THE OVERFITTING IN PROBABILISTIC LATENT SEMANTIC MODELS

V. A. Leksin
Moscow Institute of Physics and Technology, Moscow, Russia
vleksin@mail.ru

K. V. Vorontsov
Computing Centre of the Russian Academy of Sciences, Moscow, Russia
voron@ccas.ru

Introduction

- **Key Observation:** If two people have similar preferences for a subset of the items in a given collection, then they are likely to have similar preferences for other items in the same collection.
- Collaborative filtering (CF) methods work by analyzing the observed preferences of a group of people in order to make predictions about each person's unobserved preferences.
- These preferences can be implicitly collected observations like the number of times a user accessed an internet site, or explicit quantifications of preference like the rating assigned by the user to a movie.

Problem statement

- Let:
  - U - set of users; R - set of items
  - $\{u_i, r_i\}_{i=1}^{l} \in U \times R$ - given sample of *co-occurrence* observations.
- The goal is to induce similarity functions on
  - users $\rho_U(u,u')$
  - items $\rho_R(r,r')$.
- The final goals: personal recommendation, prediction of user behavior, items cathegorization, similarity search, etc.

Collaborative filtering methods

- Memory based
- Relevance Models
- Clients Environment Analysis (CEA)
- Latent Class Models
  - Latent Semantic Analysis (LSA)
  - Probabilistic LSA (pLSA) or aspect model
- Matrix Factorization
- Clustering
- Transitive Associations
- Trust Inference
- Perception-based

Clients Environment Analysis (CEA)

- The final applications are:
  - recommender systems
  - direct marketing
  - personalized advertising
  - similarity search
  - similar minded people search in social networks.
- The main idea of CEA is to use the consistent similarity measures:
  - items are similar if they are used by similar users
  - users are similar if they use similar items.

Probabilistic LSA, latent profiles

- Suppose each user $u \in U$ is interested in a subset of topics from the set of topics *T.*
- *Latent profile of the user* - a vector of conditional probabilities:
  $$q_{tr} = q(t \mid r), t = 1,\ldots,|T|, \sum_{t \in T} q_{tr} = 1.$$
- *Latent profile of the item* - a vector of conditional probabilities:
  $$p_{tu} = p(t \mid u), t = 1,\ldots,|T|, \sum_{t \in T} p_{tu} = 1.$$

Bayesian model of data

- The probability of co-occurrence can be alternatively represented by two different generative models:

$$(1)\quad p(u,r)=\sum_t p(u)\,p_{tu}\,q(r\mid t,u),\quad q(r\mid t)=\frac{q_{tr}\,q(r)}{\sum_{s\in R} q_{ts}\,q(s)},$$

$$(2)\quad p(u,r)=\sum_t q(r)\,q_{tr}\,p(u\mid t,r),\quad p(u\mid t)=\frac{p_{tu}\,p(u)}{\sum_{s\in U} p_{ts}\,p(s)}$$

- Sample of *co-occurrence* observations: $D=\{u_i,r_i\}_{i=1}^{l}$.
- Maximization of the log-likelihood:

$$L(D;\{p_{tu}\},\{q_{tr}\})=\ln\prod_{i=1}^{l} p(u_i r_i)\to\max_{\{p_{tu},q_{tr}\}}.$$

The symmetric EM algorithm

Repeat until profiles converge:

