Principal components analysis

Gavin Band

Why do PCA?

PCA is good at detecting “directions” of major variation in your data. This might be:

- Population structure – subpopulations having different allele frequencies.
- Unexpected (“cryptic”) relationships.
- Artifacts such as genotyping errors, etc.

Apart from intrinsic interest, these are precisely the factors that need to be controlled for in association tests.

Performing PCA

1. Take genotype data(*)...

$$X = \overbrace{\begin{bmatrix} x_{11} & x_{12} & \cdots \\ x_{21} & x_{22} & \\ x_{31} & x_{32} & \\ \vdots & & \ddots \end{bmatrix}}^{N \text{ samples}} \left.\vphantom{\begin{bmatrix} x \\ x \\ x \\ x \end{bmatrix}}\right\} L \text{ SNPs}$$

(*) Suitably normalised – see later.

Performing PCA

1. Take genotype data[*]...

$$X = \overbrace{\begin{bmatrix} x_{11} & x_{12} & \cdots \\ x_{21} & x_{22} & \\ x_{31} & x_{32} & \\ \vdots & & \ddots \end{bmatrix}}^{N \text{ samples}} \left.\vphantom{\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}}\right\} L \text{ SNPs}$$

2. Form ‘relatedness matrix’...

$$R = \overbrace{\begin{bmatrix} r_{11} & r_{12} & r_{13} & \cdots \\ & r_{22} & r_{23} & \\ & & r_{33} & \\ & & & \ddots \\ & & & \end{bmatrix}}^{N \text{ samples}} \left.\vphantom{\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}}\right\} N \text{ samples}$$

$$R = \frac{1}{L} X^t X$$

r_{ij} = relatedness[*] between sample i and sample j.

[*] Suitably normalised – see later.

[*] With suitable normalisation:
$r_{ij} \approx 1$ if samples i and j are duplicates (or MZ twins)
$r_{ij} \approx 0$ if samples i and j are unrelated (relative to the sample.)

Performing PCA

1. Take genotype data(*)...

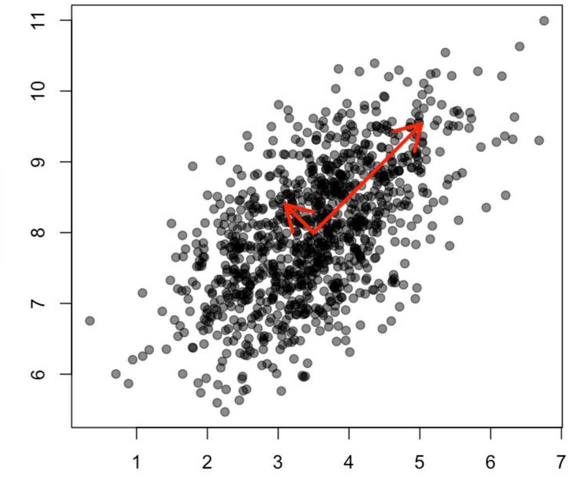
$$X = \overbrace{\begin{bmatrix} x_{11} & x_{12} & \cdots \\ x_{21} & x_{22} & \\ x_{31} & x_{32} & \\ \vdots & & \ddots \end{bmatrix}}^{N \text{ samples}} \left.\vphantom{\begin{bmatrix} a \\ a \\ a \\ a \end{bmatrix}}\right\} L \text{ SNPs}$$

2. Form ‘relatedness matrix’...

$$R = \overbrace{\begin{bmatrix} r_{11} & r_{12} & r_{13} & \cdots \\ & r_{22} & r_{23} & \\ & & r_{33} & \\ & & & \ddots \end{bmatrix}}^{N \text{ samples}} \left.\vphantom{\begin{bmatrix} a \\ a \\ a \\ a \end{bmatrix}}\right\} N \text{ samples}$$

3. Eigen-decompose it...

$$R = UDU^t$$



Eigen-decomposition picks out *directions in the data along which the variance is maximised*.

Eigenvalues represent *the variance of the data along these directions*.

