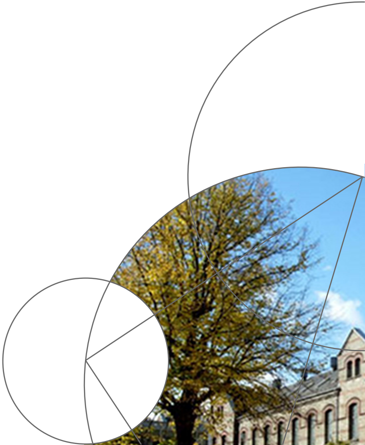
Faculty of Health Sciences

Statistical methods in bioinformatics

Integrative data analysis

Claus Thorn Ekstrøm
Biostatistics,
University of Copenhagen
E-mail: ekstrom@sund.ku.dk



Summary so far

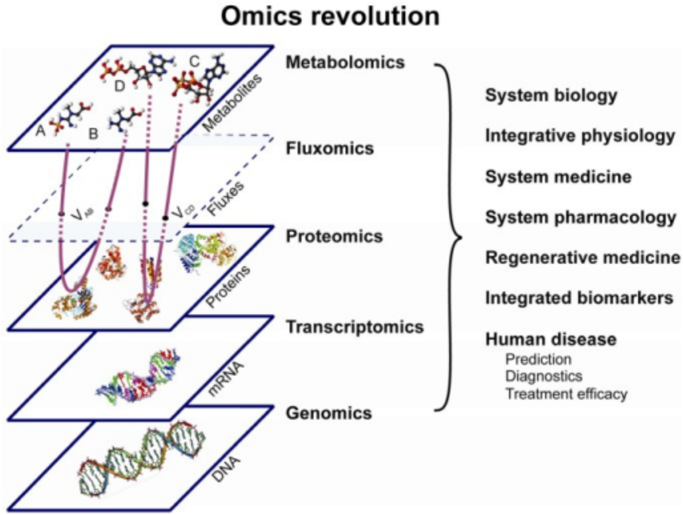
So far we have mainly considered two situations:

1. Large number of outcomes, few predictors.
2. One outcome, large number of predictors.
   - GWAS, gene expression, lasso, pca, ...
   - For example: Networks, (could swap outcome/predictors), ...

Summary so far

- General techniques
- Networks and text mining
- GWAS and genomics
- RNA

The omics revolution



Omics revolution

Metabolomics

Fluxomics

Proteomics

Transcriptomics

Genomics

System biology

Integrative physiology

System medicine

System pharmacology

Regenerative medicine

Integrated biomarkers

Human disease
- Prediction
- Diagnostics
- Treatment efficacy

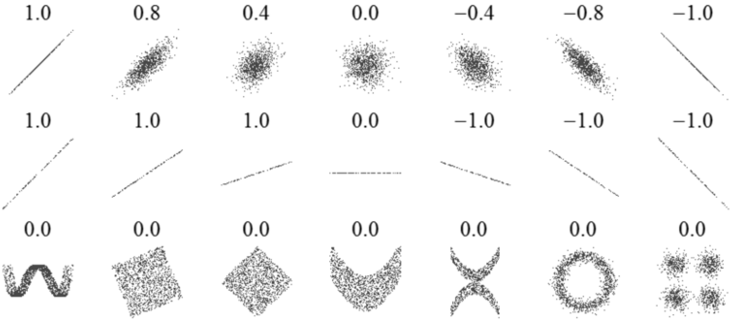
Revisiting correlation

The Pearson correlation between to quantitative variables, X, and Y is

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum_{i=1}^{n}(x_i - \bar{x})^2)(\sum_{i=1}^{n}(y_i - \bar{y})^2)}}$$

Measures the *linear* relationship between X and Y.

Revisiting correlation

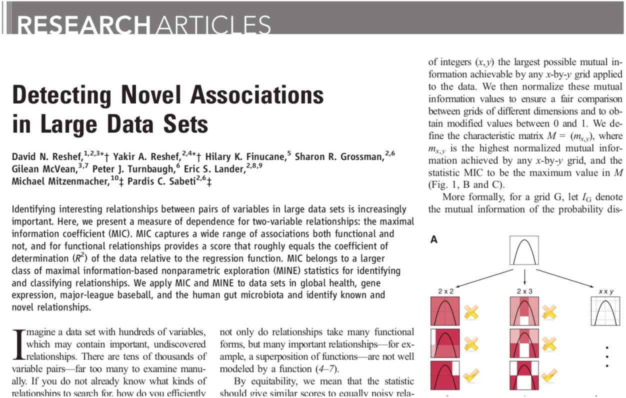


Next generation correlation = MIC ?

Can we do something more advanced than simple correlations? Maximum information correlation

Next generation correlation = MIC ?

