Multivariate models for dimension reduction and biomarker selection in omics data

Kim-Anh Lê Cao

The University of Queensland Diamantina Institute
Brisbane, Australia
Challenges

Close interaction between statisticians, bioinformaticians and molecular biologists

- Understand the biological problem
- Try answer complex biological questions
- Irrelevant (noisy) variables
- $n \ll p$ and $n$ very small
  \rightarrow \text{limited statistical validation}
- Biological validation needed
- Keep up with new technologies
- Anticipate computational issues
Linear multivariate approaches

Linear multivariate approaches use latent variables (e.g. variables that are not directly observed to reduce the dimensionality of the data).
A large number of observable variables are aggregated in a model to represent an underlying concept, making it easier to understand the data.

- Dimension reduction
  → project the data in a smaller subspace
- To handle multicollinear, irrelevant, missing variables
- To capture experimental and biological variation
Outline of this talk and research questions

1. Principal Component Analysis
   - The workhorse for linear multivariate statistical analysis
   - Detection and correction of batch effects
   - Variable selection

2. One data set
   - Looking for a linear combination of multiple biomarkers from a single experiment
   - Improving the prediction of the model

3. Integration of multiple data sets
   - Looking for a linear combination of multiple ‘omics biomarkers from multiple ‘omics experiments
   - Improving the prediction of the model
Principal Component Analysis

PCA: the workhorse for linear multivariate statistical analysis is an (almost) compulsory first step in exploratory data analysis to:

- Understand the underlying data structure
- Identify bias, experimental errors, batch effects

Insightful graphical outputs to:

- to visualise the samples in a smaller subspace (components ‘scores’)
- to visualise the relationship between variables ... and samples (correlation circles)
Maximize the variance

In PCA, we think of the variance as the 'information' contained in the data. We replace original variables by artificial variables (principal components) which explain as much information as possible from the original data. Toy example with data ($n = 3 \times p = 3$):

\[
\text{Cov(data)} = \begin{pmatrix}
1.3437 & -0.16015 & 0.1864 \\
-0.01601 & 0.6192 & -0.1266 \\
0.1864 & -0.1266 & 1.4855
\end{pmatrix}
\]

We have $1.3437 + 0.6192 + 1.4855 = 3.4484$

Replace the data by the principal components, which are orthogonal to each other (covariance = 0).

\[
\text{Cov(pc1, pc2, pc3)} = \begin{pmatrix}
1.6513 & 0 & 0 \\
0 & 1.2202 & 0 \\
0 & 0 & 0.5769
\end{pmatrix}
\]

We have $1.6513 + 1.2202 + 0.5769 = 3.4484$
PCA as an iterative algorithm

1. Seek for the **first PC** for which the variance is maximised
2. **Project** the data points onto the first PC
3. Seek for the **second PC** that is **orthogonal** to the first one
4. **Project** the residual data points onto the second PC
5. And so on until we explain enough variance in the data
A linear combination of variables

Seek for the best directions in the data that account for most of the variability. Objective function:

$$\max \ var(Xv)$$

where $$\frac{\|v\|}{=1}$$

Each principal component $$u$$ is a linear combinations of the original variables:

$$u = v_1x^1 + v_2x^2 + \cdots + v_px^p$$

- $$X$$ is a $$n \times p$$ data matrix with $$\{x^1, \ldots, x^p\}$$ the $$p$$ variable profiles.
- $$u$$ is the first principal component with max. variance
- $$\{v_1, \ldots, v_p\}$$ are the weights in the linear combination
The data are projected into a smaller subspace

- Each principal component is orthogonal to each other to ensure that no redundant information is extracted.
- The new PCs form a smaller subspace of dimension $<< p$.
- Each value in the principal component corresponds to a score for each sample → this is how we project each sample into a new subspace spanned by the principal components → approximate representation of the data points in a lower dimensional space → summarize the information related to the variance
Computing PCA

Several ways of solving PCA

- **Eigenvalue decomposition**: the old way, does not work well in high dimension $Sv = \lambda v; u = \tilde{X}v$

  $S =$ variance covariance matrix or correlation matrix if $\tilde{X}$ is scaled

- **NIPALS** algorithm: long to compute but works for missing values and enables least squares regression framework (useful for variable selection)

- **Singular Value Decomposition** (SVD): the easiest and fastest, implemented in most softwares.
PCA is simply a matrix decomposition

\[
X = \tilde{U} \Lambda \tilde{V}^T
\]

- \( \Lambda \) diagonal matrix with \( \sqrt{\lambda_h} \)
- \( U = \tilde{U} \Lambda \), \( U \) contains the PCs \( u_h \)
- \( V \) contains the loading vectors \( v_h \)
- \( h = 1..H \)

The variance of a principal component \( u_1 \) is equal to its associated eigenvalue \( \lambda_1 \), etc.. The obtained eigenvalues \( \lambda_h \) are decreasing.
Tuning PCA

How many principal components to choose to summarize most of the information?

