Making sense of sparse rating data in collaborative filtering via topographic organization of user preference patterns

Gabriela Polčicová

Institute of Informatics and Software Engineering

Faculty of Informatics and Information Technologies

Slovak University of Technology

Ilkovičova 3, SK-84216 Bratislava, Slovakia

tel.: +421 2 60291 728

fax: +421 2 654 205 87

polcicova@fiit.stuba.sk

Peter Tiňo

School of Computer Science

University of Birmingham, Birmingham, B15 2TT, UK

tel.: +44 121 414 8558

fax: +44 121 414 4281

P.Tino@cs.bham.ac.uk

[Contributed Article]

Abstract

We introduce topographic versions of two latent class models (LCM) for collaborative filtering. Latent classes are topologically organized on a square grid. Topographic organization of latent classes makes orientation in rating/preference patterns captured by the latent classes easier and more systematic. The variation in film rating patterns is modelled by multinomial and binomial distributions with varying independence assumptions. In the first stage of topographic LCM construction, self-organizing maps with neural field organized according to the LCM topology are employed. We apply our system to a large collection of user ratings for films. The system can provide useful visualization plots unveiling user preference patterns buried in the data, without loosing potential to be a good recommender model. It appears that multinomial distribution is most adequate if the model is regularized by tight grid topologies. Since we deal with probabilistic models of the data, we can readily use tools from probability and information theories to interpret and visualize information extracted by our system.

Keywords

Collaborative filtering; Self-organizing maps, Topographic ordering; Latent space models

1 Introduction

The amount of available information is steadily increasing. We can be easily overloaded with information of different nature and quality. When deciding which book to read, which film to watch, or which web-site to visit, people often rely on advise given by other people (Hofmann, 2001). This is possible only inside a small communities, where people know other peoples' interests. In many situations we would like to automate the process of sharing evaluations and making recommendations among people that potentially do not know each other but have some “common tastes”. One method addressing this problem is collaborative filtering (CF). Recommendations in CF are produced by leveraging the existing user preferences (ratings/profiles).

There are two main approaches to CF, namely memory-based and model-based. In the former approach, in order to recommend a new item to a particular user, we take into account ratings from people with “similar” interests. The latter approach uses probabilistic modeling (e.g. probabilistic latent class models (LCM)) to infer new recommendations. The main advantage of the model-based approach is that it is able to automatically discover preference patterns in user profile data without suffering the flaw of memory based approaches – the inability to account for the fact that one person can be a reliable recommender for another person on a *subset* of items, but not necessarily for all possible items (Hofmann, 2001).

While much work has focused on designing accurate and fast algorithms for rating predictions, relatively few attempts have been made to implement systems for understanding and visualization of principal user preference patterns/rating trends buried inside potentially very large and sparse body of collaborative filtering data. For example, (Igo et al., 2002) designed a recommender system that makes recommendations in real-time and uses multidimensional visualization to visualize the recommender systems’ results. Shared Wisdom through the Amalgamation of Many Interpretations (SWAMI) by (Fisher et al., 2000) is a framework for building and studying collaborative filtering systems that contains a visualization component. The visualization component supports creation of informative pictures for both developers and end users by enabling e.g. views of sparse user rating matrix and structure in user/rating-item correlation matrices. Such methods, however, are not model-based, i.e. they do not represent hidden preference trends in the data based on a consistent visualization-driven model formulation of user ratings. In this paper we introduce one possible approach to model based visualization for understanding collaborative filtering data. Namely, we propose to make use of preference patterns extracted by LCM with a large number of latent classes endowed with a two-dimensional topological organization. The purpose of topological organization of latent classes is twofold: (1) it enables visualization of common interest/taste patterns in an easily accessible way, (2) it constrains the model so that employing a large number of latent centers (needed for detailed data analysis) does not result in a strongly overfitting model.

The paper is organized as follows. In section 2 we describe latent class models for user ratings and in Section 3 we endow their latent space with a grid topology. Section 4 is devoted to Expectation-Maximization algorithm for training the models. The experiments are described in section 5. The results are presented in section 6. Section 7 contains example visualization plots. In section 8 we discuss the experimental findings and examine sensitivity of the models to variations in the construction parameters. The paper is concluded in section 9 by summarazing key results of this study.

2 Latent Class Models for User Ratings

In this section we briefly describe latent class approach to modeling user ratings introduced in (Hofmann, 2001; Hofmann and Puzicha, 1999). There are three sets we will work with: the set $\mathcal{U}$ of users, the set of films (in general - the set of items), $\mathcal{Y}$, and the set $\mathcal{V}$ of rating values that are used by users to evaluate films.

We would like to predict the rating $v_{u,y} \in \mathcal{V}$ given by a user $u \in \mathcal{U}$ to a film $y \in \mathcal{Y}$. With each triplet $(u, y, v_{u,y})$ we associate a latent variable (class) $z_{u,y} \in \mathcal{Z} = \{1, 2, \ldots, K\}$ that "explains" why the user u rates the film y by $v_{u,y}$. Triplets $(u, y, v_{u,y})$ are either actually observed or just hypothetical entities (in most of the cases).

The latent variables $z \in \mathcal{Z}$ index "abstract" classes of users in two types of models:

- **Type I** – given a film y, all users from class z tend to adopt the same (class-specific) rating pattern expressed through conditional distribution $P(v|y,z)$ over evaluations from $\mathcal{V}$. Given a user u and a film y, the probability of vote v is modeled as

$$P(v|y,u) = \sum_{z \in \mathcal{Z}} P(v|y,z) P(z|u), \tag{1}$$

where $P(z|u)$ is the probability that the user u "participates" in class z.

- **Type II** – all users from class z tend to adopt the same preferences over the [rating, film] pairs (v,y). Here we predict the rating in conjunction with a selection of films (Hofmann, 2001). Given a user u, the probability of a pair (v,y) is modeled as

$$P(v,y|u) = \sum_{z \in \mathcal{Z}} P(v,y|z) P(z|u). \tag{2}$$

Actually, classes z are not meant to represent a clear-cut clustering of users according to their voting preferences, but rather they express "common interest patterns" among the users found in the ratings (Hofmann, 2001). $P(z|u)$ then represents to what extend the user u participates in the common interest pattern z.

Given a set of observation triplets, free parameters of the model, $P(z|u)$ and $P(v|y,z)$, are determined by an expectation-maximization procedure outlined in (Hofmann, 2001).

3 Introducing a Topology into Latent Class Models

Many variants of such topographic representations of data patterns can be found in the machine learning literature. Perhaps the most famous example is the Kohonen self-organizing map (SOM) (Kohonen, 1982, 1990). More recently, a statistically principled reformulations and extensions of SOMs appeared in e.g. (Bishop, 1998; Kaban and Girolami, 2001). Of particular interest to us is the link between SOMs and vector quantization through noisy communication channels established in (Buhmann, 1993; Hofmann and Buhmann, 1998;

Luttrell, 1989). Briefly, in the information theoretic interpretation of SOM, the topological organization of codebook vectors (that correspond to nodes (classes) on the latent grid) emerges through non-uniformity of the channel noise: to minimize the average quantization error at the receiving end of the communication channel, the codebook vectors that are more likely to be corrupted into each other should represent "similar" data patterns, i.e. should lie "close" to each other in the data space.

We endow the latent classes with a topographic organization. Latent classes are organized on the grid topology that places latent classes on an easily readable two dimensional grid, where similar classes tend to lie close to each other.

Topology is introduced into the latent space via the channel noise methodology (Hofmann, 2000). We place latent classes on a regular two-dimensional grid in $[-1, 1]^2$. Chanel noise is then expressed through the neighborhood function

$$P(z_2|z_1) = \frac{\exp\left(\frac{-\|z_1-z_2\|^2}{2\sigma^2}\right)}{\sum_{z\in\mathcal{Z}}\exp\left(\frac{-\|z_1-z\|^2}{2\sigma^2}\right)}. \tag{3}$$

For latent classes z_1 and z_2 lying close to each other[1] on the grid, the probability of corrupting one into the other is high. The parameter $\sigma > 0$ determines "specificity" of the topological neighborhood for class z_1: low values of σ correspond to sharply peaked localized transition probabilities concentrated on close neighbors of z_1, while large values of σ induce general broad neighborhoods spanning large areas of the latent grid.

It is convenient to work with two copies $\mathcal{Z}_Y$ and $\mathcal{Z}_Z$ of the latent space $\mathcal{Z}$. For each user $u \in \mathcal{U}$, the film-conditional ratings v (type I) or pairs (v, y) (type II) are generated as follows:

1. randomly generate a latent class index $z_Y \in \mathcal{Z}_Y$ by sampling the user-conditional probability distribution $P(\cdot|u)$ on $\mathcal{Z}_Y$.

[1] with respect to the metric induced by the Euclidean norm $\|\cdot\|$

2. instead of using the class z_Y to index the "common interest patterns" $P(v|y, z_Y)$ (type I) or $P(v, y|z_Y)$ (type II), we transmit the class identification z_Y through a noisy communication channel, and receive (a possibly different) class index $z_Z \in \mathcal{Z}_Z$ with probability $P(z_Z|z_Y)$.

3. randomly generate a film-conditional rating v with probability $P(v|y, z_Z)$ (type I) or a pair (v, y) with probability $P(v, y|z_Z)$ (type II).

The models for user ratings have now the following form (see eqs. (1), (2))

$$\begin{aligned} P(z_Z|u) &= \sum_{z_Y \in \mathcal{Z}_Y} P(z_Z|z_Y)P(z_Y|u), && (4) \\ P(v|y,u) &= \sum_{z_Z \in \mathcal{Z}_Z} P(v|y,z_Z)P(z_Z|u) \qquad \text{[type I]} && (5) \\ P(v,y|u) &= \sum_{z_Z \in \mathcal{Z}_Z} P(v,y|z_Z)P(z_Z|u) \qquad \text{[type II]} && (6) \end{aligned}$$

4 Parameter Estimation

Given a set $\mathcal{D} = \{(u_1, y_1, v_1), ..., (u_N, y_N, v_N)\}$ of N observation triplets $(u, y, v_{u,y})$, the log likelihood of the data $\mathcal{D}$ is

$$\begin{aligned} \mathcal{L} &= \sum_u \sum_{y \in \mathcal{Y}_u} \log P(v_{u,y}|y,u) \qquad \text{[type I]} && (7) \\ \mathcal{L} &= \sum_u \sum_{y \in \mathcal{Y}_u} \log P(v_{u,y}, y|u) \qquad \text{[type II]}, && (8) \end{aligned}$$

where $\mathcal{Y}_u$ is the set of films evaluated by the user u,

$$\mathcal{Y}_u = \{y \in \mathcal{Y} |\ (u, y, v_{u,y}) \in \mathcal{D}\}. \qquad (9)$$

Following (Hofmann, 2001), we denote by $\rho(v, y, z_Z)$ the probabilities $P(v|y, u)$ and $P(v, y|u)$ in type I and type II models, respectively.

