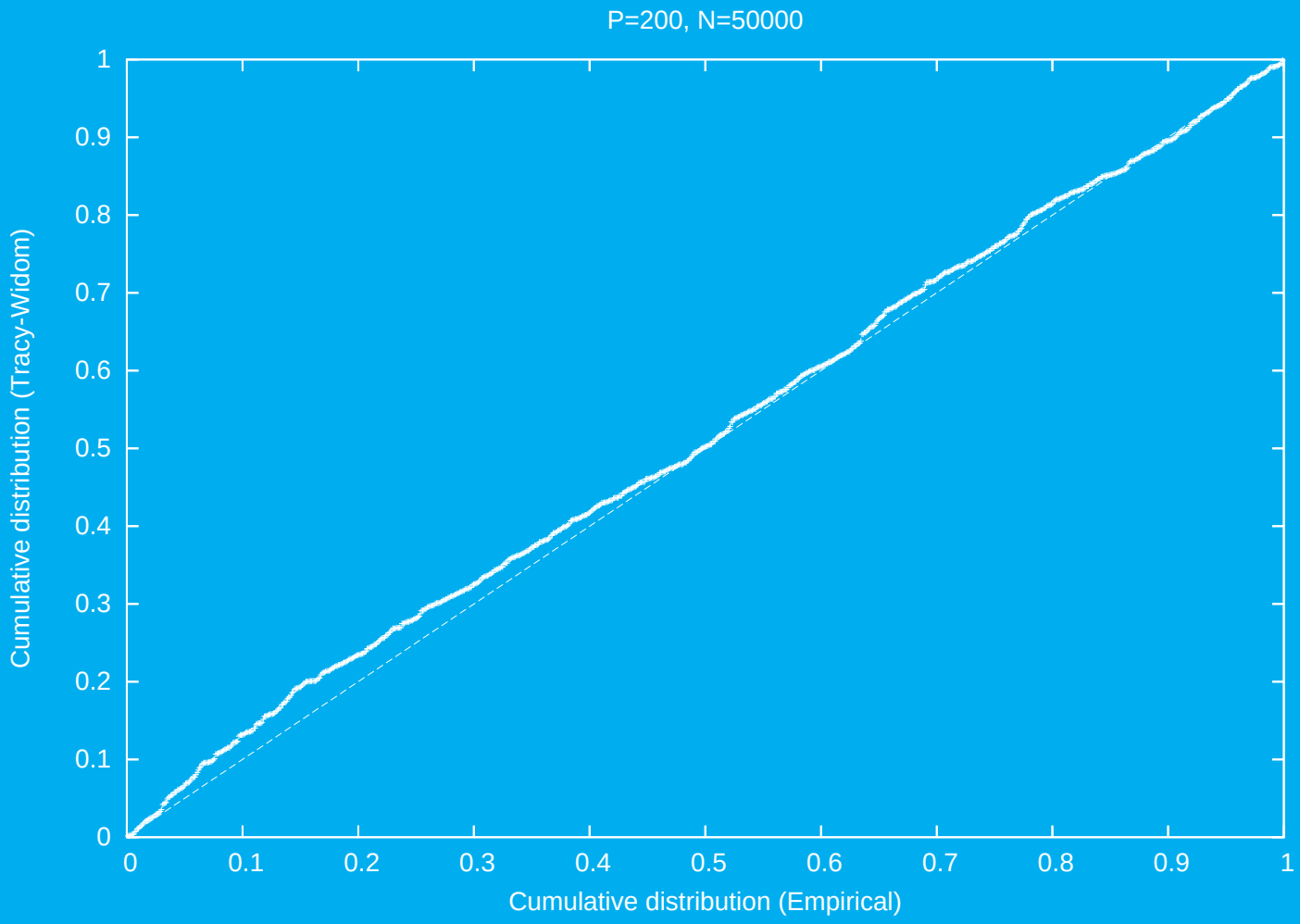
Some modern theory

Wishart matrix: Cells of M are standard normal.
Joint distribution of eigenvalues known since 1939
For λ_1: we have (Johnstone, 2001):
There exist explicit constants $\mu(p,n), \sigma(p,n)$
such that as $p, n \to \infty$

$$\frac{\lambda_1 - \mu}{\sigma} \to TW$$

in distribution.
TW is the *Tracy-Widom* distribution. (Tracy and Widom 1994).


