# Two-way analysis of high-dimensional collinear data

Ilkka Huopaniemi[1] Tommi Suvitaival[1] Janne Nikkilä[1,2]
Matej Orešič[3] Samuel Kaski[1]

1 Department of Information and Computer Science, Helsinki University of Technology, Finland
2 Department of Basic Veterinary Sciences, (Division of Microbiology and Epidemiology), Faculty of Veterinary Medicine, University of Helsinki, Finland
3 VTT Technical Research Centre of Finland, Espoo, Finland

September 9, 2009

TEKNILLINEN KORKEAKOULU
HELSINKI UNIVERSITY OF TECHNOLOGY

# Outline

Introduction: Two-way experimental setups

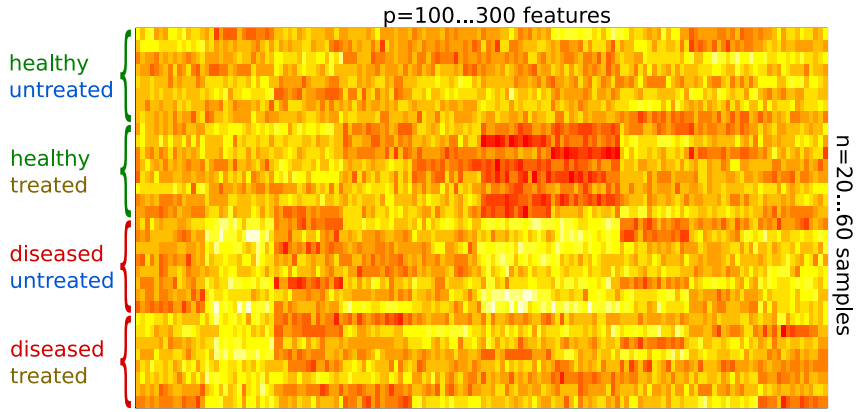
Bayesian Multi-way model
- Factor analysis
- Solving small n, large p problem by clustering approach
- Two-way covariate information as population priors

Results
- Performance with small sample size generated data
- Healthy-diseased, male-female metabolomics example
- Time-series metabolomics data

Conclusions

# Biological experiment with a two-way experimental setup

- Metabolomics: Diseases affect the concentrations of metabolites=chemical compounds
- Sample(metabolic profile) = concentrations of $> 100$ metabolites
- Estimate: Disease effect, treatment effect, interaction (cure?)
- Small $n$, large $p$
- Lot of such bioinformatics datasets exist



# Two-way experimental setups

- **One-way methods**: diseased-healthy comparison, only find biomarkers for disease
- **Additional covariates**: Medical treatment, gender, time-series, **multi-way**
- **Interaction effect** of covariates
- Find features with differential concentration, confidence intervals

# Challenges:

- **Small sample size, high dimensionality** ($n \ll p$), problems with traditional methods
- Noise, False positives
- **Collinearity**: **groups** of similarly behaving **correlated features** = interesting information, can help overcoming small sample-size problems

# Alternative methods

Univariate

- Pairwise t-tests
- **2-way ANOVA**

Problem: False discovery rate
Loss of information of similarly behaving features (collinearity)

**Multivariate** methods

- **MANOVA**, Linear discriminant analysis
- Problem: $n \ll p$, singular covariance matrix

**Dimension reduction**

- PCA + 2-way ANOVA
- PCA +2-way MANOVA
- ASCA, 5050 MANOVA

Supervised 1-way learning approaches in **small** $n$**, large** $p$:

- Partial least squares
- Covariance matrix regularization, sparsity
- Bayesian factor regression models

# Alternative methods

Univariate

- Pairwise t-tests
- **2-way ANOVA**

Problem: False discovery rate
Loss of information of similarly behaving features (collinearity)

**Multivariate** methods

- **MANOVA**, Linear discriminant analysis
- Problem: $n \ll p$, singular covariance matrix

