Conjugate-Prior-Regularized Multinomial PLSA for Collaborative Filtering

*Marcus Klasson, *Stefan Ingi Adalbjörnsson, †Johan Swärd, and *Søren Vang Andersen

*Department of Mathematics, Lund University, Sweden, and †Department of Mathematical Statistics, Lund University, Sweden

Abstract

We consider the over-fitting problem for multinomial **probabilistic Latent Semantic Analysis** (pLSA) in collaborative filtering using **maximum a posteriori** (MAP) approach based on conjugate priors as regularization approach. Our proposal ensure that complexity of each step remains the same as compared to the unregularized method. In the numerical section, we show that the regularization method and training scheme yields an improvement on commonly used data sets, as compared to previously proposed heuristics.

Collaborative Filtering & PLSA

In recommender systems, **Collaborative Filtering** (CF) is an approach where personalized recommendations are based on what products other consumers with similar preferences have purchased. We consider using the multinomial version of pLSA, which is applicable to e.g. binary and categorical data sets, to recommend movies. The following notation is used

- Ratings $r_{u,i}$ stored in $\mathbf{R} \in \mathbb{R}^{m \times n}$
- Set of users $\mathcal{U} = \{u_1, \ldots, u_m\}$
- Set of items $\mathcal{I} = \{i_1, \ldots, i_n\}$

We introduce latent states $z \in \{z_1, \ldots, z_K\}$ and define the pLSA model

$$P(r|u,i;\theta) = \sum_z P(r|i,z)P(z|u).$$

The training algorithm for pLSA is **Expectation-Maximization** (EM), which optimizes a lower bound on the data log-likelihood.

The three main challenges with multinomial pLSA in CF is

- Over-fitting
- Sparsity of the data
- Computational cost of the training

We generalize the model by introducing a conjugate prior on the parameters to mitigate over-fitting and better handle the data sparsity.

