Unsupervised and Semi-Supervised Learning

Nataliya Sokolovska

Sorbonne University
Paris, France

Master 2 BIM
January, 25, 2019

Outline

- Patients Stratification and Methods of Personalized Medicine
- An application: Obesity stratification based on metagenomics
  - Weight Loss Prediction
- Some (Fancy) Clustering Methods
  - Some widely used Clustering Methods
  - Spectral Clustering
  - Biclustering
  - Robust Clustering
- Semi-Supervised Learning
- Canonical Correlation: Correlation between Sets of Variables

Patients Stratification and Methods of Personalized Medicine

An application: Obesity stratification based on metagenomics

Some (Fancy) Clustering Methods

Semi-Supervised Learning

Canonical Correlation: Correlation between Sets of Variables

What is Metagenomics?

- Metagenome
  - can be defined as the ensemble of the microbes from a given ecological niche
- Metagenomics
  - allows to characterize composition, properties, and dynamics of a microbiome by studying the metagenome

Obesity stratification based on metagenomics

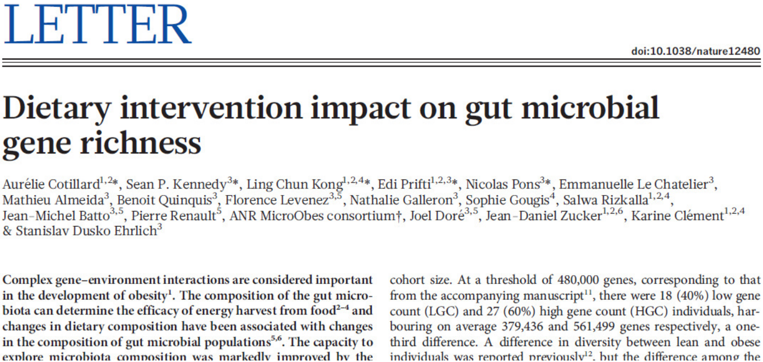
Adapted from Nicolas Pons, Ecole NGS INRA, Lyon, january 2012

□ 100 trillion microorganisms ; 10-fold more cells than the human body; 2 kg of mass!

□ Interface between food and epithelium

□ In contact with the 1st pool of immune cells and the 2^{nd} pool of neural cells of the body

MicroObese Study



LETTER

doi:10.1038/nature12480

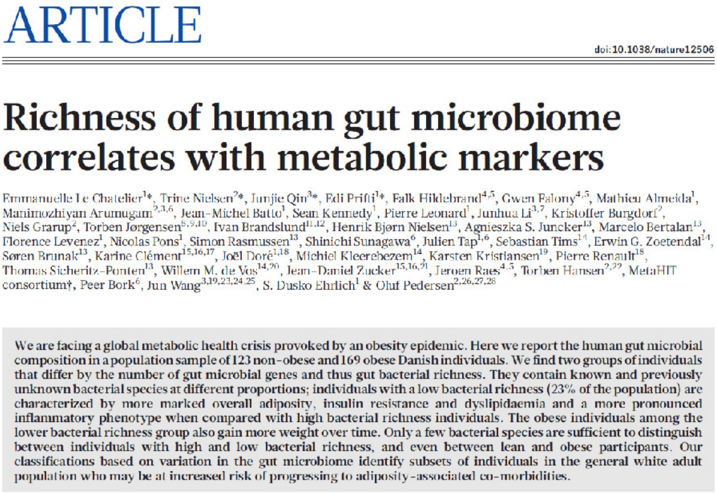
Dietary intervention impact on gut microbial gene richness

Aurélie Cotillard[1,2]*, Sean P. Kennedy[3]*, Ling Chun Kong[1,2,4]*, Edi Prifti[1,2,3]*, Nicolas Pons[3]*, Emmanuelle Le Chatelier[3], Mathieu Almeida[3], Benoit Quinquis[3], Florence Levenez[3,5], Nathalie Galleron[3], Sophie Gougis[4], Salwa Rizkalla[1,2,4], Jean-Michel Batto[3,5], Pierre Renault[5], ANR MicroObes consortium†, Joel Doré[3,5], Jean-Daniel Zucker[1,2,6], Karine Clément[1,2,4] & Stanislav Dusko Ehrlich[3]

Complex gene–environment interactions are considered important in the development of obesity[1]. The composition of the gut microbiota can determine the efficacy of energy harvest from food[2–4] and changes in dietary composition have been associated with changes in the composition of gut microbial populations[5,6]. The capacity to explore microbiota composition was markedly improved by the

cohort size. At a threshold of 480,000 genes, corresponding to that from the accompanying manuscript[11], there were 18 (40%) low gene count (LGC) and 27 (60%) high gene count (HGC) individuals, harbouring on average 379,436 and 561,499 genes respectively, a one-third difference. A difference in diversity between lean and obese individuals was reported previously[12], but the difference among the

MetaHIT Study



ARTICLE

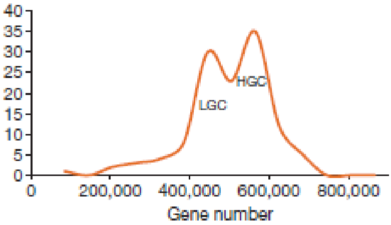
doi:10.1038/nature12506

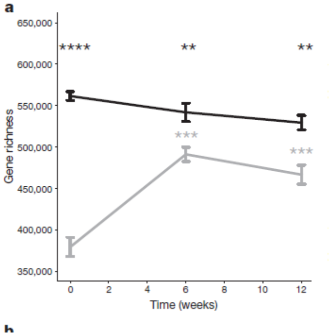
Richness of human gut microbiome correlates with metabolic markers