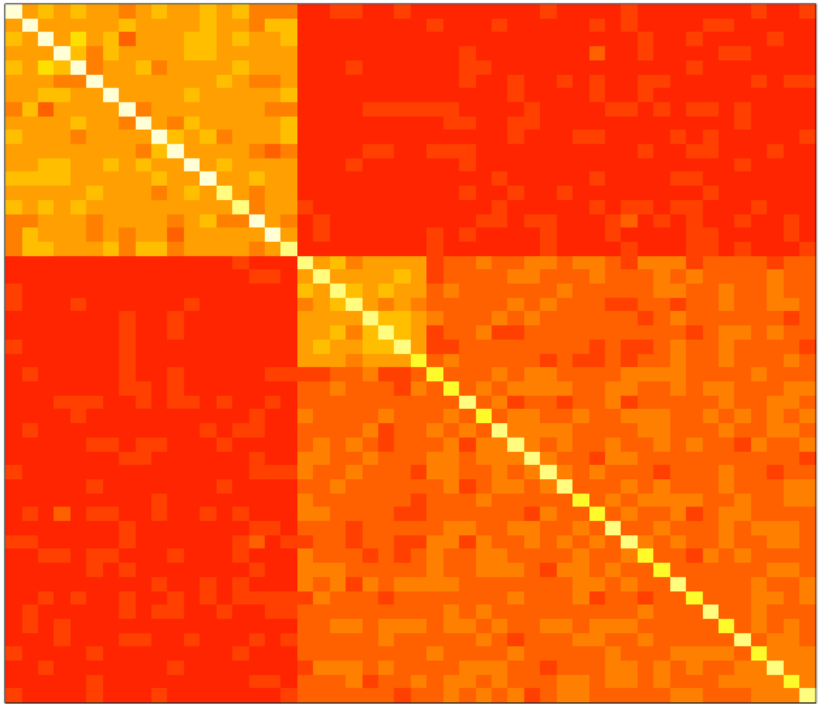
You can do this in R! E.g:

```
> R = 1/L * (t(X) %*% X)
> V = eigen(R)$vectors
> plot( V[,1], V[,2] )
```

Example

(Simulated data, N=50 individuals, L=1000 SNPs)

Relatedness matrix R

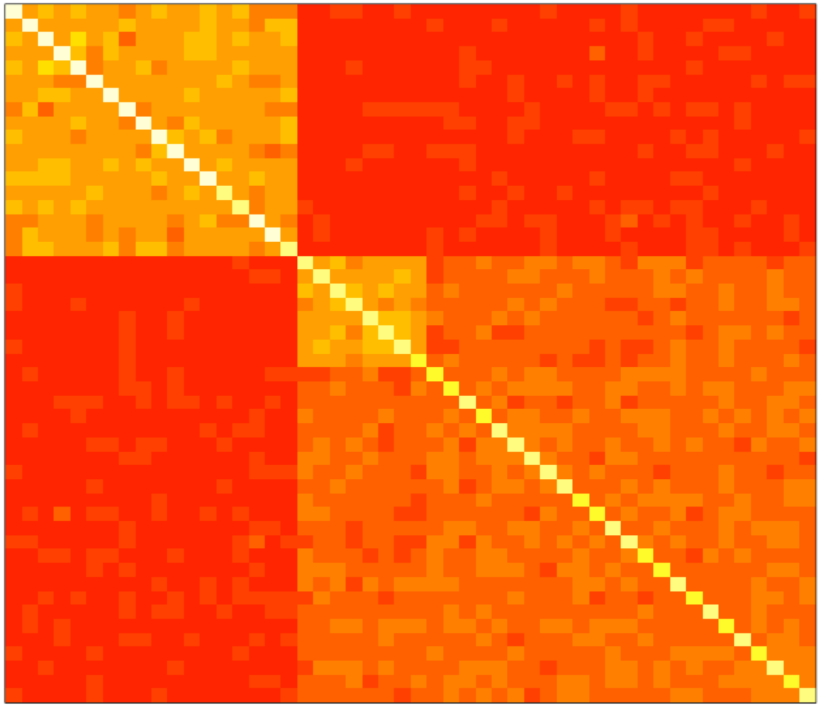


```
> R = (1/1000) %*% (t(X) * X)
```

Example

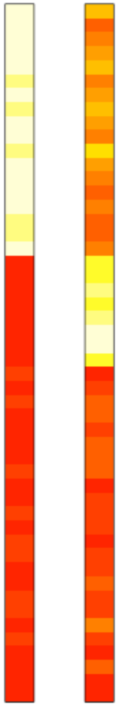
(Simulated data, 50 individuals, 1000 SNPs)