Note: we can have as many components as the rank of the matrix $X$

- Look at proportion of explained variance
- Look at the screeplot of eigenvalues. Any elbow?
- Look at sample plot. Makes sense?
- Some stat tests exist to estimate the ‘intrinsic’ dimension

Proportion of explained variance for the first 8 principal components:

- PC1: 0.5876
- PC2: 0.1587
- PC3: 0.0996
- PC4: 0.0848
- PC5: 0.0159
- PC6: 0.0098
- PC7: 0.0052
- PC8: 0.0051
PCA is useful to visualise batch effects

‘Batch effects are sub-groups of measurements that have qualitatively different behaviour across conditions and are unrelated to the biological or scientific variables in a study.

For example, batch effects may occur if a subset of experiments was run on Monday and another set on Tuesday, if two technicians were responsible for different subsets of the experiments, or if two different lots of reagents, chips or instruments were used’.

(Leek et al. (2010), Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* )
Cell lines case study (www.stemformatics.org, AIBN)

Five microarray data set studies including $n = 82$ samples, $p = 9K$ genes and $K = 3$ groups of cell lines including 12 fibroblasts, 22 hPSC and 48 hiPSC. This is a complex case of cross-platform comparison where we expect to see a batch effect (the study).
PCA in action

What we would really like to see is a separation between the different types of cell lines. PCA is a good exploratory method to highlight the sources of variation in the data and highlight batch effects if present.
PCA to visualise batch effect correction

Remove known batch effects
- Combat: uses empirical Bayesian framework
- Linear (mixed) model

Remove unknown batch effects
- sva: estimates surrogate variables for unknown sources of variation
- YuGene: works directly on the distribution of each sample

combat: Johnson WE. et al (2007), Biostatistics;
sva: Leek JT and Storey JD. (2007), PLoS Genetics;
YuGene: Lê Cao KA, Rohart F et al. (2014), Genomics.
Summary on PCA

- PCA is a matrix decomposition technique that allows dimension reduction

- **Run a PCA first** to understand the sources of variation in your data
  
  Note: I illustrated the batch effect on the very topical issue of meta-analysis. But batch effects happen very often!
  
  → What did Alec say? **Experimental design is important!**

- If clear **batch effect**: try to remove/ accommodate for it

- If no clear separation between your groups:
  
  → Biological question may not be related to the highest variance in the data, try **ICA** instead
  
  → Mostly due to a large amount of **noisy variables** that need to be removed
The curse of dimensionality

Fact: in high throughput experiments, data are noisy or irrelevant. They make a PCA analyse difficult to understand.

→ clearer signal if some of the variable weights \( \{v_1, \ldots, v_p\} \) set to 0 for the 'irrelevant' variables:

\[
u = 0 \times x^1 + v_2 x^2 + \cdots + 0 \times x^p
\]

Important weights = important contribution to define the principal components.
The curse of dimensionality

Since the ’90, many statisticians have been trying to address the issue of high (and now ‘ultra high’) dimension.

Let’s face it, classical statistics cannot cope!

- Univariate analysis: multiple testing to address # false positives
- Machine Learning approaches
- Parsimonious statistical models to make valid inferences → variable selection
sparse Principal Component Analysis

Aim: sparse loading vectors to remove irrelevant variables which determine the principal components. The idea is to apply LASSO (least absolute shrinkage and selection operator, Tibshirani, R. (1996)) or some soft-thresholding approach in a PCA framework.

(least squares regression framework by using the close link between SVD and low rank matrix approximations).

→ obtain sparse loading vectors, with very few non-zero elements
→ perform internal variable selection

Parameters to tune

1. The number of principal components (as seen earlier)
2. The number of variables to select. → tricky!

At this point we can already sense that separating hPSC vs hiPSC is going to be a challenge
Correlation circles enable to understand the relationship

- between variables (correlation structure)
- between variables and samples
Linear Discriminant analysis seeks for a linear combination of features which characterizes or separates two or more classes of objects or events.

