Introduction to Statistical Genetics

Max Turgeon

STAT 4690–Applied Multivariate Analysis

Overview i

- We will look at three papers that use PCA in slightly different ways:
    1. Price *et al*. "Principal components analysis corrects for stratification in genome-wide association studies." *Nature genetics* (2006).
    2. Leek & Storey. "Capturing heterogeneity in gene expression studies by surrogate variable analysis." *PLoS genetics* (2007).
    3. Gao *et al*. "A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms." *Genetic epidemiology* (2008).

Overview ii

- The main purpose of this lecture is to:
  - Introduce you to important concepts in applied statistics (e.g. confounding and multiple testing).
  - Give you a sense of the versatility of PCA.
  - Give an overview of the interplay between theoretical, methodological and applied research in statistics.
- All three papers can be found on UM Learn (or online).

Introduction to Genetics

DNA

- Long molecule, double-stranded, made of four types of nucleotides:
  - **T**hymine
  - **C**ytosine
  - **G**uanine
  - **A**denine
- Nucleotides are paired:
  - A-T and C-G
- This pairing allows *replication*:
  - DNA molecule opens up
  - From complimentarity, we can reconstruct two molecules.

Central Dogma

- Explains how DNA leads to proteins
- DNA $\Longrightarrow$ RNA $\Longrightarrow$ Protein
  - **Transcription** and **translation**
  - $(T, C, G, A) \Longrightarrow (U, C, G, A)$
  - Codon (i.e. triple) $\Longrightarrow$ Amino acid
- **Gene**: sequence of nucleotides that encodes a protein
  - Other gene products are possible: microRNA, tRNA, etc.

Genetic variation

- Random mutations
- After fertilization, a zygote has a copy of each chromosome from each parent
    - *Assortment* is random
- Before that, at meiosis, there is *recombination*
- At the population level:
    - Population bottleneck
    - Founder effect
    - Natural selection
- The most studied genetic variation: *Single Nucleotide Polymorphism* (SNP)
    - A location in the genome where in the population we observe at least different nucleotides

Some vocabulary

- **Allele**: Sequence observed at a specific location
  - One basepair for SNP
  - Can be longer
- **Minor/Major Allele**: Least/Most observed allele in a population
- **MAF**: Minor Allele Frequency
  - Frequency at which the minor allele is observed in the population
  - *Population specific*
- **Phenotype**: Observable characteristic or trait

Gene Expression

- All cells have the same DNA, but they produce different proteins.
- Same cell type, under different conditions, can also produce different proteins.
- Different mechanisms:
    - Transcription factors
    - Epigenetics

Population Stratification

High-throughput technologies

- Since the mid-2000s, SNP data is routinely collected at hundreds of thousands, or even millions, of genetic loci.
- There are two basic types of technologies:
  1. *Micro-arrays*: Designed to identify the allele at pre-selected loci
  2. *Next-generation sequencing*: Sequence large portions of DNA.
- The data is similar: high-dimensional data (i.e. more variables than observations).

Genome-Wide Association Studies

- **GWAS**: Every genetic measurement is tested for association with a single (or a few) *phenotype* of interest.
- *Goal*: Find genetic locations with evidence of causal effect on disease of interest
    - Or at least genetic locations that inherited together with causal locus
- Two main challenges:
    - Multiple testing (we'll come back to it)
    - Population stratification (i.e. confounding)

Confounding

- **Confounder**: common cause of both the exposure and outcome of interest
  - E.g. Obesity is a cause of diabetes and cardiovascular diseases.
- Failure to adjust for confounding can lead to spurious correlations
- Three main methods for confounder adjustment:
  - Randomisation
  - Regression model
  - Weighting

Population stratification as a confounder

- Because of migration patterns and natural selection, some alleles are preferentially selected in certain populations
  - E.g. *LCT* gene and lactose intolerance.
- **Population stratification**: "allele frequency differences between cases and controls due to systematic ancestry differences" (Price et al)
- If a given allele and the phenotype of interest are more prevalent in a certain population, this may give rise to spurious correlation.
- **Major problem**: Population stratification is very hard (if not impossible) to measure accurately.
- *Solution*: Estimate it from the collected genetic data.

