Detecting similar high-dimensional responses to experimental factors from human and model organism

Tommi Suvitaival 1  Ilkka Huopaniemi 1,* Matej Orešič 2
Samuel Kaski 1,3

1Helsinki Institute for Information Technology HIIT, Aalto University School of Science
firstname.lastname@aalto.fi
research.ics.tkk.fi/mi

* Currently at
Mount Sinai School of Medicine

2VTT Technical Research Centre of Finland
firstname.lastname@vtt.fi
sysbio.vtt.fi

3Helsinki Institute for Information Technology HIIT, University of Helsinki

December 16, 2011 - NIPS PM
Introduction

- data fusion from multi-species experiments
  - no paired samples/variables
  - hierarchical Bayesian model

- responses to experimental factors
Motivation

Human studies → Identification of similar responses → Predictive models for human patients

Identification of similar responses → Model organism studies

Predictive models for human patients → Improved predictive models for human patients
Introduction: Experimental design

- samples of the data set divided into categories by experimental factors (covariates)
  - known/partially known/unknown
Introduction: Multi-way decomposition

- covariate effect tells how much the mean of the population deviates from the baseline
for univariate data, covariate effects are traditionally estimated by analysis of variance (ANOVA):

\[ x_{(a,b)} = \mu + \alpha_a + \beta_b + (\alpha \beta)_{ab} + \varepsilon \]
Introduction: Univariate two-way decomposition (2)

<table>
<thead>
<tr>
<th>covariates</th>
<th>data space:</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>100...300 metabolites</td>
</tr>
<tr>
<td>$b$</td>
<td></td>
</tr>
</tbody>
</table>

- healthy
  - untreated
  - treated
- diseased
  - untreated
  - treated

\[
x_{(a,b)} = \mu + \alpha_a + \beta_b + (\alpha\beta)_{ab} + \varepsilon
\]

- $x$: observation
- $\mu$: grand mean
- $\varepsilon$: residual
- $a$, $b$: covariate realizations
- $\alpha_a$, $\beta_b$: main effects of covariate realizations $a$ and $b$, respectively
- $(\alpha\beta)_{ab}$: interaction effect of covariate realizations $a$ and $b$
Bayesian model for high-dimensional multi-way data\textsuperscript{a}

- dimensionality reduction by clustering variables
- multi-way modeling by category-specific latent variables ($\alpha$, $\beta$ and ($\alpha \beta$)) that generate the sample-specific latent variables ($x^{\text{lat}}$)

\textsuperscript{a}Two-Way Analysis of High-Dimensional Collinear Data, Huopaniemi et al., ECML'09
Alignment of unknown time covariate\textsuperscript{b}

- the time series are aligned by a hidden Markov model (HMM)
- samples are assigned to a category ("developmental states") according to the HMM alignment
- also additional information to division diseased patients into several stages of the disease (levels of covariate \textsuperscript{b})

\textsuperscript{b}Graphical Multi-Way Models, Huopaniemi et al., ECML’10
One data set: Starting point

- high-dimensional data
- additional multi-level categorical information about each sample (covariates)
- time series
One data set: Clustering

Organism X

covariate $b$

- healthy
- diseased

data space $X$

time series ( ): varying lengths, unknown alignments

- cluster 1
- cluster 2
- cluster 3

- assign collinear variables into clusters
One data set: Latent variables

Organism X

covariate $b$

healthy

diseased

data space $X$

time series ( ): varying lengths, unknown alignments

▶ represent a cluster of variables by a latent variable
One data set: Multi-way decomposition

Organism X

covariate $b$

data space $X$

healthy

diseased

time series ( ): varying lengths, unknown alignments

- explain the dependency of the latent variables on covariates (categorical information) with covariate effects
- covariate effect: how much is a cluster moved from the baseline due to a specific realization of the covariates
Generalization to two data sets (1)

- same \textit{multi-way} experiment design
  - similar disease in both species
  - time series of observations from healthy/diseased patients
Generalization to two data sets (2)

nothing in common except the covariate structure
Research question

- identify responses to the covariates (disease, time)
  - shared across the two species, and
  - specific to either species

Organism X

- no matched variables, different dimensionalities
- data space X
- covariate b
- healthy
- diseased
- no paired samples

Organism Y

- data space Y
covariate b
- healthy
- diseased
time series ( ):
- varying lengths, unknown alignments
Multi-way decomposition of two data sets

Organism X

- Data space X
- Covariate b
- Healthy
- Diseased

Organism Y

- Data space Y
- Covariate b
- Healthy
- Diseased

No matched variables, different dimensionalities

Time effect: a = 1 2 3 4 5

Disease effect: b = 1 2 3 4 5

Time series (varying lengths, unknown alignments)
Matching

Organism X

no matched variables, different dimensionalities

data space

covariate b

healthy

diseased

no paired samples

time effect

disease effect

matching clusters based on their profiles

Organism Y

data space

covariate b

healthy

diseased

time series (varying lengths, unknown alignments)

healthy

diseased

covariate b

data space

matching clusters based on their profiles
Experiments

1. toy data
2. matching of lipid groups between two subsets of samples
3. matching of lipid and metabolite groups
Experiments: Toy data

Generated responses

Estimated main response $\alpha$:
HMM-aligned experimental factor $a$

Estimated interaction response $(\alpha\beta)$:
Interaction of HMM-aligned experimental factor $a$ and experimental factor $b$
Experiments: Matching of lipid groups, ground truth

**Match 1:**
triglycerides

**Match 2:**
glycerophosphocholines

**Match 3:**
glycerophosphocholines

▶ two subsets of samples from one data set were aligned
▶ results in line with the ground truth
Experiments: Matching of lipid and metabolite groups

Table: The best-matched pair of a lipid and a metabolite cluster.

<table>
<thead>
<tr>
<th>Lipids</th>
<th>Metabolites</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPCho(14:0/18:2)</td>
<td>X4.7.10.13.16.19.Docosahexaenoic.acid</td>
</tr>
<tr>
<td>GPCho(18:2/16:1)</td>
<td>X9.Octadecenoic.acid..Z.</td>
</tr>
<tr>
<td>GPCho(16:0/20:5)</td>
<td>Hexadecanoic.acid</td>
</tr>
<tr>
<td></td>
<td>Phosphoric.acid</td>
</tr>
</tbody>
</table>

- three first of the listed metabolites are fatty acids
- fatty acids are building blocks for glycerophosphocholine (GPCho) lipids
Conclusions (1)

- we can decompose data sets into shared and specific responses
- the decomposition is done by finding responses to covariates
- only the covariate structure is assumed to be similar in the data sets

Detecting similar high-dimensional responses to experimental factors from human and model organism

Acknowledgements. T.S., I.H. and S.K belong to the Finnish Centre of Excellence in Adaptive Informatics Research. The work was funded by Tekes MASI program and by Tekes Multibio project. I.H. is funded by the Graduate School of Computer Science and Engineering. S.K. is partially supported by EU FP7 NoE PASCAL2, ICT 216886.
Conclusions (2)

- cross-species studies
  - no 1-to-1 matching of observations
- studies from different hospitals
  - disjoint sets of patients
  - different measurement technologies

Detecting similar high-dimensional responses to experimental factors from human and model organism

Acknowledgements. T.S., I.H. and S.K belong to the Finnish Centre of Excellence in Adaptive Informatics Research. The work was funded by Tekes MASI program and by Tekes Multibio project. I.H. is funded by the Graduate School of Computer Science and Engineering. S.K. is partially supported by EU FP7 NoE PASCAL2, ICT 216886.