To fit model parameters $P(z_Y|u)$ and $\rho(v, y, z_Z)$ to the data $\mathcal{D}$, we use Expectation-Maximization (EM) algoritm (Dempster, 1977) that maximizes likelihood $\mathcal{L}$. The EM algorithm is a standard algorithm for maximum likelihood estimation in latent variable models. It iterates two steps - Expectation (E) and Maximization (M) - until convergence. Detailed derivations of update equations are presented in Appendix B-C.

4.1 E-step

In the E-step, the algorithm computes the expected values of latent variables using the current values of the model parameters:

$$P(z_Y|\ y,u,v) = \frac{P(z_Y|u)\sum_{z_Z}\rho(v,y,z_Z)P(z_Z|z_Y)}{\sum_{z'_Y}P(z'_Y|u)\sum_{z_Z}\rho(v,y,z_Z)P(z_Z|z'_Y)}, \tag{10}$$

$$P(z_Z|\ y,u,v) = \frac{\rho(v,y,z_Z)\sum_{z_Y}P(z_Z|z_Y)P(z_Y|u)}{\sum_{z'_Z}\rho(v,y,z'_Z)\sum_{z_Y}P(z'_Z|z_Y)P(z_Y|u)}. \tag{11}$$

4.2 M-step

In the M-step, the algorithm re-estimates the model parameters by maximizing the expected complete data log-likelihood evaluated in the E-step. To derive the update equations, we need to determine the types of distributions for $P(z_Y|u)$ and $\rho(v,y,z_Z)$. It is natural to assume multinomial $P(z_Y|u)$. However, for $\rho(v,y,z_Z)$ we use either multinomial, or binomial distribution. Multinomial distribution simply models probabilities of occurrence of ratings v. In the other hand, binomial distribution respects the ordering of rating values v (i.e. it takes into account that ratings 4 and 5 are closer to each other than 1 and 5) and imposes the assumption of unimodal rating distribution.

For models of type I ($\rho(v,y,z_Z) = P(v|y,z_Z)$) we simply model $P(v|y,u)$ as a multinomial or binomial distribution. For models of type II ($\rho(v,y,z_Z) = P(v,y|z_Z)$) we cannot directly use binomial distribution that respects ordering, because there is no ordering on the set of pairs (v,y). Thus, we assume conditional independence of v and y, given abstract class z_Z, a widely used assumption in latent space modeling: $P(v, y|z_Z) = P(v|z_Z)P(y|z_Z)$. We use either multinomial, or binomial distribution for the rating probability $P(v|z_Z)$. In both cases $P(y|z_Z)$ is multinomially distributed. In summary, we work with the following models:

- type I
  - multinomial $P(v|y, z_Z)$ (I-Mult),
  - binomial $P(v|y, z_Z)$ (I-Bin).
- type II
  - joint multinomial $P(v, y|z_Z)$ (II-Mult),
  - $P(v, y|z_Z) = P(v|z_Z)P(y|z_Z)$, where both $P(y|z_Z)$ and $P(v|z_Z)$ are multinomials (II-IndM),
  - $P(v, y|z_Z) = P(v|z_Z)P(y|z_Z)$, where $P(y|z_Z)$ is multinomially and $P(v|z_Z)$ is binomially distributed (II-IndB).

Update equation for $P(z_Y|u)$ is the same for all types of models:

$$P(z_Y|u) = \frac{\sum_{y\in\mathcal{Y}_u} P(z_Y|\ y, u, v_{u,y})}{|\mathcal{Y}_u|}. \tag{12}$$

4.2.1 Type I and II – Multinomial $\rho(v, y, z_Z)$

When $\rho(v, y, z_Z)$ is multinomially distributed, the update equations for $\rho(v, y, z_Z)$ are

$$\rho(v, y, z_Z) \;=\; P(v|\ y, z_Z) = \frac{\sum_{u\in\mathcal{U}_{v,y}} P(z_Z|\ y, u, v)}{\sum_{v'} \sum_{u\in\mathcal{U}_{v',y}} P(z_Z|\ y, u, v')} \qquad \text{[type I]} \tag{13}$$

$$\rho(v, y, z_Z) \;=\; P(v, y|\ z_Z) = \frac{\sum_{u\in\mathcal{U}_{v,y}} P(z_Z|\ y, u, v)}{\sum_{v',y'} \sum_{u\in\mathcal{U}_{v',y'}} P(z_Z|\ y', u, v')} \quad \text{[type II]} \tag{14}$$

where $\mathcal{U}_{v,y}$ is the set of users that evaluated film y with rating v,

$$\mathcal{U}_{v,y} = \{u \in \mathcal{U}|\ (u, y, v) \in \mathcal{D}\}. \tag{15}$$

4.2.2 Type II with Conditional Independence

Assume $P(v, y|z_Z) = P(v|z_Z)P(y|z_Z)$, with both $P(y|z_Z)$ and $P(v|z_Z)$ multinomially distributed. Update equations are given by:

$$P(y|z_Z) = \frac{\sum_{u \in \mathcal{U}_y} P(z_Z|\ y, u, v_{u,y})}{\sum_{y'} \sum_{u \in \mathcal{U}_{y'}} P(z_Z|\ y', u, v_{u,y'})}, \tag{16}$$

$$P(v|z_Z) = \frac{\sum_y \sum_{u \in \mathcal{U}_{y,v}} P(z_Z|\ y, u, v)}{\sum_{v'} \sum_y \sum_{u \in \mathcal{U}_{yv'}} P(z_Z|\ y, u, v')}. \tag{17}$$

4.2.3 Type II – Binomial $P(v|z_Z)$

When $P(v, y|z_Z) = P(v|z_Z)P(y|z_Z)$ and $P(v|z_Z)$ is binomially and $P(y|z_Z)$ multinomially distributed, the update equation for $P(y|z_Z)$ is the same as in eq. (16).

$P(v|z_Z) = \binom{V}{v} {p_{z_Z}}^v (1 - p_{z_Z})^{V-v}$ is a binomial distribution (see appendix A) with mean $p_{z_Z} \cdot V$ and shape parameter p_{z_Z}. Rating values v come form $\mathcal{V} = \{1, 2, ...|\mathcal{V}|\}$, and so $V = |\mathcal{V}|$. Update equation for parameter p_z is given by[2]:

$$p_z = \frac{\sum_y \sum_{u \in \mathcal{U}_y} P(z_Z|\ y, u, v_{u,y}) v_{u,y}}{V\ \sum_y \sum_{u \in \mathcal{U}_y} P(z_Z|\ y, u, v_{u,y})}. \tag{18}$$

4.2.4 Type I – Binomial $P(v|y, z_Z)$

If $P(v|y, z_Z)$ is binomially distributed, then parameter $p_{z,y}$ of the distribution is updated according to:

$$p_{z,y} = \frac{\sum_{u \in \mathcal{U}_y} P(z_Z|\ y, u, v_{u,y}) v_{u,y}}{V \sum_{u \in \mathcal{U}_y} P(z_Z|\ y, u, v_{u,y})}. \tag{19}$$

[2] To keep the notation readable, from now on, we will not write the explicit reference to the copy of $\mathcal{Z}_Z$ of the latent space $\mathcal{Z}$ when referring to the shape parameter p_{z_Z} corresponding to the latent class z_Z, i.e. we write p_z instead of p_{z_Z}.

4.3 Topographic initialization with SOM

It is well-known that the EM algorithm can be strongly sensitive to initialization of the model parameters. Successful application of latent-space modeling with maximum-likelihood parameter estimation via EM is therefore dependent on the initial position in the parameter space. Ideally we would like the parameters of our LCM to be initialized in a relatively fast, non-probabilistic manner, while respecting the imposed topology of latent classes in LCM. We propose to do so by running SOM on a data set $\mathcal{R}$ of user ratings across films $\mathcal{Y}$ derived from the data $\mathcal{D} = \{(u_1, y_1, v_1), ..., (u_N, y_N, v_N)\}$ of observation triplets (u_n, y_n, v_n). The number of nodes in SOM is equal to the number of latent classes, K, and the grid topology of SOM mimics the topology of latent classes in LCM induced by the channel noise. Denote by $|\mathcal{Y}|$ size of the film set $\mathcal{Y}$. For each user u that produced a rating in $\mathcal{D}$, the set $\mathcal{R}$ contains a $|\mathcal{Y}|$-dimensional vector $\mathbf{v}_u = (v_{u,1}, v_{u,2}, ..., v_{u,|\mathcal{Y}|})$ representing ratings by u across films $y \in \mathcal{Y}$. We assume that ratings $v \in \mathcal{V}$ are positive numbers. When user u does not rate film y, we set by default $v_{u,y} = 0$. Since the users typically vote only on a small fraction of the films, the rating data is usually sparse and so the distance between the codebook prototype vectors of SOM and the data points is computed only on the observed values[3] $v_{u,y} \neq 0$ in $\mathcal{R}$.

Units in the neural field of SOM are considered nodes of a grid. Each node is associated with a $|\mathcal{Y}|$-dimensional weight vector. Topologically close nodes are connected by arc of length 1. The distance between nodes i and j in the neural field is equal to the length of the shortest path from i to j. We used exponentially shrinking Gaussian neighborhood function.

After training the SOM, the user conditional latent priors $P(z|u)$ in LCM are estimated

[3] An alternative is to use the default value 0 (or e.g. mid-point in the rating scheme) in place of missing ratings and compute the distances using all the dimensions. Because of the data sparseness, this approach can introduce a significant bias towards the default rating value.

by consulting memberships of users u to clusters defined nodes $z \in \mathcal{Z} = \{1, 2, ..., K\}$ of the SOM. The binary memberships are further "softened" by the following transformation: $P(z|u) = A$, if the user u belongs to the cluster defined by the node z of SOM; $P(z|u) = (1 - A)/(K - 1)$, otherwise. The parameter A is set by postulating that if u belongs to the cluster z, $P(z|u)$ should be $B > 1$ times higher than $P(z'|u)$ for all the other z'. Hence, $A = B/(K - 1 + B)$.

The rating patterns $\rho(v, y, z) = P(v|y, z)$ in type I models are estimated by calculating for each film y and latent class z the empirical distribution of ratings for film y by the users belonging to the class z. Due to the data sparseness, we perform a smoothing of the empirical estimates by applying Laplace correction ((Jelinek, 1998)):

$$P(v|y, z) = \frac{N(v, y, z) + m}{m|\mathcal{V}| + \sum_{v' \in \mathcal{V}} N(v', y, z)}, \tag{20}$$

where m is a positive number and $N(v, y, z)$ is the number of times in the data set $\mathcal{D}$ that users belonging to the cluster (SOM center) z rated the film y by v. Usually $m = 1$, or $m = |\mathcal{V}|^{-1}$. Here we use the latter choice. The parameter m can be viewed in the Dirichlet prior interpretation for the multinomial distribution $P(v|u, z)$ as the effective number of times each rating value $v \in \mathcal{V}$ was used to rate the film y by the collection of users from class z *prior* to evaluations collected in our data set $\mathcal{D}$.

