Comparison of Component Models in Analysing Dialectal Features

Antti Leino[1], Saara Hyvönen[2]

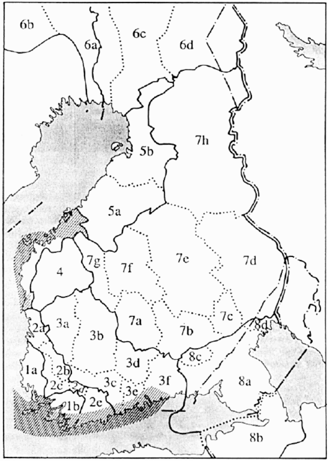
5 August 2008

[1,2] **Department of Computer Science**
[1] **Research Institute for the Languages of Finland**
[2] **Fonecta Ltd**

Background

Finnish Dialects

- Traditional view, relatively unchanged for the past half century
- Western dialects
  1. Southwestern dialects
  2. Mid-Southwestern dialects
  3. Tavastian dialects
  4. Southern Ostrobothnian dialect
  5. Central and Northern Ostrobothnian dialects
  6. Northernmost dialects
- Eastern dialects
  7. Savonian dialects
  8. Southeastern dialects



Corpora

- Traditional Finnish dialectology based largely on morphological and phonological features

- Lauri Kettunen, *Suomen murteet IIIA: murrekartasto* 'Finnish Dialects IIIA: Dialect Atlas', 1940
  - 213 maps
  - Out of print for decades, still widely used
  - Computer corpus (Embleton – Wheeler 1997)

- Lexical variation: *Suomen murteiden sanakirja* 'Dictionary of Finnish Dialects'
  - Ongoing project at the Research Institute for the Languages of Finland: Vol. I 1985, Vol. VIII 2008, Vol. XX c. 2040
  - Here: distribution maps for c. 5 500 articles
  - Used earlier by Hyvönen et al. (2007)

Objectives

- Earlier work (Hyvönen et al. 2007) indicated that clustering is not ideal for dialect features

- Component analysis works better
  - But which one? So many to choose from

- Compare five different methods to these two corpora
  - Factor Analysis
  - Non-negative Matrix Factorisation
  - Aspect Bernoulli
  - Independent Component Analysis
  - Principal Component Analysis

Methods
Overview

- All these can be thought of as latent variable models
- Aim to find a small number of latent variables / factors / components / aspects that explain the data
- Ideally, factors interpretable in terms of dialect regions
- Each factor can be visualised as a choropleth map, with a colour slide between the extremes

Methods

Factor Analysis

- Aims to find a small set of factors which explain the data
- Factors hopefully interpretable
- Here one would expect some correspondence between factors and dialects
- Implementation by Trujillo-Ortiz et al. (2006)

Methods

Non-negative Matrix Factorisation

- Aims to find a small set of *non-negative* factors which explain the data in a non-negative way, e.g.
  - Kainuu could be explained using Savonian and Ostrobothnian components with weights 1 and 2 ('1 part Savonian and 2 parts Ostrobothnian')
  - A certain dialect word or feature could be half Tavastian and half Southwestern
- Non-negativity intuitively appealing: what does it mean if a word is '-0.5 Tavastian'?
- Berry et al. (2007)

Methods

Aspect Bernoulli

- Designed for binary data
- Interpretation can be given in terms of probabilities
  - E.g. a municipality / dialect word / dialect feature is 83 % Savonian
- Designed to deal with noisy data
  - But as seen later on, too much is too much
- Kaban et al. (2004)

Methods

Independent Component Analysis

- Aims to find statistically independent components
- Most often used to separate signals with lots of measurements and a few measurement points
  - E.g. the coctail party problem: 5 people talking, 5 microphones: separate speech signals
- Here we consider dialect words / features as the signal
- PCA as a pre-processing step: we use only a small number of principal components (otherwise components too localized)
- Hyvärinen et al. (2001)

Methods

Principal Component Analysis

- Finds the direction in the data which explains most of the variation
- Each component explains the variation left in the data after the variation explained by previous components has been removed
- Interpretation must be done bearing the previous components in mind
- Hotelling (1933)

