Multivariate statistical analyses and machine learning for metabolomics

Roy Goodacre
School of Chemistry, The University of Manchester, Sackville Street, PO Box 88, Manchester, M60 1QD

Roy.Goodacre@manchester.ac.uk
www.biospec.net
Welcome to our Home Page.

Research: The research theme in our group is predominantly directed towards developing metabolomic and proteomic technologies for the rapid, accurate characterisation of biological systems. This is achieved via a tandem analysis using analytical instrumentation that are used to produce rapid physiological metabolome, proteome or 'holistic' whole-organism (phenotypic) fingerprints of bacteria, fungi, human and animal body fluids and plant materials. In order to analyse these high dimensional multivariate data we have been very active in the development of novel chemometric and machine learning techniques. In cognate projects, we are also developing molecular methods for the phylogenetic analysis of microorganisms.

In October 2005 we will move into the Manchester Interdisciplinary Biocentre (\texttt{www.mab.ac.uk}), a £35M building designed to bring together physical scientists, engineers, computer scientists and mathematicians to attack and solve complex, multidisciplinary biological problems.

Postdocs: Dr Catherine Winder, Dr Robert Cormell, Dr Roger Jarvis, Dr Mohammad Afzaal, A.N. Other x 5,
With Collabs: Dr Seetharaman Vaidyanathan, Dr John Fletcher
Research Technicians: Jo Ellis, Steffi Schuler + A.N. Other
From Oct 1: Nicola Wood, Dong Hyun Kim.
Research Students (MSc): Nicola Wood, Sadia Rabbini, Martin Coleston, Graham Mullard.
£s, €s, $s: BBSRC, EPSRC, RSC, EU F6, DEFRA, NERC, ORS
‘Top down’ metabolomics

**Strategy:** Inductive approach to knowledge discovery via holism

- Carefully designed Data-generating experiment
- Inductive reasoning by computation
- Generation of hypotheses
- Analysis and test hypotheses

“Hiring a statistician after the data have been collected is like hiring a physician when the patient is in the morgue. He might be able to tell you what went wrong, but is unlikely to be able to fix it”

Data floods

“Data does not equal information; information does not equal knowledge; and, most importantly of all, knowledge does not equal wisdom. We have oceans of data, rivers of information, small puddles of knowledge, and the odd drop of wisdom.”

Henry Nix, 1990
Typical metabolic profile

Deconvolution

- Glucose 0.1
- Indole 0.001
- Tryptophan 1.2
- Ethanolamine 0.7
- Metabolite #88 0.9
- Metabolite #167 0.05

Name of metabolite with its concentration
## Data handling

<table>
<thead>
<tr>
<th>Objects going down in different rows</th>
<th>X-var 1</th>
<th>X-var 2</th>
<th>X-var 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1</td>
<td>Metabolite or peak 1</td>
<td>Metabolite or peak 2</td>
<td>Metabolite or peak 3</td>
</tr>
<tr>
<td>Sample 2…</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metabolite</th>
<th>Conc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glucose</td>
<td>0.1</td>
</tr>
<tr>
<td>Indole</td>
<td>0.001</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>1.2</td>
</tr>
<tr>
<td>Ethanolamine</td>
<td>0.7</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>Metabolite #88</td>
<td>0.9</td>
</tr>
<tr>
<td>Metabolite #167</td>
<td>0.05</td>
</tr>
</tbody>
</table>
## Data handling

<table>
<thead>
<tr>
<th>Objects going down in different rows</th>
<th>X-var 1</th>
<th>X-var 2</th>
<th>X-var 3</th>
<th>Y-var 1</th>
<th>Y-var 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1</td>
<td>Metabolite or peak 1</td>
<td>Metabolite or peak 2</td>
<td>Metabolite or peak 3</td>
<td>Lots of Metadata</td>
<td>Diseased or Healthy (Levels)</td>
</tr>
<tr>
<td>Sample 2…</td>
<td></td>
<td></td>
<td></td>
<td>Species</td>
<td>0 (control)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Age</td>
<td>1 (diseased)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>M/F</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>BMI</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>sampling processing etc, etc…</td>
<td></td>
</tr>
</tbody>
</table>

### Input data
- X-var 1
- X-var 2
- X-var 3

### Output data
- Y-var 1
- Y-var 2
Unsupervised learning

- System is shown a set of inputs (spectra) and then left to cluster the spectra into groups.
- This optimization procedure is usually simplification or dimensionality reduction.
Projection of the data

\[ x_1, x_2, \ldots x_i, \ldots \text{ variables} \quad \ldots, x_n \]

\[ \begin{align*}
1 \\
2 \\
3 \\
k
\end{align*} \]

samples

\[ \begin{align*}
p_1 \\
p_2 \\
\vdots \\
p_i
\end{align*} \]

loadings \((p)\)

\(\text{summarise variation in variables}\)

\[ \begin{align*}
t_1 & \geq t_2 \geq \ldots \geq t_i \\
\text{scores (t)} \\
\text{summarise variation in samples} \\
\text{uncorrelated orthogonal}
\end{align*} \]

\[ \begin{align*}
t_1 &= p_1 x_1 + p_2 x_2 + \ldots + p_i x_i + \ldots + p_n x_n \\
\text{scores = loadings \times data}
\end{align*} \]
PCA

- Finds structure in data
- Rotate to uncover *maximum* correlations with respect to *natural* variation

**PC1**
- 1st principal component
- Describes largest variance
- Goes through variable origin space
- \( t_{j1} \) = score for point j = distance from projection of that point onto PC1 from origin

**PC2**
- 2nd principal component
- Describes 2nd largest variance
- Goes through variable origin space
- \( t_{j2} \) = score for point j

PC1 = 1\textsuperscript{st} principal component
describes largest variance
goes through variable origin space
\( t_{j1} \) = score for point j = distance from projection of that point onto PC1 from origin

PC2 = 2\textsuperscript{nd} principal component
describes 2\textsuperscript{nd} largest variance
goes through variable origin space
\( t_{j2} \) = score for point j
The goal is to find a mathematical model that will correctly associate the inputs with the targets.