In models of type II, $\rho(v, y, z) = P(v, y|z)$ are estimated by calculating for each latent class z the empirical distribution of [rating,film] pairs (v, y) across users belonging to the cluster z. Laplace correction now reads:

$$P(v, y|z) = \frac{N(v, y, z) + m}{m|\mathcal{V}||\mathcal{Y}| + \sum_{v' \in \mathcal{V}} \sum_{y' \in \mathcal{Y}} N(v', y', z)}. \tag{21}$$

The parameter m can now be considered the effective number of times each [rating,film] pair was considered by the collection of users from class z *prior* to creating the data set

$\mathcal{D}$. Suggested value for m is now $(|\mathcal{V}||\mathcal{Y}|)^{-1}$

5 Experiments

In this section we demonstrate latent class models with grid topology of $K = 8 \times 8 = 64$ latent classes. We train models of types I and II with different distribution models for $\rho(v, y, z_Z)$, as described in Section 4.2. Specificity of the topological neighborhood, σ (see eq. (3)), was set to 0.5. Parameter B involved in smoothing memberships of users u to nodes of SOM (see section 4.3) was set to 5.

5.1 Data

We experimented with the publicly available *EachMovie dataset*[4] containing ratings for films. The data set contains ratings by $61,265$ users for 1623 films. User ratings are expressed on a 6-point scale from 0.0 to 1.0. In our experiments, the ratings are transformed to $\mathcal{V} = \{1, 2, \ldots, 6\}$. We selected a set of 100 most rated films. The number of users that rated at least one film from the selected set was $60,895$. The final number of ratings was $1,472,253$. Note that the data is still quite sparse. Out of $6,000,895$ possible ratings of 100 films by $60,895$ users, only $1,472,253$ ratings (24.5%) are observed.

5.2 Outline of the experiments

We partitioned the set of ratings into two sets – training and test sets. The training set $\mathcal{D}$ is used to train the models and visualize the data. The test set T is used for evaluation of generalization capabilities of the models *within the set of users contained in* $\mathcal{D}$[5].

[4] http://www.research.compaq.com/SRC/eachmovie/

[5] LCM studied in this paper are not *not* generative probabilistic models and hence cannot be consistently used to produce ratings for *previously unknown* users. However, they can be used to recommend

Similarly to (Breese et al., 1998; Hofmann, 2001), we applied *all but one protocol*: one randomly selected rating from each user having at least 10 ratings was assigned to the test set. The test set consists of 3.065% from all ratings, i.e. $45,136$ ratings.

The models are trained on the training set $\mathcal{D}$. After the initial SOM-based phase (see section 4.3) the cluster memberships of users are "softened" according to the scheme described in Section 4.3, models are trained with the EM algorithm (likelihood typically levelled up after 50 training iterations) and data are visualized.

We use normalized negative log likelihoods

$$NNL_{train} = -\frac{1}{|\mathcal{D}|} \sum_{(u,y,v_{uy}) \in \mathcal{D}} \log \tilde{P}(v_{uy}, y, u) \tag{22}$$

and

$$NNL_{test} = -\frac{1}{|T|} \sum_{(u,y,v_{uy}) \in T} \log \tilde{P}(v_{uy}, y, u) \tag{23}$$

of ratings on the training set $\mathcal{D}$ and test set T, respectively. Here, $\tilde{P}(v, y, u) = P(v|y, u)$ and $\tilde{P}(v, y, u) = P(v, y|u)$ for models of types I and II, respectively. Normalized negative log likelihood measures how well the probabilistic model explains the observed data. Lower values of NNL indicate better compatibility of the model with observed data.

Besides validating the probability distributions given by our models, we take a more pragmatic view and check the ability of the models to make useful recommendations on a previously unseen test set. We do so by employing five additional measures found in the literature. The measures can be divided into two categories:

- *Statistical accuracy metrics* compare the estimated and observed user ratings:
  - *mean absolute deviation* of estimated ratings $r_{u,y}$ from the observed ones $v_{u,y}$ (e.g. (Hofmann, 2001)):

unseen items (films) to users contained in the training corpus. Here we are mainly concerned with analyzing preference patterns within large collections of rating data and so, strictly speaking, we do not need generative reformulations of LCM (see e.g. (Schein et al., 2001))

$$MAD = \frac{1}{|T|} \sum_{(u,y,v_{u,y}) \in T} |r_{u,y} - v_{u,y}|, \tag{24}$$

where $r_{u,y}$ is the most probable rating under the model,

$$r_{u,y} = \arg\max_{v \in \mathcal{V}} P(v|y,u). \tag{25}$$

For model of type II, the probability $P(v|y,u)$ is computed by

$$P(v|y,u) = \frac{P(v,y|u)}{\sum_{v \in \mathcal{V}} P(v,y|u)}. \tag{26}$$

- *ratio of correctly predicted votes* (Hofmann, 2001):

  $$CPV = \frac{1}{|T|} \sum_{(u,y,v_{u,y}) \in T} E_0(u,y,v_{u,y}), \tag{27}$$

  where

  $$E_0(u,y,v_{u,y}) \begin{cases} 1 & \text{if } r_{u,y} = v_{u,y}, \\ 0 & \text{otherwise.} \end{cases} \tag{28}$$

  This measure, termed in (Hofmann, 2001) Prediction accuracy, calculates the ratio of *exactly* correct rating predictions. A drawback of CPV is that it equally penalizes situations of 'near miss', such as $v_{u,y} = 1$, $r_{u,y} = 2$, and cases of obviously wrong estimates, e.g. $v_{u,y} = 1$, $r_{u,y} = 5$. Whereas in the context of collaborative filtering the former case would still constitute a useful information, the latter case is clearly undesirable. The next category of measures, used e.g. in (Billsus, 1998), takes a more pragmatic view when assessing models used for collaborative filtering.

- *Decision-support accuracy metrics* evaluate effectivity of the method in helping the user select high-quality items. Collaborative filtering is a binary operation - user either decides to select an item, or not. When users select items with e.g. estimated rating greater than 4, it is not important whether the estimated rating is 1, 2 or 3 (Good et al., 1999).

Let T_P be the set of observations with positive *observed* ratings, i.e. observations $(u, y, v_{u,y}) \in T$ having $v_{u,y} \geq t$, where t is a threshold. T_R is the set of observations with positive *estimated* ratings, i.e. observations $(u, y, v_{u,y}) \in T$ for which $\sum_{r \geq t} P(r|y, u) > \sum_{r<t} P(r|u, y)$.

We set the threshold value t to 4. Viewing collaborative filtering as a binary classification problem (recommend/reject an item) suggests the following three measures:

– *Precision* is the percentage of recommendations that are relevant (Billsus, 1998; Soboroff and Nicholas, 1999),

$$\text{Precision} = \frac{|T_R \cap T_P|}{|T_R|}. \tag{29}$$

– *Recall* is the percentage of positive observations that will indeed be recommended (Soboroff and Nicholas, 1999),

$$\text{Recall} = \frac{|T_R \cap T_P|}{|T_P|}, \tag{30}$$

– *F-measure* is often used as it is easy to optimize either of the two previous measures separately (Billsus, 1998),

$$\text{F-measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{31}$$

6 Results

Results for five different noise models introduced in section 4.2 are presented in Table 1. To study the influence of introducing grid topology on latent classes (beneficial for the purposes of visualization), we also constructed LCM with the same number of latent classes $K = 64$,

	NNL_{train}	NNL_{test}	MAD	CPV	Precision	Recall	F
I-Bin	1.39	1.45	1.27	0.31	0.86	0.73	0.79
I-Mult	1.27	1.42	1.21	0.32	0.86	0.75	0.80
II-IndB	5.66	5.64	1.40	0.28	0.85	0.70	0.77
II-IndM	5.56	5.54	1.37	0.29	0.86	0.66	0.66
II-Mult	5.53	5.53	1.31	0.31	0.86	0.68	0.76

Table 1: Evaluation of models with grid latent class topology.

	NNL_{train}	NNL_{test}	MAD	CPV	Precision	Recall	F
I-Bin	1.25	1.46	1.21	0.32	0.87	0.76	0.81
I-Mult	0.97	2.37	1.31	0.29	0.85	0.75	0.80
II-IndB	5.14	7.76	1.76	0.24	0.85	0.56	0.67
II-IndM	4.94	10.43	1.51	0.26	0.84	0.65	0.73
II-Mult	4.90	11.25	1.40	0.27	0.85	0.66	0.74

Table 2: Evaluation of models without topology on latent classes.

but without constraining the latent classes by any topology. This situation is equivalent to setting the ‘neighborhood width’ σ to a very small number. Such topology-free models are initialized using K-means clustering. Unlike in SOM, K-means clustering does not constrain the codebook vectors by any neighborhood topological structure. Results for topology-free models are shown in Table 2.

In general, the NNL values for models of type I on training data are smaller for multinomial distribution than for the binomial distribution. For topology-free models, binomial distribution beats multinomial distribution on test data. This indicates that binomial distribution better regularizes the models with no topology by introducing less degrees of freedom (free parameters) and by imposing a unimodal structure on ordered rating values

v. The tight grid topology is a strong regularization factor in itself, enabling the more flexible multinomial rating distribution to be efficiently used. The overall advantage of using the binomial distribution is reflected by similar NNL on both training and test sets.

Results for models of type II tell us the same story. The multinomial models are more flexible than the binomial ones, and so lead to lower NNL on the training data. The price to be paid, as evidenced by the test set results, is that for topology-free models, overfitting of the training set is much worse for multinomial models than for the binomial ones.

Comparing the models as recommender systems, by concentrating on recommendation decisions themselves, and not their probabilities, leaves the models on a more levelled footing. The difference between models with topology-constrained latent classes and those with unconstrained latent space is less pronounced. On average, the models distribute the probability mass on the 'correct side' of the rating scale $\mathcal{V}$ (as evidenced by MAD and precision/recall/F values). However, constrained models are more conservative in putting too much probability mass on a particular rating based on evidence from training data. This leads to better NNL values on test data. The average absolute deviation of estimated ratings from the observed ones is in most cases in a tolerable range 1.2–1.4. CPV values are quite low, because MAD is greater than one, but as explained earlier, for purposes of collaborative filtering, the most relevant measures are the decision-support accuracy measures such as precision, recall and F-measure.

To summarize, from the practical point of view of using our models as recommender systems, introduction of topology into the latent space (desirable for visualization purposes) does not harm the models' recommendation performance. Furthermore, when evaluating the systems as probabilistic models of data, slightly better NNL values of unconstrained models on the training set are achieved only at the price of strong overfitting of the training data.

