UNIVERSITEIT GENT

Principal Bicorrelation Analysis

Thas, Mattiello

Introduction
PBA
Case Study
Conclusions

Integrated statistical analysis of three data sources for the detection of chemical fingerprint features

Olivier Thas[1,2], Federico Mattiello[1]

[1]Department of Mathematical Modelling, Statistics and Bioinformatics
Ghent University
[2]Centre for Statistical and Survey Methodology, School of Mathematics and Applied Statistics, University of Wollongong, NSW 2522, Australia

September 25th, 2012

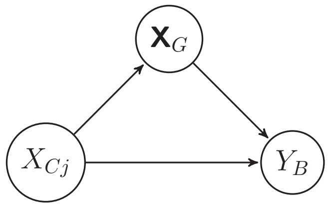
UNIVERSITEIT GENT

Principal Bicorrelation Analysis

Thas, Mattiello

Introduction
PBA
Case Study
Conclusions

Introduction
The "magic triangle"



X_{Cj}: fingerprint feature (FF) $j = 1, \ldots, n_C$
Y_B: a single bio-assay (BA)
$\mathbf{X}_G$: the $n \times n_G$ gene expression (GE) matrix.

UNIVERSITEIT GENT

Principal Bicorrelation Analysis

Thas, Mattiello

Introduction
PBA
Case Study
Conclusions

Outline

1. Introduction
2. Principal Bicorrelation Analysis
3. Case Study
4. Conclusions

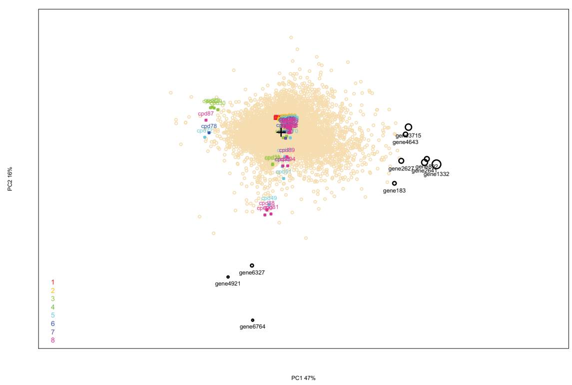
UNIVERSITEIT GENT

Principal Bicorrelation Analysis

Thas, Mattiello

Introduction
PBA
Case Study
Conclusions

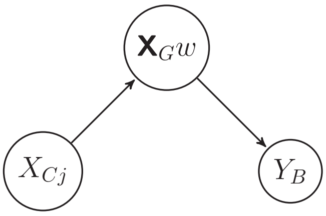
Gene Expression Matrix
Spectral Map: an exploratory tool



Introduction

Central idea

The role of GE data may be explored by a linear combination of genes (here the columns of $\mathbf{X}_G$), say $\mathbf{X}_G w$, so that it maximises simultaneously its correlation with the FF j and with the BA.



The construct $\mathbf{X}_G w$ is referred to as the gene signature.

Principal Bicorrelation Analysis

maximisation problem

In particular we are looking for the vector w that maximises the weighted sum of squared correlations (*bicorrelation*):

$$R_j^2(w) = \tau \text{Cor}\{X_{Cj}, \mathbf{X}_G w\}^2 + (1-\tau)\text{Cor}\{Y_B, \mathbf{X}_G w\}^2,$$
$$= \tau w^\top \mathbf{X}_G^\top X_{Cj} X_{Cj}^\top \mathbf{X}_G w + (1-\tau) w^\top \mathbf{X}_G^\top Y_B Y_B^\top \mathbf{X}_G w$$

where $0 < \tau < 1$ is a user-defined weight.

The maximisation is subject to the constraints:

- $\|w\|_2 = 1$, if the desired solution need not to be sparse (i.e. the *dense* solution)
- $\|w\|_2 = 1$ *and* $\|w\|_1 \leq k$ if a *sparse* solution is desired (k = tuning parameter).

