Advances in Neuroinformatics

Aapo Hyvärinen
Neuroinformatics Team

Mission:
- Develop statistical data analysis methods, with focus on
  - Unsupervised machine learning methods
  - Neuroscience applications
- Non-Gaussianity a central theoretical framework

Members:
- Aapo Hyvärinen, leader
- Patrik Hoyer, co-leader (until 8/2013, started own company)
- 2-4 postdocs, 2-4 PhD students
- From 2012, partly in CoE of Inverse Problems Research
Highlight 1: Causal analysis

- Passive observation vs. interventions
  - Completely passively observed data (our LiNGAM from 2006)
  - Experiment with (optimal?) interventions
    (Hyttinen, Eberhardt, Hoyer, *JMLR*, 2012, 2013a)

- Causality in fMRI, jointly with Stephen Smith
  - Oxford Centre for Functional Imaging of the Human Brain
  - Developer of simulated data for comparing algorithms
  - Our tailor-made methods (*JMLR*, 2013b)
    - Have best performance on simulated data
    - Are particularly simple variants of LiNGAM

![Graph showing causal analysis results]
Highlight 2: Testing independent components

- In independent component analysis, testing almost inexistent
  - Components could be local minima, or random effects

- We developed a method which uses a proper null hypothesis and the theory of classical hypothesis testing
  - Do ICA on multiple datasets (e.g. subjects), and see if you get the same component in more than one data set
  - Applications in MEG
    (NeuroImage, 2011):

- Application on fMRI needed further theory (Frontiers in Human Neuroscience, 2013)
Highlight 3: Practical brain imaging data analysis

- Decoding brain state from MEG (*NeuroImage*, 2013)
  - Optimal combination of ICA with classification methods
  - Must use nonlinear classification
- Two-person neuroscience: measuring interacting subjects
  - Riitta Hari's ERC AdG for constructing a system of two MEG scanners with video connection
  - Extremely challenging, still ongoing
- Analysing nonstationary dynamics
  - Result of sabbatical at ATR, Japan, in 2013, a leading centre in brain imaging
Future

- Co-leader Patrik Hoyer left academia
  - Group size reduced
  - Causal analysis given less emphasis
- New planned project: Modelling spontaneous brain activity
  - Very popular topic in brain imaging
  - But: our approach is to model the computations happening in the brain
    - Theoretical neuroscience instead of brain imaging
Data mining: theory and applications

Aristides Gionis

18 March, 2014
2011 vs. 2013

prof. Heikki Mannila
prof. Panagiotis Papapetrou
Dr. Kai Puolamäki
prof. Aristides Gionis
Dr. Nikolaj Tatti
Dr. Michael Mathioudakis
Dr. Jeffrey Lijffijt
Dr. Esa Junntila
Dr. Markus Ojala
Dr. Niko Vuokko
Dr. Sami Hanhijärvi

Academy of Finland
Stockholm U
KU Leuven
U of Toronto
Yahoo! Research
Research labs / Industry
University
research activities

• foundations in pattern discovery
  • statistical significance of patterns

• sequence analysis
  • episodes, segmentation, surprising events

• applications
  • biology, paleontology, linguistics, ...
selected publication venues  
(2012-2014)

• TODS 2014
• 2 x DMKD 2014
• 3 x DMKD 2013
• 2 x ECML PKDD 2013
• ACM Transactions on Applied Perception 2013
• Proceedings of the Royal Society B 2012
• International Journal of Data Mining and Bioinformatics 2012
• VLDB 2012
research highlights
comparison and exploration of event sequences

- Jefrey Lijffijt, PhD dissertation, Dec 2013
  - best doctoral dissertation in the Aalto school of Science in 2013

- data: event sequences
  - DNA, texts, sensor readings

- problems:
  - are two data sets equivalent with respect to pattern X?
  - are there parts of the data different from the whole?
  - which set of granularities to use when looking for patterns?
are there parts of the data that are different?

