Introduction to dimension reduction: applications and statistical methodology

Jeanine Houwing-Duistermaat[1,2,3], Said el Bouhaddani[4], Zhujie Gu[1,5]

[1] Department of Mathematics, Radboud University, Nijmegen, NL
[2] Department of Statistics, University of Leeds, UK
[3] Department of Statistical Sciences, University of Bologna, IT [3] Department of Data Science and Biostatistics, UMC Utrecht, NL
[4] Medical Research Council Biostatistics Unit, University of Cambridge, UK

CNC, August 2023

Radboud University

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

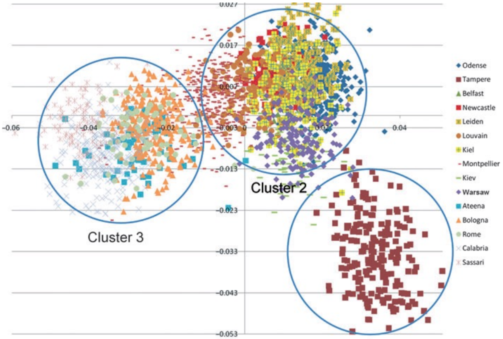
Funded by
the European Union
NextGenerationEU

Genome wide association studies

- First genome wide datasets contained 300K single nuclear polymorphisms (SNPs)
    - These SNPs have small correlations ($r^2 < 0.8$)
    - Single point analysis were performed (300K p-values).
- Nowadays: GWAS, gene expression, proteomics, metabolomics, epigenetics etc for the same individual.
- Single point analyses per dataset is standard while there is within and between datasets correlation.
- For specific questions, multivariate analysis, i.e. principal components analysis, is used.

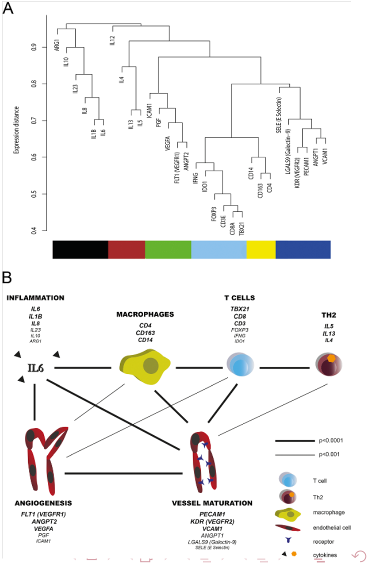
Population structure: GEHA

- Beekman et al. Aging Cell (2013) aimed to identify genomic regions for human longevity using 6000 SNPs genotyped in 2000 sibling pairs across Europe.
- Difference in distribution of SNPs between European countries?
- Principal components were calculated and plotted:



Gene clusters in gene expression network analyses

- Punt et al. Molecular Cancer (2015) aimed to identify gene pathways involved in cervical cancer and their relationships with secondary outcomes.
- Gene expressions from 42 genes of 52 samples.
- Calculation of principal components of gene clusters (eigengenes) to associate gene cluster with outcomes.



Polygenic risk scores

- Genome wide association studies for human diseases provide effect sizes for SNPs
- These results are used to build a linear combination of SNPs
- Training sets are used to determine thresholds
- Personalized risk prediction
- For omics features: He Li compares this approach with multivariate PLS approaches discussed this morning (Methods for High Dimensional and Big Data, Session 4, Thursday at 10:30)

Down Syndrome (dataset used in exercises)

- Down Syndrome (DS) is a genetic disorder caused by the presence of a third copy of chromosome 21.
- Associated with intellectual disability and ‘accelerated aging’
  - premature skin wrinkling, greying of hair, early menopause, early declining immune function, and early onset of Alzheimer’s disease.
- Methylation and glycomics profiles are associated with human longevity: Epigentic clock and GlycoAge Test (markers for biological age)
- These datasets are measured and analysed in DS study of 29 trios of DS, sibling (SB) and their mother (MA).

Methylation and Down syndrome

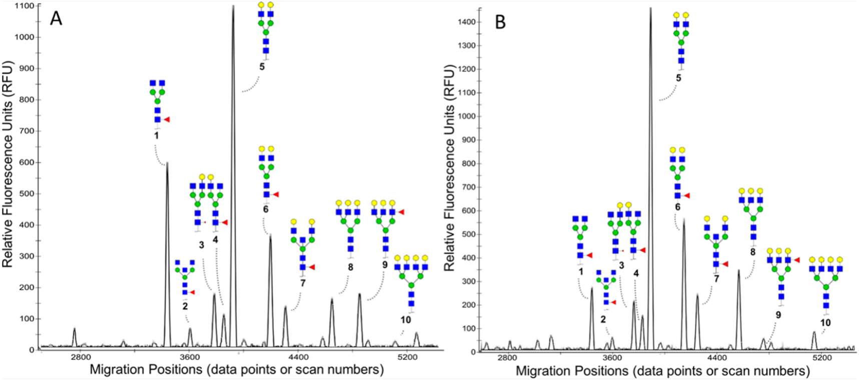
- Statistical method used in DS study:
  - Methylation CpG sites are arranged in BOPs (blocks of probes).
  - Sliding window approach across BOP comparing methylation of DS with their siblings for sets of 3 CpG sites, FDR corrections.
  - Differential methylated regions were mapped to genes.
- Results: embryonic development (HOXA family), haematological (RUNX1 and EBF4), neuronal (NCAM1) development. regulation of chromatin structure (PRDM8, KDM2B, TET1).
- We will only consider chromosome 21 comprising 3322 CpG sites and RUNX1 gene.