Can we do something more advanced than simple correlations? Maximum information correlation



RESEARCH ARTICLES

Detecting Novel Associations in Large Data Sets

David N. Reshef,[1,2,3*†] Yakir A. Reshef,[2,4*†] Hilary K. Finucane,[5] Sharon R. Grossman,[2,6] Gilean McVean,[3,7] Peter J. Turnbaugh,[6] Eric S. Lander,[2,8,9] Michael Mitzenmacher,[10‡] Pardis C. Sabeti[2,6‡]

Identifying interesting relationships between pairs of variables in large data sets is increasingly important. Here, we present a measure of dependence for two-variable relationships: the maximal information coefficient (MIC). MIC captures a wide range of associations both functional and not, and for functional relationships provides a score that roughly equals the coefficient of determination (R^2) of the data relative to the regression function. MIC belongs to a larger class of maximal information-based nonparametric exploration (MINE) statistics for identifying and classifying relationships. We apply MIC and MINE to data sets in global health, gene expression, major-league baseball, and the human gut microbiota and identify known and novel relationships.

Imagine a data set with hundreds of variables, which may contain important, undiscovered relationships. There are tens of thousands of variable pairs—far too many to examine manually. If you do not already know what kinds of relationships to search for, how do you efficiently

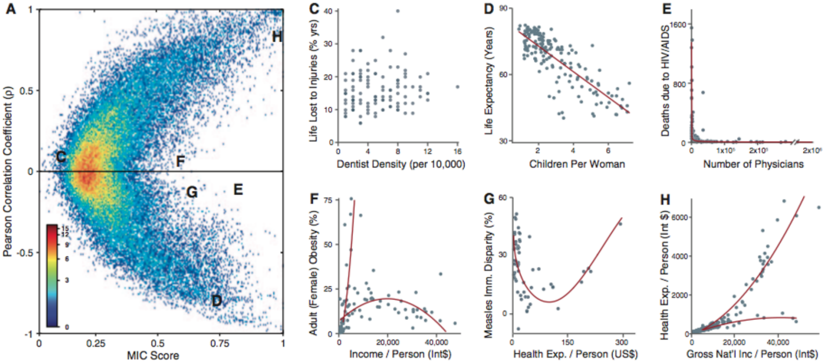
not only do relationships take many functional forms, but many important relationships—for example, a superposition of functions—are not well modeled by a function (*4*–*7*).

By equitability, we mean that the statistic should give similar scores to equally noisy rela-

of integers (x,y) the largest possible mutual information achievable by any x-by-y grid applied to the data. We then normalize these mutual information values to ensure a fair comparison between grids of different dimensions and to obtain modified values between 0 and 1. We define the characteristic matrix $M = (m_{x,y})$, where $m_{x,y}$ is the highest normalized mutual information achieved by any x-by-y grid, and the statistic MIC to be the maximum value in M (Fig. 1, B and C).

More formally, for a grid G, let I_G denote the mutual information of the probability dis-

Example — from MIC paper



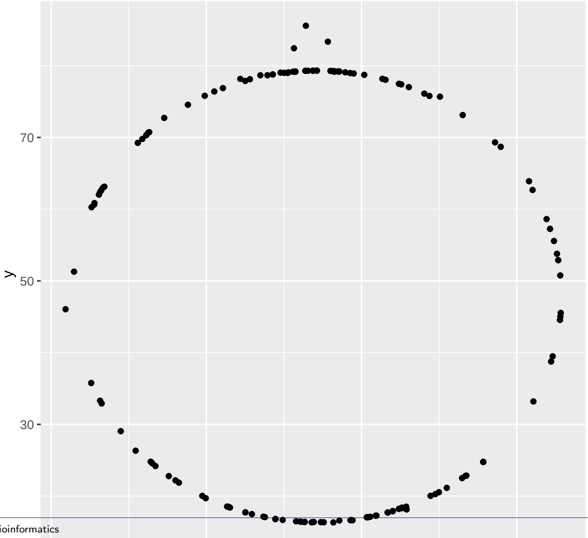
dCor — distance correlation matrix

Produces a measure of variable dependence: From 0 (corresponds to statistical independence) to 1 (no noise).

- Produces number between 0 and 1
- Can have different dimensions (but requires same N)
- Can detect both linear and non-linear dependence
- Approximates standard Pearson correlation coefficient when relationship is roughly linear.

dCor

```
> library("energy")       # Pearson cor: -0.068
> cor(x,y); dcor(x, y)  # dcor = 0.2291
```



Computing dCor

Compute the distance correlation between $X \in R_k^N$ and $Y \in R_j^N$.

1. Compute matrix of Euclidian distances between N cases for X and Y.
2. Perform double centering for each matrix
3. Multiply the matrices element-wise and compute sum.
4. Divide by N^2 (ie, compute average).
5. Take square root. This is the distance covariance.
6. Variances can be computed for each matrix against itself.
7. The distance correlation is computed similarly to the Pearson correlation.