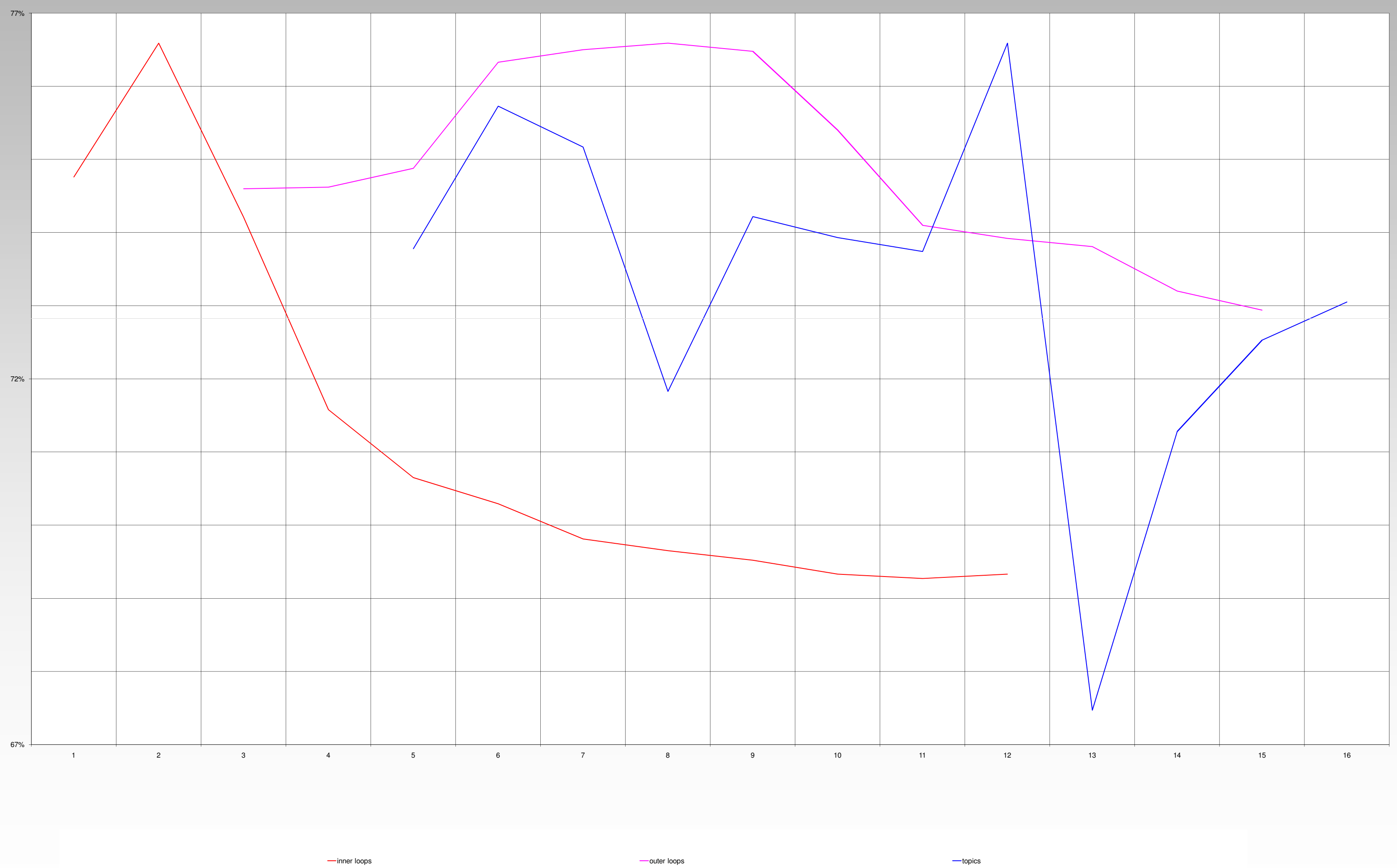
- Optimize p_{tu} for fixed q_{tr}:
  - E-step: $H_{tr}(u) = p_{tu} q(r \mid t) \Big/ \sum_{s} p_{su} q(r \mid s)$ - hidden variables
  - M-step: $p_{tu} = \sum_{r:(u,r)\in D} H_{tr}(u) \Big/ \sum_{r:(u,r)\in D} 1$ - latent profiles
- Optimize q_{tr} for fixed p_{tu}:
  - E-step: $H^{*}_{tu}(r) = q_{tr} p(u \mid t) \Big/ \sum_{s'} q_{s'r} p(u \mid s')$ - hidden variables
  - M-step: $q_{tr} = \sum_{u:(u,r)\in D} H^{*}_{tu}(r) \Big/ \sum_{u:(u,r)\in D} 1$ - latent profiles

Experiments

- Log file of clicks on documents returned by the search machine Yandex:
  - 1024 most visited web sites as items
  - 7292 most active users
  - The latent profile size has been fixed as T=12
  - The meaning of topics has not been fixed a priory.
- Classified subsample:
  - 400 web sites classified into 12 classes.
- The profile quality criterion:
  - a number of labeled sites such that the position of the maximum in their profile coincides with the most frequent position of the maximum over the class.