7 Visualization

In this section we visualize the "common interest patterns" found by models with latent classes organized in the grid topology. To describe the latent organization, it is convenient to label classes on the *grid* according to their "chessboard position": columns are represented by letters and rows by numbers. For example, A1 stands for the bottom left latent class on the grid.

Models of type I and II model different distributions and thus they are suitable for different type of visualization.

7.1 Models of Type II

Models of type II are appropriate in cases where $P(y|u) = \sum_{v,z_Z} P(v, y|z_Z)P(z_Z|u)$ is an interesting quantity to model, irrespective of the actual vote (Hofmann, 2001). Similarly, these models are suitable for visualizing the most probable films for each abstract class, i.e. films with largest $P(y|z) = \sum_v P(v, y|z)$. For each latent class on the grid, we present 5 most probable films.

In order to understand to what degree are the films in particular classes similar or dissimilar, in addition to film names, we show genre codes from Internet Movie Database[6]. The film genres were not explicitly used in training the models. The genres are represented by abbreviations shown in Table 3. There are 17 genres which we order into a template `[A,V,N,L,C,P,D,F,Y,H,M,U,R,S,T,W,E]`. Each film is represented a string created from the template, where we substitute '-' for the genres into which the film is not categorized. Note that one film is usually categorized into more than one genre. As an example, a film categorized as *action*, *crime* and *thriller* will be represented by `[A,-,-,-,C,-,-,-,-,-,-,-,-,-,T,-,-]`.

[6] http://www.imdb.com

Abbreviation	Genre
A	Action
V	Adventure
N	Animation
L	Classic
C	Comedy
P	Crime

Abbreviation	Genre
D	Drama
F	Family
Y	Fantasy
H	Horror
M	Musical
U	Mystery

Abbreviation	Genre
R	Romance
S	Sci-Fi
T	Thriller
W	War
E	Western

Table 3: Genre abbreviations.

Consider the *II-IndB* model as an example. Tables with genre codes and names of the 5 most probable films for each latent class are presented in Tables 4 and 5. There are clear patterns in genre codes for the most probable films associated with latent classes. For example, class A2 contains *action thrillers* and movies associated to D2 are *dramas*. Note the topological organization of the latent classes: genre patterns are similar for classes that are close to each other. For example, adjacent classes E4, E5 contain *romantic comedies* and classes A2, A3 consist of *action thrillers*.

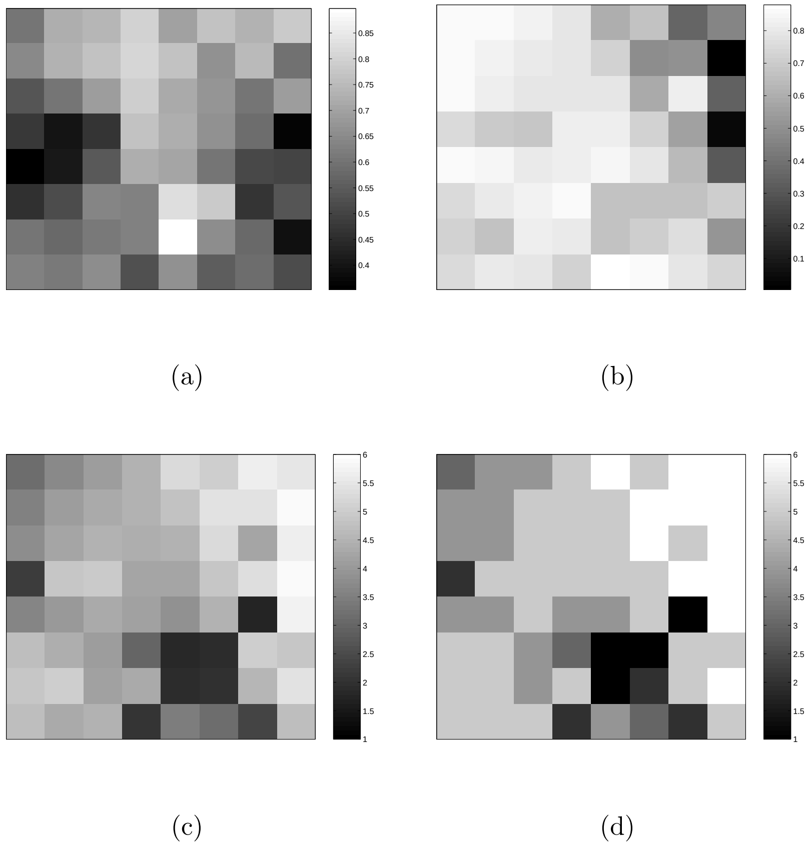
Figures 1 (a) and (b) show normalized entropies

$$H[P(y|z)] = \sum_{y\in\mathcal{Y}} P(y|z) \log_{|\mathcal{Y}|} P(y|z) \tag{32}$$

and

$$H[P(v|z)] = \sum_{v\in\mathcal{V}} P(v|z) \log_{|\mathcal{V}|} P(v|z) \tag{33}$$

of the class-conditional film and rating distributions $P(y|z)$ and $P(v|z)$, respectively. In the case of joint multinomial model for films and ratings, the rating and film distributions can be obtained by marginalization. Shown are also the means and modes of the rating distributions $P(v|z)$ (Figures 1 (c) and (b), respectively). It is interesting that the low-rating region E2–E3 and F2–F3 (mostly *action comedies and dramas*) has high film entropy



(a) (b)

(c) (d)

Figure 1: Model of type II – GTop II-IndB. Shown are the entropies $H[P(y|z)]$ (a) and $H[P(v|z)]$ (b) as well as means (c) and modes (d) of the class-conditional rating distributions $P(v|z)$.

and the high-rating region H4–H7 (mostly *non-action crime and drama*) is associated with low rating entropy. It seems that the low-rating sentiment for action films in the region E2–E3 and F2–F3 is quite far reaching: many films get captured by this region (as evidenced by the high film entropy). On the other hand, there are relatively few films in the region H4–H7 (low film entropy) and the positive rating opinion is quite unified in its concentration on high values (low rating entropy).

Such visualization plots can be utilized in several ways. As an example, consider a romance oriented TV station that have recently aired films *Sleepless nights in Seattle*, *Pretty woman* and *Ghost*. All three films were highly rated. In this situation the Table 5 can be used to select further films to air. Highly rated romantic films are concentrated in abstract classes E4–E6. It may be a good idea to select films that are highly probable under those classes. Films supported by the neighboring classes could also be selected. For example, one can pick romantic comedies *Mrs. Doubtfire*, *Four weddings and the funeral*, or thrillers *The Firm*, *The Client* and/or dramas such as *Philadelphia* or *The Piano*. On the other hand, for a well-balanced program, films from different areas of the grid should be selected in batches. Such manipulations can be extended by constructing several visualization plots based on ratings from different geographical regions, times of day, etc.

7.2 Models of Type I

Models of type I are suitable for highly interactive and selective inspection of recommendation patterns in the database. For example, given the latent class z, it is possible to visualize the rating distribution $P(v|y,z)$ for each fixed film y. By inspecting the latent-class-conditional rating distributions $P(\cdot|y,z)$, we can demonstrate that "similar" films tend to have similar rating distributions. For illustration purposes, we choose four films: two are *romantic comedies* – *Ghost* and *Pretty Woman*, one film is a *criminal horror* – *Silence of the Lambs*, the last film is a *criminal drama* – *Pulp Fiction*.

As an example, figure 2 show means of $P(v|y,z)$ for the four chosen films in the *I-Bin* model. Rating patterns for two *romantic comedies* in Figures 2(a) and 2(b) are very similar. In contrast, a very different distribution is obtained for the *criminal horror* (Figure 2(c)). Not surprisingly, rating patterns in Figure 2(c) are similar to those for the *criminal drama* in Figure 2(d).

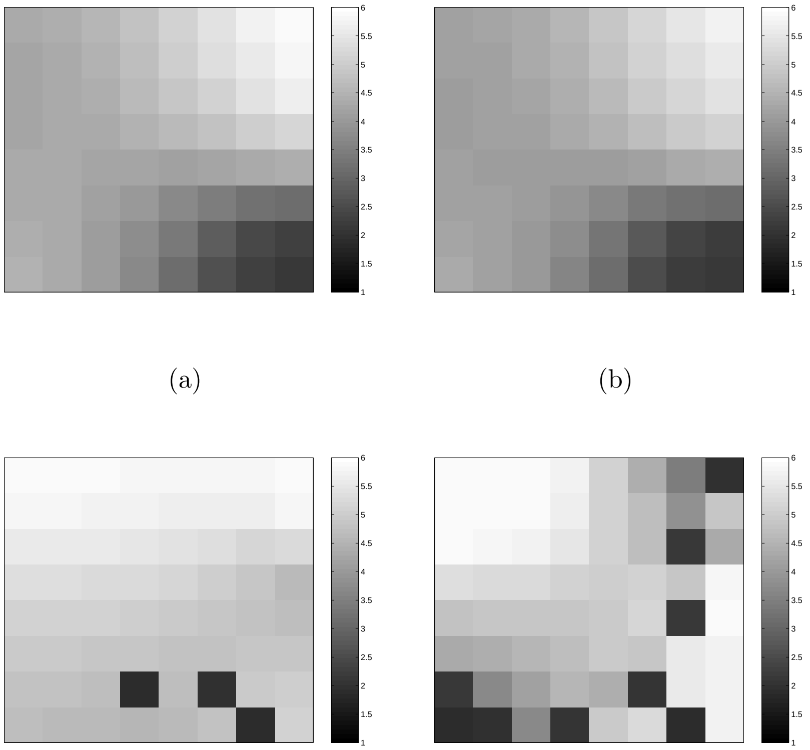
As an example of using type-I visualization plots, consider a TV station that has already selected films to air. Rating patterns of type I can help in the process of film-