Grid Search Results

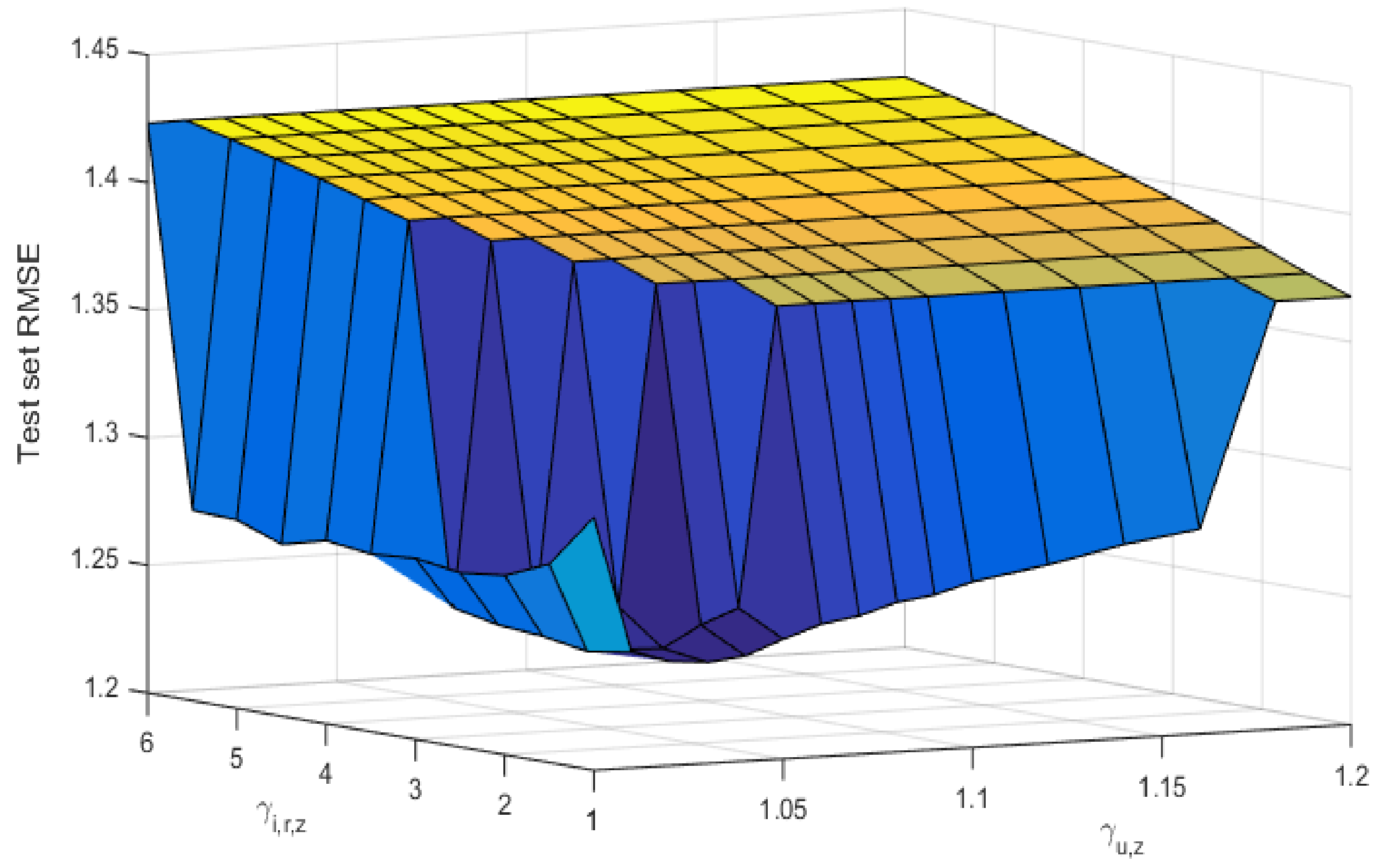

Figure 1: Grid search over RMSE with respect to $\gamma_{u,z}$ and $\gamma_{i,r,z}$ for $K = 200$.