Emmanuelle Le Chatelier[1]*, Trine Nielsen[2]*, Junjie Qin[3]*, Edi Prifti[1]*, Falk Hildebrand[4,5], Gwen Falony[4,5], Mathieu Almeida[1], Manimozhiyan Arumugam[2,3,6], Jean-Michel Batto[1], Sean Kennedy[1], Pierre Leonard[1], Junhua Li[3,7], Kristoffer Burgdorf[2], Niels Grarup[2], Torben Jørgensen[8,9,10], Ivan Brandslund[11,12], Henrik Bjørn Nielsen[13], Agnieszka S. Juncker[13], Marcelo Bertalan[13], Florence Levenez[1], Nicolas Pons[1], Simon Rasmussen[13], Shinichi Sunagawa[6], Julien Tap[1,6], Sebastian Tims[14], Erwin G. Zoetendal[14], Søren Brunak[13], Karine Clément[15,16,17], Joël Doré[1,18], Michiel Kleerebezem[14], Karsten Kristiansen[19], Pierre Renault[18], Thomas Sicheritz-Ponten[13], Willem M. de Vos[14,20], Jean-Daniel Zucker[15,16,21], Jeroen Raes[4,5], Torben Hansen[2,22], MetaHIT consortium†, Peer Bork[6], Jun Wang[3,19,23,24,25], S. Dusko Ehrlich[1] & Oluf Pedersen[2,26,27,28]

We are facing a global metabolic health crisis provoked by an obesity epidemic. Here we report the human gut microbial composition in a population sample of 123 non-obese and 169 obese Danish individuals. We find two groups of individuals that differ by the number of gut microbial genes and thus gut bacterial richness. They contain known and previously unknown bacterial species at different proportions; individuals with a low bacterial richness (23% of the population) are characterized by more marked overall adiposity, insulin resistance and dyslipidaemia and a more pronounced inflammatory phenotype when compared with high bacterial richness individuals. The obese individuals among the lower bacterial richness group also gain more weight over time. Only a few bacterial species are sufficient to distinguish between individuals with high and low bacterial richness, and even between lean and obese participants. Our classifications based on variation in the gut microbiome identify subsets of individuals in the general white adult population who may be at increased risk of progressing to adiposity-associated co-morbidities.

MicroObese Study





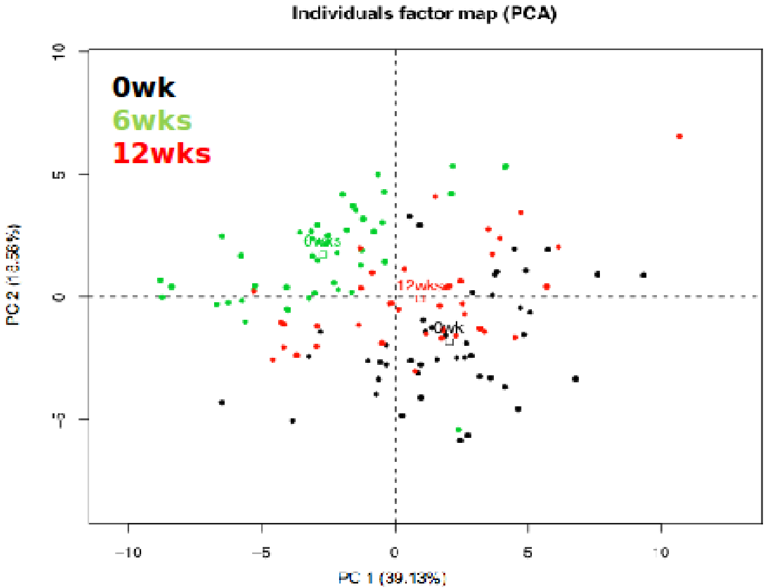
Obesity stratification based on metagenomics

- Gut microbial gene richness can influence the outcome of a dietary intervention
- A quantitative metagenomic analysis stratified patients into two groups: a group with low gene count (LGC) and a high gene count (HGC) group
- The LGC individuals appeared to have increased blood triglycerides, higher insulin-resistance and low-grade inflammation, and therefore the gene richness is strongly associated with obesity-driven diseases.
- The individuals from a low gene count group seemed to have an increased risk to develop obesity-related cardiometabolic risk compared to the patients from the high gene count group.

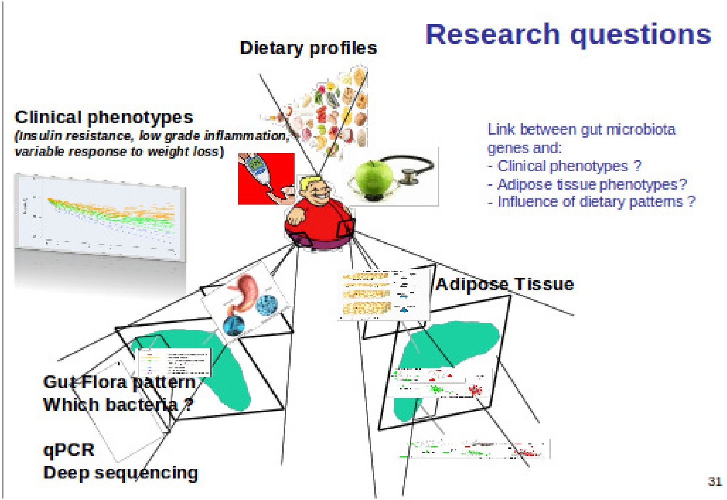
Stratification of Dutch individuals

- E. Le Chatelier et al., 2011 conducted a similar study with Dutch individuals, and made a similar conclusion: there is a hope that a diet can be used to induce a permanent change of gut microbiota, and that treatment should be phenotype-specific.
- A particular diet is able to increase the gene richness: an increase of genes was observed with the LGC patients after a 6-weeks energy-restricted diet

PCA: example on real data



Weight Loss Prediction



Research questions

Dietary profiles

Clinical phenotypes

(Insulin resistance, low grade inflammation, variable response to weight loss)

