Use of latent semantic analysis at Otec

presentation at L3S

Hannover

15 May 2006

Jan van Bruggen

Outline

- Background
- LSA in a nutshell
- Applications
  - Positioning and APL
  - Question answering and community formation

Background

- Otec development program
  - Learning networks
  - Bottom-up / data-driven
- Methodology
  - IT tools and technique
  - Psychological research methods
- LSA
  - Data driven
  - Latent variables explain data

Latent semantic analysis is like ...

- Principal component analysis / factor analysis
  - Symmetric matrix **M -** correlations
  - Factor analysis: insert communalities
  - Eigenvalue en eigenvectors
  - **M = U Λ U'**
  - **Λ** is diagonal matrix with sorted eigenvalues
  - Reproduction: remove smallest eigenvalues in **Λ** and columns and rows in **U** and **U'**
  - Rotate and interpret factor solution

Latent semantic analysis

- Asymmetric matrix (data-matrix)
  - Terms by documents
  - Word frequencies
- Singular value decomposition: **D = L S R'**
- LSA: reproduction based on a model with less dimensions
- If **M** = **DD'** then no difference with PCA

Application areas

- Document retrieval (LSI)
- Cognitive science
  - Semantics of text (Kintsch)
  - Concept learning
- Education
  - Essay rating
  - Selection and sequencing of material
  - Writing

Application domains

Document retrieval

- Large corpora
- Heterogeneous
- High-dimensional

Educational uses

- Small corpora
- Homogeneous
- Low dimensional

Methods:

Minimum size of corpus

Number of dimensions to retain

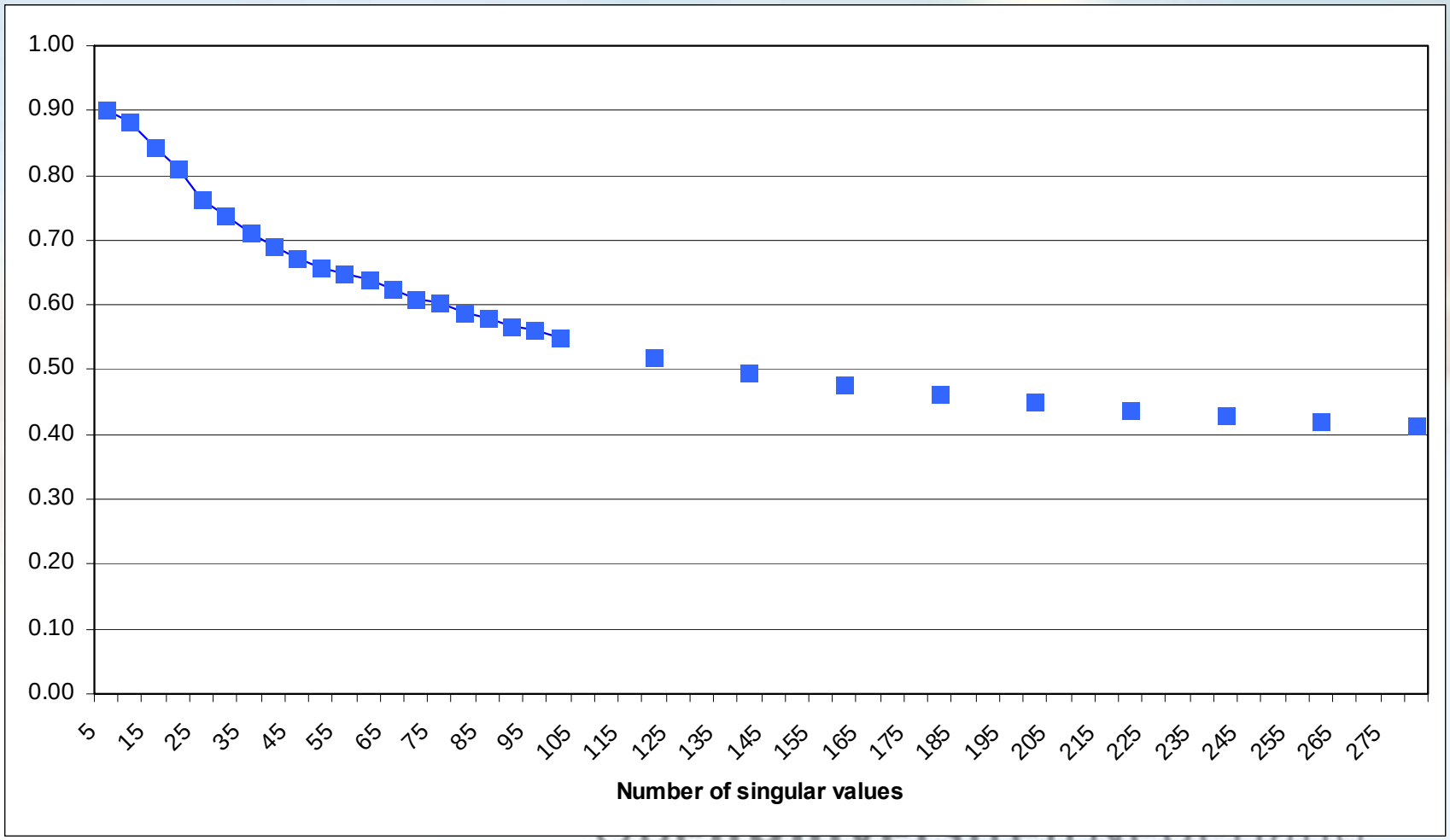
Term frequency measure

Filtering noise words

Methodological considerations

- LSA is not a statistical method
- SVD is a least squares technique
- Error rulez!
  - x = True(x) + error
  - covariance(xy) = covar(True(x)True(y))
  - correlation = cov(xy) / sqrt(varx.vary)
  - Thus: the better you reproduce the data, the worse your correlations get

Example


1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
5
15
25
35
45
55
65
75
85
95
105
115
125
135
145
155
165
175
185
195
205
215
225
235
245
255
265
275
Number of singular values

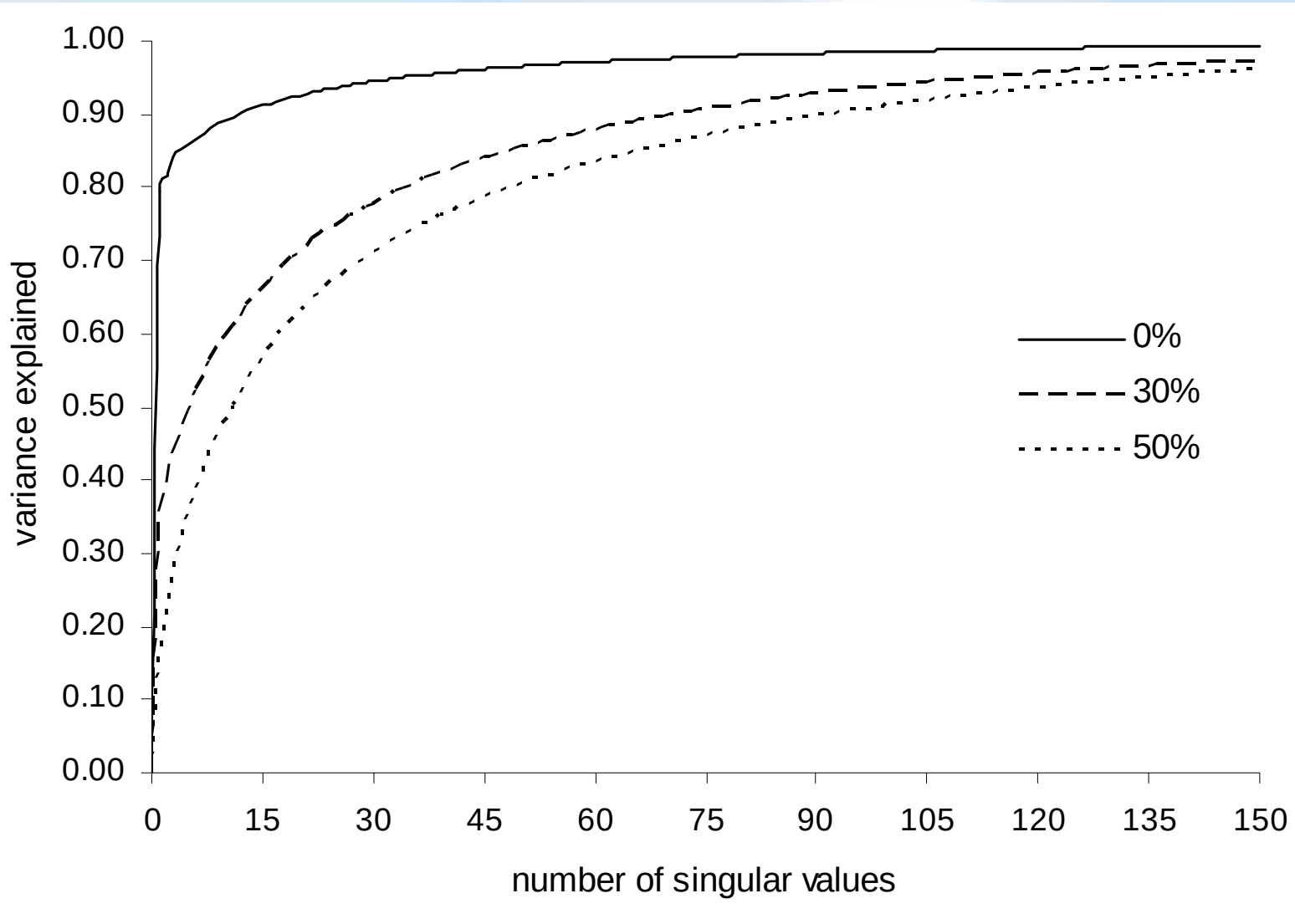
Explained variance and # SVs

Reliability limits explained variance (.80 is good!)