**Dimension reduction**

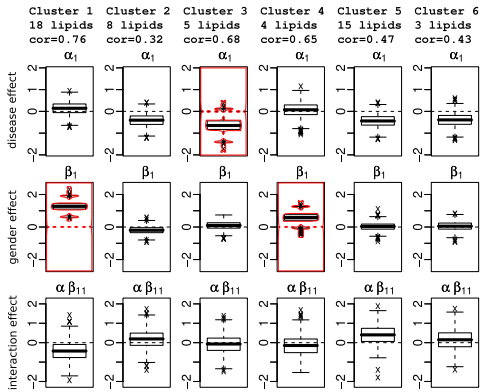
- PCA + 2-way ANOVA
- PCA +2-way MANOVA
- ASCA, 5050 MANOVA

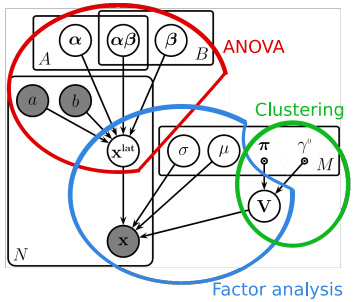
Supervised 1-way learning approaches in **small** $n$**, large** $p$:

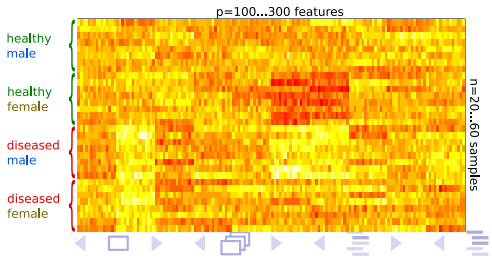
- Partial least squares
- Covariance matrix regularization, sparsity
- Bayesian factor regression models

# Task: Search statistically significant covariate-related differences in population means of features

- **Prior knowledge: Features form strongly correlated groups** (metabolites are highly correlated due to the existence of biochemical networks)
- Dimensionality reduction by clustering similarly behaving features
- Can now deal with **small $n$, large $p$**
- Solution: Fit an **ANOVA model for each correlated group** and study statistical significance of the terms

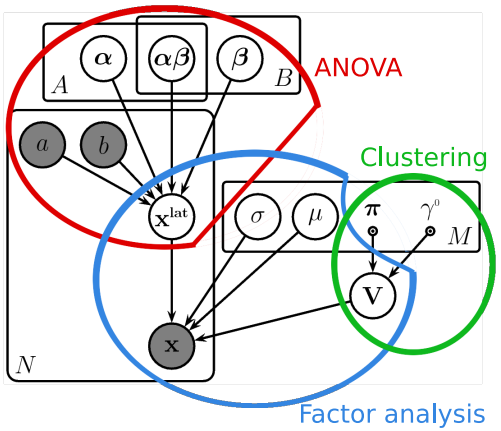


# Bayesian hierarchical model

- Extend **factor analysis**
- Small $n$, large $p$: Non-singular projection matrix by **clustering** features
- **Population-specific priors** $\alpha$, $\beta$, $\alpha\beta$ for covariate effects
- Solved by Gibbs sampling
- **Generative** modelling, remains unsupervised learning



# Classical Factor analysis

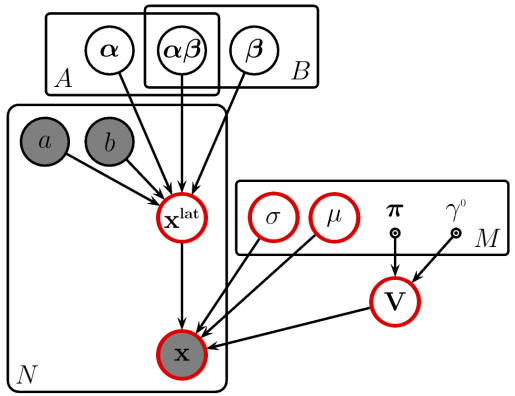
Models correlations in multivariate space, a set of features strongly correlated with each factor, closely related to PCA

$$\mathbf{x}_j^{lat} \sim \mathcal{N}(0, \mathbf{I})$$
$$\mathbf{x}_j \sim \mathcal{N}(\boldsymbol{\mu} + \mathbf{V}\mathbf{x}_j^{lat}, \boldsymbol{\Psi})$$