Computing dCor

$$(X,Y) = [(0,0),(0,1),(1,0),(1,1)]$$

Inference

What about inference?

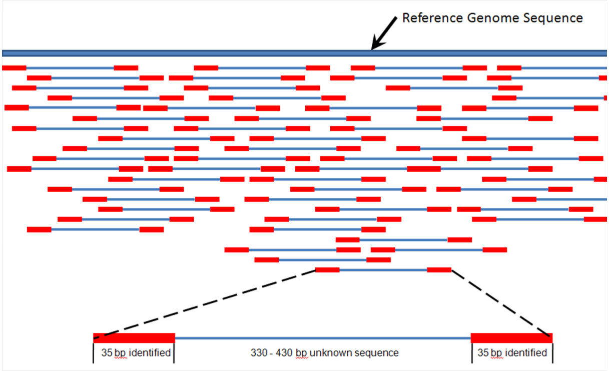
For a given pair of high-dimensional variables:

- Compute a modified version of the distance correlation.
- Use `dcorT.test()`

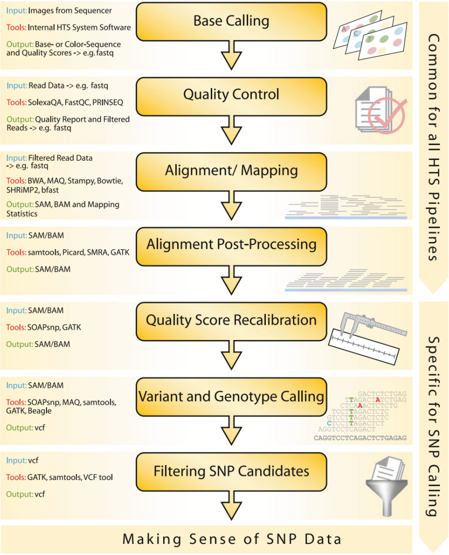
NGS / RNA-seq

Microarrays are limited in what we can find as we can only measure intensities of the probes already on the array.

High-throughput DNA sequencing methods / next-generation sequencing



Gene variant calling



Input: Images from Sequencer
Tools: Internal HTS System Software
Output: Base- or Color-Sequence and Quality Scores -> e.g. fastq

Base Calling

Input: Read Data -> e.g. fastq
Tools: SolexaQA, FastQC, PRINSEQ
Output: Quality Report and Filtered Reads -> e.g. fastq

Quality Control

Input: Filtered Read Data -> e.g. fastq
Tools: BWA, MAQ, Stampy, Bowtie, SHRiMP2, bfast
Output: SAM, BAM and Mapping Statistics

Alignment/ Mapping

Input: SAM/BAM
Tools: samtools, Picard, SMRA, GATK
Output: SAM/BAM

Alignment Post-Processing

Input: SAM/BAM
Tools: SOAPsnp, GATK
Output: SAM/BAM

Quality Score Recalibration

Input: SAM/BAM
Tools: SOAPsnp, MAQ, samtools, GATK, Beagle
Output: vcf

Variant and Genotype Calling

Input: vcf
Tools: GATK, samtools, VCF tool
Output: vcf

Filtering SNP Candidates

Making Sense of SNP Data

Common for all HTS Pipelines

Specific for SNP Calling

NGS technologies

Recall from yesterday:

1. Align sequenced fragments with reference sequence (alternatively make *de novo* assembly).
   - really a non-trivial task, but will not go into details. abundance.
2. Count the number of fragments mapping to certain regions
   - usually genes
   - The read counts linearly approximate target transcript abundance.

A large number of short DNA fragments. The reads are then used for several applications, e.g., sequence reconstruction, DNA assembly, gene expression profiling, mutation analysis.

Normalization

Number of reads are approximately proportional to length of transcript, the total number of mapped reads.

Typically considering the reads per kilobase per million reads (RPKM) or variations on this theme.

1. Count up the total reads
2. Divide by 1,000,000 ⇒ "per million" scaling factor to normalize for sequencing depth (RPM)
3. Divide the RPM values by the length of the gene, in kilobases. This gives you RPKM.

Modeling read counts

Back to the linear model?

$$\text{count}_i = \mathbf{X}\beta + \varepsilon_i$$

Assumption of continuous data for each gene. But they really are counts (discrete) and relatively infrequent.

Let N_i be total number of fragments counted in sample i, and p_i the probability that a fragment matches a particular gene of interest.
The observed number of reads for gene in sample i is

$$R_i \sim \text{Poisson}(N_i p_i)$$

Note: $\mathbb{E}(R_i) = \text{Var}(R_i) = N_i p_i$.

Modeling read counts

Wish to, say, compare two groups: cases and controls? Assume $\log(p_i = \alpha + \beta x_i)$, where x_i is 0 (controls) or 1 (cases).