Link between gut microbiota genes and:

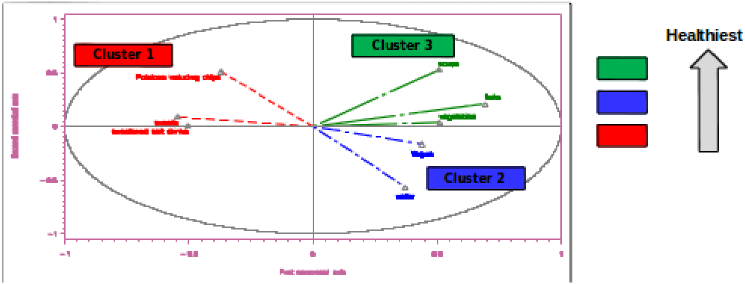
- Clinical phenotypes ?
- Adipose tissue phenotypes?
- Influence of dietary patterns ?

Adipose Tissue

Gut Flora pattern
Which bacteria ?

qPCR
Deep sequencing

Weight Loss Prediction



Patients Stratification and Methods of Personalized Medicine

An application: Obesity stratification based on metagenomics

Some (Fancy) Clustering Methods

- Some widely used Clustering Methods
- Spectral Clustering
- Biclustering
- Robust Clustering

Semi-Supervised Learning

Canonical Correlation: Correlation between Sets of Variables

The clustering problem

- Motivation: find patterns in a sea of data
- Input
  - A large number of data points
  - A measure of distance between any two points
- Output
  - Grouping (clustering) of the elements int K similarity clusters
- Clustering is useful for
  - Similarity/dissimilarity analysis
  - Dimensionality reduction

Patients Stratification and Methods of Personalized Medicine

An application: Obesity stratification based on metagenomics

Some (Fancy) Clustering Methods

- Some widely used Clustering Methods
- Spectral Clustering
- Biclustering
- Robust Clustering

Semi-Supervised Learning

Canonical Correlation: Correlation between Sets of Variables

Some widely used Clustering Methods

- Hierarchical clustering
- K-means
- Probabilistic methods of clustering (Mixtures of Gaussians, EM)

Patients Stratification and Methods of Personalized Medicine

An application: Obesity stratification based on metagenomics

Some (Fancy) Clustering Methods

- Some widely used Clustering Methods
- Spectral Clustering
- Biclustering
- Robust Clustering

Semi-Supervised Learning

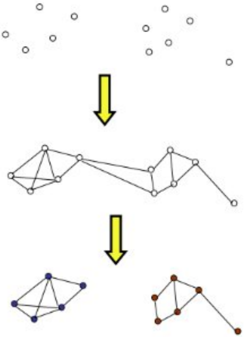
Canonical Correlation: Correlation between Sets of Variables

Spectral Clustering

- *U. von Luxburg, "A tutorial on spectral clustering", Stat. Comp., 2007*
- One of the most popular clustering algorithms
- It can be proved that under very mild conditions, spectral clustering algorithms are statistically consistent. This means that is we assume that the data has been sampled randomly according to some probability distribution from some underlying space, and if we let the sample size increase to infinity, then the results of clustering converge (these results do not necessary hold of unnormalized spectral clustering).

Graph notation and similarity graphs

If we do not have more information than similarities between data points, a nice way of representing the data is in form of **similarity graph**. The vertices represent the data points. Two vertices are connected if the similarity between the corresponding data points is positive (or larger than a certain threshold), and the edge is weighted by the similarity.



Graphs and Cluster Assumption

The problem of clustering: we want to find a partition of the graph such that the edges between different groups have a very low weight.

“Cluster assumption”: two points are likely to have the same class label if there is a path connecting them passing through regions of high density only. Or, the decision boundary should lie in regions of low density.

Graph notations

- $G = (V, E)$ is an undirected graph
- the graph is weighted: each edge between two vertices v_i and v_j has a weight $w_{ij} > 0$
- The weighted adjacency matrix W ($w_{ij} = 0$ mean that the vertices are not connected)
- Graph is undirected, $w_{ij} = w_{ji}$
- The degree of a vertex v_i is defined as $d_i = \sum_{j=1}^{n} w_{ij}$
- The degree matrix D

Graph notations Cont'd

- A subset of vertices A
- Two ways of measuring the size of A
  - $|A|$ – the number of vertices in A
  - $vol(A) = \sum_{i \in A} d_{ij}$ – measure the size of A by the weights of its edges
- a subset A is connected is any two vertices in A cab be joined by a path such that all intermediate points also lie in A.

Different similarity graphs (used in Spectral Clustering)

There are several popular constructions to transform a given set of data points into a graph. Most of them lead to a sparse representation ⇒ computational advantages.

- The ϵ-neighborhood graph. We connect all points whose pairwise distances are smaller than ϵ. Usually considered as an unweighted graph.
- k-nearest neighbor graphs. We connect vertex v_i with vertex v_j if v_j is among the k nearest neighbors of v_i.
- The fully connected graph. We connect all points with positive similarity with each other, and we weight the edges by s_{ij}. The graph should model the local neighborhood relationships. An example of similarity function is the Gaussian similarity function $s(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2})$. The parameters σ controls the width of the neighborhoods.

Graph Laplacians

- The main tool for spectral clustering are graph Laplacian matrices
- In the literature, there is no unique convention which matrix exactly is called “graph Laplacian”
- The unnormalized graph Laplacian matrix is defined as

$$L = D - W$$

.