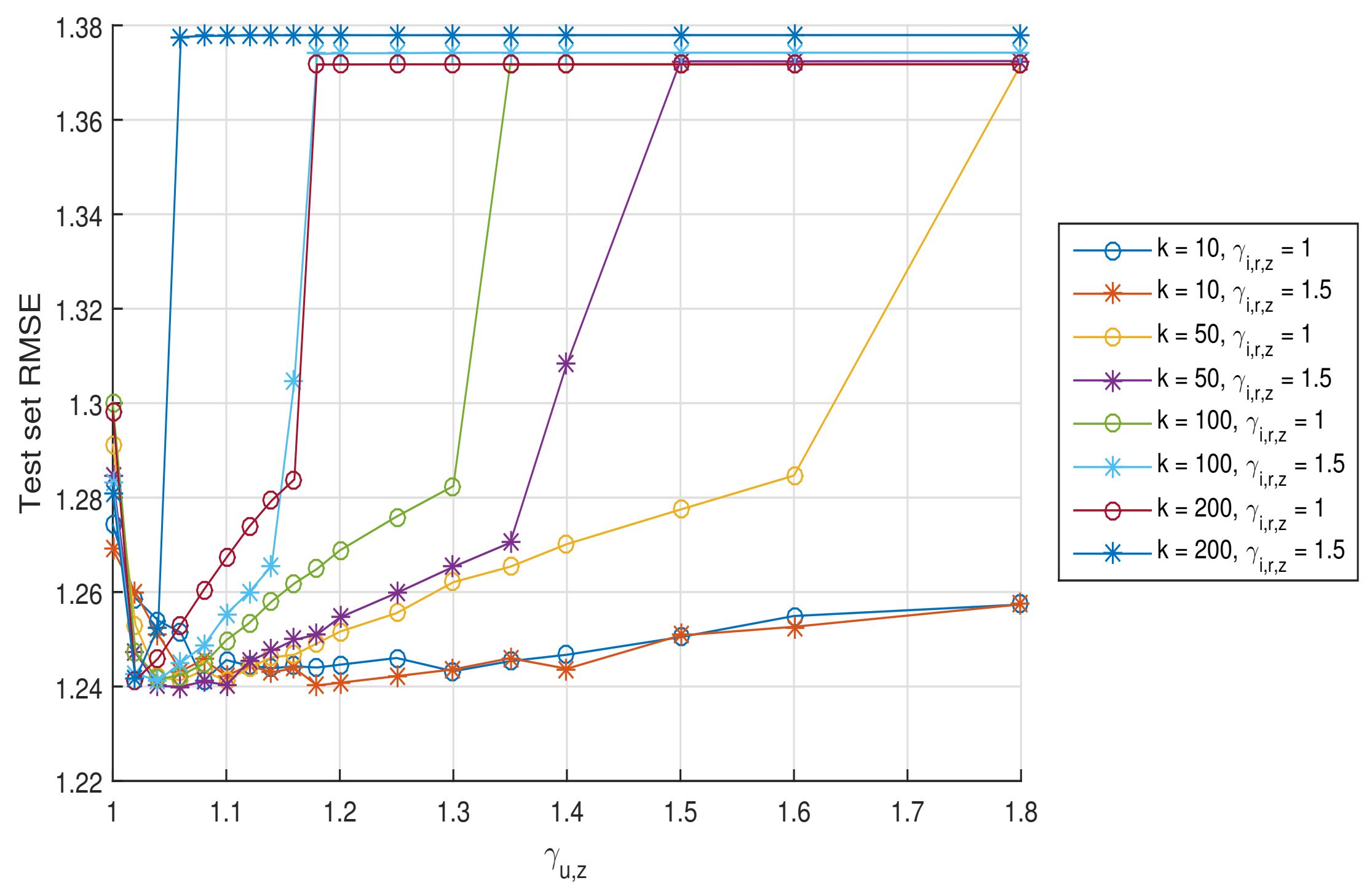

Figure 2: RMSE with respect to $\gamma_{u,z}$, different state sizes K and $\gamma_{i,r,z} = 1.0$ and 1.5.

Optimization Scheme

The lower bound ℓ is formed by incorporating **variational probability distributions** $Q(z|\mathcal{D};\theta)$ to the log-likelihood, where $\mathcal{D}$ denotes the observations $\langle u, i, r \rangle$:

$$\ell(\theta;\mathcal{D}) = \sum_{\mathcal{D}} \sum_z Q(z|\mathcal{D};\theta) \log \frac{P(r|i,z)P(z|u)}{Q(z|\mathcal{D};\theta)}.$$

In MAP estimation, we update the log-posterior instead of the log-likelihood. The multinomial distribution's conjugate prior is the Dirichlet distribution, which is proportional to

$$P(\theta) \propto \prod_z \left(\prod_{i,r} P(r|i,z)^{\gamma_{i,r,z}-1} \prod_u P(z|u)^{\gamma_{u,z}-1} \right).$$

The bound we are optimizing is

$$\log P(\theta|\mathcal{D}) = \ell(\theta;\mathcal{D}) + \log P(\theta),$$

by alternating between

E-step:

$$Q^*(z'|u,i,r;\theta) = \frac{P(r|i,z')P(z'|u)}{\sum_z (P(r|i,z)P(z|u)}$$

M-step:

$$P(z'|u') = \frac{\sum_{i,r} Q^*(z'|u',i,r;\theta) + (\gamma_{u',z'} - 1)}{\sum_z \sum_{i,r} Q^*(z|u',i,r;\theta) + (\gamma_{u',z} - 1)}$$

$$P(r'|i',z') = \frac{\sum_u Q^*(z'|u,i',r';\theta) + (\gamma_{i',r',z'} - 1)}{\sum_{u,r} Q^*(z'|u,i',r;\theta) + (\gamma_{i',r,z'} - 1)}$$

Note that setting $\gamma_{u,z} = \gamma_{i,r,z} = 1$ gives the standard EM algorithm.

Hyperparameter Search

We used grid searches for finding suitable prior hyperparameters $\gamma_{u,z}$ and $\gamma_{i,r,z}$ by evaluating the performance on RMSE. These hyperparameter values are selected as

$$\gamma_{\{\cdot\}} = n_{\{\cdot\}} + 1,$$

where n can be interpreted as an additional set of artificial data points to regularize the parameter estimates. From Figure 1 we found that adjusting $\gamma_{u,z}$ for the latent states had better impact on decreasing the prediction error than $\gamma_{i,r,z}$. Figure 2 shows that multiple state sizes have the potential of performing almost equally well for some of the selected hyperparameter values.

Numerical Evaluation

Data sets: EachMovie and MovieLens 1M.

Metrics: RMSE and MAE.

Data sets were split with the **leave-one-out** method. Rating $\hat{r}$ given user u and unseen item i was predicted with the expected value,

$$\hat{r}_{u,i} = \mathrm{E}[r|u,i] = \sum_r r \sum_z P(r|i,z)P(z|u).$$

We compare the conjugate-prior-regularized pLSA to a standard pLSA with an early stopping (ES) condition and the Pop item-average estimator.

Numerical Results

We used $\gamma_{u,z} = 1.08$, $\gamma_{u,z} = 1.5$ and $K = 50$ for our proposed pLSA and $K = 200$ for pLSA with ES in all experiments. We examined the methods with 10 cross-validations and averaging the resulting prediction errors from each test set.