Relatedness matrix R



Eigenvectors

v_1 v_2

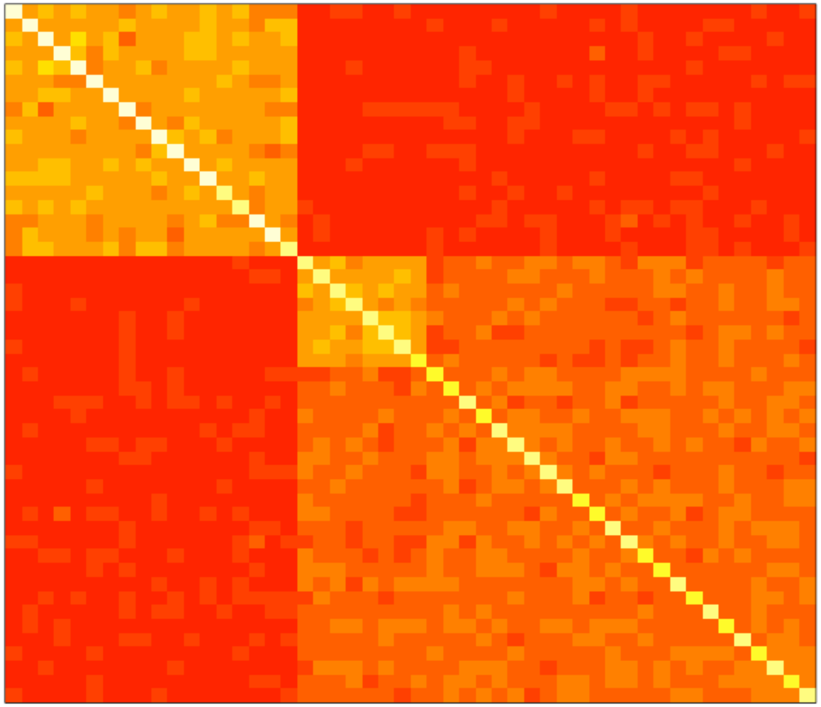


```
> V = eigen(R)$vectors
```

Example

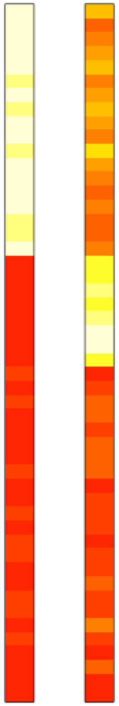
(Simulated data, 50 individuals, 1000 SNPs)