Principal Bicorrelation Analysis

maximisation problem

A note on the two correlations involved in $R_j^2(w)$:

- Cor $\{Y_B, \mathbf{X}_G w\}^2$: Both Y_B and $\mathbf{X}_G w$ are continuous variables;
- Cor $\{X_{Cj}, \mathbf{X}_G w\}^2$: The FF X_{Cj} is a binary indicator and $\mathbf{X}_G w$ is a continuous variable; hence, the Pearson correlation perhaps is not the best choice (see next slide)

A potential problem: it is easier to find a w to make Cor $\{X_{Cj}, \mathbf{X}_G w\}^2$ large, than it is to make Cor $\{Y_B, \mathbf{X}_G w\}^2$ large.

Our solution: assign weights $0 < \tau < 1$ and $1 - \tau$ to the two correlations (with τ small).

Principal Bicorrelation Analysis

maximisation problem

Since the FF X_{Cj} is a binary indicator, the Pearson correlation Cor $\{X_{Cj}, \mathbf{X}_G w\}^2$ may not be the most approriate measure for association.

We have also evaluated the Ranked Pointwise Biserial coefficient (a mean rank difference),

$$\rho_{rpb} = \frac{2}{n}(\bar{r}_1 - \bar{r}_0),$$

where $\bar{r}_p = \frac{1}{n_p}\sum_{i:y_i=p} R\{x_i\}$, $n_p = \#\{y_i = p\}$, $p = 0$ or 1 and $R\{\cdot\}$ is the rank operator.

But we will show only results for the first one for better interpretation.

Principal Bicorrelation Analysis
maximisation problem

This w may be obtained from the SVD of the $2 \times n_G$ matrix

$$M := \begin{bmatrix} \tau \text{Cor}\{X_{Cj}, \mathbf{X}_G\} \\ (1-\tau)\text{Cor}\{Y_B, \mathbf{X}_G\} \end{bmatrix} \propto \begin{bmatrix} \tau X_{Cj}^\top \mathbf{X}_G \\ (1-\tau) Y_B^\top \mathbf{X}_G \end{bmatrix}^{a},$$

- The dense solution: w is equal to the first right singular vector of M;
- The sparse solution: w is calculated from M by applying the methods described in D. Witten *et al.* (2009) for sparse PCA; an iterative algorithm with $u^\top X v$ as objective function (plus constraints).

[a] in the Pearson correlation case, see E. Bair *et al.* (2004) for details

PBA: interesting Plots 1
Description

Plot of Cor $\{Y_B, \mathbf{X}_G \hat{w}_1\}^2$ *vs* Cor $\{X_{Cj}, \mathbf{X}_G \hat{w}_1\}^2$ for all FFs can be used to select FFs for which $\mathbf{X}_G \hat{w}_1$ is:

- highly correlated with both the BA and FF j
- extremely low correlated with both the BA and FF j, but for which correlation along the direct path FF-BA is high.

Note that in this plot each dot is a FF.

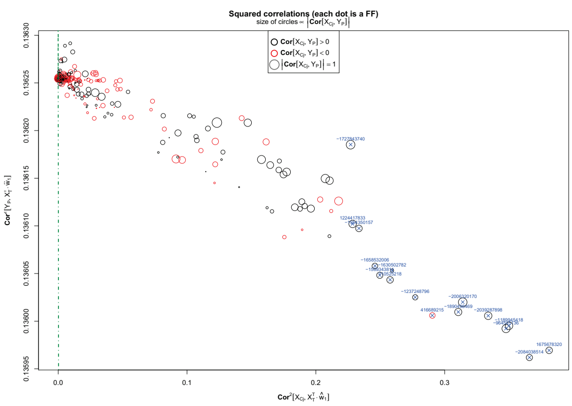
Principal Bicorrelation Analysis
maximisation problem

Note 1 all data matrices have been divided by their own first singular value to balance the contribution of each source of data, as in *Multiple Factor Analysis*.

Note 2 the vector w that maximises $R_j^2(w)$ is denoted by $\hat{w}$;

Note 3 just like in a PCA, more than one gene signature can be obtained: *i.e.* $\hat{w}_1, \hat{w}_2, \ldots$.