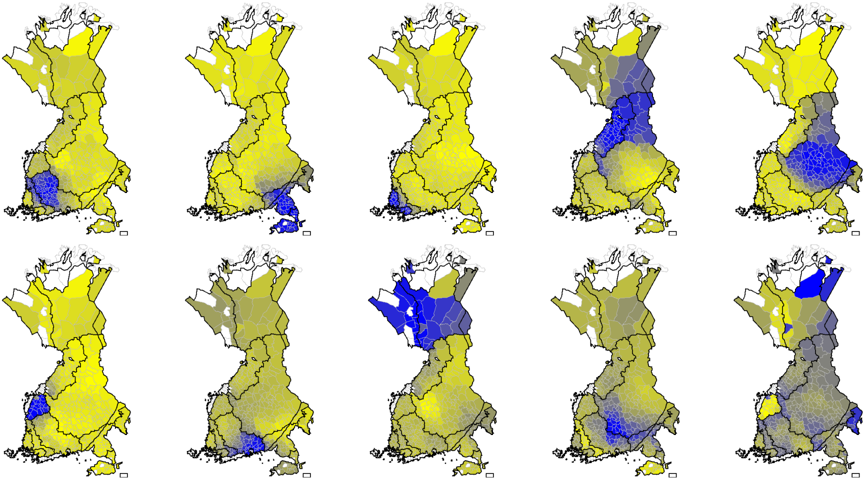
Series I: Dialect Atlas

Overview

- Relatively clean data: no significant gaps
  - Exception: northernmost Finland
- Ten-component run for each method
  - For PCA, first ten components
- Most methods manage this without too many problems
- Some differences