- **multiple testing**

- **challenge**: provide accurate correction without randomization/simulation

- **computational question**: given a Bernoulli process that runs for \( n \) steps, what is the probability that in any subsequence of length \( m \), there are \( k \) or more events?

- **thesis introduces upper-bound that works well in practice**
finding informative window lengths

- [Lijffijt, Papapetrou, Puolamäki, PKDD 2012]
- many sequence algorithms use *sliding windows*
- how to *choose* window lengths?
- treat as an *optimization problem*
- pick a set of window lengths that *explains most of the variability* in statistics over all possible window lengths
fast sequence segmentation using log-linear models

• [Tatti, DMKD 2013]

(a) Sequence

(b) Cost of segmentation

(a) speedup vs. sequence length (b) speedup vs. # of segments

Data mining: theory and applications
future directions
new research directions

• graph mining and social network analysis
• analysis of information networks
• analysis of evolving networks
• smart cities
Recent paper

- given a graph with weights on the nodes, find dense and heavy subgraph
- solutions using submodular function maximization and semidefinite programming
- applications in finding events in cities

(a) 01.06.12 Primavera sound music festival
(b) 18.09.12 festival of the Poblenou neighborhood
(c) 31.10.12 Halloween
Regression models for data streams with missing values

Indrė Žliobaitė
Postdoctoral Researcher
Problem setting

- Predictive modelling for streaming data
  - Data arrives and needs to be mined in real time
  - Real valued inputs, real valued target variable
  - Linear regression models
Examples of streaming data

- Sensor data (monitoring)
- Transactional data (events)
- Web data (user generated content)
Predictive modelling for streaming data
- data arrives and needs to be mined in real time
- real valued inputs, real valued target variable
- linear regression models

During operation predictive models can be regularly updated with recent data

**Problem:** massively missing input data, while predictions are needed continuously

**Our approach:** make predictive models robust to missing data, use simple mean imputation
Possible solutions

Case deletion → :( No predictions

Models on subspaces → :( Computationally infeasible, 2r models

For making predictions:

Imputing missing values

Single imputation → :( Biased estimates

Model based imputation → :( Computationally infeasible
Predictions by different linear models

What makes a predictive model robust to missing data?
Analysis of the expected error

Expected MSE of a linear model

\[ E[MSE_p] = (1 - p)E[MSE_0] + p - p(1 - p)\beta^T (C - I)\beta \]

- MSE grows linearly with number of missing inputs
- Quadratically
- Deterioration Index D

If D = 0 inputs are treated as independent

We can make use of dependency in inputs to ensure sub-linear MSE growth
Illustrative example

- Data: $x_1 = x_2 = x_3 = x_4 = y \sim N(0,1)$

**Independent:** $y = x_1$

**PCA:** $y = 0.25x_1 + 0.25x_2 + 0.25x_3 + 0.25x_4$

**Overfitted:** $y = 2x_1 - 1.5x_2 + x_3 - 0.5x_4$
Theoretically optimal model

\[ \hat{\beta}_{ROB} = \left( (1 - p)X^T X + pnI \right)^{-1} X^T y \]

prior probability of missing values

minimizes MSE given prior probability of missing values

\[ \hat{\beta}_{RR} = (X^T X + \lambda I)^{-1} X^T y \]

is similar to regularized regression
Illustrative example

\[ \text{theoretical } MSE_p \]

\[ \text{probability of missing } (p) \]

- **Overfitted**
- **Independent**
- **PCA**
- **ROB**
- **naive**

\[\text{Independent: } y = x_1\]
\[\text{PCA: } y = 0.25x_1 + 0.25x_2 + 0.25x_3 + 0.25x_4\]
\[\text{Overfitted: } y = 2x_1 - 1.5x_2 + x_3 - 0.5x_4\]
\[\text{ROB regression: different model for each value of } p\]
Addressing data stream challenges