- The normalized Laplacian

$$L = D^{-1/2}(D - W)D^{-1/2}$$

Properties of L

- For every vector $f \in \mathbb{R}^n$ we have

$$f'Lf = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij}(f_i - f_j)^2$$

- L is symmetric and positive semi-definite
- The smallest eigenvalue of L is 0, the corresponding eigenvector is the constant one vector $\mathbb{1}$.
- L has n non-negative, real-valued eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$

Unnormalized Spectral Clustering

- Input: Similarity matrix $S \in \mathbb{R}^{n \times n}$, number k of clusters to construct
    - Construct a similarity graph; W is its weighted adjacency matrix
    - Compute the unnormalized Laplacian L
    - Compute the first k eigenvectors $v_1, \ldots, v_k$ of L.
    - Let $V \in \mathbb{R}^{n \times k}$ be the matrix containing the vectors $v_1, \ldots, v_k$ as columns
    - For $i = 1, \ldots, n$, let $y_i \in \mathbb{R}^k$ be the vector corresponding to the i-th row of V
    - Cluster the points $(y_i)_{i=1,\ldots,n} \in \mathbb{R}^k$ with the k-means algorithm into clusters $C_1, \ldots, C_k$
- Output: Clusters $A_1, \ldots, A_k$.

Normalized Spectral Clustering (Shi and Malik, 2000)

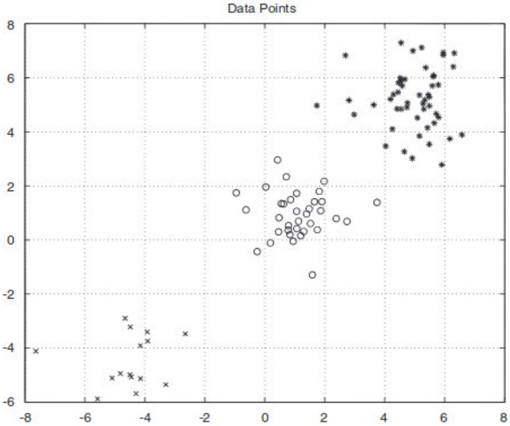
- Input: Similarity matrix $S \in \mathbb{R}^{n \times n}$, number k of clusters to construct
  - Construct a similarity graph; W is its weighted adjacency matrix
  - Compute the unnormalized Laplacian L
  - Compute the first k eigenvectors $v_1, \dots, v_k$ of the generalized eigenproblem $Lv = \lambda Dv$.
  - Let $V \in \mathbb{R}^{n \times k}$ be the matrix containing the vectors $v_1, \dots, v_k$ as columns
  - For $i = 1, \dots, n$, let $y_i \in \mathbb{R}^k$ be the vector corresponding to the i-th row of V
  - Cluster the points $(y_i)_{i=1,\dots,n} \in \mathbb{R}^k$ with the k-means algorithm into clusters $C_1, \dots, C_k$
- Output: Clusters $A_1, \dots, A_k$.

Normalized spectral clustering (Ng, Jordan, and Weiss, 2002)

- Input: Similarity matrix $S \in \mathbb{R}^{n \times n}$, number k of clusters to construct
  - Construct a similarity graph; W is its weighted adjacency matrix
  - Compute the normalized Laplacian L_{sym}
  - Compute the first k eigenvectors $v_1, \dots, v_k$ of L_{sym}.
  - From the matrix $U \in \mathbb{R}^{n \times k}$ from V by normalizing the row sums to have norm 1, that $u_{ij} = v_{ij} / (\sum_k v_{ik}^2)^{1/2}$
  - For $i = 1, \dots, n$, let $y_i \in \mathbb{R}^k$ be the vector corresponding to the i-th row of V
  - Cluster the points $(y_i)_{i=1,\dots,n} \in \mathbb{R}^k$ with the k-means algorithm into clusters $C_1, \dots, C_k$
- Output: Clusters $A_1, \dots, A_k$.

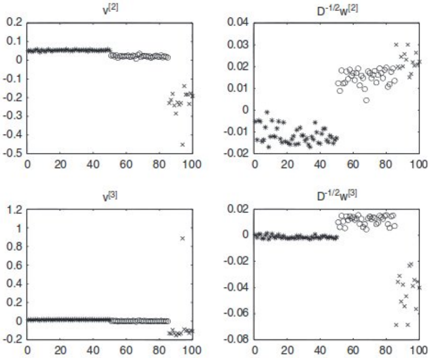
Experiments on simulated data

Higham et al., Spectral clustering and its use in bioinformatics. Journal of Computational and Applied Mathematics, 2007



Experiments on simulated data Cont'd

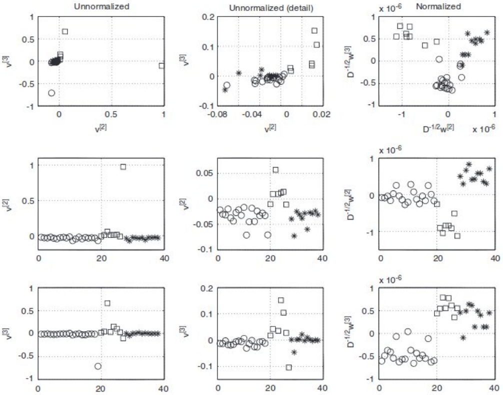
Higham et al., Spectral clustering and its use in bioinformatics. Journal of Computational and Applied Mathematics, 2007



Components of the second and third eigenvectors for the data. Left unnormalized. Right normalized.

Experiments on real data

Higham et al., Spectral clustering and its use in bioinformatics. Journal of Computational and Applied Mathematics, 2007



Leukaemia: ALL-B (circles), ALL-T (squares), AML (stars). Upper line: scatter plots of the second versus third eigenvectors. Middle line: components of the second singular vectors. Lower line: components of the third singular vectors.

Graph cut point of view

If data are given as a similarity graph, the problem can be restated

- we want to find a partition of the graph such that the edges between different groups have a very low weight
- and the edges within a group have high weights

For two disjoint subsets A and B, we define

$$cut(A, B) = \sum_{i \in A, j \in B} w_{ij}$$

.