Tracy-Widom distribution
Probability density
Argument
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-6
-4
-2
0
2
4
6


P=200, N=50000
Cumulative distribution (Tracy-Widom)
Cumulative distribution (Empirical)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Effective number of markers

Our matrices are *similar* to Wishart matrices.
(especially if markers have LD) *not Wishart*
Idea: Think of our covariance matrix as from a Wishart
BUT:

- Variance σ^2.of Gaussian in each cell unknown
- Number of markers N unknown.

Estimate N, σ.

A moments estimator:

$$\hat{N} = \frac{(p+1)(\Sigma_i \lambda_i)^2}{\left((p-1)\,\Sigma_i \lambda_i^2\right) - (\Sigma_i \lambda_i)^2} \tag{1}$$

$$\hat{\sigma}^2 = \frac{\Sigma_i \lambda_i}{(p-1)\hat{N}}$$

This works better than max likelihood.

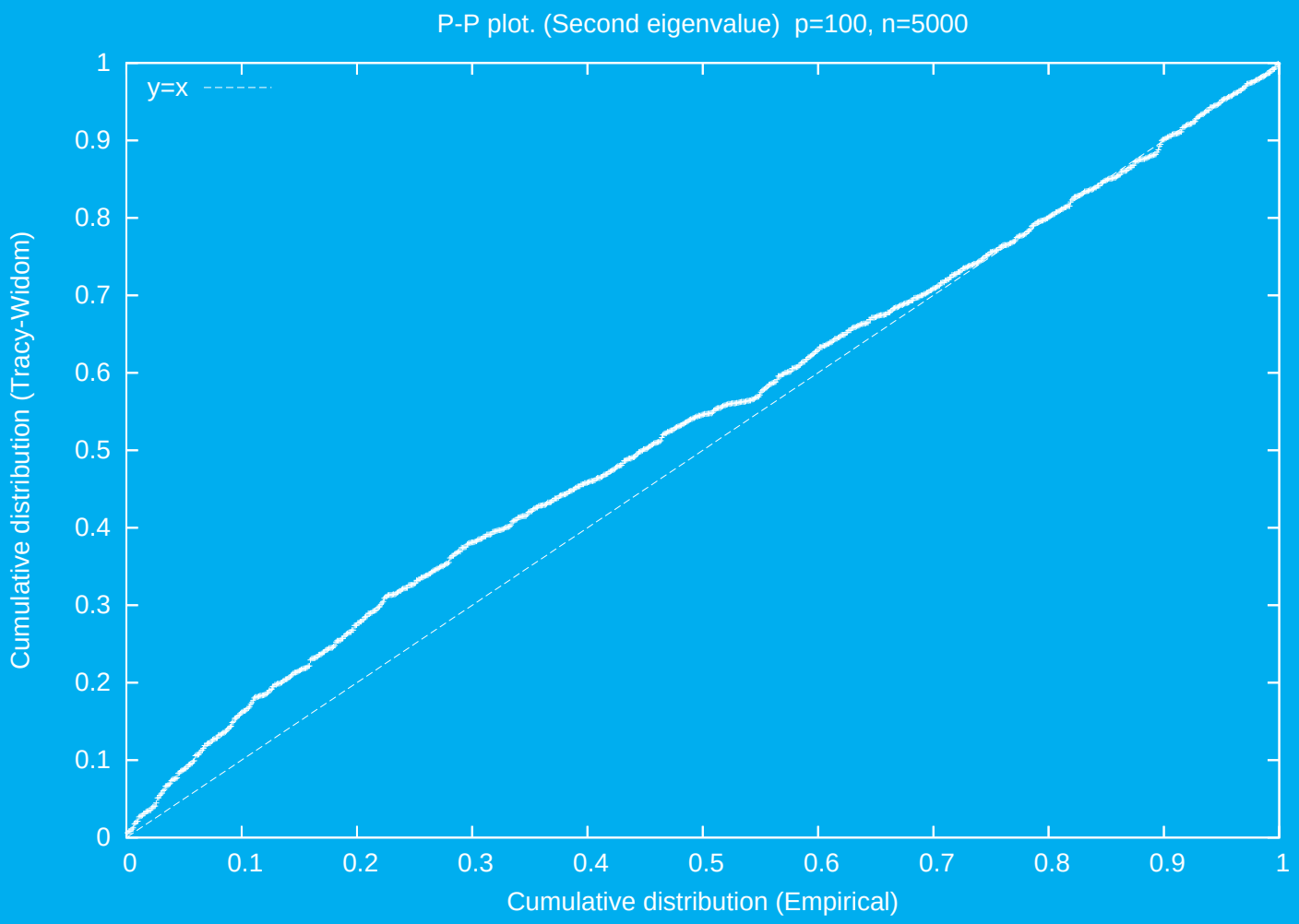
Testing additional values

Just ignore eigenvalues already found to be significant

Run test on remaining eigenvalues

This is conservative (Johnstone, 2001)

Works *extremely well*


P-P plot. (Second eigenvalue) p=100, n=5000
y=x
Cumulative distribution (Tracy-Widom)
Cumulative distribution (Empirical)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