- Data evolves over time
  - not only data distribution
  - but also *how* data is missing

![Graph showing probability of missing data over time](image-url)
Online adaptive ROB algorithm

new observation $x$ arrives, predict $\hat{y} = x\beta$

true target value $y$ arrives

update missing value estimate $p \leftarrow \gamma \frac{m}{r} + (1 - \gamma)p$

If no missing values

update covariance estimate and model

$S_t = S_{t-1} + xx^T - p(xx^T - I)$

$\beta_t = \beta_{t-1} + S_t^{-1}(y - x^T\beta_{t-1}) - S_t^{-1}p(xx^T - I)\beta_{t-1}$
Summary

- We developed
  - an optimization criteria (MSE) for regression being robust to massively missing data
  - a corresponding regression model
  - an algorithm for online operation on streaming data (recursive updates)
Modeling inter-linguistic relationships and language evolution

Roman Yangarber
Algodan
March 2014

University of Helsinki, Finland
Uralic Language Family

- Fin-Ugr
- Samoyed
- Uralic

Branches:
- Baltic
  - Finnish
  - Estonian
- Fin-Ugr
  - Perm'
  - Volga
  - Ob'
  - Hungarian
  - Hanty
  - Komi
  - Mansi
  - Udmurt

Note: The diagram shows the relationships between different language groups within the Uralic family.
Uralic Language Family

Uralic tree
Data sources

Data is arranged in *Cognate Sets*: set of genetically-related words, from different languages in the language family

→ ... Raw data sample
→ ... Aligned data sample
Central Principle

- **Regularity of sound change:**
  - Sound change is conditioned only on its phonetic environment, not on any other factor.
  - Sound change is deterministically conditioned

- **NB:** different from, e.g., biological sequence alignment, where mutations are sporadic.
Example sound change: German vs. Germanic

<table>
<thead>
<tr>
<th>Germanic t</th>
<th>English</th>
<th>German</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>two</td>
<td>zwei</td>
<td>zwei</td>
<td>stone</td>
<td>Stein</td>
</tr>
<tr>
<td>ten</td>
<td>zehn</td>
<td>zu</td>
<td>star</td>
<td>Stern</td>
</tr>
<tr>
<td>to</td>
<td>zähle-n</td>
<td>zu</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tell</td>
<td>Zahn</td>
<td>Zähre</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tooth</td>
<td>ziehe-n</td>
<td>Zähre</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tear</td>
<td>Zagel</td>
<td>Herz</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tow</td>
<td>Zagel</td>
<td>Herz</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tail</td>
<td>Zagel</td>
<td>Herz</td>
<td></td>
<td></td>
</tr>
<tr>
<td>heart</td>
<td>Zagel</td>
<td>Herz</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tail</td>
<td>Zagel</td>
<td>Herz</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>tip</td>
<td>Zipf-el</td>
<td>...</td>
<td>dead</td>
<td>tot</td>
</tr>
<tr>
<td>tide</td>
<td>Zeit</td>
<td>...</td>
<td>door</td>
<td>Türk</td>
</tr>
<tr>
<td>timber</td>
<td>Zimmer</td>
<td>...</td>
<td>do</td>
<td>tu-n</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>under</td>
<td>unter</td>
</tr>
</tbody>
</table>
Example sound change: German vs. Germanic

- There are “exceptions” to rules
  - “regular” exceptions?
  - rare/occasional exceptions?
  → probabilistic modeling
    → MDL
    code most of the data with rules, then code the exceptions.
Principal Tasks

Long-term goal: Determine the origin of everything

- Find cognate sets (from raw language data)
  - difficult to model semantics...
- Find sound-by-sound alignment of all related words
- Find rules of sound correspondence

- Reconstruct philogenetic trees
- Reconstruct proto-forms → at root and internal nodes of the philogeny

- Model borrowing across languages / families
- Model timing → anchor data on absolute time scale
Components

Model

Alignment
Pairwise distances

Rules

Imputation

Data

Gold-standard philogenetic tree(s)

1-1 n-n N-D
Context-based

Projection + pop-genetics model

Normalized compression distance

Normalized edit distance

UPGMA NeighborJoin CompLearn NeighborNet

Philogenetic trees & networks
Problem formulation

Dual problem:
A find the *globally best alignment* for the complete data, and
B find the *rules of correspondence*

Chicken and egg...