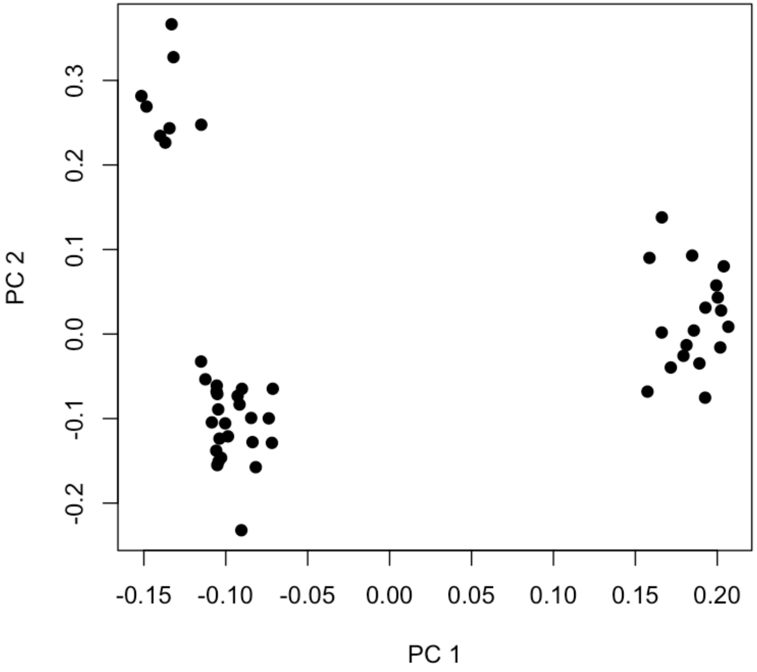
Relatedness matrix R



Eigenvectors

v_1 v_2



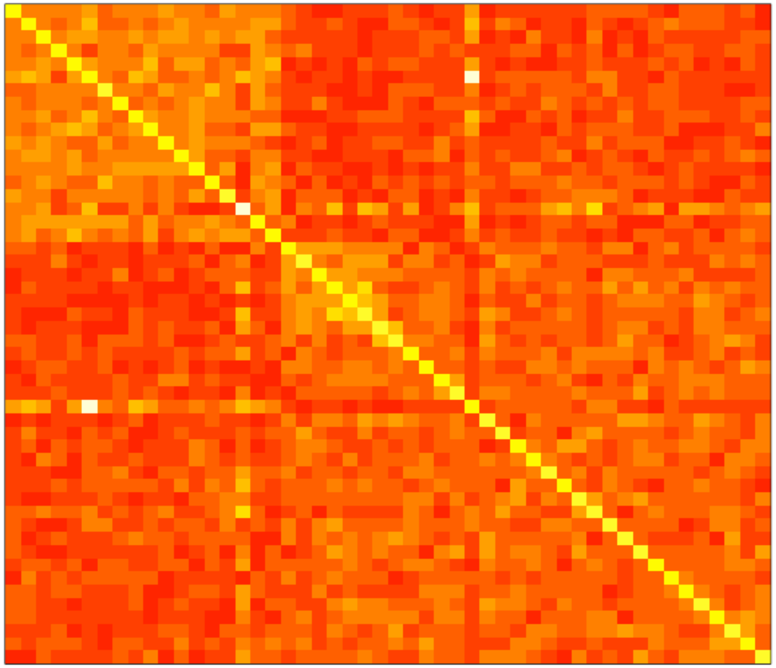


```
> plot( V[,1], V[,2] )
```

Caution!

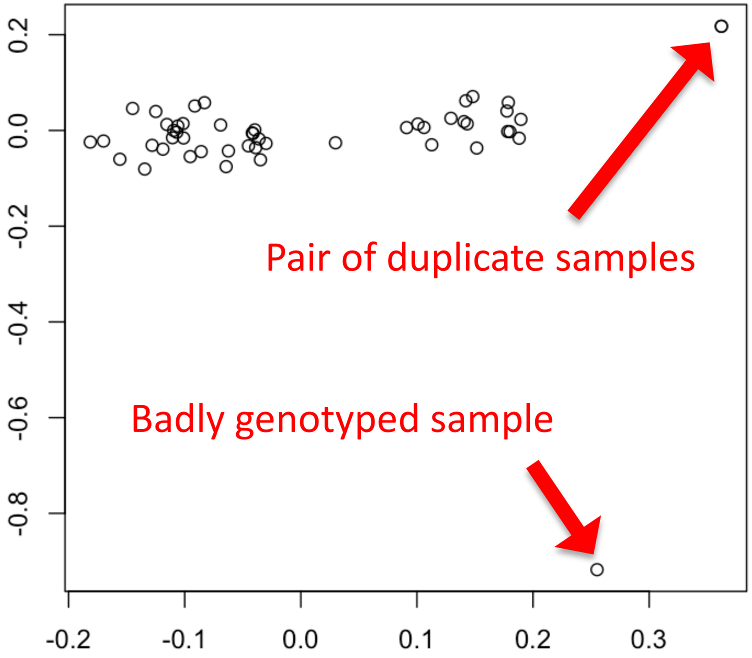
PCA picks up *any* source of variation

duplicate



duplicate

Badly genotyped sample



Pair of duplicate samples

Badly genotyped sample

Relatedness or why scale by f(1-f)

At a SNP with frequency *f* in a ‘base’ population.

What is the probability of seeing these alleles in two haplotypes drawn from the population?