Generalized linear model (Poisson regression):

$$\log(\mathbb{E}(R_i)) = \underbrace{\log(N_i)}_{\text{Not interesting}} + \alpha + \beta x_i$$

Hypothesis of no differential expression between the groups

$$H_0 : \beta = 0$$

```
glm(reads ~ group + offset(N), data=DF, family="poisson")
```

Can extend the model to Generalized linear mixed effect (Poisson mixed effect model) to account for additional sources of variation.

Modeling read counts

Overdispersion can be a problem.

Recall the assumption from the Poisson distribution:

$\mathbb{E}(R_i) = \mathrm{Var}(R_i) = N_i p_i$

Modeling read counts

Overdispersion can be a problem.
Recall the assumption from the Poisson distribution:
$\mathbb{E}(R_i) = \mathrm{Var}(R_i) = N_i p_i$

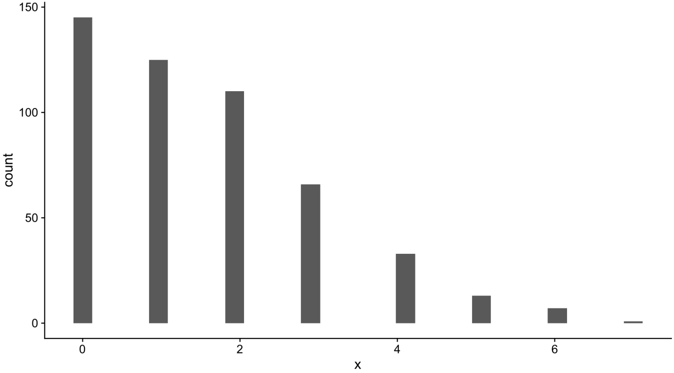
Alternatives:

- Use a Poisson regresion with overdispersion, i.e., where $\mathrm{Var}(R_i) = \sigma\mathbb{E}(R_i)$.
- Use another distribution — for example a negative binomial distribution — to describe the read counts.

```
glm(reads ~ group + offset(N), data=DF,
    family="quasipoisson")
```

Zero-inflation models

The dispersion problem in Poisson/NB models is often caused by zero-inflation.

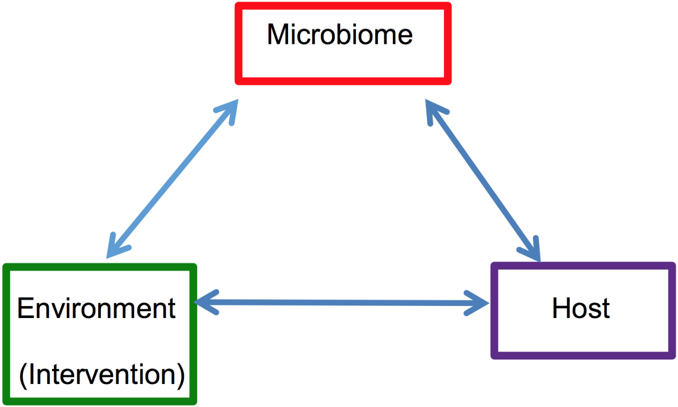


Zero-inflation models

Useful in situations like:

- RNA sequence reads
- Microbiome data (abundance counts or percentages)
- (Some) mixture modeling

Example: microbiome data



Example: microbiome data, abundance

Individuals analysis of operational taxonomic units (OTUs)

```
        T1   T2    T3    T4    T5    T6    T7    T8    T9   T10
Sam1  0.00 0.00 16.72 28.52  0.00  4.74 22.69  0.00 11.81 15.53
Sam2 24.10 7.69  0.00  0.00 16.59  0.00  0.00  6.61 20.26 24.76
Sam3 12.99 0.00 36.00  0.00 18.22 12.24  0.00  8.84  0.00 11.71
Sam4 10.33 7.15  8.28 23.03  4.12  3.66  0.00 21.77  6.36 15.31
Sam5  4.47 5.66 13.77 15.24  0.00 31.41 23.38  0.00  6.07  0.00
```

Two types of zeroes!
Compositional data.

Zero-inflation models

Often two-part models:

$$Y_i \sim \begin{cases} \delta_0 & \text{if } \pi_i \\ F_i & \text{if } (1-\pi_i) \end{cases},$$

where δ_0 is a point-mass in zero, and π_i is a mixture probability.

Mixture model with two components:

- A model for the mixture component.
- A conditional model for the data *given* that it is not zero.

Different interpretations

- Zero-inflated models: A standard distribution, F_i, and an excess of zeroes, δ_0.
- Hurdle models: A standard distribution *which does not contain zeroes*, F_i, and a number of zeroes, δ_0.

Different interpretation and view of contamination.