Approach *in tandem*
Baseline: Initial simplifications

- **Pairwise** alignment: only two languages at a time, “source:target”
  - $\rightarrow$ N-dimensional alignment, $N > 2$ languages

- **1-1** alignment: one source symbol may correspond to only one target symbol—or to empty symbol $\epsilon$ (marked “."),
  - $\rightarrow$ Align n-n symbols (2x2)

- Ignore context
  - $\rightarrow$ Model how the *Context* conditions the changes

- Symbols/sounds are treated as **ATOMS**
  - $\rightarrow$ Symbols/sounds analyzed as *vectors of distinctive features*
Problem formulation

Alignment → ... Complete data
Rules → in baseline model: simply the **counts** of events

How do we know which rules are better?

(recall, in baseline: rules are 1×1 alignments)
Rules: high entropy
Rules: low entropy
Extend the baseline model to a 2x2 model: correspondences of up to two symbols on both sides.

The set of admissible *kinds* of events becomes:

\[ K = \left\{ \begin{array}{ll} \text{(# : #)} & \text{(σ : .)} \\ \text{(· : τ)} & \text{(σ : τ)} \end{array} \right\} \]
Extend the baseline model to a 2x2 model: correspondences of up to two symbols on both sides.

The set of admissible *kinds* of events becomes:

\[
K = \left\{ \begin{array}{ccc}
(# : #) & (\sigma : .) & (\sigma \sigma' : .) \\
(. : \tau) & (\sigma : \tau) & (\sigma \sigma' : \tau) \\
(. : \tau \tau') & (\sigma : \tau \tau') & (\sigma \sigma' : \tau \tau')
\end{array} \right\}
\]
Align more than two languages: e.g., Finnish : Estonian : Mordva

\begin{tabular}{cccccccc}
y & h & d & e & k & s & ä & n \\
\hline
\end{tabular}

\begin{tabular}{cccccccc}
\ddot{u} & h & e & k & s & a & .
\hline
\end{tabular}

\begin{tabular}{cccccccc}
v & e & \chi & . & k & s & a & .
\hline
\end{tabular}

Model each 3D event as three pairwise events. Some examples are incomplete – missing data in one language:

\begin{tabular}{cccccc}
h & a & a & m & u
\hline
\end{tabular}

\begin{tabular}{cccccc}
\_ & \_ & \_ & \_ & \_ & \_ \\
\hline
\_ & \_ & \_ & \_ & \_ & \_ \\
\_ & \_ & \_ & \_ & \_ & \_ \\
\end{tabular}

\begin{tabular}{cccccc}
\check{c} & a & m & a
\hline
\end{tabular}
3-D Model
Building decision trees
Performance: Compression rates

Aligning Finnish with Estonian

Compressed size (bytes) vs. Data set size

- GZip
- BZip2
- Two-part code
- 2x2-boundaries
- Context-0
**Language Distance**

**Sanity check:** Use alignment to measure inter-language distances

- Cost for different language pairs $C(a, b)$ are not comparable
- *Normalised Compression Distance* (Cilibrasi & Vitanyi, 2005)

$$\delta(a, b) = \frac{C(a, b) - \min(C(a, a), C(b, b))}{\max(C(a, a), C(b, b))}$$

Align all languages in StarLing *pairwise*, e.g., using two-part 1x1 model.

→ ...
Table: Pairwise normalised compression distances for Finno-Ugric sub-family of Uralic, StarLing data.