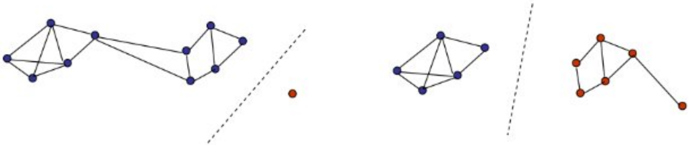
Given a similarity graph with adjacency matrix W, the simplest and most directed way to construct a partition is to solve the mincut problem: choose the partition $A_1, \ldots, A_k$ which minimizes

$$\sum_{i=1}^{K} cut(A_i, \bar{A}_i)$$

.

Graph cut point of view Cont'D

Sensitive to outliers!



What we get

What we want

Graph cut point of view Cont'D

Intuition: clusters should be reasonably large groups of points

- $RatioCut(A_1, \ldots, A_k) = \sum_{i=1}^{k} \frac{cut(A_i, \bar{A}_i)}{|A_i|}$
- $NCut(A_1, \ldots, A_k) = \sum_{i=1}^{k} \frac{cut(A_i, \bar{A}_i)}{vol(A_i)}$

- The problem is NP-hard.
- Approximate the solution

Random walks point of view

Spectral clustering can be explained via random walks.

- A random walk on a graph is a stochastic process which randomly jumps from vertex to vertex.
- A partition with a low cut will also have the property that the random walk does not have many opportunities to jump between clusters
- Formally, the transition probability of jumping in one step from vertex i to vertex j is proportional to the edge weight w_{ij}, and is given by $p_{ij} = w_{ij}/d_i$. The transition matrix $P = (p_{ij})$ of the random walk is thus defined by

$$P = D^{-1}W$$

Patients Stratification and Methods of Personalized Medicine

An application: Obesity stratification based on metagenomics

Some (Fancy) Clustering Methods

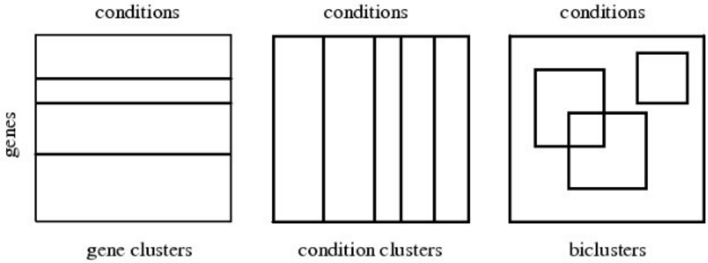
- Some widely used Clustering Methods
- Spectral Clustering
- Biclustering
- Robust Clustering

Semi-Supervised Learning

Canonical Correlation: Correlation between Sets of Variables

Biclustering

- Simultaneous clustering of both rows and columns of a data matrix
- Identifies groups of genes with similar/coherent expression patterns under a specific subset of conditions



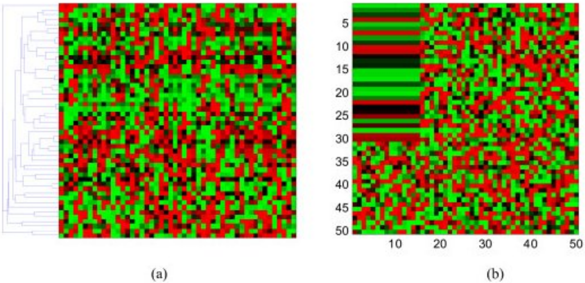
Why biclustering and not just clustering?

Biclustering is the key technique to use when

- Only a small number of the genes participates in a cellular process of interest
- An interesting cellular process is active only in a subset of the conditions
- A single gene may participate in multiple pathways that may or not be co-active under all conditions

Biclustering: motivation

Gan et al., Discovering biclusters in gene expression data based on high-dimensional linear geometries, BMC Bioinformatics 2008



(a)

(b)

An illustrative example where conventional clustering fails but biclustering works: (a) A data matrix, which appears random visually even after hierarchical clustering. (b) A hidden pattern embedded in the data would be uncovered if we permute the rows or columns appropriately.

The Cheng-Church Algorithm (2000)

The algorithm of Cheng and Church is

- a simple, greedy approach towards finding maximal sized biclusters satisfying a certain condition
- The input is a matrix $A = (a_{ij})$
- The rows represent genes
- The columns represent conditions

- The algorithm attempts to find a submatrix B, representing a bicluster.
- The quality of B as a bicluster is measures using the Residue score.

The Cheng-Church Algorithm Cont'd

The residue score of an element

- $a_{ij} = a_{ij} - a_{i,J} - a_{I,j} + a_{I,J}$.

The mean squared residue score for the sub-matrix $A_{I,J}$ is then

$$H(I,J) = \sum_{i \in I, j \in J} (a_{ij} - a_{i,J} - a_{I,j} + a_{I,J})^2 / (|I||J|)$$

.

- I and J are row and column subsets representing a sub-matrix
- $a_{I,j} = \sum_{i \in I}(a_{ij})/|I|$ (sub-matrix column j average)
- $a_{i,J} = \sum_{j \in J}(a_{ij})/|J|$ (sub-matrix column i average)
- $a_{I,J} = \sum_{i \in I, j \in J}(a_{ij})/(|I||J|)$ (the entire sub-matrix average)

The Cheng-Church Algorithm Cont'd

δ-biclusters

- The most natural goal would be to find a bicuster minimizing the mean squared residue score
- It is easy to see that the mean squared residue score is 0 iff the submatrix satisfies the assumption
- It would yield trivial (one gene and one condition) biclusters, and would in general prefer small biclusters
- Therefore, we define $A_{I,J}$ as a δ-bicluster if $H(I,J) \leq \delta$, and try to find larger biclusters

The Cheng-Church Algorithm Cont'd

Finding δ-biclusters

- Given A and δ, finding the largest δ-bicluster is NP-hard.