		Individual 2		
Individual 1		Allele A	Allele B	*frequency*
	Allele A			f
	Allele B			1-f
	frequency	f	1-f	

Relatedness or why scale by f(1-f)

At a SNP with frequency *f* in a 'base' population.

What is the probability of seeing these alleles in two haplotypes drawn from the population?

		Individual 2		
Individual 1		Allele A	Allele B	*frequency*
	Allele A	f^2	$f(1-f)$	f
	Allele B	$f(1-f)$	$(1-f)^2$	1-f
	frequency	f	1-f	
	Correlation = 0			

"Unrelated" individuals

Alleles drawn independently

Relatedness or why scale by f(1-f)

At a SNP with frequency *f* in a ‘base’ population.

What is the probability of seeing these alleles in two haplotypes drawn from the population?

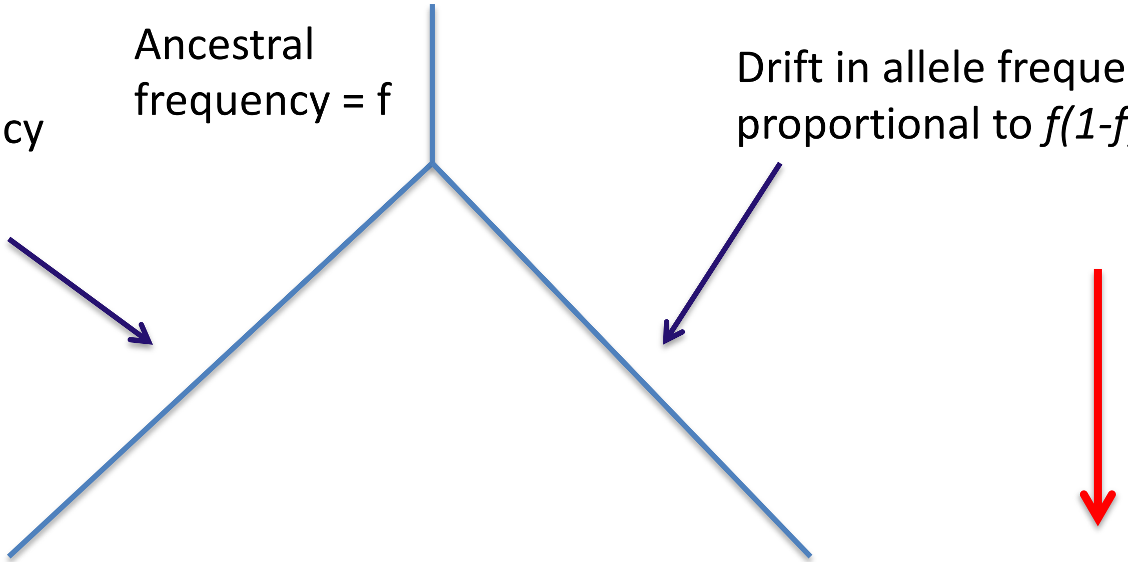
		Individual 2		
		Allele 1	Allele 2	total
Individual 1	Allele 1	$rf + (1-r)f^2$	$(1-r)f(1-f)$	f
	Allele 2	$(1-r)f(1-f)$	$r(1-f)+(1-r)(1-f)^2$	1-f
		f	1-f	
	Correlation = r			

Individuals with relatedness *r*

Alleles co-inherited ”identical by descent” with probability r

Relatedness and population history – a heuristic explanation

Ancestral population



Ancestral frequency = f

Drift in allele frequency proportional to *f(1-f)*

Drift in allele frequency proportional to *f(1-f)*

Time

Population 1
SNP frequency $f_1 = f + \epsilon_1$

Population 2
SNP frequency $f_2 = f + \epsilon_2$

So $\frac{f_i - f}{\sqrt{f(1-f)}}$ = the amount of drift in population i, similar across all variants