Possibilities in R

- Zero-inflated models: A standard distribution, F_i, and an excess of zeroes, δ_0.
- Hurdle models: A standard distribution *which does not contain zeroes*, F_i, and a number of zeroes, δ_0.

Different interpretation and view of contamination.

The compositional problem

Truly overdispersed Dirichlet-multinomial data:

- Multiple testing problem.
- OTU's are not independent (when looking at relative abundance).
- Constraints. Negative correlation.

Analysis of composition of microbiomes (ANCOM)

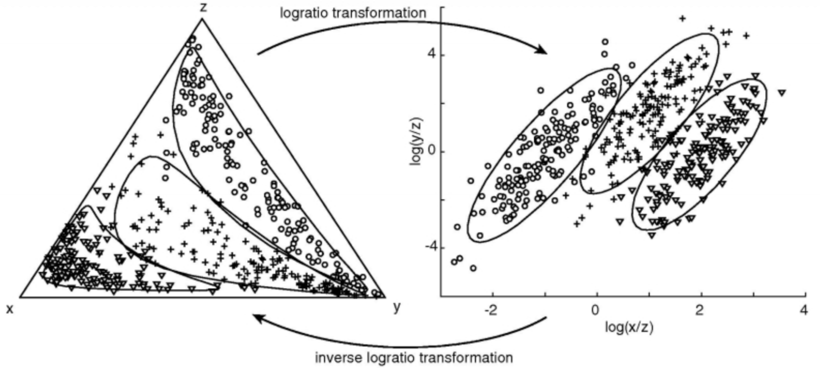
Aitchison's solution to the compositional data problem. Transform data from Δ_{N-1} to R^{N-1} using the log-ratio transformation, e.g., for $(X+Y+Z=1)$ we use

$$V = \log(X/Z), W = \log(Y/Z)$$

Inverse log-ratio transform

$$X = \frac{\exp(V)}{\exp(V)+\exp(W)+1}, Y = \frac{\exp(W)}{\exp(V)+\exp(W)+1}, Z = \frac{1}{\exp(V)+\exp(W)+1}$$

The transform



The isometric-log-ratio (ilr) transformation

1. Represent a composition as a real vector
2. Coordinates in an *orthogonal* system
3. Use function `ilr()` from the `compositions` package.
4. Interpretation of the results may be difficult, since there is no one-to-one relation between the original parts
5. Can be analyzed using multivariate analysis tools

Integrative data analysis

Integrative data analysis: analysis of data from multiple sources (aka Multi-Omics analysis).

Typically several high-dimensional datasets. Analysing each of them by itself could be problematic.

How can we combine them?

- Data pooling
- Multi-step methods
- Simultaneous analysis

No golden standard!

Data pooling

Large dataset from different sources — on the same type of experiment.

Data pooling

Large dataset from different sources — on the same type of experiment.

Not really a “problem”.

- If we only have summary statistics then to meta-analysis
- If we have raw data then merge the datasets and do the analysis we would do on each of them.

Statistical model

$$Y_i = \mathbf{X}\beta + \text{source}_i\gamma + \varepsilon_i$$

Data pooling

Large dataset from different sources — on the same type of experiment.
Not really a “problem”.

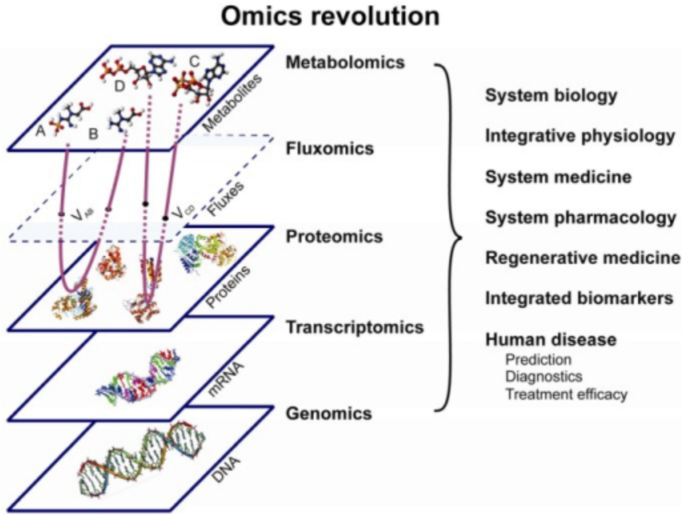
- If we only have summary statistics then to meta-analysis
- If we have raw data then merge the datasets and do the analysis we would do on each of them.

Statistical model

$$Y_i = \mathbf{X}\beta + \text{source}_i\gamma + \varepsilon_i$$

- Increased statistical power
- Increased sample heterogeneity

Integrative data analysis



High-dimensional

So far we have considered two situations:

1. Large number of outcomes, few predictors.
    - Gene expression
2. One outcome, large number of predictors.
    - GWAS, gene expression

Simultaneous analysis of multiple outcomes

How can we handle multiple outcomes?