EIGENSTRAT i

- Price et al. (2006) proposed a method to adjust for population stratification in GWASs.
- Essentially, the population stratification is estimated using the principal components of the genetic data.
- More precisely, let G be the $n \times p$ matrix of genotypes
  - The (i, j)-th entry g_{ij} is the value at the j-th locus for the i-th sample.
  - $g_{ij} \in \{0, 1, 2\}$ counts the number of copies of the minor allele.

EIGENSTRAT ii

- Create matrix X by normalizing G
  - Subtract the mean
  - Divide by binomial standard deviation $\sqrt{p_j(1-p_j)}$.
- Select first k eigenvalues of the covariance matrix of X.
- Adjust for confounding by including the PCs into a regression model.

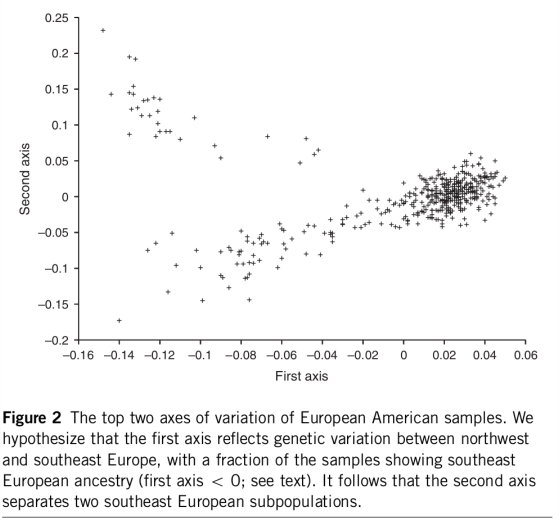


Figure 2 The top two axes of variation of European American samples. We hypothesize that the first axis reflects genetic variation between northwest and southeast Europe, with a fraction of the samples showing southeast European ancestry (first axis < 0; see text). It follows that the second axis separates two southeast European subpopulations.

Figure 1

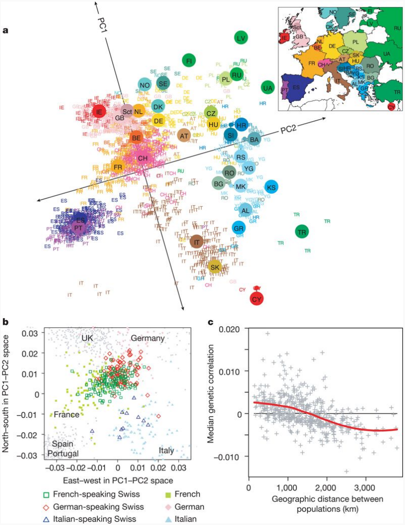


Figure 2

Novembre *et al*. “Genes mirror geography within Europe.”

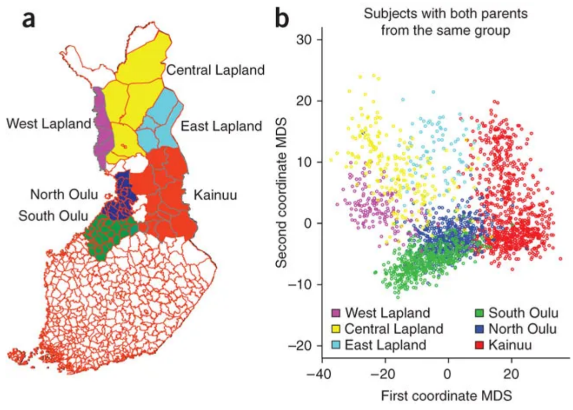


Figure 3

Sabatti *et al*. "Genome-wide association analysis of metabolic traits in a birth cohort from a founder population."

Further comments

- There is a vast literature around how to use PCA to account for population stratification
  - How many PCs to retain.
  - Theoretical justification.
  - Power analysis.
  - How granular can you get.
- *Note*: This is not how 23andMe and AncestryDNA estimate your ethnicity!
- PCA can also be used to estimate under population substructures in your data.
  - E.g. Cryptic relatedness

Adjusting for Unwanted Variation

Gene expression studies i

- As for SNP data, gene expression is nowadays measured using one of two high-throughput technology:
    - Micro-arrays
    - Next-generation sequencing
- What is measured in these experiments is the (relative) **abundance** of RNA products.
    - It can be hard to measure protein products (but see proteomics)
    - We may also be interested in other gene products
- You can think of micro-array data as *continuous*; sequencing data as *counts*

Gene expression studies ii

- There are essentially two type of analyses:
  - Association between gene expression and SNP (i.e. eQTL)
  - Association between gene expression and phenotype
- You can think of these two approaches as related to *transcription* and *translation*, respectively.