$$\frac{\sum x^2}{n} - M^2 \qquad \sum S^2 = \sum x^2$$

Approximate variance by SSQ(sv)
Determine bandwidth for number of SVs
Still filtering (stopping) is often needed

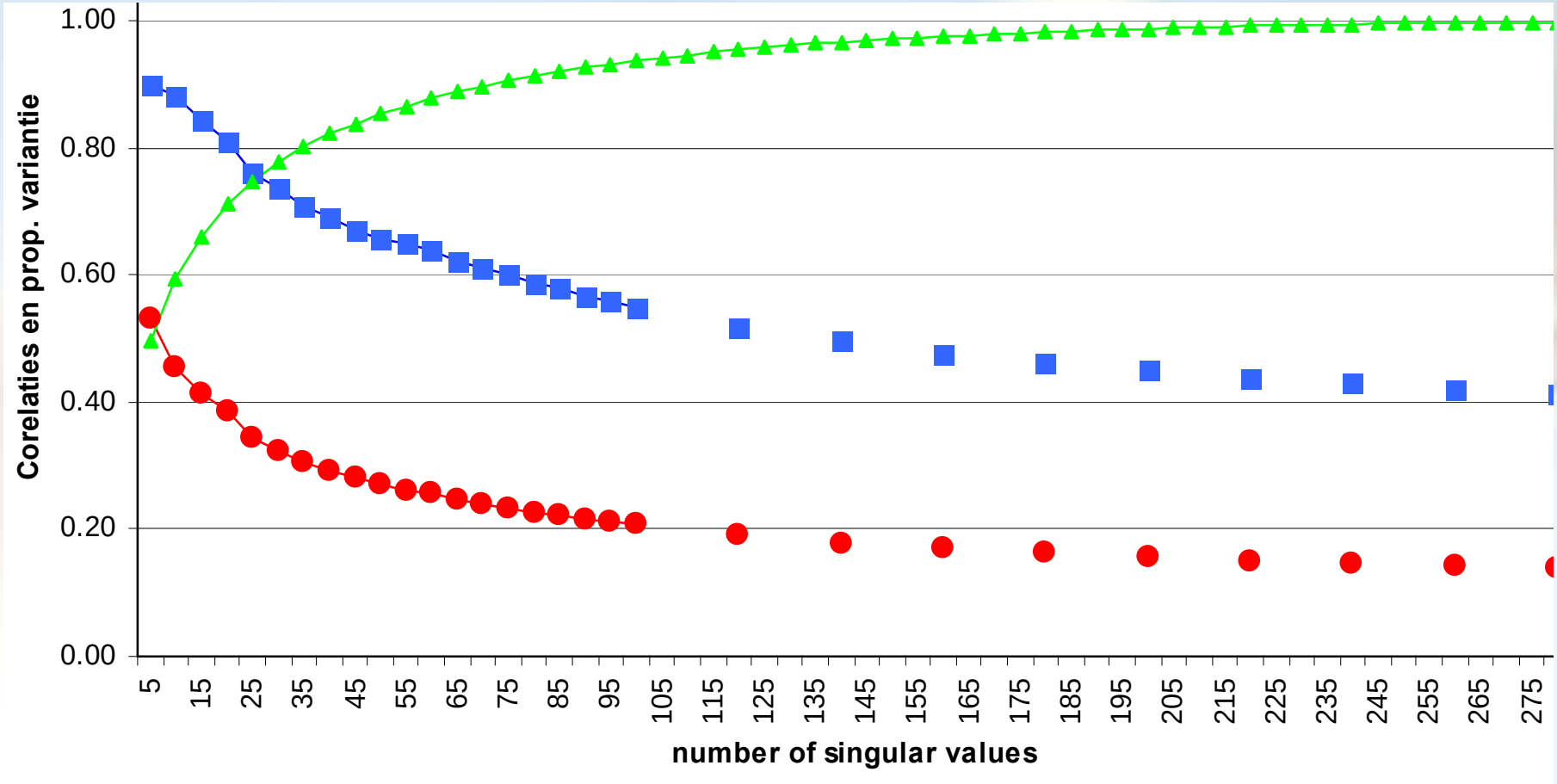
Effect of stopping



Positioning

- Positioning and APL
  - The APL problem
  - Equivalence of outcomes
- Content as a proxy for outcomes
- Performance:
  - Discriminate like an expert
  - Explain sufficient variance
- Trial