Univariate statistical model

$$Y_i = \mathbf{X}_i\beta + \varepsilon_i,\ \varepsilon_i \sim N(0,\sigma^2)$$

But we have *M* of those (one for each outcome)
Multivariate version:

$$Y_{mi} = \mathbf{X}_{mi}\beta_m + \varepsilon_{mi},\ \varepsilon_i \sim N(0,\sigma_m^2)$$

Simultaneous analysis of multiple outcomes

How can we handle multiple outcomes?

Univariate statistical model

$$Y_i = \mathbf{X}_i\beta + \varepsilon_i,\ \varepsilon_i \sim N(0,\sigma^2)$$

But we have *M* of those (one for each outcome)
Multivariate version:

$$Y_{mi} = \mathbf{X}_{mi}\beta_m + \varepsilon_{mi},\ \varepsilon_i \sim N(0,\sigma_m^2)$$

“Stack them” and analyze them using the methods we have already seen. Note we have variance hetergeneity!

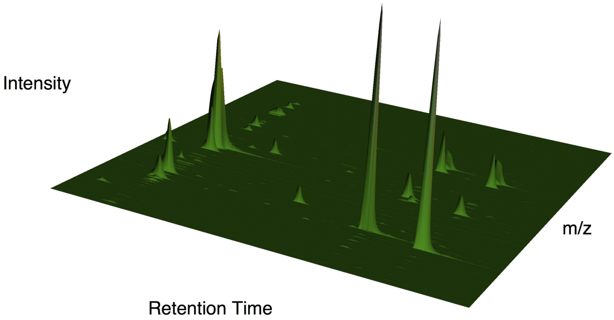
“Real” multivariate outcomes

```
gene1  1.1 0.3 0.2 -.4 1.4 1.0 ...
gene2  0.3 2.3 1.2 -.9 -.4 -.1 ...
gene3  2.0 0.0 0.0 0.2 -.2 -.2 ...
   .
   .
geneN  1.1 0.4 0.1 -.3 0.4 0.0 ...
```

Now imagine we have measurements over time.

Each individual provides a longitudinal profile of measurements.

LCMS metabolite data



Analysis of longitudinal data

Univariate statistical model

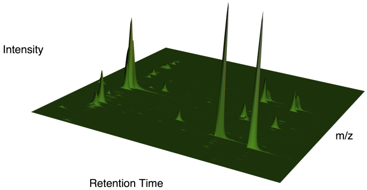
$$Y_i = \mathbf{X}_i\beta + \varepsilon_i,\ \varepsilon_i \sim N(0,\sigma^2)$$

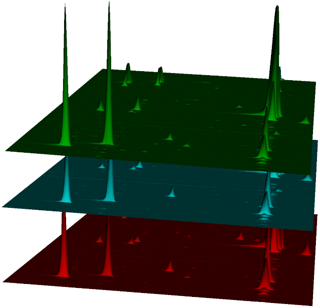
A generalized linear mixed effect model (GLMM / mixed model / random effect model) be used to extend the GLM to accommodate longitudinal measurements.

However, not really suited for super-large dataset.

Critical with multiple testing

Data types





- Measured by LC-MS
- 3-D data structure
- Regions of interest
- $\mathbf{Y} \in \mathbb{R}^{r \times k \times n}$

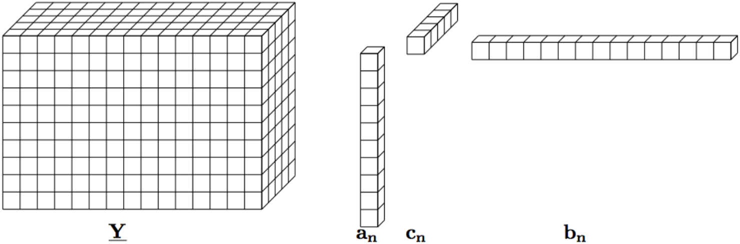
Dimension reduction

- Approximate $\mathbf{Y} \in \mathbb{R}^{r \times k \times n}$ such that

$$\mathbf{Y} \approx \sum_{i=1}^{c} \mathbf{A}_i \otimes \mathbf{B}_i \otimes \mathbf{C}_i$$

  where $\mathbf{A} \in \mathbb{R}^{k \times c}$, $\mathbf{B} \in \mathbb{R}^{r \times c}$ and $\mathbf{C} \in \mathbb{R}^{n \times c}$ and where c is the number of components.
- Important that c is fairly accurate. Chosen empirically.
- **A** and **B** can be interpreted as basis functions for retention time and m/z values
- **C** is the mixing matrix, representing the scaling of **A** and **B** needed to reconstruct the original data.