Sources of variation

- Leek & Storey are interested in the second type (i.e. GE and phenotype).
- Their model starts by identifying three main sources of variation:
    - **Modeled** variation: This is the variation coming from the variables you measured and included in your model. The phenotype of interest goes here.
    - **Unmodeled** variation: This is the variation coming from variables that you may or may not have measured, but in any case they are not included in the model. These variables typically affect more than one gene.
    - **Random** variation: This is the gene-specific *error* term, and it is assumed to be independent between genes.

Two models i

- Let X_{ij} be the gene expression value at gene i from individual j.
    - **Note**: The indices are in the opposite order of what we typically see!
- Let Y_j be the primary variable of interest for individual j.
- Let $\mathbf{G}_\ell = (G_{\ell 1}, \ldots, G_{\ell n})$ be the ℓ-th unmodeled source of variation.
- Following their breakdown of sources of variation, they posit two models:

Two models ii

1. The first one only contains the primary variable:

   $$X_{ij} = \mu_i + f_i(Y_j) + \varepsilon_{ij},$$

   where μ_i is a gene-specific mean, f_i is a gene-specific function modeling the relationship between X_{ij} and Y_j, and ε_{ij} is an error term with mean zero.
2. The second one also contains the variables $\mathbf{G}_\ell$:

   $$X_{ij} = \mu_i + f_i(Y_j) + \sum_{\ell=1}^{L} \gamma_{\ell i} G_{\ell j} + \tilde{\varepsilon}_{ij},$$

   where $\gamma_{\ell i}$ are the linear regression coefficients for the variables $\mathbf{G}_\ell$, and $\tilde{\varepsilon_{ij}}$ is a different error term, also with mean zero.

A few comments i

- In the models above, there is only one variable of interest, but the approach can easily be extended to incorporate more variables of interest.
- The functions f_i are there for *generality.* This could be the identity function (i.e. simple linear regression), they could be a gene-specific transformation of the variable of interest (e.g. $\log$), or they could be something more complex like a spline or fractional polynomial..
- The variables $\mathbf{G}_\ell$ are typically unobserved, and therefore we cannot estimate them from the data without adding any constraint.

A few comments ii

    - We could replace them and their coefficients $\gamma_{\ell i}$ by an orthogonal transformation and still get the same model.
- The constraint we will add is that they are *orthogonal*.
- The (latent) variables we will estimate are called **surrogate variables**.

SVA algorithm

1. Fit the first model (i.e. without $\mathbf{G}_\ell$) to get estimate $\hat{\mu}_i$ and $\hat{f}_i(Y_j)$.
2. Create the residual matrix $\mathbb{R}$, where the (i, j)-th entry is $R_{ij} = X_{ij} - \hat{\mu}_i - \hat{f}_i(Y_j)$.
3. Perform PCA on $\mathbb{R}$ and retain the first k principal components using an algorithm of your choice (SVA suggests using a resampling technique).
4. Refit the first model but add the estimated principal components as new covariates.

Why go through all this trouble? i

- In the previous paper we discussed, the authors extracted the information on the confounders using PCA directly on the data.
    - This worked well because the major source of variation in SNP data was due to population substructure
- In gene expression studies, the variable of interests are also driving a good proportion of the variation.
    - Therefore, we need to be able to distinguish between the variation we care about and the variation we do not care about.

Why go through all this trouble? ii

- By modeling this extra variation, we achieve two main goals:
    - Increase the variation explained and therefore power.
    - Adjust for confounding by unmodeled sources of variation.