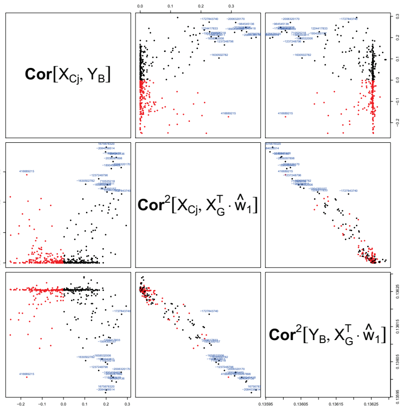
PBA: interesting Plots 1
Example 1 (sparse Pearson method, $\tau = 0.1$)



PBA: interesting Plots 1

Example 2 (sparse Pearson method, $\tau = 0.1$)

Interaction between the 3 correlations:



PBA: interesting Plots 2

Example (sparse Pearson method, $\tau = 0.1$)

UNIVERSITEIT GENT

Principal Bicorrelation Analysis

Thas, Mattiello

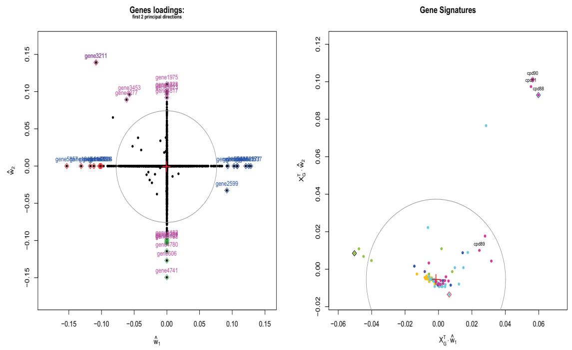
Introduction

PBA

Case Study

Conclusions

Loadings and scores for the FF with the highest (Pearson-sparse) $R_j^2(\hat{w}_1)$ value.



PBA: interesting Plots 2

Description

UNIVERSITEIT GENT

Principal Bicorrelation Analysis

Thas, Mattiello

Introduction

PBA

Case Study

Conclusions

Scatter plots:

- $\hat{w}_2$ *vs* $\hat{w}_1$:
  Scatter plot of the gene loadings of the first two dimensions can be used to select genes that play an important role in the association with both FF j and the BA.
- $\mathbf{X}_G\hat{w}_2$ *vs* $\mathbf{X}_G\hat{w}_1$: Scatter plot of the compound scores on the first two gene signatures can be used to select compounds that activate the gene signature(s) that is responsible for the association between FF j and the BA.

PBA: interesting Plots 3

Description

UNIVERSITEIT GENT

Principal Bicorrelation Analysis

Thas, Mattiello

Introduction

PBA

Case Study

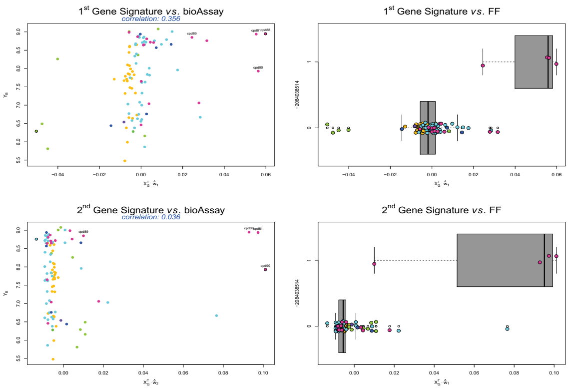
Conclusions

- Y_B *vs* $\mathbf{X}_G\hat{w}_i$ $(i = 1, 2)$:
  Scatter plots of BA versus the compound scores of the gene signatures to select compounds that have extreme scores for both BA and gene signature.
- $\mathbf{X}_G\hat{w}_i$ *vs* X_{Cj} $(i = 1, 2)$:
  Boxplots of the compound scores of the gene signatures for each of the two FF phases (absent/present) to select compounds with gene signatures that discriminate between the two FF phases.