----C--------S---	AV----D------ST--	A---C--------S---	------D-------T--	A------------ST--	A-----D-------T--	------D----------	AV--C-------R-T--
A---C------------	----C------------	----C--FY--------	A-----D-------T--	AV-------H---ST--	AV------Y----S---	-V----D---------E	AV-----------S---
AV-------H-U-S---	AV------------T--	------D-------T--	A-----D-------T--	AV---------U--T--	AV------------T--	--N-C--FY-M------	AV--CP--------T--
----C------------	----C-----------E	A-----D-------T--	------D-------T--	A-------------T--	AV--C-------R-T--	A-----D-------T--	A----P--Y-----T--
----C------------	A-------------T--	A---CP--Y--------	A---CP--Y--------	----C-D----------	----C--FY--------	AV--C-------R-T--	A-----D-------T--
AV------------T--	A-----D-------T--	A-----D-------T--	------D-------T--	A-------------T--	----C-D----------	-V----D---------E	A-----D--------W-
AV------Y----S---	AV------Y----S---	AV-----------S---	----C--F----R----	----C-DF---------	AV---------U--T--	------D----------	----C-D----------
AV----D------ST--	------D--H-------	A-----D-------T--	A---CP--Y--------	-VN----F--M------	AV-------H---ST--	--N----F--M-R----	------D----------
AV---P--Y---R-T--	AV--CP--------T--	AV---P--------T--	A-----D-------T--	A------------ST--	A-----D--------W-	--N-C--FY-M------	-----P---H----T--
A-----D-------T--	AV---P--------T--	-V--C------------	A-----D-------T--	----C-D-Y---R-T--	A------------ST--	AV---------U--T--	------D----------
AV---P--Y---R-T--	AV--CP--------T--	A-----D-------T--	------D-------T--	----C-------R----	------D--------W-	A----P--Y-----T--	----C-D----U--T--
A----P--Y-----T--	A-----D-------T--	A-----D-------T--	AV-------H---ST--	----C-D-----R----	----C-D----------	----C--F---------	-----P---H----T--
AV--CP--------T--	AV--C-------R-T--	--N-C--FY-M------	--N----F--M-R----	------D-------T--	AV---------U--T--	----C------------	-----PD----------
----C------------	--N-C--FY-M------	AV--C-------R-T--	A------------ST--	----C-D-Y---R-T--	-----P---H----T--	AV---P--Y---R-T--	------D------ST--
AV------Y----S---	--N----F--M-R----	AV--CP--------T--	-VN----F--M------	----C-DF---------	A-----D--------W-	AV----D------ST--	A---CPD----------
----C------------	------D----------	------D----------	AV-------H---ST--	----C-D-----R----	----C-------R----	----C-DFY--------	------D--------W-
-V--C------------	-----PD----------	-----P---H----T--	----C-D----------	----C-------R----	------D----------	A----P------R-T--	-----PD----------
-----PD----------	-V----D---------E	-----PD----------	A-------------T--	A-----D-------T--	------D----------	-----P-----U--T--	------D----------
AV---P--Y---R-T--	------D----------	------D----------	A------------ST--	----C-D-Y---R-T--	A---CPD----------	----C-D----U--T--	-----P-----U--T--
--N-C--FY-M------	-----P---H----T--	AV---------U--T--	-VN----F--M------	----C-------R----	------D--------W-	-----P---H----T--	-----P---H----T--
A----P--Y-----T--	A----P--Y-----T--	-V----D---------E	----C-D----------	----C-D-----R----	------D-----R----	----C------------	-----P-----U--T--
AV--C-------R-T--	AV--C-------R-T--	------D----------	AV-------H---ST--	----C-------R----	----C-------R----	-V--C------------	----C-D-----R----
-V----D---------E	-V----D---------E	--N-C--FY-M------	AV---------U--T--	----C-DF---------	A---CPD----------	A----P------R-T--	----CP--------T--
-----PD----------	------D----------	AV---------U--T--	A------------ST--	----C-------R----	----C-------R----	----C------------	-----PD----------
AV---P--Y---R-T--	--N-C--FY-M------	-----PD----------	A-------------T--	-----------U--T--	------D----------	----C--------S---	------D-----R----
--N-C--FY--------	A-------------T--	AV---P--------T--	----C--F---------	------D--H-------	----C--FY--------	----C-D-----R----	----C-------R----
A-------------T--	------D------ST--	------D------ST--	----C-------R----	----C-DFY--------	A---C------------	------D------ST--	------D-----R----
AV------------T--	A-----D-------T--	A-------------T--	----C-D-Y---R-T--	AV----D------ST--	AV----D------ST--	-----PD----------	-----PD----------
A-----D-------T--	--N-C--FY--------	AV-----FY--------	----C-DF---------	----C------------	A----P-------S---	------D-----R----	------D------ST--
A------------ST--	----C-------R----	A---CPD----------	------D-------T--	A----P------R-T--	----C-----------E	----CP--------T--	A----PD----------
A------------ST--	A-----D-------T--	A-------------T--	A----PD----------	A-----D--------W-	----C------------	----C-D-----R----	AV-L----Y----S---
A-----D-------T--	A------------ST--	AV------------T--	------D------ST--	----C-D-----R----	A-------------T--	------D-----R----	----CP--------T--
AV------------T--	------D----------	A-------------T--	------D----------	AV-----FY--------	----C--F---------	----C-------R----	AV-L----Y----S---
A-----D-------T--	--N-C--FY--------	A-----D-------T--	-----PD----------	A-----D-------T--	A---CP--------T--	--N-C--FY--------	--N-C--FY--------
A-----D-------T--	A-----D-------T--	----C--F---------	A-----D-------T--	----C-------R----	A-----D-------T--	----C-------R----	-----PD----------
----C-------R----	----C-------RS---	AV------------T--	------D-----R----	AV------------T--	AV------------T--	----C-D-------T--	AV-----------S---
A------------ST--	----C--F---------	A-----D-------T--	----C-------R----	A------------ST--	A------------ST--	----C-------RS---	----C--FY-M------
----C-------RS---	----C--FY-M------	AV------------T--	----C-D-----R----	A-------------T--	A-----D-------T--	A-----D-------T--	AV-L----Y----S---
A-----D-------T--	AV------------T--	----C-------R----	----CP--------T--	----C-------R----	----C-------R----	----C-------R----	------D-Y---R----
A-----D-------T--	----C-------R----	A-------------T--	----C-D-------T--	----C-------R----	A-------------T--	A------------ST--	AV-L----Y----S---

Table 4: Model of type II – GTop II-IndB: genre codes of the 5 most probable films in each latent class.

Coneheads Beverly Hills C Congo Ace Ventura: Wh Addams Family V	Waterworld Addams Family V Cliffhanger City Slickers I The Specialist	Demolition Man The Santa Claus Disclosure The Net The Mask	Disclosure The Net Outbreak The Firm The Mask	Terminator 2: J Jurassic Park The Fugitive Speed Forrest Gump	The Net Stargate Cliffhanger True Lies The Santa Claus	Apollo 13 Dances With Wol Aladdin Clear and Prese True Lies	True Lies Star Trek: Gene Die Hard: With Batman (1989) Clear and Prese
Cliffhanger Stargate Waterworld Batman Forever The Net	Outbreak Stargate Interview with Die Hard: With Goldeneye	Crimson Tide Star Trek: Gene Outbreak Goldeneye Dumb and Dumber	The Firm While You Were The Mask Crimson Tide Outbreak	Speed Mrs. Doubtfire The Lion King Terminator 2: J Ghost	Forrest Gump The Fugitive Jurassic Park Braveheart Terminator 2: J	Dances With Wol Apollo 13 Beauty and the Aladdin The Fugitive	Braveheart Forrest Gump The Shawshank R The Silence of Apollo 13
Batman Forever Batman (1989) Die Hard: With Ace Ventura: Pe Stargate	Die Hard: With Clear and Prese True Lies Aladdin Beauty and the	Clear and Prese Crimson Tide Aladdin True Lies Die Hard: With	The Firm Jurassic Park Beauty and the Terminator 2: J The Lion King	Pretty Woman Sleepless in Se The Firm Ghost Mrs. Doubtfire	Schindler’s Lis Forrest Gump The Fugitive The Silence of Braveheart	Batman (1989) Home Alone Addams Family V Batman Forever Waterworld	Seven The Silence of Pulp Fiction 12 Monkeys Get Shorty
Ace Ventura: Pe Dumb and Dumber Pulp Fiction Batman Forever Aladdin	Apollo 13 Pulp Fiction Dances With Wol The Shawshank R The Silence of	The Shawshank R The Silence of Pulp Fiction Apollo 13 The Fugitive	Jurassic Park Forrest Gump Speed Terminator 2: J The Lion King	Sleepless in Se Four Weddings a In the Line of Ghost Pretty Woman	Four Weddings a Philadelphia Quiz Show Get Shorty Schindler’s Lis	Babe Natural Born Ki The Usual Suspe Seven The Silence of	Schindler’s Lis Pulp Fiction The Shawshank R The Usual Suspe The Silence of
Batman (1989) True Lies Dances With Wol Pulp Fiction Batman Forever	Batman (1989) True Lies Dances With Wol Apollo 13 Aladdin	Dances With Wol Apollo 13 Aladdin The Fugitive Pulp Fiction	Forrest Gump Jurassic Park The Fugitive Terminator 2: J Speed	Sleepless in Se Dave Mrs. Doubtfire Pretty Woman The Client	The Piano Four Weddings a Get Shorty Clueless Quiz Show	Ace Ventura: Wh Dumb and Dumber Natural Born Ki Ace Ventura: Pe Coneheads	The Usual Suspe Sense and Sensi Fargo Dead Man Walkin Leaving Las Veg
Toy Story Broken Arrow Mission: Imposs Twister Independence Da	Broken Arrow 12 Monkeys Twister Toy Story The Birdcage	Goldeneye 12 Monkeys Broken Arrow Jumanji Get Shorty	Home Alone Pretty Woman Ghost Mrs. Doubtfire The Firm	Interview with Babe Waterworld Addams Family V Natural Born Ki	The Santa Claus Beverly Hills C Waterworld Judge Dredd City Slickers I	Sense and Sensi 12 Monkeys Dead Man Walkin Leaving Las Veg Fargo	The Birdcage Leaving Las Veg Dead Man Walkin 12 Monkeys Heat
Independence Da Twister Mission: Imposs The Rock Eraser	The Rock Independence Da Mr. Holland’s O Toy Story Twister	Broken Arrow Mission: Imposs The River Wild Twister Father of the B	Heat 12 Monkeys Mr. Holland’s O Dead Man Walkin The Rock	Braveheart The American Pr Jumanji Crimson Tide French Kiss	Happy Gilmore Broken Arrow Father of the B Rumble in the B Twister	Sense and Sensi Leaving Las Veg The Birdcage Toy Story The Truth about	Star Wars Fargo Return of the J Toy Story Dead Man Walkin
Grumpier Old Me Independence Da The Nutty Profe Eraser Twister	The Nutty Profe Father of the B Willy Wonka and Executive Decis Grumpier Old Me	Executive Decis Eraser Mission: Imposs Sabrina The River Wild	Leaving Las Veg The Birdcage Sense and Sensi Fargo The Cable Guy	Mission: Imposs Independence Da The River Wild Sabrina The Truth about	Mission: Imposs Independence Da Twister Sabrina The River Wild	The Cable Guy The Nutty Profe Twister Grumpier Old Me Independence Da	Star Trek: Firs Willy Wonka and Return of the J Phenomenon Star Wars

Table 5: Model of type II – GTop II-IndB: names of the 5 most probable films for each latent class.



(a)

(b)

(c)

(d)

Figure 2: Model of type I – GTop I-Bin. Shown are the means of the class-and-film-conditional rating distributions $P(v|y,z)$ for the following films: Ghost (a) Pretty Woman (b) The Silence of the Lambs (c) and Pulp Fiction (d).

preview selection. One can compare rating pattern of the currently broadcasted film with rating patterns of films scheduled to be aired in the near future. It may be appropriate to select for preview a film with rating pattern similar to that of the currently aired film.