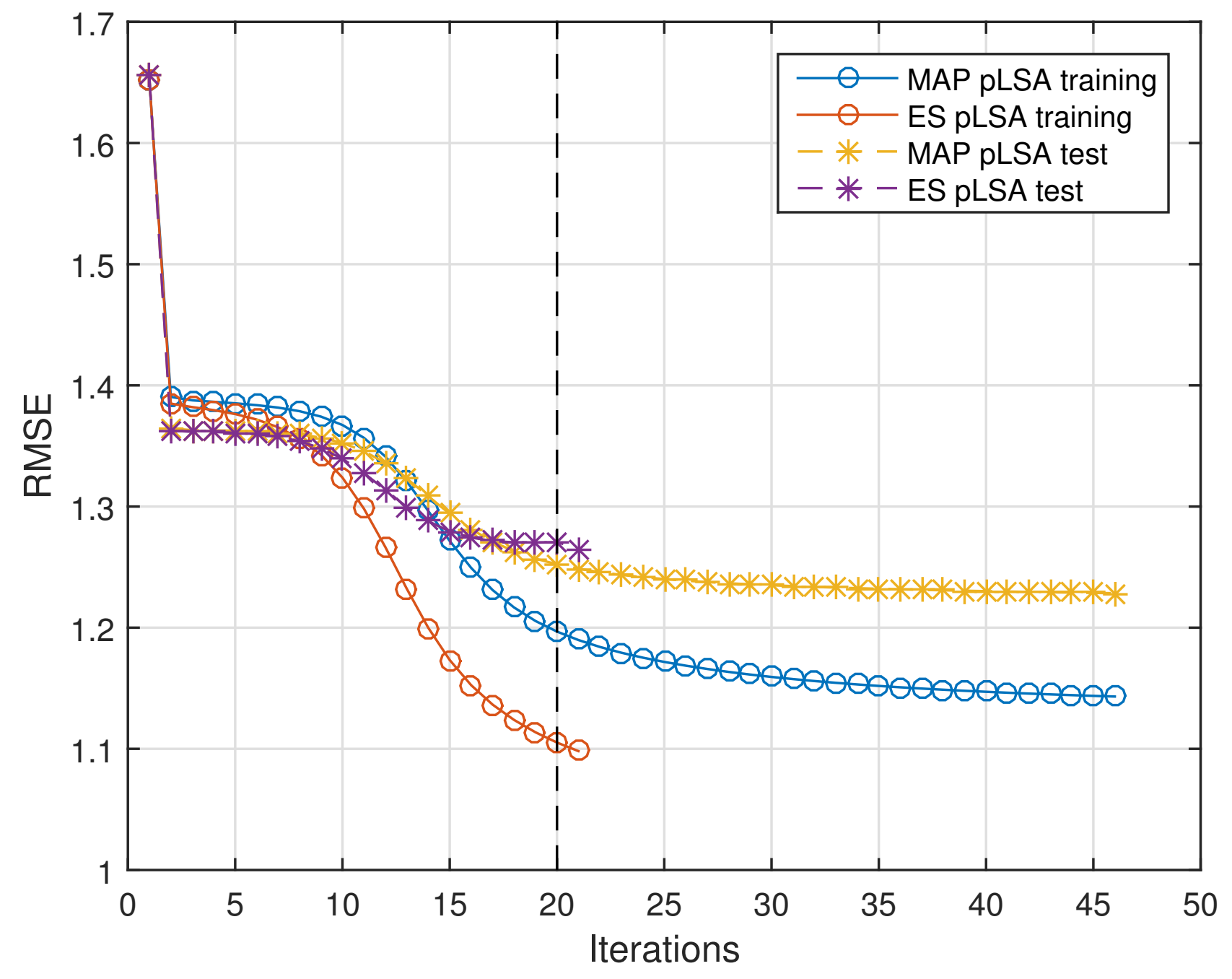

Figure 3: RMSE for training and test data from the EachMovie data set over iterations. Dashed line shows when the early stopping step occurs.

		Proposed	ES pLSA	Pop
RMSE	mean	1.2375	1.2727	1.3712
	std	0.0064	0.0070	0.0062
MAE	mean	0.9711	0.9834	1.0908
	std	0.0047	0.0052	0.0048

Table 1: Mean and standard deviation of the prediction error resulting from the EachMovie data set.

		Proposed	ES pLSA	Pop
RMSE	mean	0.9193	0.9781	0.9847
	std	0.0076	0.0176	0.0090
MAE	mean	0.7241	0.7807	0.7870
	std	0.0049	0.0184	0.0057

Table 2: Mean and standard deviation of the prediction error resulting from the MovieLens data set.

Conclusion

We have shown that conjugate-prior-regularized pLSA counteracts over-fitting in CF setups without increasing the computational complexity. For future work we would analyze its performance in an online CF setting and also other applications.