Genealogical Interpretation of PCA - Gil McVean

MATH285K - Spring 2010 *Presenter: Darren Kessner*

1 Genotype Matrix

$$Z = \left[\begin{array}{c|cccccc} & ind_1 & \cdots & ind_i & \cdots & ind_n \\ \hline snp_1 & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & & & \vdots \\ snp_s & 1 & & z_{si} & & 1 \\ \vdots & \vdots & & & \ddots & \vdots \\ snp_L & 1 & \cdots & 0 & \cdots & 0 \end{array}\right]$$

zero-center the rows $\Rightarrow$ normalized matrix X

$$X_{si} = Z_{si} - \frac{1}{n} \sum_{j=1}^{n} Z_{si}$$

2 Principal Components Analysis (PCA)

direction of maximum variance

$\Updownarrow$

maximize $|X^t v|$ on $|v| = 1$

$\Updownarrow$

maximize $v^t X X^t v$ on $v^t v = 1$

$\Updownarrow$

$$XX^t v = \lambda v$$

$\Updownarrow$

find eigenvectors of XX^t

3 PCA and Singular Value Decomposition (SVD)

In practice,

$$\begin{aligned}\text{\# of SNPs} &\approx 500,000\\ \text{\# of individuals} &\approx 1000\end{aligned}$$

So XX^t is really big. Instead, calculate eigenvectors V of X^tX:

$$X^tX = V\Lambda V^t$$

Letting $\Sigma = \Lambda^{1/2}$ and $U = XV\Sigma^{-1}$ gives us the SVD of X:

$$X = U\Sigma V^t$$

Then the columns of U are eigenvectors of XX^t:

$$\begin{aligned}XX^t &= (U\Sigma V^t)(V\Sigma U^t)\\ &= U\Lambda U^t\end{aligned}$$

4 Kinship Matrix and Pairwise Coalescence Times

The entries in the matrix X^tX can be seen as the degree of relatedness, or *kinship* between individuals.

$$M = \frac{1}{L}X^tX$$

$$\begin{aligned}E(M_{ij}) &= \frac{1}{L}\sum_{s=1}^{L} E(X_{si}X_{sj})\\ &= \frac{1}{L}\sum_{s=1}^{L} E\left[(Z_{si}-\bar{Z}_s)(Z_{sj}-\bar{Z}_s)\right]\\ &= E(Z_iZ_j) - E_k(Z_iZ_k) - E_k(Z_jZ_k) + E_{kl}(Z_kZ_l)\\ &\vdots\\ &= \frac{1}{\bar{T}}(\bar{t}_i + \bar{t}_j - \bar{t} - \bar{t}_{ij})\end{aligned}$$

The expectation $E(Z_i Z_j)$ can be considered to be the probability that i and j both share the derived allele at some locus. This happens when a mutation occurs on the branch that i and j share before splitting apart. Hence,

$$E(Z_i Z_j) = \frac{E(T_{MRCA}) - E(t_{ij})}{E(T)}$$

where $E(T)$ is the expected total branch length of the coalescence tree.

The main result is that the expected kinship matrix can be expressed in terms of expected coalescence times.

5 Example: Two Populations

We consider the example of two populations A and B, with proportions ϕ and $1-\phi$ of the samples, respectively. We suppose that the two populations have been separated for time Δ.

We can calculate the expected coalescence times, e.g.:

$$\bar{t}_{ij} = \begin{cases} 0 & \text{if } i = j \\ 1 & \text{if } i, j \text{ in same population} \\ 1 + \Delta & \text{if } i, j \text{ in different populations} \end{cases}$$

$$\bar{t}_i = \begin{cases} \phi + (1-\phi)(1+\Delta) & \text{if } i \text{ in population A} \\ (1-\phi) + \phi(1+\Delta) & \text{if } i \text{ in population B} \end{cases}$$

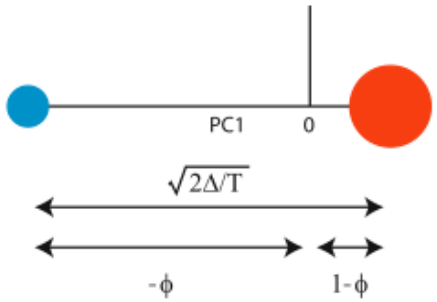
$$\bar{t} = [\phi^2 + (1-\phi)^2] + (1+\Delta)[2\phi(1-\phi)]$$

In this case, the kinship matrix has a block structure.

$$E(M) \approx \frac{1}{\bar{T}} \left[\begin{array}{cc|cc} 1+\alpha & \alpha & -\sqrt{\alpha\beta} & -\sqrt{\alpha\beta} \\ \alpha & 1+\alpha & -\sqrt{\alpha\beta} & -\sqrt{\alpha\beta} \\ \hline -\sqrt{\alpha\beta} & -\sqrt{\alpha\beta} & 1+\beta & \beta \\ -\sqrt{\alpha\beta} & -\sqrt{\alpha\beta} & \beta & 1+\beta \end{array}\right]$$

The expected projections of individuals onto the first principal component can then be expressed in terms of ϕ.

$$\text{projections} \approx \sqrt{\frac{2\Delta}{\bar{T}}}\,(1-\phi,\cdots,1-\phi,-\phi,\cdots,-\phi)$$



6 Admixture

Admixture is the mixing of two populations that have been isolated for some amount of time.

We consider an admixed individual – in this case, we can express the individual's expected projection onto the first principal component in terms of coalescence times, and with some more algebra, in terms of the admixture proportion θ_j of the individual's genome.

Expected projection of individual j on $PC1$:

$$E(y_{1j}) = \sqrt{\frac{1}{c\bar{T}}}\left[\bar{t}_{jB} - \bar{t}_{jA} + \frac{1}{2}\left(t_{AA} - t_{BB} + (1-2\phi)\,c\right)\right]$$

$$= \sqrt{\frac{c}{\bar{T}}}\,(\theta_j - \phi)$$

where $c = 2t_{AB} - t_{AA} - t_{BB}$

The final formula is obtained by calculating coalescence times based on ancestry proportion:

$$t_{jA} = \theta_j t_{AA} + (1-\theta_j)t_{AB}$$
$$t_{jB} = \theta_j t_{AB} + (1-\theta_j)t_{BB}$$

References

[1] McVean G (2009) A Genealogical Interpretation of Principal Components Analysis. PLoS Genet 5(10): e1000686.

[2] Novembre et al. (2008) Genes mirror geography within Europe. Nature 456: 98-101.