8 Discussion

In general, model-based approaches to data visualization introduce some sort of a-priori topological structure on the modeling elements of the system. The topological structure (e.g. neighborhood organization on a square grid) reflects the nature of the visualization space (e.g. computer screen). For example, SOM can be viewed as a constrained vector quantization, where quantization centers are constrained to respect the neighborhood structure (e.g. 2-dimensional grid) imposed a-priori before fitting the data. Its analog on the side of probabilistic modeling of data, Generative Topographic Mapping (GTM) (Bishop, 1998), constrains means of local Gaussian noise models (corresponding to codebook vectors in SOM) to lie on a smooth two-dimensional manifold in the high dimensional data space. Naturally, one can often obtain better modeling capabilities with less complex unconstrained models[7] (e.g. using a smaller number of freely movable codebook vectors in vector quantization, or smaller unconstrained mixtures of Gaussians in case of Gaussian mixture modeling), but at the price of loosing natural visualization predispositions of appropriately constrained models.

When the task we are facing is, for example, building a good density model for a given data set of vectorial data, without any concern for data visualization, then a suitable approach may be to use e.g. mixtures a Gaussians, with appropriately chosen number of mixture components using a model selection technique. On the other hand, for model based visualization of vectorial data, we may use many Gaussian components, but constrain them with a tight two-dimensional grid neighborhood structure. Such a constrained mixture of Gaussians may not be able to compete with appropriately constructed (probably smaller) unconstrained mixture of Gaussians on the grounds of density modeling, but it is suitable for data visualization and importantly, the tight grid topology prevents constrained models with many components (suitable for high-quality visualization) from excessively overfitting the data. The issue of data explanation vs. data prediction is covered e.g. in (Ripley, 1998).

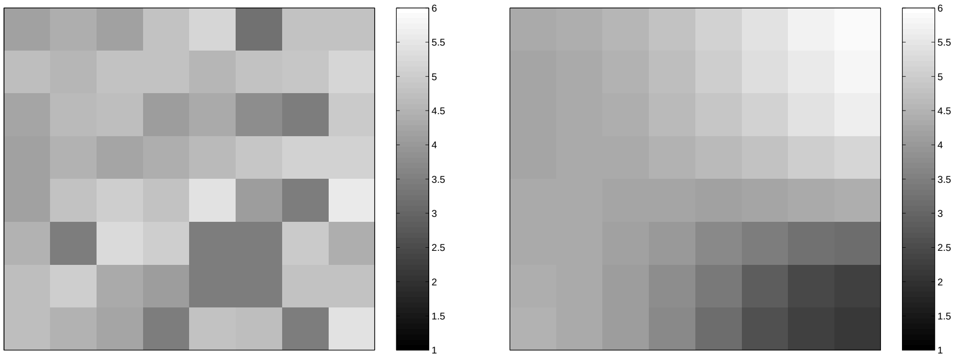
[7] When data distribution has a known structure (e.g. noisy 2-dimensional manifold embedded in a high dimensional space), appropriately constrained models (e.g. GTM with 2-dimensional latent space) will be better modeling candidates.

	NNL_{train}	NNL_{test}	MAD	CPV	Precision	Recall	F
$\sigma = 0.25$	1.34	1.44	1.25	0.31	0.86	0.75	0.80
$\sigma = 1.0$	1.47	1.50	1.32	0.29	0.84	0.71	0.77
$\sigma = 1.5$	1.53	1.54	1.33	0.29	0.84	0.71	0.77
$B = 2$	1.39	1.45	1.26	0.31	0.86	0.73	0.79
$B = 10$	1.39	1.45	1.27	0.31	0.86	0.74	0.79

Table 6: Evaluation of I-Bin with varying parameters.

Comparing constrained and unconstrained LCM for collaborative filtering as probabilistic models of data (see NNL values in tables 1, 2), we see that there is almost no overfitting in constrained models, moreover, the *test set* NNL values of constrained models are not too far from the *training set* NNL values of unconstrained models that overfit the training data. Hence, in our case, the constrained models combine a potential for model based data visualization with good quality probabilistic modeling (within the realm of LCM). They can be utilized, even though we stress that this is not the primary goal of their construction, in producing practical recommendations, as evidenced by the Precision, Recall and F-measures.

Some of the model construction parameters, such as the neighborhood width σ and smoothing parameter B in the SOM initialization phase, were set in the experiments to fixed values. It is appropriate to study sensitivity of the models to changes in those parameters. Table 6 shows results for models I-Bin with varying parameters σ and B. Models with varying σ were trained with B set to 5. Outcomes show that the smaller the σ is, i.e. the tighter is the neighborhood, the better the models fit the data. Model performance is quite stable with a slight tendency of broader neighborhoods to result in more rigid models loosing precision of the fit. Models with varying B have σ set to 0.5. Model performance is stable with respect to varying values of parameter B.



(a) (b)

Figure 3: **(a)** Means of $P(v|y, z_Z)$ for film *Ghost* after SOM-based initialization. **(b)** Means of $P(v|y, z_Z)$ for film *Ghost* after training.

We also give an example of how the rating distributions evolve during training, starting from the initial SOM-based phase. We inspected evolution of $P(z_Y|u)$ during the training process for several users u. Initially, the probability of one latent class (winner node from SOM) is B times greater than those of the other classes. As an example consider I-Bin model and user no. 2 from the training set. The mode of $P(z_Y|u)$ has moved from latent class I5 (winner in SOM) to H1. Furthermore, contrary to the SOM-based initial flat distribution $P(z_Y|u)$, except for one mode, after training, the highly probable classes (given the user u) are located in the upper-left corner of the grid around H1 (I1-I3, H1-H3, G-1-G2 and F1-F2).

Rating distributions $P(v|y, z_Z)$ in class-I models evolve from initially scattered probability mass into more coherent, topologically organized distributions. An example of evolution of means of rating distributions $P(v|y, z_Z)$ for movie *Ghost* in the I-Bin model can be seen in figure 3.

We experimented with initialising the constrained models using topology-free K-means clustering. Compared with SOM-based initialization, the NNL results on both training and test sets were consistently slightly worse. Shared topology between neural field in SOM and latent classes in LCM results in better initialization of the E-M algorithm that (by its nature) is bound to converge to a local optimum in the likelihood landscape.

As mentioned in the introduction, while much work has focused on designing competitive algorithms for rating predictions, few attempts have been made to implement systems for understanding and visualization of principal user preference patterns buried inside large and sparse collections of collaborative filtering data. Closest to our system is the *Shared Wisdom through the Amalgamation of Many Interpretations* (SWAMI) system by (Fisher et al., 2000). SWAMI is an interesting framework for building and studying collaborative filtering systems. It contains a visualization component enabling e.g. views of movie-to-movie correlation matrix based on user votes, or film-by-user matrix. Explanatory power of such figures is, however, greatly hampered by the huge dimensionality of matrices involved, as well as sparsity in collaborative filtering data collections.

The system supports simple histogram plots showing e.g. distribution of rating values across the rating dataset, number of movies with at least n votes (for varying n), etc. Histograms for individual movies, such as distribution of rating values for a selected movie, are also possible. Such plots are helpful in getting a feeling for the collaborative filtering data at hand either on a very global level (e.g.distribution of rating values across the set), or on detailed levels of individual films/users. However, to unveil interesting hidden rating patterns buried inside the set, one must operate on 'intermediate' levels of film/user groupings. To this end, the authors of SWAMI applied PCA analysis using movie-to-movie and user-to-user correlation matrices. Attempts with user-to-user correlation matrix had to be abandoned because of large dimensionality of the correlation matrix. Using full *EachMovie* dataset, they found (after applying several necessary noise reduction techniques) six groups of films corresponding to six most dominant eigenvectors of the movie-to-movie correlation matrix based on user ratings. As an example, we site the first three film groupings: *'highbrow to lowbrow'* (The Postman, Mighty Aphrodite, Richard III,...), *'classics to contemporary'* (The Bridge on the River Kwai, 20,000 Leagues Under the Sea, The Great Escape, ...), *'girls' movies to boy's movies'* (Black Beauty, A Little Princess, How to Make an American Quilt, ...). While interesting in themselves, the problem of such 'eigenfilm/eigentaste' approaches (see also (Goldberg et al., 2001)) is that the groupings are not results of a principled modeling approach in which formation of such groups is an integral part of the rating model. So it is not clear what the groupings represent and how they relate to each other. Film/user groupings in our topologically constrained LCM follow naturally from a principled model formulation aimed at explaining rating data using appropriate noise distributions. Moreover, since our visualizations are based on probabilistic models of data, many refinements leading to useful additional information are naturally possible, for example plots of means, modes, and entropies of rating distributions.

9 Conclusions

We have developed topographic versions of two latent class models for collaborative filtering. Topographic grid organization of the latent space enables us to better understand hidden patterns in large and sparse rating databases. The preference patterns can be inspected in a variety of ways. Our system is a probabilistic model of the data, and so we can readily use tools from probability and information theories to interpret and visualize trends captured by the abstract latent classes. In this respect, we have by no means exhausted the full range of possibilities. For example, besides means presented in the paper we could use modes, entropies and mutual information to detect dependencies between ratings and films in the models. Topographic organization of the latent classes makes orientation in those trends easier and more systematic.

We used different distribution models to account for (class conditional) variations in user ratings. We found that multinomial distribution is adequate if the model is regularized by the tight grid topology on the latent space. For LCM with a large number of latent classes, binomial distribution may be more appropriate, since it adds an extra degree of regularization by having fewer free parameters and by imposing a unimodal structure on ordered set of rating values.

We demonstrated our system on a large collection of user ratings for films. In the future we plan to experiment with different methods of initialization and to study variants of EM-algorithm presented in the literature (e.g. (Hofmann and Puzicha, 1999)) in order to speed up the training process. Also, the probabilistic character of the model may allow for a principled *interactive* construction of a hierarchy of topographic latent class models (see e.g. (Tino and Nabney, 2002)), where we would be able to interactively "zoom in" into interesting user groups and rating patterns.

Acknowledgements

The work was partially supported by Slovak Science Grant Agency VEGA, project No. 1/0162/03.

References

Billsus, D. & Pazzani, M. (1998). Learning Collaborative Information Filters. In *Proceedings of the International Conference on Machine Learning.* Morgan Kaufmann Publishers. Madison. 46–54.
`http://www.ics.uci.edu/ pazzani/Publications/MLC98.pdf`

Bishop, C.M., Svensén, M., & Williams, C.K.I. (1998). GTM: The generative topographic mapping. *Neural Computation*, 10(1):215–235.

Breese, J.S., Heckerman, D., & Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. *Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence*, pages 43–52.

Buhmann, J. & Kuhnel, H. (1993). Vector quantization with complexity costs. *IEEE Transactions on Information Theory*, 39(4):1133–1145.

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, 39(1):1–38.