PBA: interesting Plots 3

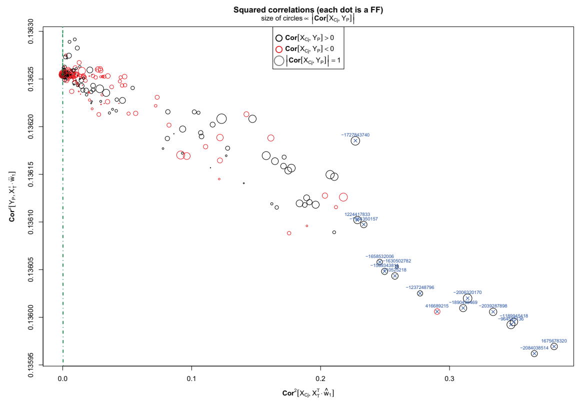
Example (sparse Pearson method, $\tau = 0.1$)

Same FF as before: highest Pearson-sparse $R_j^2(\hat{w}_1)$ value.



Onco: Bicorrelation Plots

Sparse Pearson method



Case Study

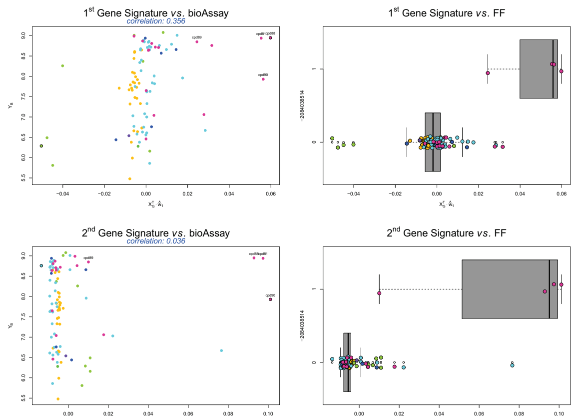
Procedure

Steps of the analysis we performed so far:

- standardise the gene expression matrix (just I/NI call filtered) as in the SMA (log-transformed, double centered, globally normalised);
- compute the maximised bicorrelation coefficient $R_j^2(\hat{w}_1)$ for all FFs, plot the related graphs, and select the top 12 FFs;
- rank FFs by their (Pearson-sparse) $R_j^2(\hat{w}_1)$ value;
- produce the interesting plots for the top FFs;
- rank gene loadings by their average $|\hat{w}_1| + |\hat{w}_2|$, computed over the selected FFs;
- rank gene signatures by their average $|\mathbf{X}_G\hat{w}_1| + |\mathbf{X}_G\hat{w}_2|$, computed over the selected FFs.

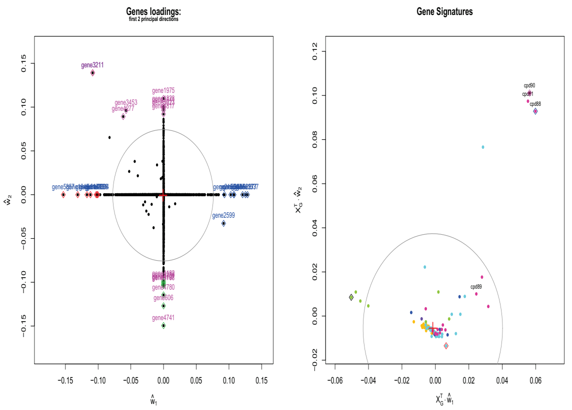
Onco: Gene Signatures vs BA

Top FF: sparse Pearson method, $\tau = 0.1$)



Onco: Gene Loadings and Signatures

Top FF: sparse Pearson method, $\tau = 0.1$)



References

- D. Witten, R. Tibshirani and T. Hastie (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10**, 3, 515–534.
- E. Bair, T. Hastie, D. Paul and R. Tibshirani (2004). Prediction by supervised principal component. *Technical Paper*, Stanford University. `http://www-stat.stanford.edu/~tibs/ftp/spca.pdf`
- Jérôme Pagès. Multiple Factor Analysis: main features and application to sensory data. *Technical Paper*, Agrocampus Rennes (Rennes Cédex, France). `http://factominer.free.fr/docs/PagesAFM.pdf`

Conclusions

- Data integration of three data sources for selecting genes/compounds/fingerprint features that play an important role in triggering the bioassay outcome
- PBA is an asymmetric method in the sense that one data source is selected as a response (here: BA)
- Descriptive plots to get insight into the associations
- Two data sources are univariate (BA and FF) and one data sourse is multivariate (GE)
  Further research: allow for multiple BA and FF simultaneously

Thanks for your attention