MSc. thesis presentation

Tommi Suvitaival

3.9.2009
Title: Bayesian Two-Way Analysis of High-Dimensional Collinear Metabolomics Data
Instructor: MSc. Ilkka Huopaniemi
Supervisor: Prof. Samuel Kaski
Contents

- Introduction to analysis of high-throughput biological data
- The focus is in metabolomics and multi-way analysis
- A new method is proposed and applied to biological data
Bioinformatics

- Bioinformatics analyses observations from biological organisms
- Analysis is performed using computational and statistical methods
- Lines of bioinformatics study genome, gene activity, protein concentration and metabolite concentration.
- Aim at gaining new knowledge on functioning of the biological system
- Often motivated by an interest in finding an explanation to a disease
Metabolomics

- A line of bioinformatics studying concentrations of small molecules, metabolites
- Metabolite is a substrate or product of a biological process that is catalysed by proteins
- Lipids are a sub-group of metabolites
- Lipids take part in many important biological processes, such as cell signaling
- Changes in lipid concentrations are related to many metabolic diseases, such as diabetes
Experiment setup in bioinformatics

- High-throughput measurements produce observations from large numbers of features

- $n < p$ problem: less samples than features in the data

- Number of samples is low due to high financial and ethical costs

- In metabolomic data, one feature corresponds to concentration of one metabolite

- One sample is a vector of features measured from one patient on one occasion
A metabolomic data set (1)

Figure: An example data matrix, where patients have two treatments.
A metabolomic data set (2)

Figure: Simulated data. Can you identify treatment effects?
Traditional solutions

- ANOVA (analysis of variance): univariate method handling one feature at a time
- MANOVA (multivariate analysis of variance): multivariate but non-functioning for $n < p$ data
Bayesian method: justification

- To deal with the $n < p$ problem
- To estimate uncertainty of the model
- To bring prior knowledge into the model
Bayesian method: clustering and multi-way analysis

- Features are clustered according to similarity
- Common treatment effects for each cluster are estimated
Bayesian method vs. a traditional approach

- The proposed model includes all three steps
- Instead of performing the steps sequentially, they are done simultaneously within the model
Bayesian method: the plate graph

Figure: The plate graph
Finnish children were screened for type 1 diabetes. The children were monitored 1 to 4 times a year. Certain antibody levels in blood were measured. These antibodies are useful in indicating the onset of the disease. It is already too late to prevent the disease at the time the antibodies emerge.
Could be detected earlier from the metabolic profile?

Around 100 children took part in a more detailed study, where lipid profiles were measured from blood serum.

53 lipids were identified.

Only 54 patients were included in analysis due to missing time points.

The Bayesian method was used to find possible predictors of the disease.
Results with a lipidomic data set (1)

Cluster 1
18 lipids
cor=0.76

Cluster 2
8 lipids
cor=0.32

Cluster 3
5 lipids
cor=0.68

Cluster 4
4 lipids
cor=0.65

Cluster 5
15 lipids
cor=0.47

Cluster 6
3 lipids
cor=0.43

**Figure:** Estimated treatment effects of a two-way data set
Results with a lipidomic data set (2)

Figure: Estimated time and time-disease interaction effect of a time series data set
Results with simulated data

Figure: Estimated treatment effects as function of sample-size