The Cheng-Church algorithm

- Employs a greedy heuristic for detecting a large bicluster
- It starts with a sub-matrix identical to the input matrix, and then proceeds with two phases
  - Iterative removal of rows/columns until $H(I, J) < \delta$
  - Iterative addition of rows/columns until no addition is possible without H exceeding δ
- The remaining sub-matrix will be declared a bicluster
- If the remaining sub-matrix is empty, then no δ-bicluster is found
- The removal of a row/column is done by choosing (in every iteration), the row/column which has the maximum contribution to the score H (in effect, the "worst" one)

The Cheng-Church Algorithm Cont'd

Finding more than one bicluster

- Note that the algorithm is completely deterministic: consecutive runs of the algorithm on the same matrix will yield the same bicluster
- To find other biclusters, the complete algorithm repeats the process after masking the bicluster found
- Masking is performed by filling the positions of the biscluster with random values
- The new random values will probably not form any recognizable pattern

The Cheng-Church Algorithm Cont'd

Shortcomings of the Cheng-Church algorithm

- The results are not assigned a statistical-significance value
- Since δ is constant, then given a large enough initial matrix, we are almost guaranteed to find a random bicluster, of arbitrary size satisfying the condition
- The greedy nature of this algorithm clearly does not guarantee the convergence to global optimal solutions
- The masking technique would seriously reduce the change to find biclusters with any overlap (these overlaps may be a natural result of a gene having more than one function)

Patients Stratification and Methods of Personalized Medicine

An application: Obesity stratification based on metagenomics

Some (Fancy) Clustering Methods

- Some widely used Clustering Methods
- Spectral Clustering
- Biclustering
- Robust Clustering

Semi-Supervised Learning

Canonical Correlation: Correlation between Sets of Variables

Partitioning around Medoids

PAM (Partitioning around Medoids) is a k-partitioning approach (Kaufman and Rousseeuw, 1990).

- The algorithm finds the representative object, medoid, which is the multidimensional version of the median
- Tries to minimize the total cost

$$\sum_{r} d(\hat{x}, x_r)$$

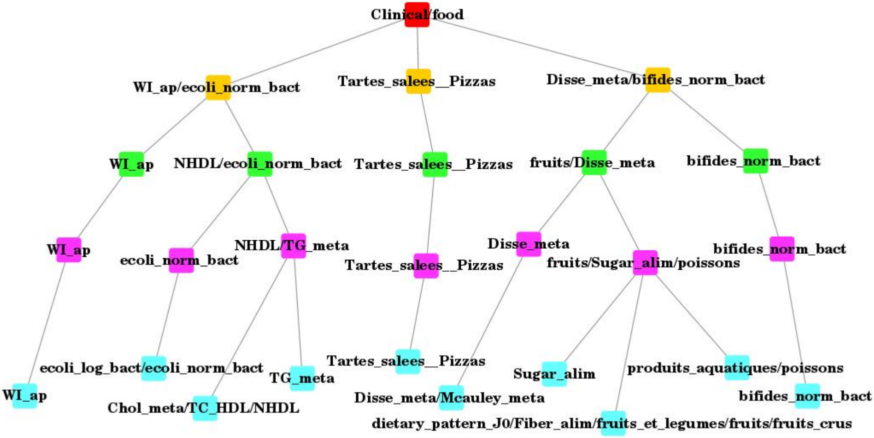
- PAM finds a local minimum for the objective function

PAM Cont'd

The PAM algorithm

```
1. Initialize: randomly select k of the n data points as the medoids
2. Associate each data point to the closest medoid. ("closest" here is defined using any valid distance metric, most commonly Euclidean distance, Manhattan distance or Minkowski distance)
for For each medoid m do
    for For each non-medoid data point o do
        3. Swap m and o and compute the total cost of the configuration
    end for
end for
4. Select the configuration with the lowest cost.
Repeat steps 2 to 4 until there is no change in the medoid.
```

A hierarchy of clinical parameters of MicrObese data



Patients Stratification and Methods of Personalized Medicine

An application: Obesity stratification based on metagenomics

Some (Fancy) Clustering Methods

Semi-Supervised Learning

Canonical Correlation: Correlation between Sets of Variables

Semi-Supervised Learning

- Traditionally: Unsupervised and supervised learning
- Semi-supervised learning: halfway between supervised and unsupervised learning
- Semi-supervised learning with constraints: “these points have (or do not have) the same target”
- A problem related to SSL was introduced by V.Vapnik: transductive learning: do prediction for the test points only

A Brief History of Semi-Supervised Learning

- Self-learning: the earliest idea of SSL
  - Use repeatedly a supervised method.
  - It starts by training on the labeled data only; then label unlabeled data, etc.

- Transductive inference
  - Vapnik's principle: When trying to solve some problem, one should not solve a more difficult problem as an intermediate step
  - No general decision rule is inferred
  - E.g. a combinatorial optimization on the labels of the test points in order to maximize the likelihood of their model

- Mixture of Gaussians
  - The likelihood of the model is maximized using the labeled and unlabeled data with the help of iterative algorithm such as Expectation-Maximization
  - Instead of mixture of Gaussians, use a mixture of multinomial distributions

A Brief History of Semi-Supervised Learning Cont'd

- Theoretical analysis
  - Learning rates exist for SSL of a mixture of two Gaussians: probability of error has an exponential convergence to the Bayes risk
- Text applications and natural language processing

When Can Semi-Supervised Learning Work?

In comparison with a supervised algorithm, can one hope to have a more accurate prediction by taking into account the unlabeled data?

- Prerequisite: the distribution of examples is relevant to the classification problem
- In a more mathematical formulation: the knowledge of $p(x)$ has to carry information that is useful in the inference of $p(y|x)$

When Can Semi-Supervised Learning Work?