The resulting combination may be used as a linear classifier, or, more commonly, for dimensionality reduction prior to classification.
**Fisher’s Linear Discriminant**

We can decompose the total variation into

\[ V = B + W \]

\((W\) the Within-groups and \(B\) the Between-groups variance)\n
The problem to solve is:

\[
\max_a \frac{a' Ba}{a' Wa} 
\]

→ maximise the Between-groups variance

→ minimise the Within-groups variance

Number of components to choose: \( \leq \min(p, K - 1) \), \( K = \# \) classes

Prediction of a new observation \( x \) is based on the discriminant score

Pros and Cons of LDA

- LDA has often been shown to produce very good classification results

- Assumptions:
  - equal variance in each group
  - independent variables to be normally distributed
  - independent variables to be linearly related

- But with too many correlated predictors, LDA does not work so well

→ regularize or introduce sparsity, using a cousin of LDA called Partial Least Squares Discriminant analysis
Partial Least Squares Discriminant Analysis seeks for the PLS components from $X$ which best explain the outcome vector $Y$.

Objective function based on the general PLS:

$$\max_{||u||=1,||v||=1} \text{cov}(Xu, Yv)$$

$Y$ is the qualitative response matrix (dummy block matrix).
sparse PLS - DA

Include soft-thresholding or LASSO penalization on the loading vectors \((v_1, \ldots, v_p)\) similar to sPCA.

Parameters to tune:

- **Number of components** often equals to \(K - 1\)
- **Number of variables to select** on each dimension → classification error, sensitivity, specificity

What is \( k \) fold cross-validation?

→ estimate how accurately a predictive model will perform in practice

- Partition the data into \( k \) equal size subsets
- Train a classifier on \( k - 1 \) subsets
- Test the classifier on the \( k^{th} \) remaining set
- Estimate classification error rate
- Average across \( k \) folds
- Repeat the process many times

Be careful of selection bias when performing variable selection during the process!
Kidney transplant study (PROOF Centre, UBC)

Genomics assay ($p = 27K$) of 40 patients with kidney transplant, with acute rejection (AR, $n_1 = 20$) or no rejection (NR, $n_2 = 20$) of the transplant.

**Training step:**

- 5 fold cross-validation (100 times) on a training set of $n_{\text{train}} = 26$
- Selection of 90 probe sets
Testing step:

- Training performed: model with 90 probes
- Genes mostly related to inflammation and innate immune responses
- External test set of $n_{test} = 14$
- Project the test samples onto the training set subspace of sPLS-DA

<table>
<thead>
<tr>
<th>Classifier</th>
<th># probes</th>
<th>Error rate</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>sPLS-DA</td>
<td>90</td>
<td>0.14</td>
<td>0.71</td>
<td>1</td>
<td>0.90</td>
</tr>
<tr>
<td>PLS-DA</td>
<td>27K (all)</td>
<td>0.21</td>
<td>0.57</td>
<td>1</td>
<td>0.82</td>
</tr>
</tbody>
</table>
Oesophageal cancer study (UQDI)

Proteomics assay ($p = 129$) of 40 patients with Barrett’s oesophagus benign ($n_1 = 20$) or oesophageal adenocarcinoma cancer ($n_2 = 20$). **Aim**: develop blood tests for detection and personalised treatment.

- sPLS-DA model and selection of 9 proteins
- Individual proteins biomarkers give a poor performance (AUC 0.46-0.79)
- A combined signature from sPLS-DA model improves the AUC up to 0.94

Acknowledgements: Benoit Gautier
Biomarker discovery when integrating multiple data sets

- Model relationships between different data sets
- **Select relevant biological entities** which are correlated across the different data sets
- Use multivariate integrative approaches: **Partial Least Squares regression** (PLS) and **Canonical Correlation Analysis** (CCA) variants
- Same idea as before: maximize covariance between latent components associated to 2 data sets at a time
sparse Generalised Canonical Correlation Analysis

- sGCCA generalizes sPLS to more than 2 data sets and maximizes the sum of pairwise covariances between two components at a time

\[
\max_{a_1, \ldots, a_J} \sum_{j,k=1, j \neq k}^{J} c_{kj} g(\text{Cov}(X_j a_j, X_k a_k))
\]

s.t. \[\tau_j ||a_j||^2 + (1 - \tau_j) \text{Var}(X_j a_j), \text{with } j = 1, \ldots, J, \text{where the } a_j \text{ are the loading vectors associated to each block } j, g(x) = |x|.

- sGCCA: variable selection is performed on each data set


Kim-Anh Lê Cao
2014 Winter School in Mathematical & Computational Biology
Asthma study (UBC, Vancouver)

Multiple assay study on asthma patients pre ($n_1 = 14$) and post ($n_2 = 14$) challenge.
Three data sets: complete blood count ($p_1 = 9$), mRNA ($p_2 = 10,863$) and metabolites ($p_3 = 146$).

Aims
- Identify a multi omic biomarker signature to differentiate asthmatic response.
- Can we improve the prediction with biomarkers of different types?