$\mathbf{x}_j$ Data vector (high-dimensional)
$\mathbf{x}_j^{lat}$ Latent variable (low-dimensional, ($< 10$) factors)
$\mathbf{V}$ Projection matrix, $\boldsymbol{\Psi}$ Noise

Problems:

- $\mathbf{V}$ not solvable if $n < p$
- Covariate information not taken into account



# Making the projection matrix invertible by clustering

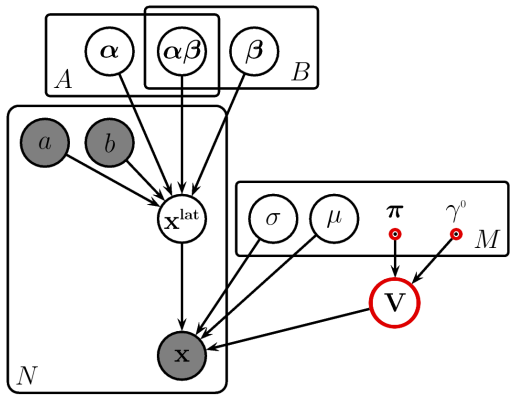
Cluster index of the feature $v_i$, non-zero element in $\mathbf{V}$

$$p(v_i = k) = \frac{\pi_k \prod_j p(x_{ji}|\mu_i + \gamma_i x_{jk}^{lat}, \sigma_i)}{\Sigma_k \pi_k \prod_j p(x_{ji}|\mu_i + \gamma_i x_{jk}^{lat}, \sigma_i)}$$

$$\mathbf{V} = \begin{bmatrix} \gamma_1 & 0 & 0 \\ 0 & 0 & \gamma_2 \\ \vdots & \vdots & \vdots \\ 0 & \gamma_j & 0 \\ 0 & \gamma_{j+1} & 0 \\ \vdots & \vdots & \vdots \end{bmatrix}$$

A correlated group of features now belongs exactly to one factor

- Number of clusters by predictive likelihood
- Features(metabolites) have different scales $\rightarrow$ Scale parameter $\gamma_i^2 \sim \text{Inv-}\chi^2$



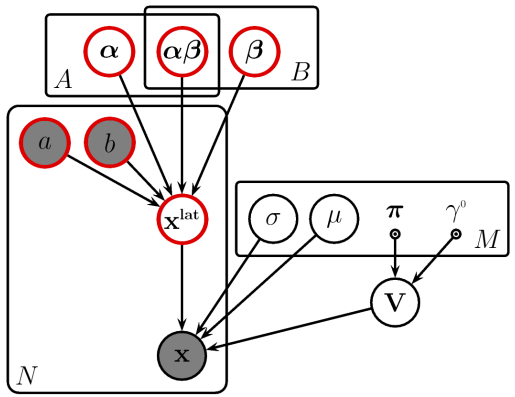
# Two-way linear model on the latent variables

Terms $\boldsymbol{\alpha}_a$, $\boldsymbol{\beta}_b$, $(\boldsymbol{\alpha\beta})_{ab} \sim \mathcal{N}(0, \mathbf{I})$ estimate the effects of **disease, treatment** and their **interaction** for **each cluster** of features.

$$\mathbf{x}_j^{lat}|_{\text{Control}} \sim \mathcal{N}(0, \mathbf{I}) \qquad \mathbf{x}_j^{lat}|_{\text{Diseased}} \sim \mathcal{N}(\boldsymbol{\alpha}_1, \mathbf{I})$$
$$\mathbf{x}_j^{lat}|_{\text{Treated}} \sim \mathcal{N}(\boldsymbol{\beta}_1, \mathbf{I}) \qquad \mathbf{x}_j^{lat}|_{\text{Diseased treated}} \sim \mathcal{N}(\boldsymbol{\alpha}_1 + \boldsymbol{\beta}_1 + (\boldsymbol{\alpha\beta})_{11}, \mathbf{I})$$