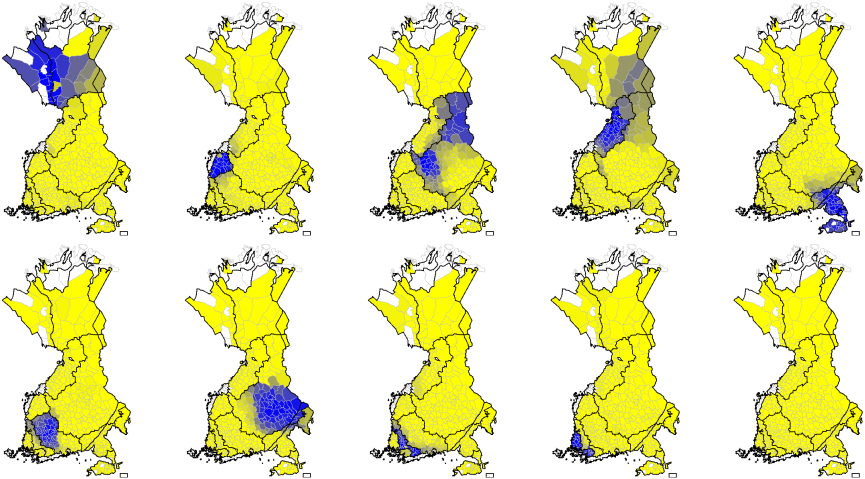
Series I: Dialect Atlas

Factor Analysis



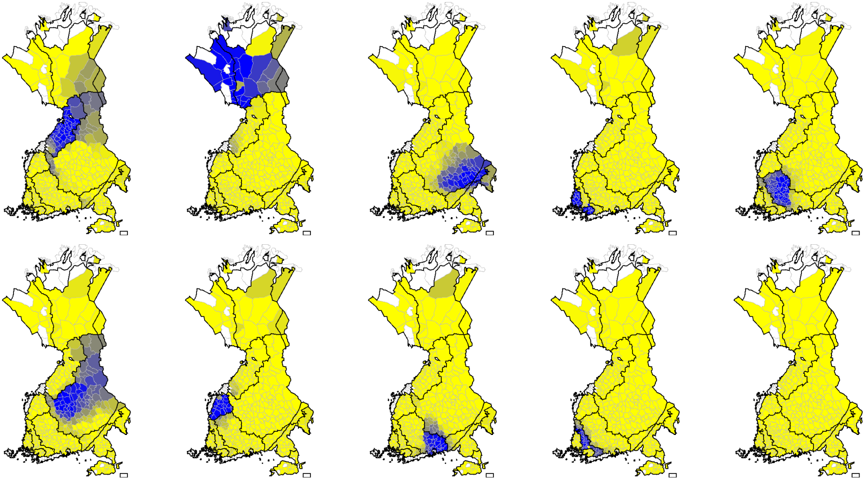
Series I: Dialect Atlas

Non-negative Matrix Factorisation



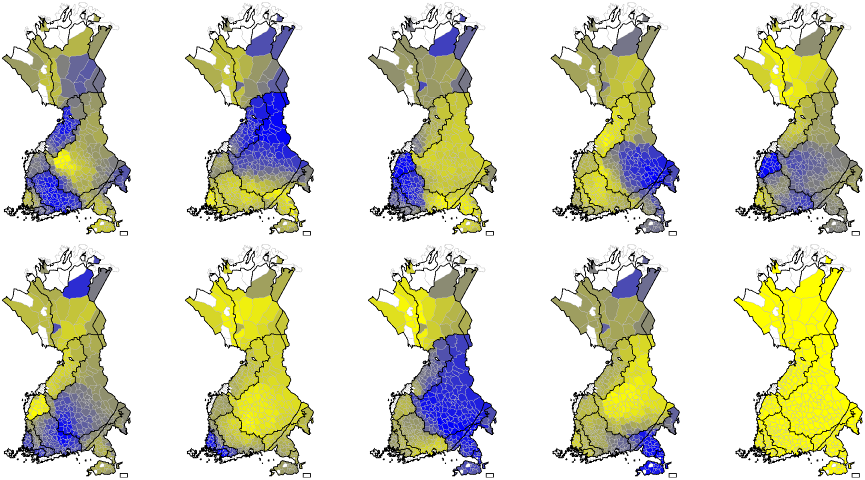
Series I: Dialect Atlas

Aspect Bernoulli



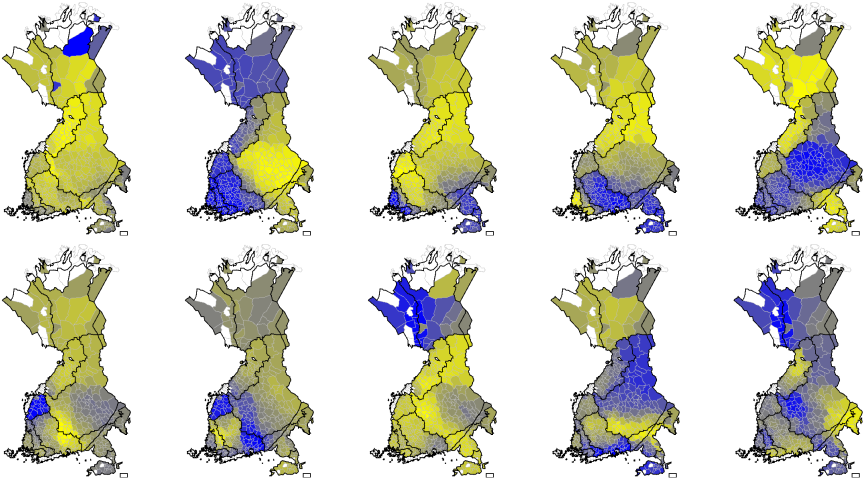
Series I: Dialect Atlas

Independent Component Analysis



Series I: Dialect Atlas

Principal Component Analysis



Series II: Dialect Dictionary

Overview

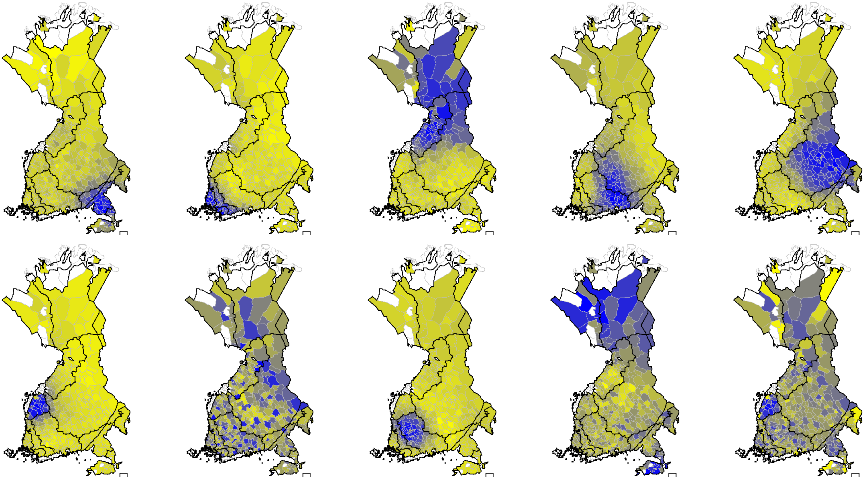
- Much more spotty data
  - Some municipalities thoroughly surveyed
  - Some far less so