Fisher, D., Hildrum, K., Hong, J., Newman, M., Thomas, M. & Vuduc, R. (2000). SWAMI: a framework for collaborative filtering algorithm development and evaluation. *SIGIR 2000*: Athens, Greece. 366-368. `http://guir.cs.berkeley.edu/projects/swami/swami-paper/paper.html`

Goldberg, D., Roeder, T., Gupta, D., & Perkins, Ch. (2001). Eigentaste: A Constant Time Collaborative Filtering Algorithm. *Information Retrieval Journal*, 4(2): 133–151.

Good, N., Schafer, J.B., Konstan, J., Borchers, A., Sarwar, B., Herlocker, J., & Riedl, J. (1999). Combining Collaborative Filtering with Personal Agents for Better Recommendations. *Proceedings of the 1999 Conference of the American Association of Artifical Intelligence* (AAAI-99). pp 439-446.

Hofmann, T. & Buhmann, J.M. (1998). Competetive learning algorithms for robust vector quantization. *IEEE Transactions on Signal Processing*, 46(6), 1665–1675.

Hofmann, T. (2000). Probmap - a probabilistic approach for mapping large document collections. *Joumal for Intelligent Data Analysis*, 4:149-164.

Hofmann, T. (2001). Learning what people (don't) want. In L. De Raedt & P. Flach, editors, *12th European Conference on Machine Learning (ECML)*, pages 214–225. Springer.

Hofmann, T., & Puzicha, J. (1999). Latent Class Models for Collaborative Filtering. *Proceedings of the International Joint Conference in Artificial Intelligence*. 688–693.

Igo Jr., F.J., Brand, M., Wittenburg, K., Wong, D. & Azuma, S. (2002). Multidimensional Visualization for Collaborative Filtering Recommender Systems. Technical Report *TR20002-39*, Mitsubishi Electric Research Laboratories, Cambridge, MA.

Jelinek, F. (19998). Statistical Methods for Speech Recognition (Language, Speech, and Communication). *MIT Press (Boston, MA)*.

Kabán, A. & Girolami, M. (2001). A combined latent class and trait model for the analysis and visualization of discrete data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(8):859–872.

Kohonen, T. (1982). Self–organizing formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1479.

Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., & Reidl, J. (1997). Grouplens: Applying collaborative filtering to Usenet news. *Communications of the ACM*, 40(3):77–87.

Luttrell, S.P. (1989). Hierarchical vector quantization. *IEE Proceedings*, 136:405–413.

Ripley, B. (1998). Statistical Theories of Model Fitting. In Ch. Bishop, editor, *Neural Networks and Machine Learning* (NATO ASI series III, Computer and System Sciences), pages 3–26. Springer Verlag.

Schein, A., Popescul, A., Ungar, L.H., & Pennock, D.M. (2001). Generative models for cold-start recommendations. Workshop on Recommender Systems at the 24th Annual International ACM SIGIR Conference, New Orleans, LA.

Soboroff, I.M., & Nicholas, C.K. (1999). Combining content and collaboration in text filtering. *In Thorsten Joachims, editor, Proceedings of the IJCAI'99 Workshop on Machine Learning in Information Filtering*, 86–91.

Tiňo, P. & Nabney, I. (2002). Hierarchical GTM: Constructing localized non-linear projection manifolds in a principled way. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):639–656.

Appendix

A Binomial Distribution

Distributions from the exponential family have the form

$$p_G(v|\theta) = \exp\{\theta v - G(\theta)\} p_0(v),$$

where G is the cumulant function. Binomial distribution

$$f(v) = \binom{V}{v} p^v (1-p)^{V-v} \tag{34}$$

is a member of the exponential family and so can be rewritten as

$$f(v) = \exp\{v \log \frac{p}{1-p} + V \log(1-p) + \log \binom{V}{v}\},$$

where

$$\theta = \log \frac{p}{1-p}, \qquad \text{and} \qquad p = \frac{e^\theta}{1+e^\theta} \qquad \text{and} \qquad 1-p = \frac{1}{1+e^\theta}.$$

Cumulant function is then:

$$G(\theta) = V \log(1+e^\theta),$$

the link function is:

$$g(\theta) = G'(\theta) = V \frac{e^\theta}{1+e^\theta} = Vp,$$

and

$$G''(\theta) = V \frac{e^\theta}{(1+e^\theta)^2}.$$

B E-step

B.1 E-step for Type I Models

We estimate

$$P(z_Y|\ y,u,v) = \frac{P(v|\ y,u,z_Y)P(z_Y|\ y,u)}{\sum_{z'_Y} P(v|\ y,u,z'_Y)P(z'_Y|\ y,u)}.$$

By model assumptions, $P(z_Y|\ y,u) = P(z_Y|u)$ and

$$\begin{aligned} P(v|\ y,u,z_Y) &= \sum_{z_Z} P(v, z_Z|\ y,u,z_Y) \\ &= \sum_{z_Z} P(v|\ y,u,z_Y,z_Z)P(z_Z|y,u,z_Y) \\ &= \sum_{z_Z} P(v|\ y,z_Z)P(z_Z|z_Y), \end{aligned}$$

and so

$$P(z_Y|\ y,u,v) = \frac{P(z_Y|u)\sum_{z_Z}\rho(v,y,z_Z)P(z_Z|z_Y)}{\sum_{z'_Y} P(z'_Y|u)\sum_{z_Z}\rho(v,y,z_Z)P(z_Z|z'_Y)}. \tag{35}$$

Analogously,

$$P(z_Z|\ y,u,v) = \frac{P(v|\ y,u,z_Z)P(z_Z|\ y,u)}{\sum_{z'_Z} P(v|\ y,u,z'_Z)P(z'_Z|\ y,u)},$$

where $P(v|\ y,u,z_Z) = P(v|\ y,z_Z)$ and

$$\begin{aligned} P(z_Z|\ y,u) &= \sum_{z_Y} P(z_Z, z_Y|\ y,u) \\ &= \sum_{z_Y} P(z_Z|\ z_Y,y,u)P(z_Y|\ y,u) \\ &= \sum_{z_Y} P(z_Z|z_Y)P(z_Y|u). \end{aligned}$$

Hence,

$$P(z_Z|\ y,u,v) = \frac{\rho(v,y,z_Z)\sum_{z_Y} P(z_Z|z_Y)P(z_Y|u)}{\sum_{z'_Z}\rho(v,y,z'_Z)\sum_{z_Y} P(z'_Z|z_Y)P(z_Y|u)}. \tag{36}$$

B.2 E-step for Type II Models

We write

$$P(z_Y|\ y,u,v) = \frac{P(v,y|\ u,z_Y)P(z_Y|\ u)}{\sum_{z'_Y} P(v,y|\ u,z'_Y)P(z'_Y|\ u)},$$

where

$$\begin{aligned} P(v,y|\ u,z_Y) &= \sum_{z_Z} P(v,y,z_Z|\ u,z_Y) \\ &= \sum_{z_Z} P(v,y|\ u,z_Y,z_Z)P(z_Z|u,z_Y) \\ &= \sum_{z_Z} P(v,y|\ z_Z)P(z_Z|z_Y), \end{aligned}$$

and so $P(z_Y|\ y,u,v)$ is updated as prescribed by eq. (35).

Analogously,

$$P(z_Z|\ y,u,v) = \frac{P(v,y|\ u,z_Z)P(z_Z|\ u)}{\sum_{z'_Z} P(v|\ y,u,z'_Z)P(z'_Z|\ y,u)},$$

where $P(v,y|\ u,z_Z) = P(v,y|\ z_Z)$ and

$$\begin{aligned} P(z_Z|\ u) &= \sum_{z_Y} P(z_Z,z_Y|\ u) \\ &= \sum_{z_Y} P(z_Z|\ z_Y,u)P(z_Y|\ u) \\ &= \sum_{z_Y} P(z_Z|z_Y)P(z_Y|u). \end{aligned}$$

Hence, (36) is the update equation for $P(z_Z|\ y,u,v)$.

C Derivation of Parameter Estimates for M-step

To derive update equations for the model free parameters $P(z_Y|u)$ and $\rho(v,y,z_Z)$ in the Expectation-Maximization (EM) framework (Dempster, 1977), we introduce (latent) indicator variables $\delta^{u,y,v}_{z_Y,z_Z}$ for couples of latent classes $(z_Y,z_Z) \in \mathcal{Z}_Y \times \mathcal{Z}_Z$: $\delta^{u,y,v}_{z_Y,z_Z} = 1$, if in the process of generating the rating v for the film y by the user u, the latent class z_Y was first selected and then transmitted through the noisy channel as the class z_Z; otherwise $\delta^{u,y,v}_{z_Y,z_Z} = 0$. The complete data log-likelihood reads:

$$\mathcal{L}_C = \sum_{u \in \mathcal{U}} \sum_{y \in \mathcal{Y}_u} \sum_{z_Y \in \mathcal{Z}_Y} \sum_{z_Z \in \mathcal{Z}_Z} \delta^{u,y,v_{u,y}}_{z_Y,z_Z} \log\{\rho(v_{u,y}, y, z_Z) P(z_Z|z_Y) P(z_Y|u)\}. \tag{37}$$

It will be more convenient to represent $\mathcal{L}_C$ as

$$\begin{aligned} \mathcal{L}_C = \sum_{n=1}^{N} \sum_{u \in \mathcal{U}} \Delta^{u_n}_{u} \sum_{y \in \mathcal{Y}} \Delta^{y_n}_{y} \sum_{v \in \mathcal{V}} \Delta^{v_n}_{v} \sum_{z_Y \in \mathcal{Z}_Y} \sum_{z_Z \in \mathcal{Z}_Z} \delta^{u,y,v}_{z_Y,z_Z} \\ \log\{\rho(v, y, z_Z) P(z_Z|z_Y) P(z_Y|u)\}, \end{aligned} \tag{38}$$

where $\Delta^{\alpha}_{a} = 1$ if and only if $a = \alpha$, otherwise $\Delta^{\alpha}_{a} = 0$. We also write

$$\Delta^{\alpha_1,\ldots,\alpha_m}_{a_1,\ldots,a_m} = \prod_{i=1}^{m} \Delta^{\alpha_i}_{a_i}.$$

C.1 M-step for Type I and II – Multinomial $\rho(v, y, z_Z)$

For multinomially distributed $\rho(v, y, z_Z)$, the M-step equations maximize the expected[8] complete data log-likelihood, extended with Lagrange multiplier terms to account for proper normalization:

$$\begin{aligned} < \mathcal{L}_C > \;=\; & \sum_{n=1}^{N} \sum_{u,y,v} \Delta^{u_n,y_n,v_n}_{u,y,v} \sum_{z_Y,z_Z} P(z_Y, z_Z|\; y, u, v) \\ & \qquad \{\log \rho(v, y, z_Z) + \log P(z_Z|z_Y) + \log P(z_Y|u)\} \\ & + \sum_{u} \lambda_u \left(\sum_{z_Y} P(z_Y|u) - 1 \right) \\ & + T(\rho), \end{aligned} \tag{39}$$

where

$$\begin{aligned} T(\rho) &= \sum_{y,z_Z} \lambda_{y,z_Z} \left(\sum_{v} P(v|\; y, z_Z) - 1 \right) \qquad \text{[type I]} \\ T(\rho) &= \sum_{z_Z} \lambda_{z_Z} \left(\sum_{v,y} P(v, y|\; z_Z) - 1 \right) \qquad \text{[type II].} \end{aligned} \tag{40}$$