Parallel factor analysis



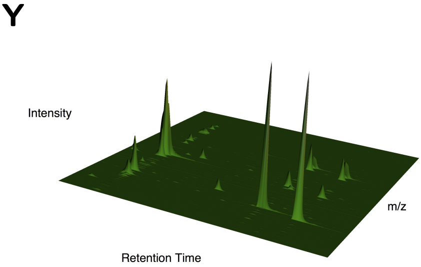
Multidimensional dimension reduction

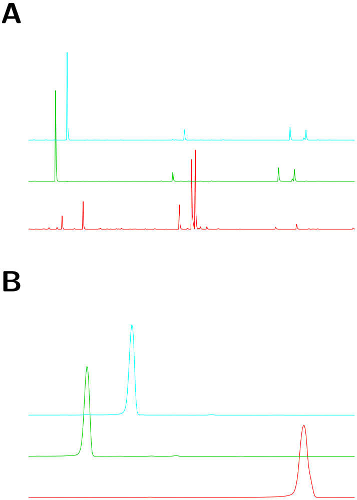
In a sense it is like PCA:

We wish to find a few simple components that can approximate the matrix well.

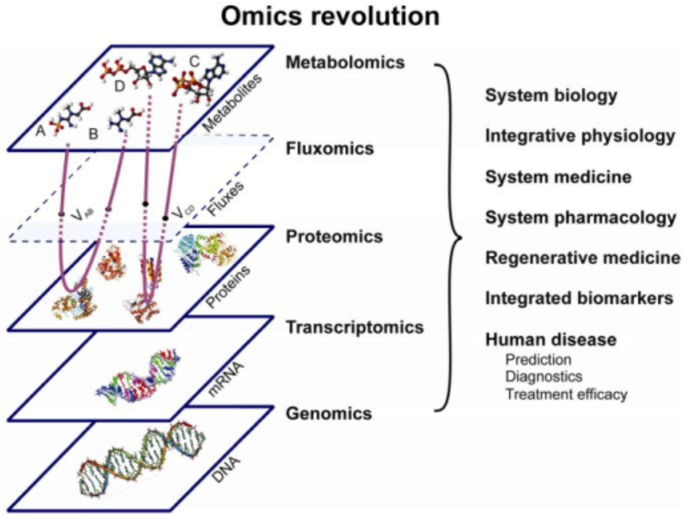
Optimal solution: We wish to find a few simple components that can approximate the matrix well *and that we can interpret!*.

PARAFAC (parallel factor analysis)





Integrative data analysis



Omics revolution

System biology

Integrative physiology

System medicine

System pharmacology

Regenerative medicine

Integrated biomarkers

Human disease

Prediction
Diagnostics
Treatment efficacy

Integrative data analysis

$$\mathit{phenotype} \longleftarrow \mathit{metabolite} \longleftarrow \mathit{geneexpression} \longleftarrow \mathit{SNP}$$

$$Y = X\beta$$

$$Z = Y\gamma$$

$$W = Z\theta$$

Could be analyzed with a multiple regression model:

$$W = Z\theta = (Y\gamma)\theta = X\beta\gamma\theta$$

What about the errors?

Background

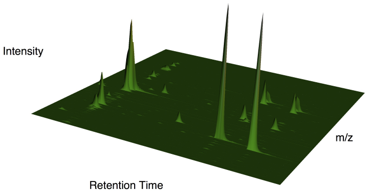
Cassava dataset 30 samples

- gene expression of 13865 genes
- metabolite profiling with LC-MS

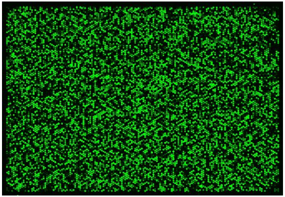
Goal

Identify new associations between gene expression and metabolites

Data types



- Measured by LC-MS
- 3-D data structure
- Regions of interest
- $\mathbf{Y} \in \mathbb{R}^{r \times k \times n}$



- Measured by DNA microarray
- 2-D data structure
- Few genes of interest
- $\mathbf{X} \in \mathbb{R}^{m \times n}$

Example

- Genes control production of metabolites
- Measure gene expression
- Measure metabolite production
- Construct a model that includes both data types
- Results directly related to the underlying biology

Method

We wish to formulate a model

$$\mathbb{E}(\mathbf{Y}) \text{ is a function of } \mathbf{X}\beta,$$

where $\mathbf{Y} \in \mathbb{R}^{r \times k \times n}$ is a 3-D tensor of spectra. $\mathbf{X} \in \mathbb{R}^{m \times n}$ is a matrix of gene expression and β is a coefficient matrix

Samples from n experiments, m genes, $m \gg n, r \times k \gg n$

Problems:

- Dimension reduction
- Variable/feature selection
- Biological interpretation
- Some kind of inference

Dimension reduction

- Approximate $\mathbf{Y} \in \mathbb{R}^{r \times k \times n}$ such that

$$\mathbf{Y} \approx \sum_{i=1}^{c} \mathbf{A}_i \otimes \mathbf{B}_i \otimes \mathbf{C}_i$$

  where $\mathbf{A} \in \mathbb{R}^{k \times c}$, $\mathbf{B} \in \mathbb{R}^{r \times c}$ and $\mathbf{C} \in \mathbb{R}^{n \times c}$ and where c is the number of components.
- Important that c is fairly accurate. Chosen empirically.
- **A** and **B** can be interpreted as basis functions for retention time and m/z values
- **C** is the mixing matrix, representing the scaling of **A** and **B** needed to reconstruct the original data.