Beyond gene expression i

- It is often very convenient to measure gene expression using whole blood samples.
- However, whole blood is a mixture of cells:
  - Lymphocytes
  - Monocytes
  - Neutrophils
- Each cell type has a different gene expression signature, so we observe a mixture.
- **Crucially**, the mixture weights can be correlated with the variable of interest

Beyond gene expression ii

    - E.g. to fight certain diseases, your blood cell type proportions will change.
- It turns out this is also an issue for DNA methylation experiments.
- SVA has been shown to be effective when trying to correct for cell-type composition bias in DNA methylation experiments (McGregor *et al*, Genome biology, 2016)

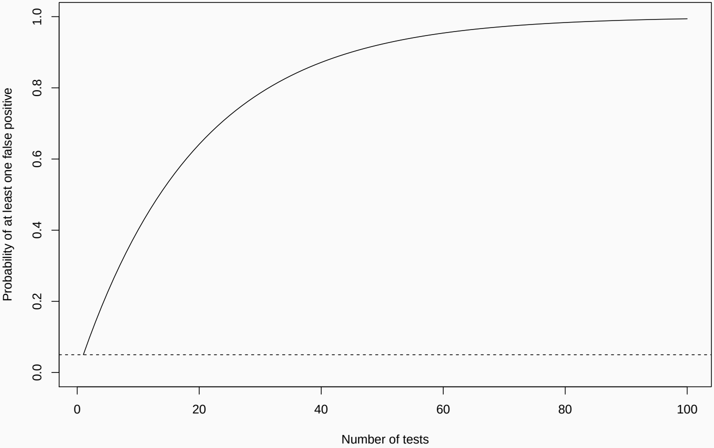
Multiple Testing

Multiple tests in statistical genetics i

- A typical statistical genetics study goes as follows:
    - We collected data using high-throughput technologies.
    - We have hundreds of thousands (or even millions) of genomic measurements on hundreds or thousands of individuals.
    - For all these measurements, we test for association with a variable of interest.
- In other words, if we have a million measurements, we perform a million tests and compute a million p-values.

Multiple tests in statistical genetics ii

- But in those tests, if we set the significance level at $\alpha = 0.05$, we expect 50,000 p-values to be significant **even if the variable of interest is associated with no genomic measurement whatsoever**.


Probability of at least one false positive
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Number of tests

Counting Type I and II errors i

- Let m be the number of hypothesis tests. Let m_0 be the number of null hypotheses that are true. Let R be the number of rejected hypotheses.

	H_0 true	H_0 false	Total
Not rejected	U	T	$m - R$
Rejected	V	S	R
	m_0	$m - m_0$	m

Counting Type I and II errors ii

- In the table above, all capital letters represent random variables.
- V is the number of Type I errors
- Therefore, if we want to control the Type I error rate with multiple tests, we want to control V.
- There are different ways of doing this:
    - **Family-wise error rate**: We want to control $P(V > 0)$.
    - **False Discovery rate**: We want to control the expected value of V/R (R is the number of “discoveries”)

Bonferroni correction

- We talked about Bonferroni CIs earlier in this course.
    - We adjusted the significance level α to α/m.
- This Bonferroni correction controls the FWER at level α.
- The main criticism about this type of correction is that it is *conservative*:
    - If we have dependence between the tests, our true Type I error rate will be lower than α.
    - In general, to control the FWER, we pay a price in terms of Type II errors, and therefore we get lower power.

Effective number of independent tests

- One way to improve the classical Bonferroni correction is to adjust the significance level to $\alpha/\tilde{m}$, with $\tilde{m} < m$.
  - But we still want to control the FWER at α, so we cannot make $\tilde{m}$ too small.
- The idea is that, when we have dependence between two tests, we are potentially performing 1.5 tests.
  - This is the *effective number of independent tests*.

Example

- Assume we have two covariates X_1, X_2 that are positively correlated $\rho = 0.5$.
- If I test X_1 against an outcome Y and reject the null hypothesis, am I
    - More likely to reject X_2 against Y than before doing the test?
    - Less likely?
    - Equally likely?

Simulation i

```r
library(mvtnorm)
n <- 25
B <- 1000
alpha <- 0.05

Sigma <- matrix(c(1, 0.5, 0.5, 1),
                ncol = 2)
```

Simulation ii

```r
results <- replicate(B, {
  X <- rmvnorm(n, sigma = Sigma)
  Y <- rnorm(n)

  test1 <- t.test(X[,1], Y)
  test2 <- t.test(X[,2], Y)

  return(c("test1" = test1$p.value,
           "test2" = test2$p.value))
})
```

Simulation iii

```r
rowMeans(results < alpha)
```

```
## test1 test2
## 0.044 0.046
```

```r
table(colSums(results < alpha))/B
```

```
##
##     0     1     2
## 0.931 0.048 0.021
```