Results


1.00
0.80
0.60
0.40
0.20
0.00
Corelaties en prop. variantie
5
15
25
35
45
55
65
75
85
95
105
115
125
135
145
155
165
175
185
195
205
215
225
235
245
255
265
275
number of singular values

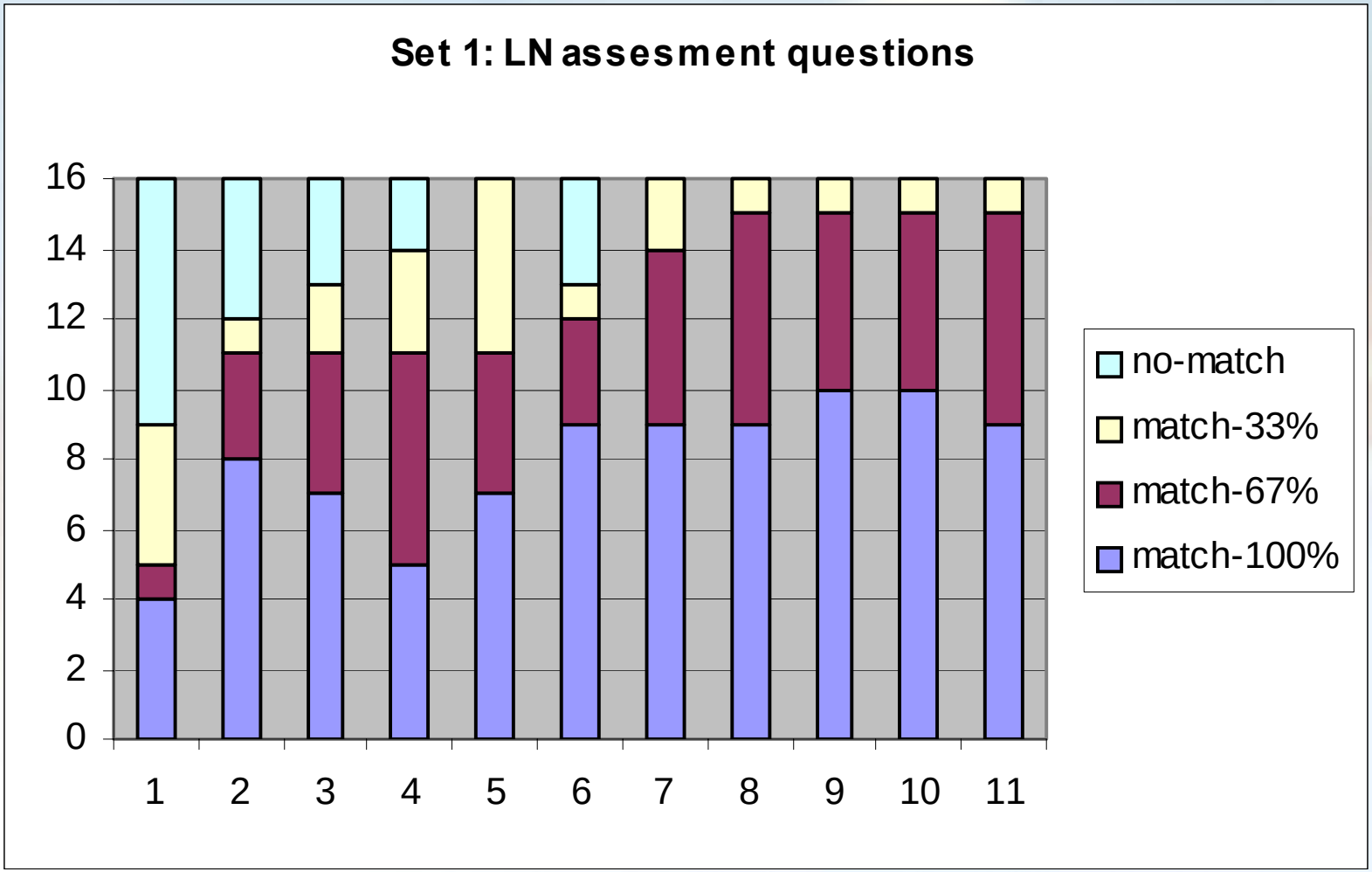
Ad-hoc community formation

- Reduce tutor load
- Question answering by peers
- Find relevant documents
- Find relevant peers
- Create ad-hoc community to answer the question

Methods

- Internet course
- Match assessment questions to three best fitting course elements
- Compare to expert match
- Stepwise optimization

Results


Set 1: LN assesment questions
16
14
12
10
8
6
4
2
0
1
2
3
4
5
6
7
8
9
10
11
no-match
match-33%
match-67%
match-100%

Conclusions

- Preliminary results encouraging
- Optimization itself can be automated
- Extend / Combine techniques:
  - Positioning project Marco Kalz
  - Recommender system Cooper