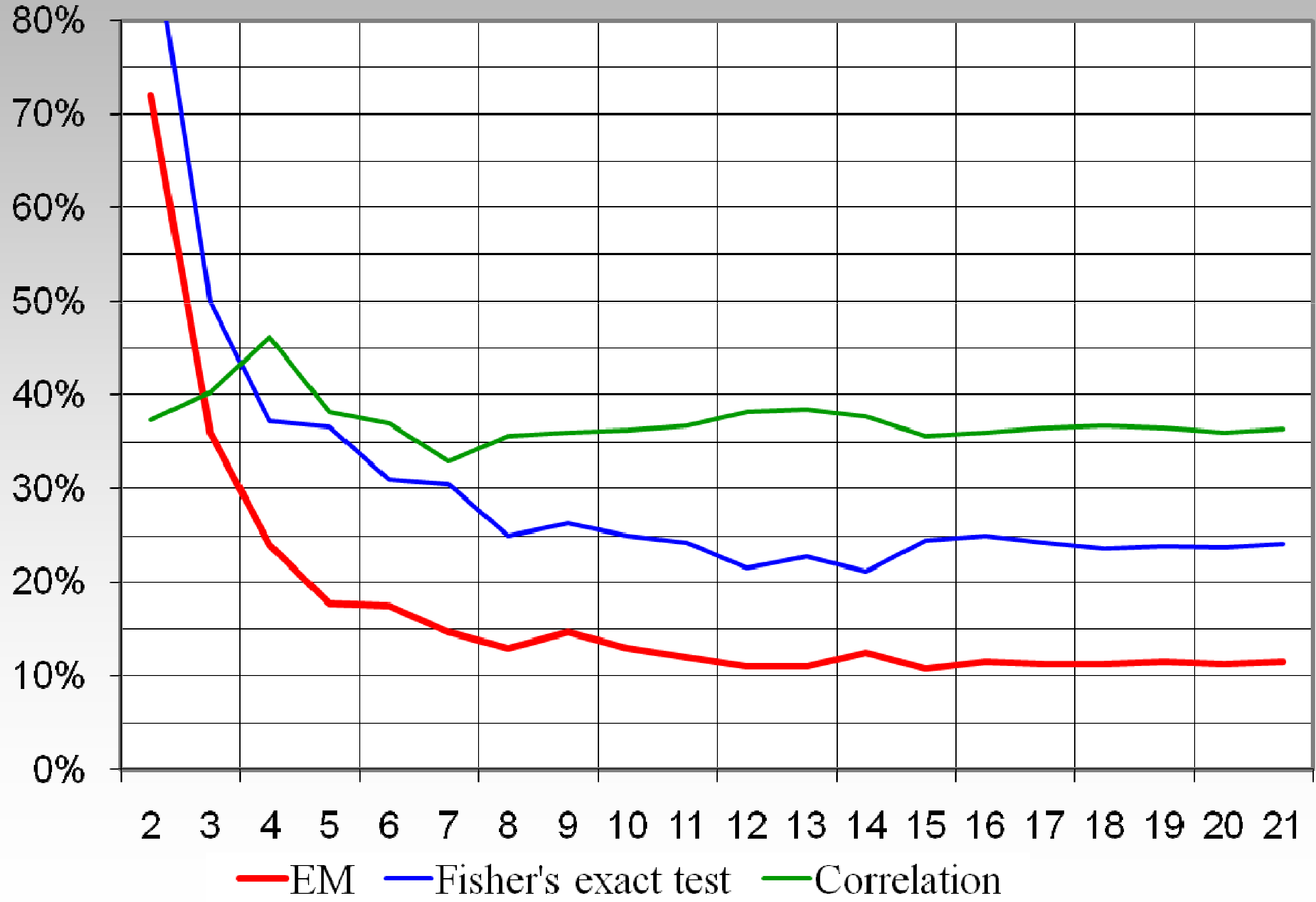
Optimization of parameters

The dependence of the number (in percents) of correctly reconstructed item profiles on three parameters of the algorithm