The four assumptions:

- Smoothness assumption: If two points x_1 and x_2 are close, then so should be the corresponding y_1 and y_2
- Cluster assumption: If points are in the same cluster, they are likely to be of the same class
- Low density separation: The decision boundary should lie in a low-density region
- The (high-dimensional) data lie (roughly) on a low-dimensional manifold

Classes of Semi-Supervised Learning Algorithm

- Generative models
    - A generative model models $p(y, x)$, and any additional information on $p(x)$ is useful
    - It can be seen as classification with additional information on the marginal density
    - It can be seen as clustering with additional information
    - Advantage: Knowledge of the structure can be incorporated

Classes of Semi-Supervised Learning Algorithm

- Generative models
  - A generative model models $p(y, x)$, and any additional information on $p(x)$ is useful
  - It can be seen as classification with additional information on the marginal density
  - It can be seen as clustering with additional information
  - Advantage: Knowledge of the structure can be incorporated

- Low-density separation: an SVM
  - The most common approach – a maximum margin algorithm such as SVM
  - The method of maximizing the margin for unlabeled as well as labeled points is called the transduction SVM
  - The corresponding problem is non-convex, and thus difficult to optimize

- Low-density separation: entropy minimization
  - Encourage the class-conditional probability $p(y|x)$ to be close to 1 or to 0 at labeled and unlabeled points

Classes of Semi-Supervised Learning Algorithm

- Graph-based methods
  - Data are represented by the nodes of a graph, the edges of which are labeled with the pairwise distances of the incident nodes
  - Most graph methods use the graph Laplacian
  - Many graph methods penalize nonsmoothness along the edges
  - Intrinsically transductive and inductive algorithms
  - Information propagation on the graph

Classes of Semi-Supervised Learning Algorithm

- Graph-based methods
  - Data are represented by the nodes of a graph, the edges of which are labeled with the pairwise distances of the incident nodes
  - Most graph methods use the graph Laplacian
  - Many graph methods penalize nonsmoothness along the edges
  - Intrinsically transductive and inductive algorithms
  - Information propagation on the graph
- Change of Representation: two-step learning
  - Change representation: perform an unsupervised step on all data, and construct a new metric
  - Ignore the unlabeled data and perform supervised learning using the new data

Hypothesis and Notations

Notations:

- X_i observation
- Y_i label
- n the number of observation pairs
- $\pi(x, y)$ the joint probability
- $\eta(y|x)$ the conditional probability
- $q(x)$ the marginal probability of observations

The hypothesis:

- The marginal probability $q(x)$ is completely known
- $\mathcal{X}$ and $\mathcal{Y}$ are finite

Semi-Supervised Probabilistic Criterion

$\{X_i, Y_i\}_{i=1}^n$ are observations and their labels

Let $g(y|x;\theta)$ be the conditional probability function, parameterized by θ. Then the standard conditional maximum likelihood estimator is defined by

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i|X_i;\theta),$$

where $\ell(y|x;\theta) = -\log g(y|x;\theta)$ denotes the negated conditional log-likelihood function.

The asymptotically optimal semi-supervised estimator $\hat{\theta}_n^s$ is defined by

$$\hat{\theta}_n^s = \arg\min_{\theta\in\Theta} \sum_{i=1}^{n} \frac{q(X_i)}{\sum_{j=1}^{n} \mathbb{1}\{X_j = X_i\}} \ell(Y_i|X_i;\theta),$$

where $q(x)$ is the marginal probability of observations.

Problem of the Covariate Shift

Covariate Shift

Let us learn an estimator from $(X_1, Y_1), \ldots, (X_n, Y_n)$, where the distribution of X_i is defined by $q_0(x)$. How to adapt the estimator if the test data X_i are distributed according to $q_1(x) \neq q_0(x)$?

- Si q_1 is known, the weights of the semi-supervised estimateur($q = q_1$) are asymototically identical to $\frac{1}{n}\frac{q_1}{q_0}(X_i)$ and the algorithm converges to

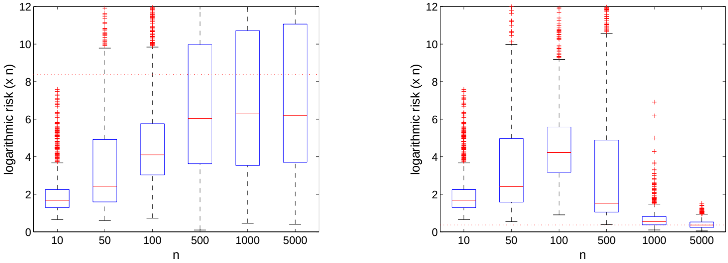
$$\theta_{1\star} = \arg\min_{\theta\in\Theta} \mathrm{E}_{\pi_1}[\ell(Y|X;\theta)] \,.$$

- The covariance matrix is smaller than the matrix of the estimator weighted by an importance ratio

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta} \sum_{i=1}^{n} \frac{q_1}{q_0}(X_i)\ell(Y_i|X_i;\theta)$$

(which is supposed to know q_0).

Experiments with logistic regression



Boxplots of the scaled excess risk as a function of the number of observations in the presence of the covariate shift.
Left: Shimodaira criterion, $n\,(\mathrm{E}_\pi[\ell(Y|X;\hat{\theta}_n)] - \mathrm{E}_\pi[\ell(Y|X;\theta_\star)])$;
Right: semi-supervised estimator, $n\,(\mathrm{E}_\pi[\ell(Y|X;\hat{\theta}_n^s)] - \mathrm{E}_\pi[\ell(Y|X;\theta_\star)])$.

Applications to real problems

In the realistic applications (binary text classification), we can not assume that the true $q(x)$ is known.

We propose an approach based on clustering.

How to "estimate $q(x)$"? The set of unlabeled data is divided into k clusters, and in the expression of the weight

$$\frac{q(X_i)}{\sum_{j=1}^{n} \mathbb{1}\{X_j = X_i\}}$$

the numerator is replaced by the empirical frequency of the cluster which contains X_i; the denominator is replaced by the number of training points which are in the same cluster as X_i.