*a*(healthy,diseased) and *b*(treated,untreated) are the observed covariates

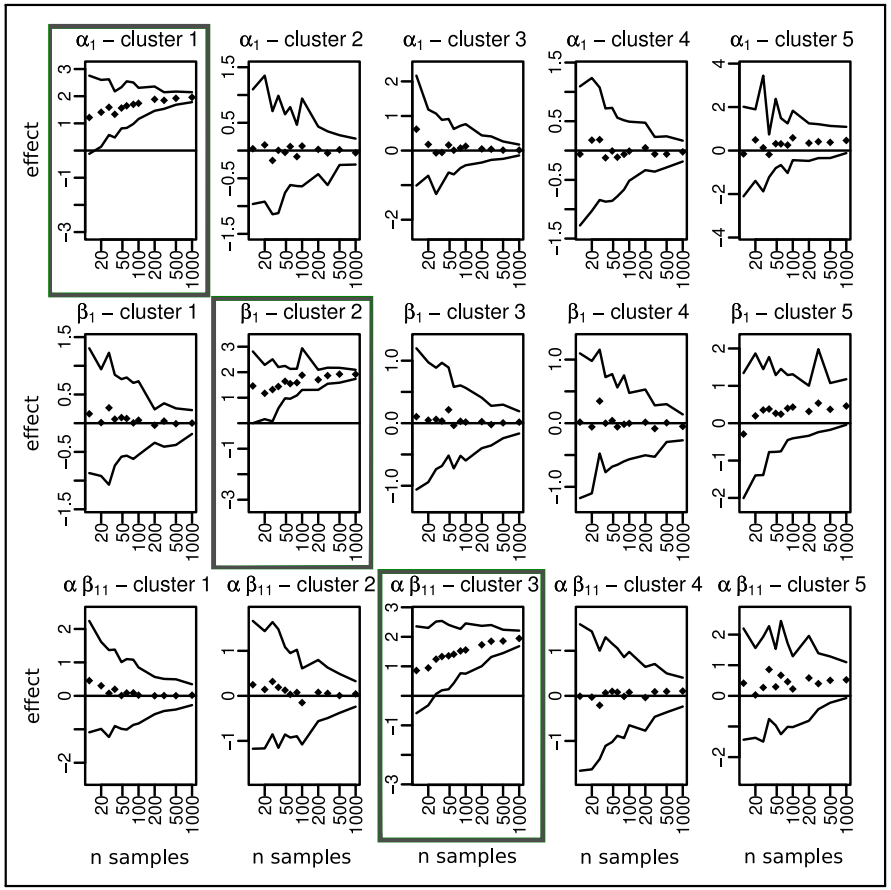
$$\mathbf{x}_j \sim \mathcal{N}(\boldsymbol{\mu} + \mathbf{V}\mathbf{x}_j^{lat}, \boldsymbol{\Psi})$$




Data centered at a control group mean with $\boldsymbol{\mu}$ (e.g. Healthy untreated)

# Generated data

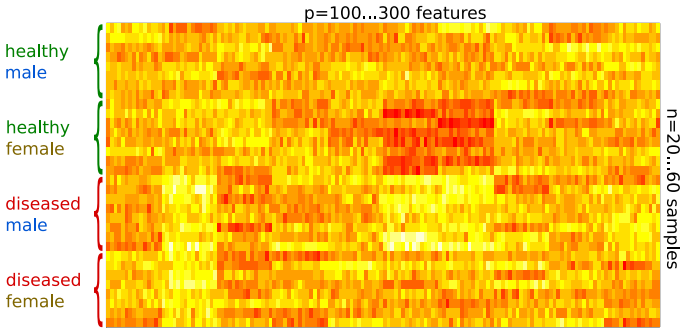
- 200-dimensional data
- Posterior of effects as a function of sample-size
- The generated **effects** are **found** already with **small sample-size**
- No significant false positives