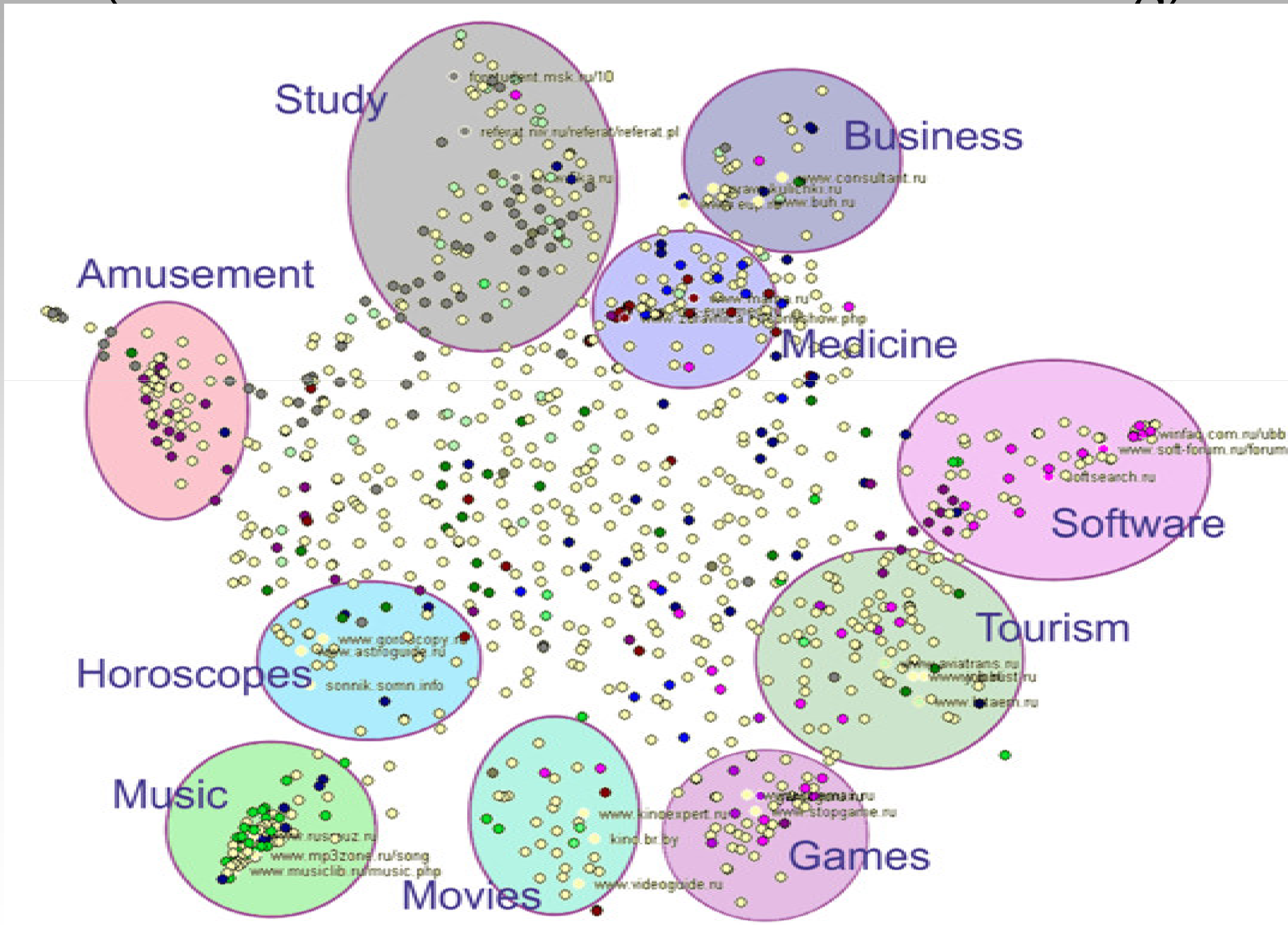


Comparison with other algorithms

The dependence of the number (in percents) of misclassified items on the parameter *k* in *k*NN algorithm for three types of metrics



Similarity map
(the result of Multidimensional Scaling)



Conclusions

- The robust Euclidean distance between profiles is a much more adequate distance measure between items if compared with standard techniques.
- The meaning of topics has not been fixed a priory. Nevertheless the latent profiles estimated by the algorithm turned out to be well interpretable.
- MDS groups web sites of similar subject matter into clusters. The sites belonging to the same cluster usually have the maximal profile component in the same position.
- Excessive optimization is redundant and can lead to overfitting.

References

[1] **Customer Environment Analysis** // Forecsys. — 2005. http://www.forecsys.ru/english/cea.php.

[2] *T. Hofmann.* **Latent Semantic Models for Collaborative Filtering** // ACM Transactions on Information Systems, Vol. 22, No. 1, 2004, Pp. 89–115.

[3] *X. Jin, Y. Zhou, B. Mobasher.* **Web Usage Mining Based on Probabilistic Latent Semantic Analysis** // Proc. of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining. — 2004. — P. 197–205.

[4] *V. A. Leksin, K. V. Vorontsov*. **The client environment analysis: the reconstruction of latent profiles and similarity estimation of users and items** // Russian conf. Mathematical Methods of Pattern Recognition (MMPO-13). — Moscow: MAKS Press, 2007. — Pp. 488–491 (in russian).

[6] *K. V. Vorontsov, K. V. Rudakov, V. A. Leksin, A. N. Efimov* // **Web Usage Mining based on web users and web sites similarity measures** // Artificial Intelligence. — Donetsk, 2006. — Pp. 285–288 (in russian).