Glycomics and Down syndrome

- Glycosylation modifies proteins by introducing variability independently from their DNA sequence.
- N-glycan patterns depend on genetic, physiological, and environmental factors and are stable within an individual.
- Patterns change with pregnancy, diseases and ageing.
- In DS study we have DSA-FACE profiles with 10 peaks.
- Comparing DS and SB yielded significant glycans: H3N4F1 (peak 1), H4N4F1 (peak 3), H5N4 (peak 5).

DSA-FACE profiles

A: DS, B: SB (Note y-axes do not have same range)



Data integration in DS study

- Used approaches do not model correlation within and between datasets.
- Our goal is to identify a methylation pattern (methylation component) associated with a glycosylation component.
- It would be of interest whether such a component can classify DS and SB.
- We will use following datasets:
  - DS, SB and MA.
  - 10 glycan peaks normalised.
  - methylation at CpG sites at chromosome 21 corrected for cell counts.

The neglect of theoretical statistics" (Fisher, 1922)

- *"A quantity of data, which usually by its mere bulk is incapable of entering the mind, is to be replaced by relatively few quantities which shall adequately represent the whole, or which, in other words, shall contain as much as possible, ideally the whole, of the* **relevant information** *contained in the original data."*
- A reduction $R : \mathbb{R}^p \to \mathbb{R}^q$, $q \leq p$, of $X$is sufficient for Y if $Y|X \sim Y|R(X)$,
- For example for linear regression model $Y = \beta^T X + \epsilon$, we have $R(X) = \beta^T X$.

Analysis of a complex statistical variable into principal components (Hotelling, 1933)

- In linear regression X is not random,
- For X is random: A reduction $R : \mathbb{R}^p \to \mathbb{R}^q$, $q \leq p$, of $X$is sufficient for Y if $Y \perp\!\!\!\perp X | R(X)$.
- We have a set of variables x_j, which we represent collectively in the vector $X = (x_j)$, j$\in \{1, ..., r\}$.
- *"It is natural to ask whether some more fundamental set of independent variables exists, perhaps fewer in number than the $[x's]$, which determine the values the $[x's]$ will take."*

Hotelling, 1933, continued

- *"In order to go as far as may reasonably be possible in a given case in expressing $[X]$ in terms of a smaller number of components, an orderly procedure is required for selecting the components in the order of the* **definiteness of their existence**, *or of* **their importance for our purposes**, *and rejecting any which prove to be of little importance, or which are not clearly defined by the data."*
- The underlying idea is "information is variance". Thus we need to select x_i which explains most variance.

Latent variable models or probabilistic components

- Define $q < r$ latent variables v and the following model

$$X_i = \mu + \theta v_i + \epsilon_i, \quad i \in \{1, ..., n\} \quad \text{with}$$

  μ and ϵ_i size $r \times 1$, v_i size $q \times 1$ and Θ size $r \times q$,v_i and $\epsilon_i \sim N(0, V)$ and $N(0, \Delta)$ respectively.

- Fisher lecture by Cook, 2007:

$$X \perp\!\!\!\perp v | \Theta^T \Delta^{-1} X$$

  i.e. $\Theta^T \Delta^{-1} X$ is sufficient statistic.

- Tipping and Bishop (1999): If $\Delta = \delta^2 I_r$ and $V = I_q$, the r latent variables v are first r eigenvectors.
- Different structure results in other models.
- Note that if we aim to model Y based on X we may want to use Y to select the best linear combination of x_i.

Outline course

In this course we will use algorithmic versions of methods. Deriving corresponding probabilistic models is ongoing research.

- Principal components
- Practical I
- Partial least squares
- O2PLS
- Break
- Practical II
- Bioinformatics and final remarks
- Hands on: Analysis of methylation and glycomics datasets in Down syndrome trios.

Funding: ALMArie CURIE 2021 - SUPER line financed using the resources of Ministerial Decree 737/2021, Italy and European Union - NextGenerationEU

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Funded by
the European Union
NextGenerationEU