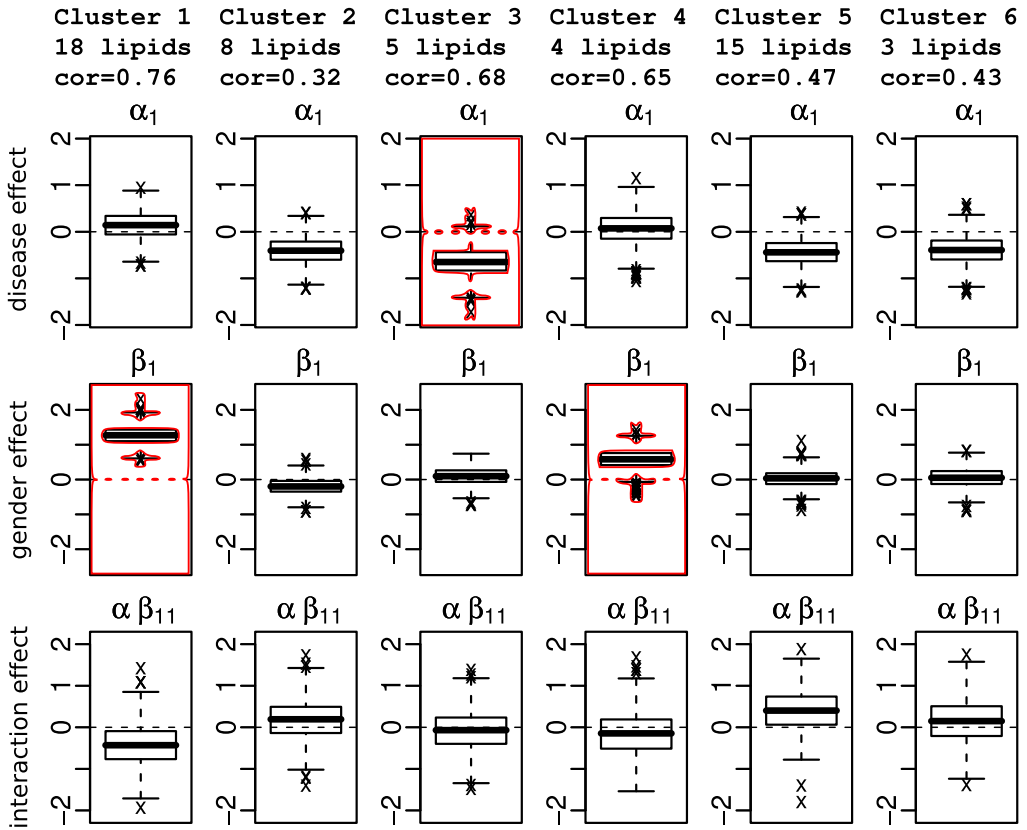



A posterior distribution above(below) zero implies a statistically significant positive(negative) effect

# Healthy-diseased, male-female lipidomics example

All the relevant information of the dataset **Inferred from posterior** of $\alpha_a$, $\beta_b$, $(\alpha\beta)_{ab}$