- Ten-component run for each method
  - For PCA, first ten components

- Different issues with different methods than in the Dialect Atlas data
- All methods have at least one ‘noise’ component

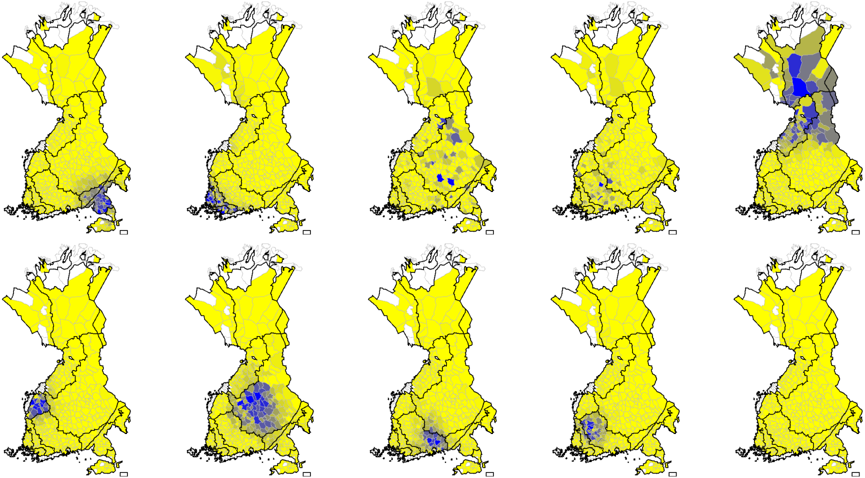
Series II: Dialect Dictionary

Factor Analysis



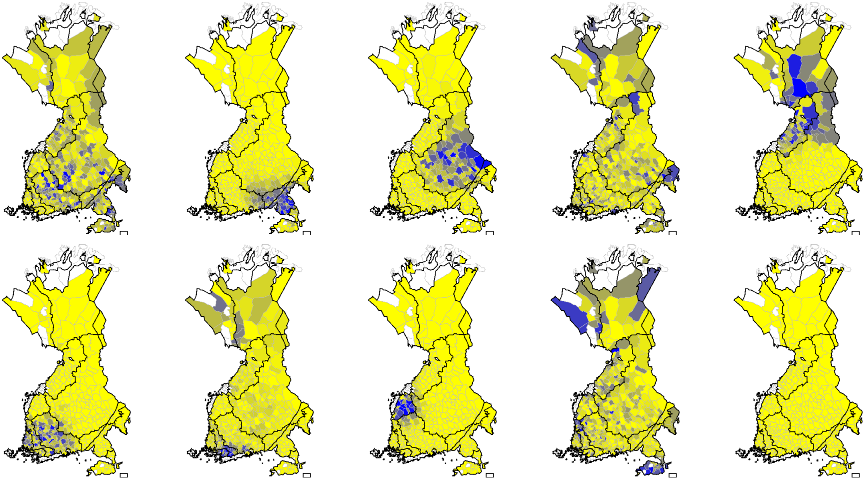
Series II: Dialect Dictionary

Non-negative Matrix Factorisation



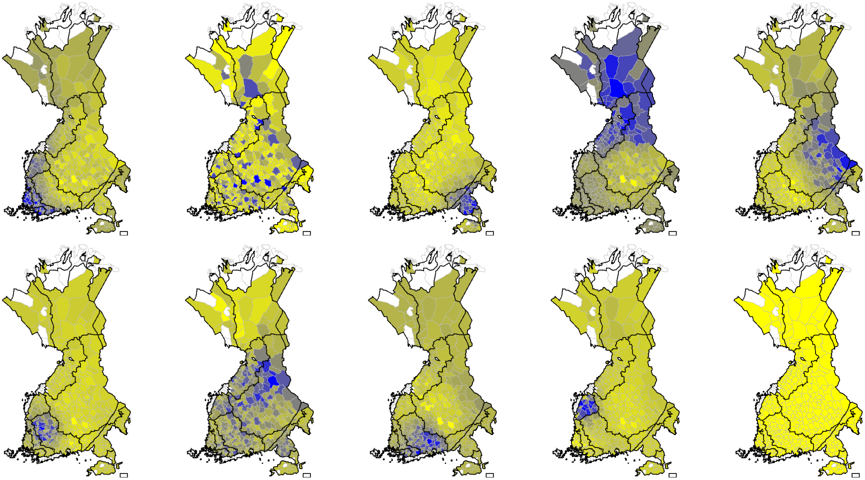
Series II: Dialect Dictionary

Aspect Bernoulli



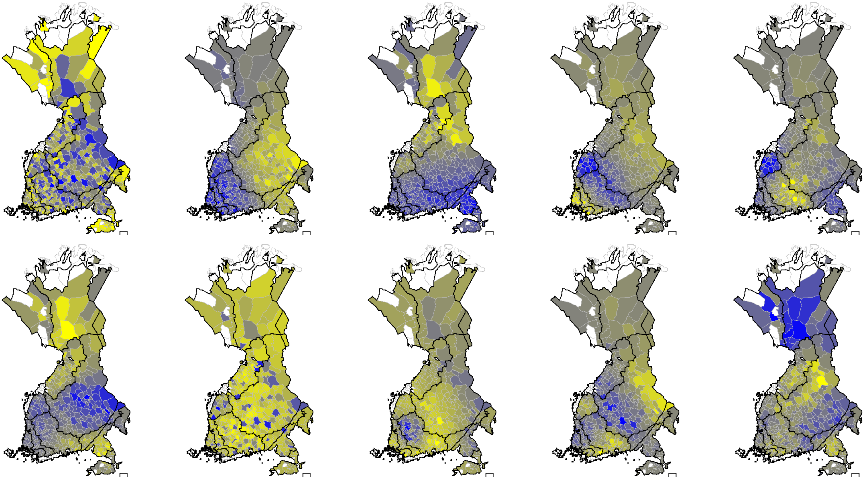
Series II: Dialect Dictionary

Independent Component Analysis



Series II: Dialect Dictionary

Principal Component Analysis



Conclusions

- Two different issues with data
  - Amount of variables (≈ features, words)
  - Amount of noise (≈ spottiness of data)
- For reasonably clean data, Non-negative Matrix Factorisation and Aspect Bernoulli work well
- For large number of variables Independent Component Analysis works
- Factor Analysis is a good compromise: not the best, but works for both cases
- Principal Component Analysis is very different from the others – use with care

References

Berry, Michael W. – Browne, Murray – Langville, Amy N. – Pauca, Paul V. – Plemmons, Robert J. 2007: Algorithms and applications for approximate nonnegative matrix factorization. Computational Statistics & Data Analysis 52(1): 155–173.