Usually achieved by minimising the error between the target and the model's response (output).
Discriminant function analysis
(aka, canonical variates analysis)

- Uses uncorrelated inputs
  \textit{a priori} information

- Projection based on:
  - Minimises within group variance
  - Maximises between group variance

- Test by projection of ‘unknown’ samples

- Statistical significance: $\chi^2$ confidence limits
Peat

- 6 groups
- Circles = 95% $\chi^2$ confidence limits
- Arrows represent outlier samples that were from upper horizon of the peat depth profile

Target encoding for PLS, ANNs, etc…

- Usually binary encoded:

<table>
<thead>
<tr>
<th>Known bacteria</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>New isolates</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Easy look up table
Supervised methods are powerful…

- Learn from experience
- Generalise from previous examples to new ones
- Perform pattern recognition on complex multivariate data.
- Make errors
  - usually because of badly chosen data
  - tanks from trees…
What have I measured that is important

Consider healthy vs. diseased

- 2 class problem
- Collect metabolite data from both classes:
  - Encompass biological variation
    - Even distribution
  - Try to keep number of samples the same
    - Statistics are easier / more rigorous

→ next do inductive bit...
Easy strategies first

- Stare and compare!
  - Always wise to actually look at the data…
- Difference spectra
  - $\text{Avg(} \text{diseased} \text{)} - \text{Avg(} \text{healthy} \text{)}$
- Univariate analyses
  - Analysis of variance (ANOVA)
  - T-test, etc
Use multivariate analyses and inspect loadings

<table>
<thead>
<tr>
<th></th>
<th>Variables →</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

loads (p)

scores (t)
Types of loadings plots

- Unsupervised
  - PCA
  - SOFM
- Supervised: discriminatory
  - DFA (CVA)
  - PLS-DA
- Supervised: regression
  - PLS
- Supervised: tree-based
  - CART

Types of loadings plots

- Unsupervised
  - PCA
  - SOFM
- Supervised: discriminatory
  - DFA (CVA)
  - PLS-DA
- Supervised: regression
  - PLS
- Supervised: tree-based
  - CART

But not always that easy…

Univariate
\( p < 0.01 \) t-test
\( p \) not signif.
Modest 250 metabolites

- 2 class problem
  - Healthy vs. Diseased
- To use or not to use?
- Yes / No, 250 times = $2^{250}$ or $1.8 \times 10^{75}$
- PC does say $10 \times 10^6$ orderings every second it would still take $> 3 \times 10^{62}$ years
Houston we have a problem…

Complex NP problem → No algorithm → Global optimal solution

Complex NP problem → Good solution

Evolutionary computation
Explanatory machine learning

Inputs
[metabolome data]

Target output(s)
[categorical or quantitative]

Quantitative levels

Metabolites

Explanatory mathematical transformation

desired responses paired with samples
Genetic algorithm

- Chromosome with $n$ genes selects which of $n$ inputs to use in DA, MLR, PLS or ANN

inputs: 1 2 $\cdots$ $i$ $\cdots$ $n-1$ $n$
genesis: 1 1 0 1 0 1 0
filter: ✓ ✓ ✗ ✓ ✓ ✓ ✗
GA: *in silico* evolution

- Population of solutions
  - Mutation
  - Crossover
- Natural selection
- Assess fitness function of population
Overall conclusions on hypothesis generation

- Many approaches can be used ranging from
  - Simple → stare and compare, ANOVA etc
  - Multivariate → loadings: PCA, DFA, PLS
  - More complex → evolutionary computation

- Need to design experiment carefully

- Hypotheses might not always be correct
  - But a good place to start when initial knowledge is non-existent.
PyChem v2.0.0 Beta
(http://pychem.org.uk)

- Standalone WinXP graphical application for MVA
- Preprocessing algorithms incorporated
- Many standard chemometrics algorithms
  - PCA, HCA, DFA, PLSR, PLS-DA.
  - Feature selection using a variety of genetic algorithm tools
- Coded in Python (http://python.org/) the software is both FREE and OPEN SOURCE. Source code can be downloaded at http://sourceforge.net/projects/pychem/.
- Can be used to analyse continuous or discrete data such as transcriptomic, metabolomic (vibrational spectra, GC-MS, LC-MS etc…)

Written by Dr Roger Jarvis
$m/z$ 117 vs. $m/z$ 136

$m/z$ 117 vs. $m/z$ 109

Caffeine

Chlorogenic acid
On (stamp) collecting data

“Science is built up with facts, as a house is with stones. But a collection of facts is no more a science than a heap of stones is a house”

Jules Henri Poincaré (1854-1912) *La Science et l'hypothèse*

_Having the parts list doesn’t mean that you know how an organism works_