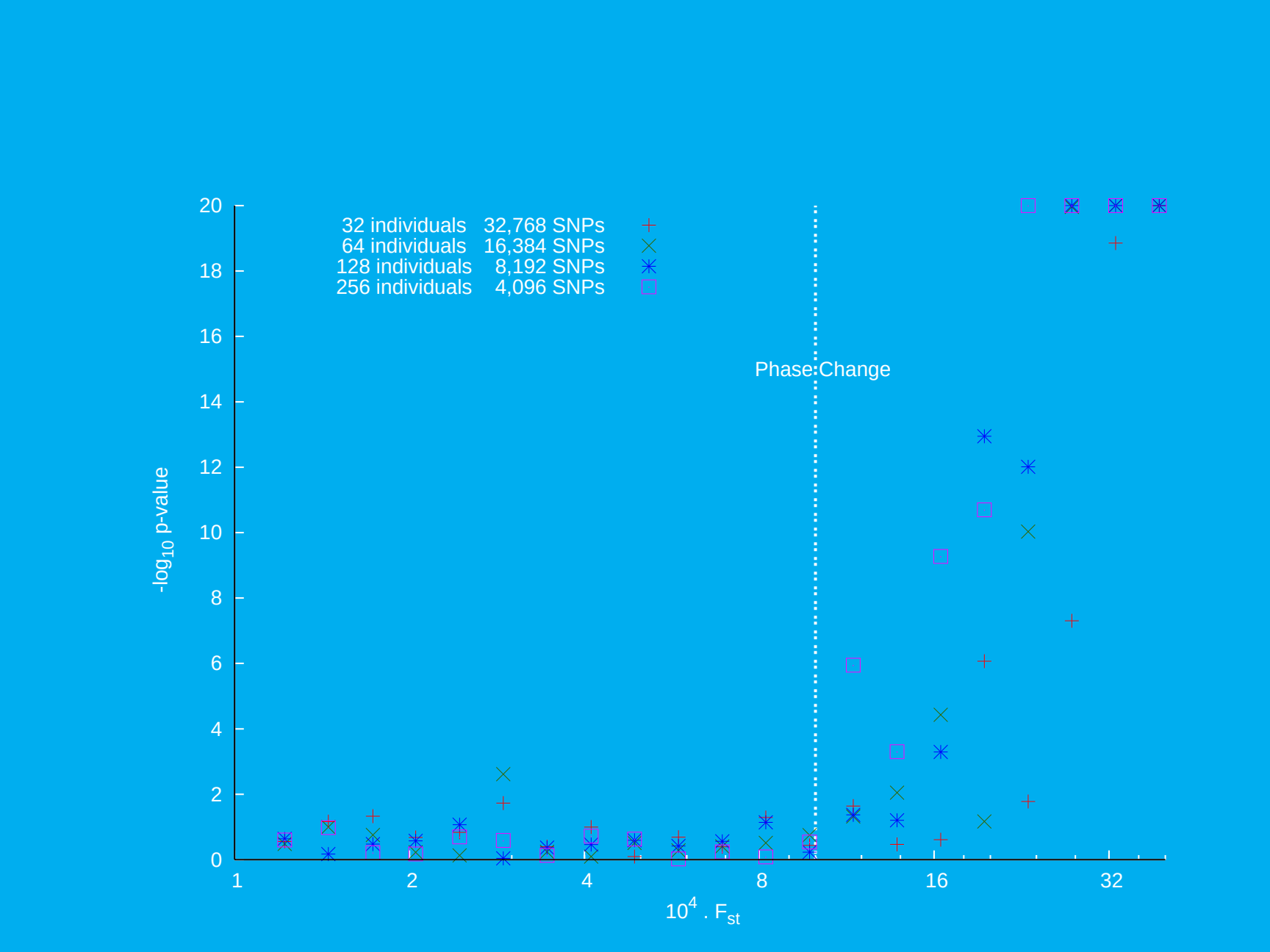
The BBP Phase Change

How much data do we need to detect population structure?
Let l_1 be the lead eigenvalue of *theoretical* covariance (rest of eigenvalues 1)
Baik et. al. (2005) prove that for a *complex* Wishart: Set $n/p = \gamma^2$. L_1 largest eigenvalue of sample covariance.
If $l_1 < 1 + 1/\gamma$ *then as* $p, n \to \infty$ L_1 *tends in distribution to the same distribution as when* $l_1 = 1$
If $l_1 > 1 + 1/\gamma$, *then the TW-statistic becomes unbounded.*
Change in character for small, large l_1.
Conjecture (Baik, Ben Arous, Péché (2005)) : True for real Wishart too. Partially proved by Baik and Silverman (2006).

Two equal-size populations with time divergence from root τ
$\tau = F_{ST}$

$$l_1 = 1 + p\tau$$

Phase transition when $\tau = 1/\sqrt{np}$.
We fix $np = 2^{20}$ Critical value: $\tau = 2^{-10}$


32 individuals 32,768 SNPs
64 individuals 16,384 SNPs
128 individuals 8,192 SNPs
256 individuals 4,096 SNPs
Phase Change
-log10 p-value
10^4 . Fst
0
2
4
6
8
10
12
14
16
18
20
1
2
4
8
16
32

Seems to work very well. Note the sensitivity with large data sets:
If $n = 100,000$ (independent), $p = 1000$
$\tau = .001$ very easy to detect.

Detectable pop. structure will be present on most large datasets.

Other (Bayesian) models

Ancient frequency P.
Pop freq. in K populations:

$$\mathbf{p} = (p_1, p_2, \ldots, p_K)$$

Conditional on P:

p_i has mean P

Covariance $P(1-P)B$

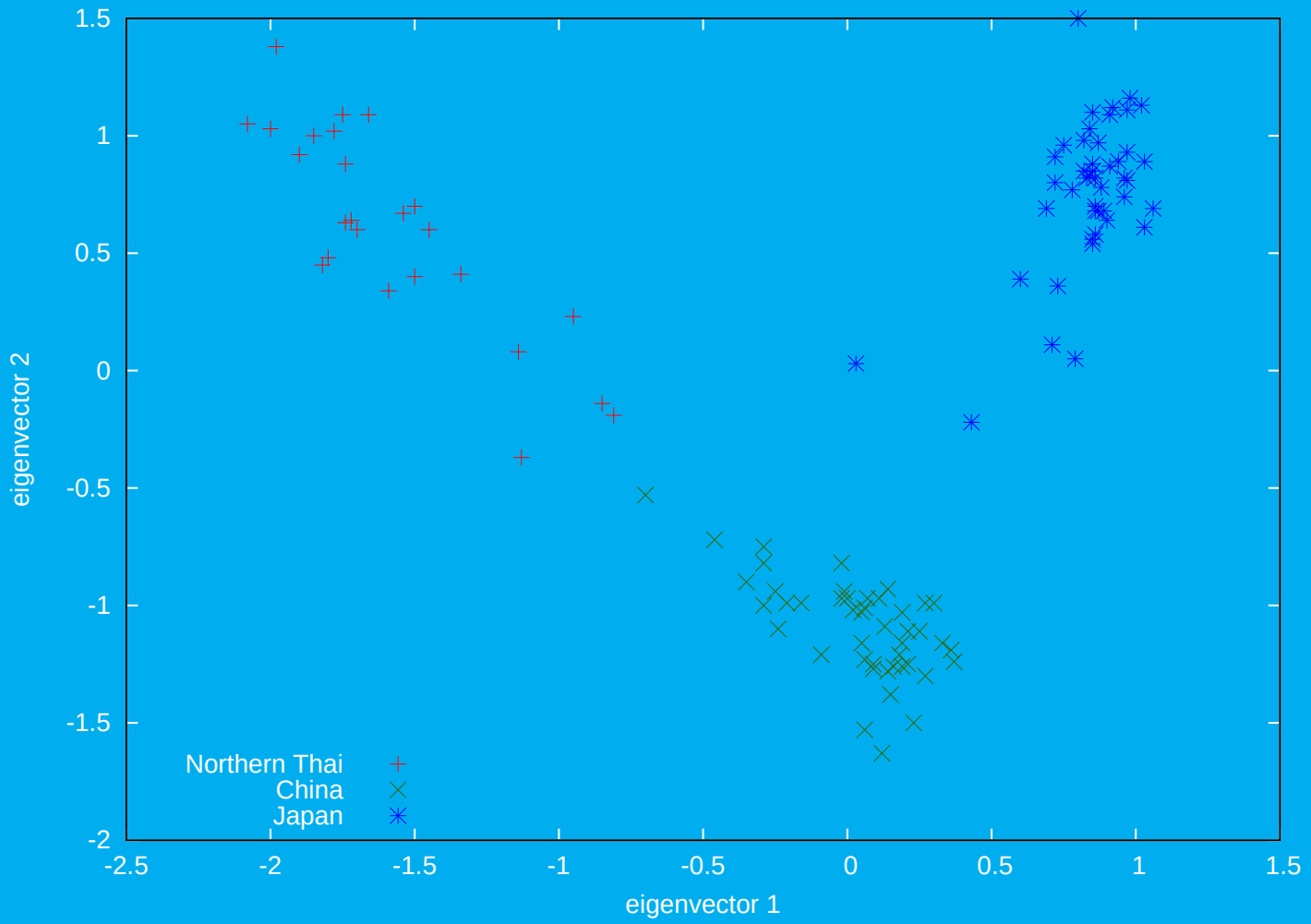
Nicholson et al., STRUCTURE ...