Modelling

Since the mixing matrix **C** is the scaling of the basis functions, gene expressions highly associated with **C** are likely to have an effect of the peaks in **Y**.

Make c models

$$C_i = \mathbf{X}\beta_i + \varepsilon_i \quad \text{for} \quad i = 1, \ldots, c$$

with β_i subject to some restrictions

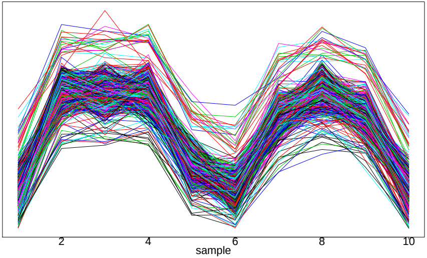
Restrictions can be LASSO, OSCAR, elastic net, …according to the purpose of the analysis

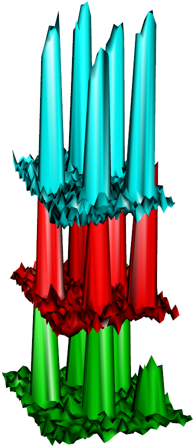
Results from each model gives information about each of the c components.

Simulations

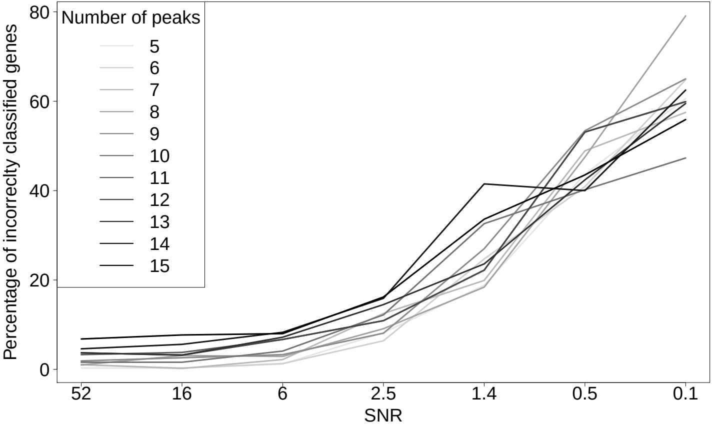
- 10 'biological' replicates
- 2 treatments
- 1000 genes
- 5-15 metabolites, controlled by as many genes
- 300 runs for each combination

Simulations

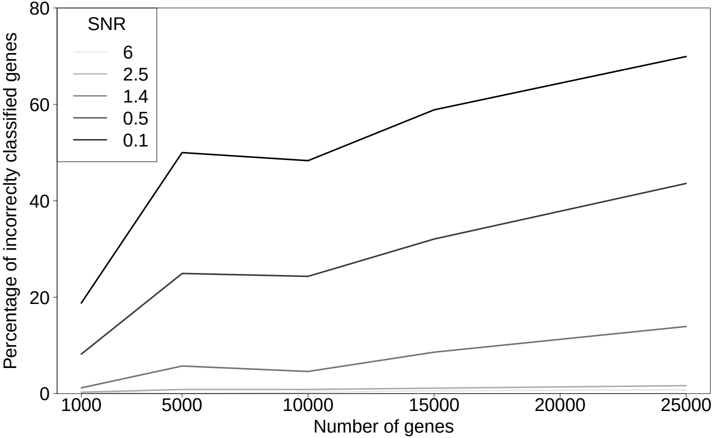




Simulation results



Simulation results



Application — Cassava

- Cassava dataset 30 samples, 13865 genes

Application — Cassava

- Cassava dataset 30 samples, 13865 genes
- Three compound↔gene relationships found.
  - Linamarin, a well-known compound in Cassava
  - Gene coupled to several CYP79 enzymes [catalyst in the synthesis of Linamarin in Cassava]
  - Final peak quite likely Lotaustralin and ... no clue

Application — Cassava

- Cassava dataset 30 samples, 13865 genes
- Three compound↔gene relationships found.
  - Linamarin, a well-known compound in Cassava
  - Gene coupled to several CYP79 enzymes [catalyst in the synthesis of Linamarin in Cassava]
  - Final peak quite likely Lotaustralin and ... no clue
- But ... Cassava is different