Patients Stratification and Methods of Personalized Medicine

An application: Obesity stratification based on metagenomics

Some (Fancy) Clustering Methods

Semi-Supervised Learning

Canonical Correlation: Correlation between Sets of Variables

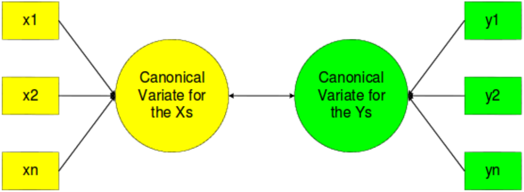
Canonical Correlation Analysis: Motivation

- Canonical correlations analysis (CCA) is an exploratory statistical method to highlight correlations between two data sets acquired on the same experimental units

- CCA is most appropriate when a researcher desires to examine the relationship between two variable set
- The method was first introduced by Harold Hotelling in 1936

Canonical Correlation Analysis: How?

- X and Y are matrices of order $n \times p$ and $n \times q$
- The columns correspond to variables and the rows correspond to experimental units (patients)



Canonical Correlation Analysis: How?

- Find two vectors *a* and *b* that maximize the correlation between the linear combinations

$$U = a_1 X^1 + a_2 X^2 + \cdots + a_p X^p$$
$$V = b_1 Y^1 + b_2 Y^2 + \cdots + b_q Y^q$$

- The problem consists in solving

$$\rho = cor(U, V) = \max_{a,b} cor(Xa, Yb)$$

Canonical Correlation Analysis: How?

- Find two vectors a and b that maximize the correlation between the linear combinations

$$U = a_1 X^1 + a_2 X^2 + \cdots + a_p X^p$$
$$V = b_1 Y^1 + b_2 Y^2 + \cdots + b_q Y^q$$

- The problem consists in solving

$$\rho = cor(U, V) = \max_{a,b} cor(Xa, Yb)$$

Canonical correlations ρ are the positive square roots of the eigenvalues λ of $P_X P_Y$ ($\rho = \sqrt{\lambda}$), where

$$P_X = X(X^T X)^{-1} X^T$$
$$P_Y = Y(Y^T Y)^{-1} Y^T$$

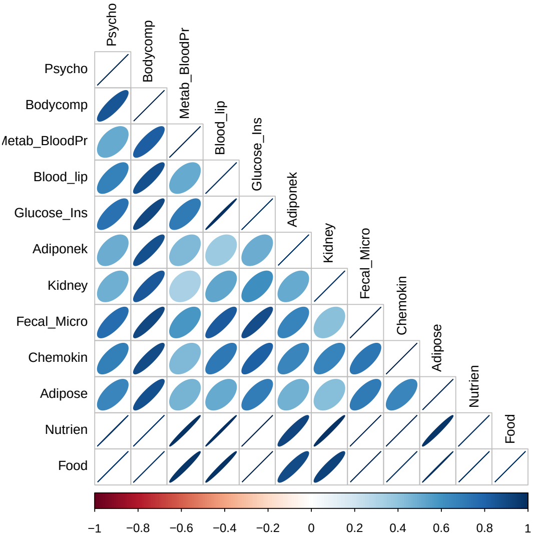
How to Interpret the Results?

- Consider canonical correlation values
  - The canonical correlation coefficient is the Pearson relationship between the two synthetic variables on a given canonical function. Because of the scaling created by the standardized weights in the linear equations, this value cannot be negative and only ranges from 0 to 1.
- Consider coefficients
  - Visualization of the results of canonical correlation is usually through bar plots of the coefficients of the two sets of variables for the pairs of canonical variates showing significant correlation.

Example: 12 sets of features

1. PA, psychological, and three factor eating questionnaires
2. Body composition
3. Metabolic rate and blood pressure
4. Blood lipids
5. Glucose homeostasis and insulin sensibility
6. Adiponekines
7. Kidney function
8. Fecal microbiota abundance, qPCR
9. Systemic inflammation and chemokines
10. Adipose tissue macrophage markers
11. Nutrient intake
12. Food intake

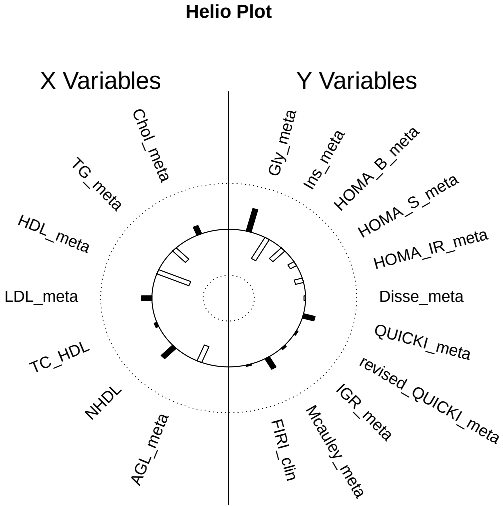
Canonical Correlation Values



How to interpret the results?

- Structure coefficients are critical for deciding what variables are useful for the model
- Bar plots of the coefficients of the two sets of variables for the pairs of canonical variates showing significant correlation.
- Coefficients increase in importance when the observed variables in the model increase in their correlation with each other

Canonical Correlation Example: Blood lipids/Glucose homeostasis and insulin sensibility


Helio Plot
X Variables
Y Variables
Chol_meta
TG_meta
HDL_meta
LDL_meta
TC_HDL
NHDL
AGL_meta
Gly_meta
Ins_meta
HOMA_B_meta
HOMA_S_meta
HOMA_IR_meta
Disse_meta
QUICKI_meta
revised_QUICKI_meta
IGR_meta
Mcauley_meta
FIRI_clin