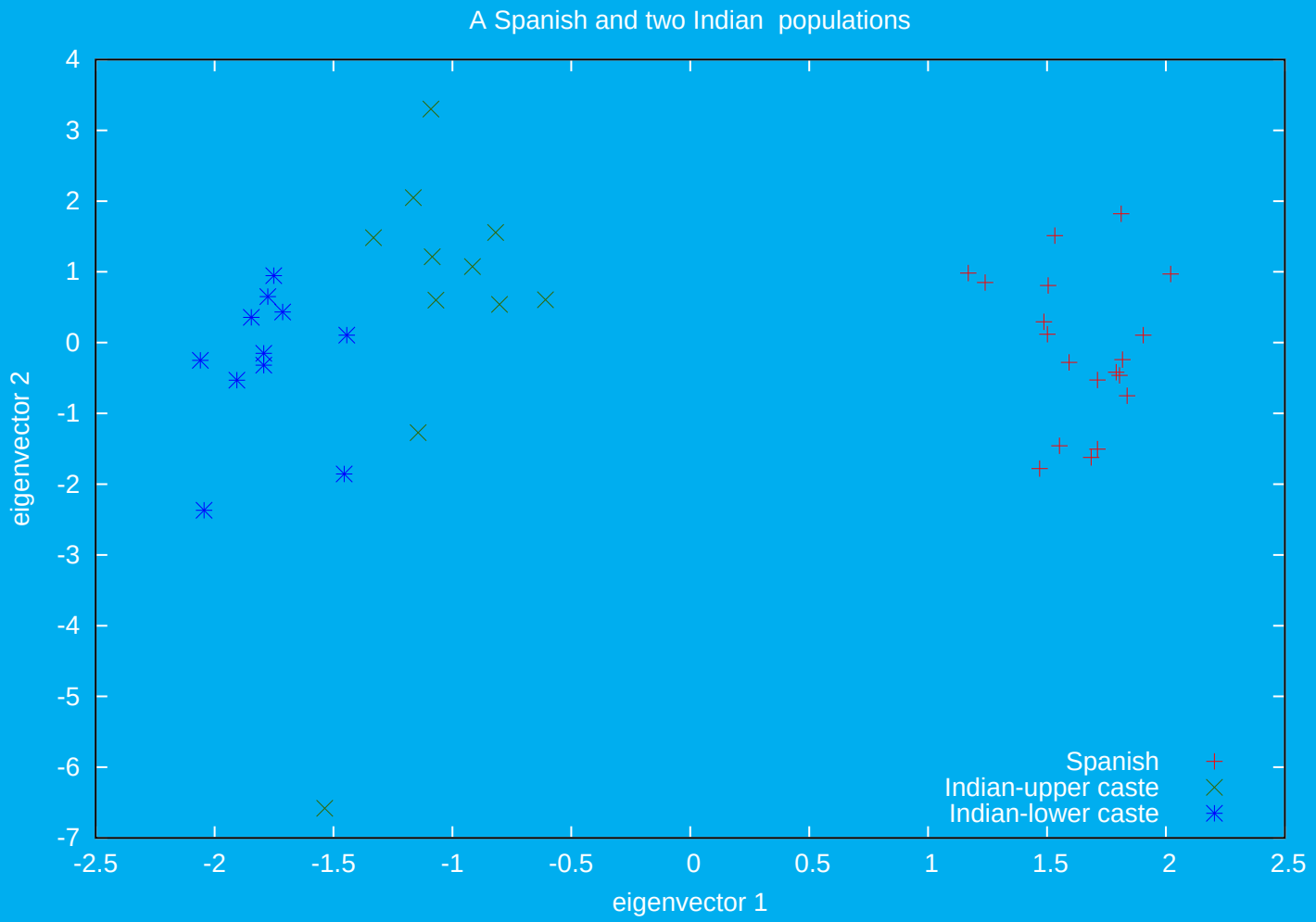
Sample covariance:

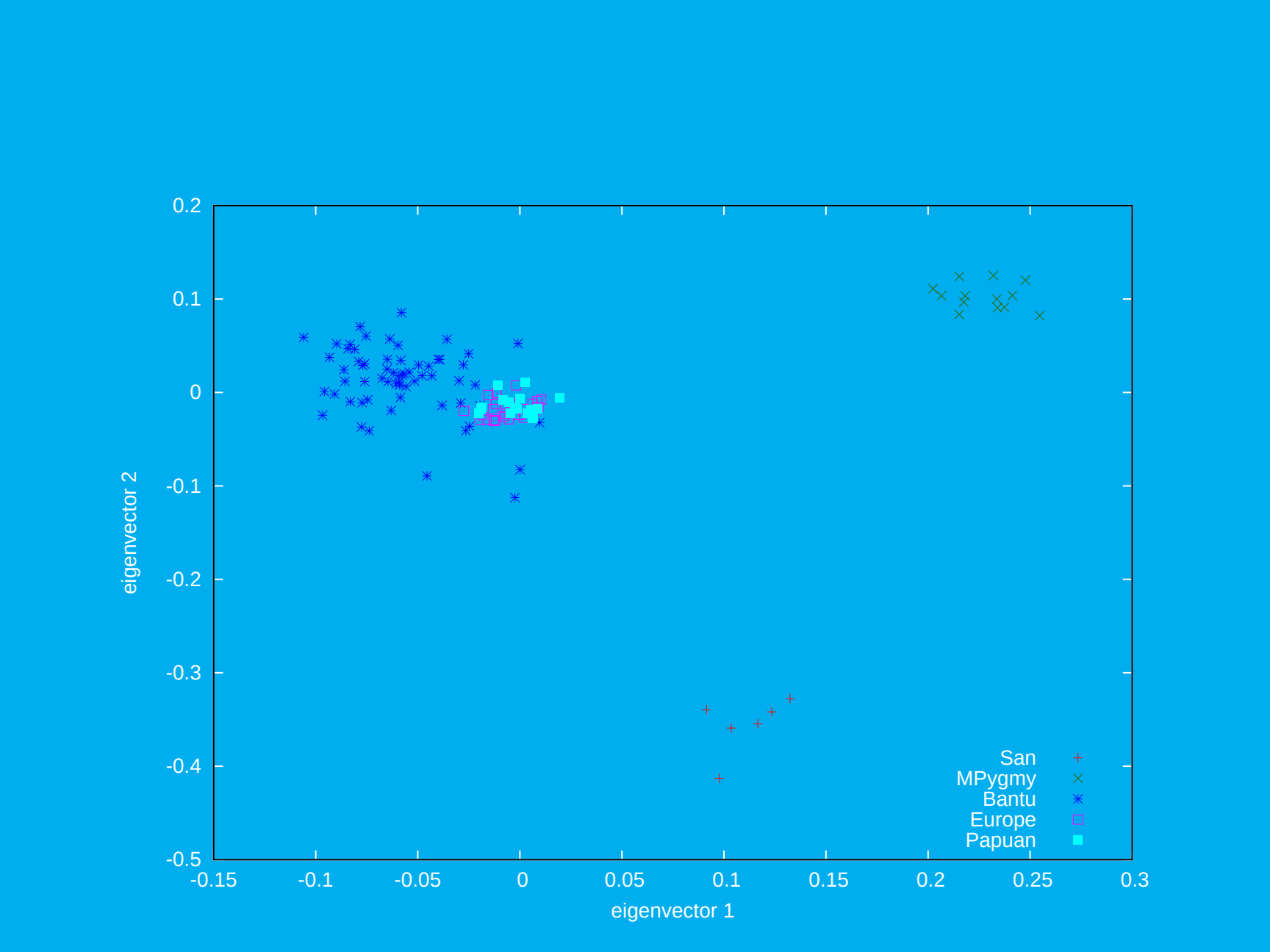
$(K-1)$ 'large' eigenvalues

1 zero

Assumptions analyzed by Johnstone (2001)


1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
eigenvector 2
eigenvector 1
Northern Thai
China
Japan


A Spanish and two Indian populations
eigenvector 2
eigenvector 1
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-7
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
Spanish
Indian-upper caste
Indian-lower caste


0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.15
-0.1
-0.05
0
0.05
0.1
0.15
0.2
0.25
0.3
eigenvector 1
eigenvector 2
San
MPygmy
Bantu
Europe
Papuan

Missing data

Missing data is problematic

- Sample handling different for different pops.
- Some missing data genuine
  (pop. dependent deletions)
- Informative missingness (Clayton)
- In particular worse quality DNA $\Longrightarrow$
  hets often called missing.

Acknowledgments

Thanks to:

- Alkes Price and David Reich
- Alan Edelman (MIT) and group
  for advice, references, preprint
- Plamen Koev for Tracy-Widom tables.
- Craig Tracy and Jinho Baik for latest on
  random matrix theory.