Acknowledgements: Amrit Singh
Method
sGCCA components score predict the class of the samples per data set.
→ Use the average score to get an ‘Ensemble’ prediction

Results
‘Ensemble’ prediction outperforms each data set’s prediction
**Results:** Selection of 4 cells, 30 genes and 16 metabolites, some of which are very relevant to Asthma.

**Cells:** Relative Neutrophils; Relative Eosinophils; T reg; T cells

**Genes:**

"low-affinity" receptor for IgE, an antibody isotype involved in allergy; expressed on B cells, macrophages and **eosinophils**

- "TEX33"
- "HSPA14"
- "IGFL4"
- "MYO10"
- "SCUBE1"
- "SCUBE3"
- "UCHL1"
- "UPK3A"
- "CLEC4D"
- "IL1R2"
- "NLRC4"
- "SH2D4B"
- "DOCK4"
- "IP3RA2"
- "KIF20A"
- "OSGDL"
- "PROM19"
- "TREM9"
- "TPST1"
- "FER1L4"
- "FZD7"
- "LSAMP"
- "MYL4"
- "RPL13AP20"
- "SCML1"
- "TSNAXIP1"

**Marker of activated T-cell**

Up-regulated by sputum **neutrophils** after low antioxidant diet; PMID:19715394

Decreased in individuals with asthma compared to normal; PMCD: PMC3185200

Kim-Anh Lê Cao

2014 Winter School in Mathematical & Computational Biology
It is all about mixOmics

mixOmics is an R package implementing multivariate methods such as PCA, ICA, PLS, PLS-DA, CCA (work in progress), the sparse variants we have developed (and much more).
To put it in a nutshell

- Multivariate linear methods enables to answer a wide range of biological questions
  - data exploration
  - classification
  - integration of multiple data sets
- Least squares framework for variable selection

Future of mixOmics

- Time course modelling
- Meta analysis / multi group analysis
- Several workshops coming up! (Toulouse 2014, Brisbane 2015, Auckland 2015)
# Acknowledgements

## mixOmics development

<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sébastien Déjean</td>
<td>Univ. Toulouse</td>
</tr>
<tr>
<td>Ignacio González</td>
<td>Univ. Toulouse</td>
</tr>
<tr>
<td>Xin Yi Chua</td>
<td>QFAB</td>
</tr>
<tr>
<td>Benoit Gauthier</td>
<td>UQDI</td>
</tr>
</tbody>
</table>

## Methods development

<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amrit Singh</td>
<td>UBC, Vancouver</td>
</tr>
<tr>
<td>Florian Rohart</td>
<td>AIBN, UQ</td>
</tr>
<tr>
<td>Jasmin Straube</td>
<td>QFAB</td>
</tr>
<tr>
<td>Kathy Ruggiero</td>
<td>Univ. Auckland</td>
</tr>
<tr>
<td>Christèle Robert</td>
<td>INRA Toulouse</td>
</tr>
</tbody>
</table>

## Data providers

<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oliver Günther</td>
<td>Kidney PROOF, Vancouver</td>
</tr>
<tr>
<td>Michelle Hill</td>
<td>EAC UQDI</td>
</tr>
<tr>
<td>Alok Shah</td>
<td>EAC UQDI</td>
</tr>
<tr>
<td>Christine Wells</td>
<td>Stemformatics AIBN, Univ. Glasgow</td>
</tr>
</tbody>
</table>
Intro
PCA: the workhorse
Variable selection
for single ...
and multiple ’omics
Conclu

Questions?

Statistics for frightened bioresearchers
A short course organised by UQDI and IMB
10:30am - 12pm Monday
28 July, 1 September, 29 September, 3 November
Room 2007, Translational Research Institute, 37 Kent Street, Princess Alexandra Hospital, Woolloongabba, QLD 4102 *
Large Seminar Room, Institute for Molecular Bioscience, Queensland Biosciences Precinct, The University of Queensland, St Lucia Campus.
You are invited to a short course in biostatistics
Four presentations over four months will cover
1. Summarizing and presenting data
2. Hypothesis testing and statistical inference
3. Comparing two groups
4. Comparing more than two groups
With a focus on real world examples, notes and resources you can take home, and quick exercises to test your understanding of the material, this short course will cover many of the essential elements of biostatistics and problems that commonly arise in analysis in the lab.

Further information and registration on
www.di.uq.edu.au/statistics

Free registration:
www.di.uq.edu.au/statistics
k.lecao@uq.edu.au
http://www.math.univ-toulouse.fr/~biostat/mixOmics

Kim-Anh Lê Cao
2014 Winter School in Mathematical & Computational Biology