Embleton, Sheila – Wheeler, Eric S. 1997: Finnish dialect atlas for quantitative studies. Journal of Quantitative Linguistics 4(1–3): 99–102.

Hotelling, Harold 1933: Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology 24: 417–41, 498–520.

Hyvärinen, Aapo – Karhunen, Juha – Oja, Erkki 2001: Independent Component Analysis. John Wiley & Sons.

Hyvönen, Saara – Leino, Antti – Salmenkivi, Marko 2007: Multivariate analysis of Finnish dialect data – an overview of lexical variation. Literary and Linguistic Computing 22(3): 271–290.

Kaban, Ata – Bingham, Ella – Hirsimäki, Teemu 2004: Learning to read between the lines: The aspect Bernoulli model. Proceedings of the 4th SIAM International Conference on Data Mining 462–466

Trujillo-Ortiz, A. – Hernandez-Walls, R. – Castro-Perez, A. – Rodriguez-Ceja, M. – Melendez-Sanchez, A.L. – del-Angel-Bustos, E. – Melo-Rosales, M. – Vega-Rodriguez, B. – Moreno-Medina, C. – Ramirez-Valdez, A. – D'Olivo-Cordero, J.P. – Espinosa-Chaurand, L.D. – Beltran-Flores, G.L. 2006: ANFACTPC: Factor Analysis by the Principal Components Method. A MATLAB file. http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=10601.