[8] with respect to posterior distribution of hidden variables $\delta^{u,y,v}_{z_Y,z_Z}$ on $\mathcal{Z}_Y \times \mathcal{Z}_Z$

By setting $\frac{\partial < \mathcal{L}_C >}{\partial \rho(v,y,z_Z)} = 0$, determining the corresponding Lagrange multiplier, and realizing that $\sum_{z_Y} P(z_Y, z_Z|\ y,u,v) = P(z_Z|\ y,u,v)$, we arrive at the update equation for $\rho(v,y,z_Z)$:

$$\rho(v,y,z_Z) \;=\; P(v|\ y,z_Z) = \frac{\sum_{u\in\mathcal{U}_{v,y}} P(z_Z|\ y,u,v)}{\sum_{v'}\sum_{u\in\mathcal{U}_{v',y}} P(z_Z|\ y,u,v')} \qquad \text{[type I]} \qquad (41)$$

$$\rho(v,y,z_Z) \;=\; P(v,y|\ z_Z) = \frac{\sum_{u\in\mathcal{U}_{v,y}} P(z_Z|\ y,u,v)}{\sum_{v',y'}\sum_{u\in\mathcal{U}_{v',y'}} P(z_Z|\ y',u,v')} \qquad \text{[type II]} \qquad (42)$$

where $\mathcal{U}_{v,y}$ is the set of users that evaluated film y with rating v,

$$\mathcal{U}_{v,y} = \{u \in \mathcal{U}|\ (u,y,v) \in \mathcal{D}\}. \qquad (43)$$

Update equation for $P(z_Y|u)$ is derived by setting $\frac{\partial < \mathcal{L}_C >}{\partial P(z_Y|u)} = 0$, determining the Lagrange multiplier λ_u, and realizing that $\sum_{z_Z} P(z_Y, z_Z|\ y,u,v_{u,y}) = P(z_Y|\ y,u,v_{u,y})$ and $\sum_{y\in\mathcal{Y}_u}\sum_{z'_Y} P(z'_Y|\ y,u,v) = |\mathcal{Y}_u|$, where $|\mathcal{Y}_u|$ is the number of films evaluated by the user u:

$$P(z_Y|u) = \frac{\sum_{y\in\mathcal{Y}_u} P(z_Y|\ y,u,v_{u,y})}{|\mathcal{Y}_u|}. \qquad (44)$$

C.2 M-step for Type II with Conditional Independence

For $P(v,y|z_Z) = P(v|z_Z)P(y|z_Z)$, with both $P(y|z_Z)$ and $P(v|z_Z)$ multinomially distributed, the expected complete data log likelihood can be expressed as

$$\begin{aligned} < \mathcal{L}_C > \;=\; & \sum_{n=1}^{N}\sum_{u,y,v} \Delta_{u,y,v}^{u_n,y_n,v_n} \sum_{z_Y,z_Z} P(z_Y, z_Z|\ y,u,v) \\ & \{\log P(v|z_Z) + \log P(y|z_Z) \\ & + \log P(z_Z|z_Y) + \log P(z_Y|u)\} \\ & + \sum_u \lambda_u \left(\sum_{z_Y} P(z_Y|u) - 1 \right) \\ & + \sum_{z_Z} \lambda_{z_Z}^{v} \left(\sum_v P(v|z_Z) - 1 \right) \end{aligned}$$

$$+\sum_{z_Z} \lambda^y_{z_Z} \left(\sum_y P(y|z_Z) - 1 \right) \tag{45}$$

Maximizing $< \mathcal{L}_C >$ with respect to $P(v|z_Z)$:

$$\frac{\partial < \mathcal{L}_C >}{\partial P(v|z_Z)} = \sum_y \sum_{u \in \mathcal{U}_{y,v}} \sum_{z_Y} P(z_Y, z_Z|\ y, u, v_{u,y}) \frac{1}{P(v|z_Z)} + \lambda^v_{z_Z} = 0, \tag{46}$$

where U_{yv} is a set of users that rated film y with rating v. Then

$$P(v|z_Z) = -\frac{\sum_y \sum_{u \in \mathcal{U}_{y,v}} P(z_Z|\ y, u, v)}{\lambda^v_{z_Z}} \tag{47}$$

Substituting $P(v|z_Z)$ back to the constraint we get

$$\lambda^v_{z_Z} = -\sum_v \sum_y \sum_{u \in \mathcal{U}_{y,v}} P(z_Z|\ y, u, v). \tag{48}$$

and so

$$P(v|z_Z) = \frac{\sum_y \sum_{u \in \mathcal{U}_{y,v}} P(z_Z|\ y, u, v)}{\sum_{v'} \sum_y \sum_{u \in \mathcal{U}_{yv'}} P(z_Z|\ y, u, v')}. \tag{49}$$

Maximizing $< \mathcal{L}_C >$ with respect to $P(y|z_Z)$:

$$\frac{\partial < \mathcal{L}_C >}{\partial P(y|z_Z)} = \sum_{u \in \mathcal{U}_y} \sum_{z_Y} P(z_Y, z_Z|\ y, u, v_{u,y}) \frac{1}{P(y|z_Z)} + \lambda_{z_Z} = 0, \tag{50}$$

where U_y is a set of users that rated film y. Then

$$P(y|z_Z) = -\frac{\sum_{u \in \mathcal{U}_y} P(z_Z|\ y, u, v_{u,y})}{\lambda_{z_Z}}. \tag{51}$$

Similarly to $P(v|z_Z)$, we get

$$P(y|z_Z) = \frac{\sum_{u \in \mathcal{U}_y} P(z_Z|\ y, u, v_{u,y})}{\sum_{y'} \sum_{u \in \mathcal{U}_{y'}} P(z_Z|\ y', u, v_{u,y'})}. \tag{52}$$

C.3 M-step for Type II – Binomial $P(v|z_Z)$

Now

$$P(v|z_Z) = \exp\{v\ \theta_z - G(\theta_z) + \log p_0(v)\}.$$

The expected complete data log likelihood for model of type II is:

$$\begin{aligned}
< \mathcal{L}_C > \;=\;& \sum_{n=1}^{N}\sum_{u,y,v}\Delta_{u,y,v}^{u_n,y_n,v_n}\sum_{z_Y,z_Z} P(z_Y,z_Z|\;y,u,v)\\
& \{[v\;\theta_z - G(\theta_z) + \log p_0(v)] + \log P(y|z_Z)\\
& + \log P(z_Z|z_Y) + \log P(z_Y|u)\}\\
& + \sum_u \lambda_u \left(\sum_{z_Y} P(z_Y|u) - 1\right)\\
& + \sum_{z_Z} \lambda_{z_Z}\left(\sum_y P(y|z_Z) - 1\right) \qquad (53)
\end{aligned}$$

Maximizing $< \mathcal{L}_C >$ with respect to the parameter θ_z amounts to

$$\frac{\partial < \mathcal{L}_C >}{\partial\;\theta_z} = \sum_y \sum_{u\in\mathcal{U}_y}\sum_{z_Y} P(z_Y,z_Z|\;y,u,v_{u,y})[v_{u,y} - g(\theta_z)] = 0. \qquad (54)$$

After realizing that $\sum_{z_Y} P(z_Y,z_Z|\;y,u,v_{u,y}) = P(z_Z|\;y,u,v_{u,y})$ and substituting for $g(\theta_z)$ we get:

$$\begin{aligned}
\sum_y\sum_{u\in\mathcal{U}_y} P(z_Z|\;y,u,v_{u,y})[v_{u,y} - V\frac{e^{\theta_z}}{1+e^{\theta_z}}] &= 0\\
\sum_y\sum_{u\in\mathcal{U}_y} P(z_Z|\;y,u,v_{u,y})v_{u,y} &= V\frac{e^{\theta_z}}{1+e^{\theta_z}}\sum_y\sum_{u\in\mathcal{U}_y} P(z_Z|\;y,u,v_{u,y})\\
\frac{\sum_y\sum_{u\in\mathcal{U}_y} P(z_Z|\;y,u,v_{u,y})v_{u,y}}{V\;\sum_y\sum_{u\in\mathcal{U}_y} P(z_Z|\;y,u,v_{u,y})} &= \frac{e^{\theta_z}}{1+e^{\theta_z}} = p_z. \qquad (55)
\end{aligned}$$

Maximization of $< \mathcal{L}_C >$ with respect to $P(y|z_Z)$ is analogous to (50) – (52):

$$P(y|z_Z) = \frac{\sum_{u\in\mathcal{U}_y} P(z_Z|\;y,u,v_{u,y})}{\sum_{y'}\sum_{u\in\mathcal{U}_y} P(z_Z|\;y',u,v_{u,y'})}. \qquad (56)$$

C.4 M-step for Type I – Binomial $P(v|y,z_Z)$

Binomial distribution $P(v|y,z_Z)$ can be written in the functional form

$$P(v|y,z_Z) = \exp\left\{v\;\theta_{z,y} - G(\theta_{z,y}) + \log\binom{V}{v}\right\}. \qquad (57)$$

The expected complete data log likelihood for model of type I is

$$\begin{aligned} < \mathcal{L}_C > \; = \; & \sum_{n=1}^{N} \sum_{u,y,v} \Delta_{u,y,v}^{u_n,y_n,v_n} \sum_{z_Y,z_Z} P(z_Y, z_Z | \; y, u, v) \\ & \{[v \; \theta_{z,y} - G(\theta_{z,y}) + \log p_0(v)] \\ & + \log P(z_Z|z_Y) + \log P(z_Y|u)\} \\ & + \sum_u \lambda_u \left(\sum_{z_Y} P(z_Y|u) - 1 \right) \end{aligned} \qquad (58)$$

Maximizing $< \mathcal{L}_C >$ with respect to parameter $\theta_{z,y}$,

$$\frac{\partial < \mathcal{L}_C >}{\partial \; \theta_{z,y}} = \sum_{u \in \mathcal{U}_y} \sum_{z_Y} P(z_Y, z_Z | \; y, u, v_{u,y})[v_{u,y} - g(\theta_{z,y})] = 0. \qquad (59)$$

we get

$$p_{z,y} = \frac{\sum_{u \in \mathcal{U}_y} P(z_Z| \; y, u, v_{u,y}) v_{u,y}}{V \sum_{u \in \mathcal{U}_y} P(z_Z| \; y, u, v_{u,y})}. \qquad (60)$$