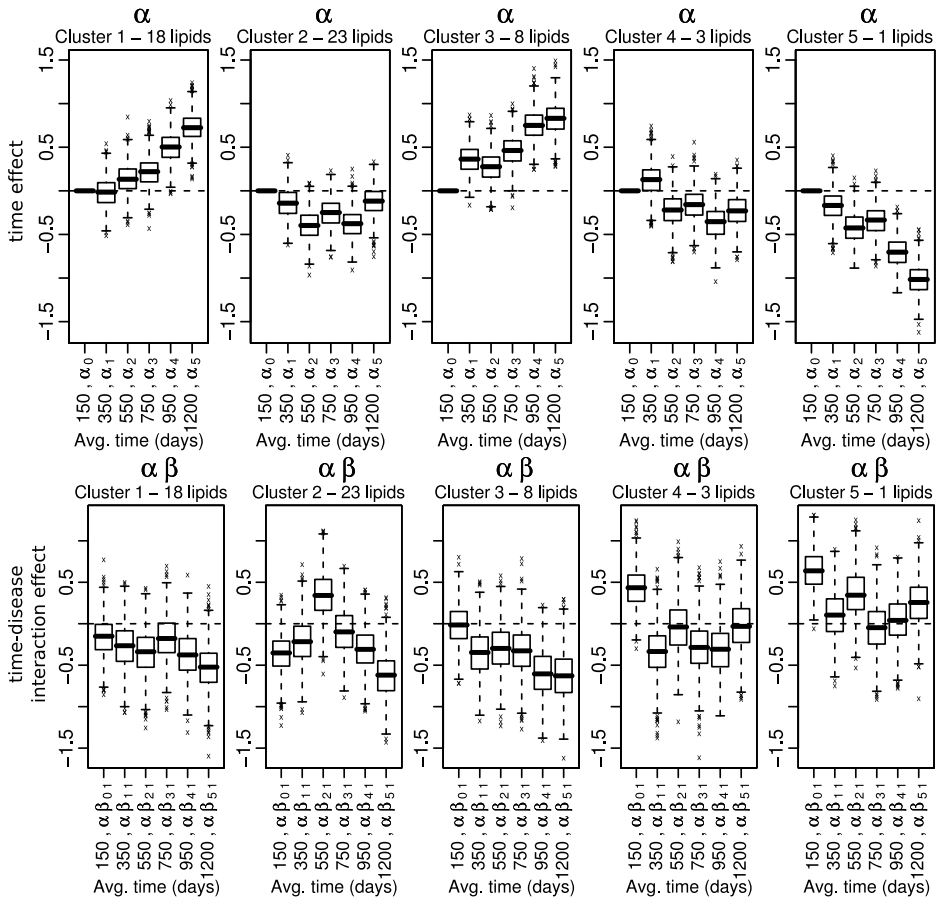



A posterior distribution above(below) zero implies a statistically significant positive(negative) effect

# Healthy-diseased, time-series metabolomics data

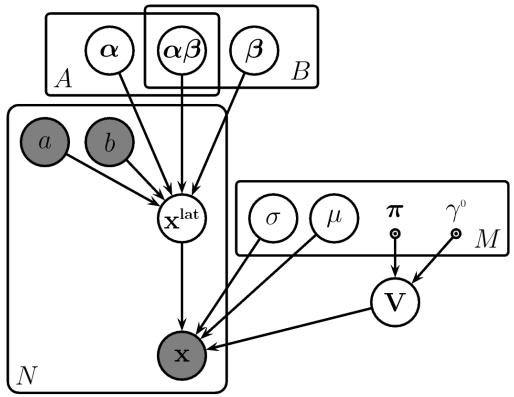
- Common time-behavior
- Interaction of time and disease




A posterior distribution above(below) zero implies a statistically significant positive(negative) effect

# Conclusions

- **Bayesian machine learning** approach for multivariate **two(multi)-way analysis**



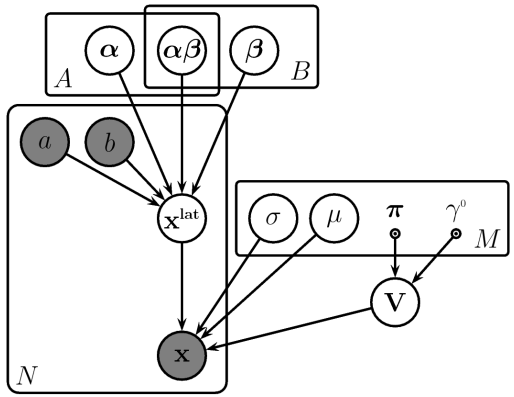

  - The methods **clusters correlated features** and estimates a **two-way ANOVA-model** for each cluster
  - The covariate effects are successfully found even with **small sample-sizes with high dimensionality**
  - Generalizes to other data types with **correlated groups of features** (common in bioinformatics)

# Acknowledgements


- Tekes(Finnish Technology Agency) MASI program
- AIRC (Adaptive Informatics Research Centre, TKK)
- HIIT (Helsinki Institute for Information Technology)
- Graduate School of Computer Science and Engineering, TKK
- EU FP7 NoE PASCAL2, ICT 216886

# Thank you, questions?

- **Bayesian machine learning** approach for multivariate **two(multi)-way analysis**




  - The methods **clusters correlated features** and estimates a **two-way ANOVA-model** for each cluster
  - The covariate effects are successfully found even with **small sample-sizes with high dimensionality**
  - Generalizes to other data types with **correlated groups of features** (common in bioinformatics)

http://www.cis.hut.fi/projects/mi/
ilkka.huopaniemi@tkk.fi