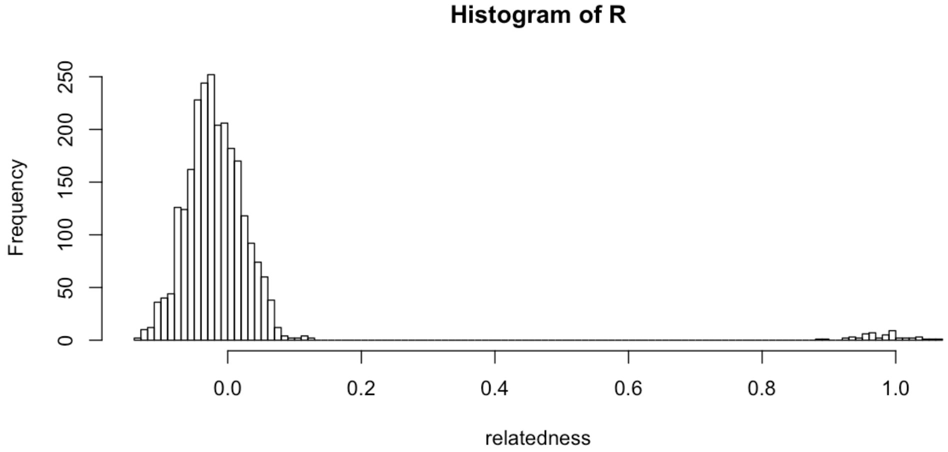
Relatedness

$$r_{ij} = \frac{1}{L} \sum_{\text{SNPs}} \frac{(g_i - 2f)(g_j - 2f)}{2f(1-f)}$$

Or: mean centre rows of X and divide by standard deviation, and compute as before:

$$R = \frac{1}{L} X^t X$$

Because f comes from the sample (not an ancestral population), ½r_{ij} is almost the same as a *kinship coefficient,* but is relative to the sample, not an ancestral population.



Association testing

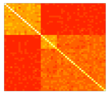
Without controlling for structure:

Outcome ~ baseline + genotype

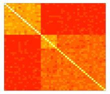
Traditional approaches control for structure using a number of principal components.:

Outcome ~ baseline + genotype + PC_1 + PC_2 + ...

The most recent *mixed model* approach includes the whole relatedness matrix to control for structure:

Outcome ~ baseline + genotype + 

Association testing with linear mixed models

`Outcome ~ baseline + genotype +` 

This is a bit like including all the PCs in a single regression, but constrained to explain a proportional amount of residual variation. In some circumstances it's been shown to control for structure better than using principal components directly. For example see *"Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis", IMSGC & WTCCC2, Nature 2011.* Or play with it at `http://www.well.ox.ac.uk/wtccc2/ms.`

However – these are *linear* models and some caveats remain in their use for case/control studies.

Summary

- PCA good at picking up sources of variation in datasets, including genetic datasets.

- Any form of variation can be picked up – population structure, but also cohort or plate effects, genotyping error, sample duplication.

- *This is what we want* when controlling for structure / unwanted variation in an association test.

Software for performing PCA

- Plink (v1.9 or above)

`http://www.cog-genomics.org/plink2`

- EIGENSOFT

`http://genetics.med.harvard.edu/reich/Reich_Lab/Software.html`

- Or use R!

Software for mixed model analysis

- GCTA

http://genetics.med.harvard.edu/reich/Reich_Lab/Software.html

- FastLMM

http://research.microsoft.com/en-us/um/redmond/projects/mscompbio/fastlmm/

- MMM

http://www.helsinki.fi/~mjxpirin/download.html

- GEMMA

http://www.xzlab.org/software.html

Recommended reading

- “*Population Structure and Eigenanalysis*”, Patterson N, Price AL, Reich D, PLoS Genetics (2006). (The “SmartPCA” paper)

- “*Population Structure and Cryptic Relatedness in Genetic Association Studies*”, Astle W. and Balding DJ, Statistical Science (2009).

- "*Reconciling the analysis of IBD and IBS in complex trait studies*”, Powell JE, Visscher PM, Goddard ME Nat. Rev. Genetics (2010).

- “*A Genealogical Interpretation of Principal Components Analysis*”, Gil McVean, PLoS Genetics (2009).

- “*Interpreting principal component analyses of spatial population genetic variation*”, John Novembre and Matthew Stephens, Nature Genetics (2008).

- “*Advantages and pitfalls in the application of mixed-model association methods”,* Yang et al, Nature Genetics (2014)