Accurate and scalable social recommendation using mixed-membership stochastic block models

Antonia Godoy-Lorite[a,1], Roger Guimerà[a,b], Cristopher Moore[c], and Marta Sales-Pardo[a]

[a]Departament d'Enginyeria Química, Universitat Rovira i Virgili, 43007 Tarragona, Catalonia, Spain; [b]Institució Catalana de Recerca i Estudis Avançats, 08010 Barcelona, Catalonia, Spain; and [c]Santa Fe Institute, Santa Fe, NM 87501

Edited by Edoardo Airoldi, Harvard University, and accepted by Editorial Board Member Adrian E. Raftery October 18, 2016 (received for review April 21, 2016)

With increasing amounts of information available, modeling and predicting user preferences—for books or articles, for example—are becoming more important. We present a collaborative filtering model, with an associated scalable algorithm, that makes accurate predictions of users' ratings. Like previous approaches, we assume that there are groups of users and of items and that the rating a user gives an item is determined by their respective group memberships. However, we allow each user and each item to belong simultaneously to mixtures of different groups and, unlike many popular approaches such as matrix factorization, we do not assume that users in each group prefer a single group of items. In particular, we do not assume that ratings depend linearly on a measure of similarity, but allow probability distributions of ratings to depend freely on the user's and item's groups. The resulting overlapping groups and predicted ratings can be inferred with an expectation-maximization algorithm whose running time scales linearly with the number of observed ratings. Our approach enables us to predict user preferences in large datasets and is considerably more accurate than the current algorithms for such large datasets.

recommender systems | stochastic block model | collaborative filtering | social recommendation | scalable algorithm

The goal of recommender systems is to predict what movies we are going to like, what books we are going to purchase, or even who we might be interested in dating. The rapidly growing amount of data on item reviews, ratings, and purchases from a growing number of online platforms holds the promise to facilitate the development of more informed models for recommendation. At the same time, however, it poses the challenge of developing algorithms that can handle such large amounts of data accurately and efficiently.

A plausible expectation when developing recommendation algorithms is that similar users relate to similar items in similar ways; e.g., they purchase similar items and give the same item similar ratings. This means that we can use the rating history of a set of users to make recommendations, even without knowing anything about the characteristics of users or items. This is the basic underlying assumption of collaborative filtering, one of the most common approaches in recommender systems (1). However, most research in recommender systems has focused on the development of scalable algorithms, often at the price of implicitly using models that are overly simplistic or unrealistic. For example, matrix factorization and latent feature approaches assume that users and items live in an abstract low-dimensional space, but whether such a space is expressive enough to accommodate the rich variety of user behaviors is rarely discussed. As a result, many current approaches have significantly lower accuracies than inference approaches based on models of user preferences that are socially more realistic (2). On the other hand, these more realistic approaches do not scale well with dataset size, which makes them unpractical for large datasets.

Here, we develop a model and algorithm for predicting user ratings based on explicit probabilistic hypotheses about user behavior. As in some previous approaches, we assume that there are groups of users and of items and that the rating a user assigns to an item is determined probabilistically by their group memberships. However, we do not assign users and items to a single group; instead, we allow each user and each item to belong to mixtures of different groups (3, 4). In addition, unlike standard matrix factorization, we do not assume that ratings depend linearly on a measure of similarity between users and items; instead, we allow each pair of groups to have any probability distribution of ratings. We combine these elements to form a generative model, which assigns a precise probability to each possible rating. Fortunately, the inference problem for this model can be solved very efficiently: We give an expectation-maximization algorithm whose running time, per iteration, scales linearly with the number of observed ratings and converges rapidly.

We show that our approach consistently outperforms state-of-the-art recommendation algorithms, often by a large margin. In addition, our probabilistic predictions are better calibrated to real data in the frequentist sense (5), generating distributions of ratings that are statistically similar to real data. Moreover, because our model has a clear probabilistic interpretation, it can deal naturally with some situations that are challenging for other approaches, such as the cold start problem. We argue that our approach may also be suitable for other areas where matrix factorization is increasingly used such as image reconstruction, textual data mining, cluster analysis, or pattern discovery (6–10).

Significance

Recommendation systems are designed to predict users' preferences and provide them with recommendations for items such as books or movies that suit their needs. Recent developments show that some probabilistic models for user preferences yield better predictions than latent feature models such as matrix factorization. However, it has not been possible to use them in real-world datasets because they are not computationally efficient. We have developed a rigorous probabilistic model that outperforms leading approaches for recommendation and whose parameters can be fitted efficiently with an algorithm whose running time scales linearly with the size of the dataset. This model and inference algorithm open the door to more approaches to recommendation and to other problems where matrix factorization is currently used.

Author contributions: A.G.-L., R.G., C.M., and M.S.-P. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. E.A. is a Guest Editor invited by the Editorial Board.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. Email: antonia.godoy@urv.cat.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1606316113/-/DCSupplemental.

A Mixed-Membership Block Model with Metadata

Our approach begins with the mixed-membership stochastic block model (MMSBM), which has been used to model networks. As in the original MMSBM (3) and related models (11), we assume that each node in the bipartite graph of users and items belongs to a mixture of groups. However, unlike refs. 3 and 11, we do not assume that these group memberships affect the presence or absence of a link. Instead, we take the set of links as given and attempt to predict the ratings. We do this with an MMSBM-like model where the rating a user gives an item is drawn from a probability distribution that depends on their group memberships.

First we set down some notation. We have N users and M items and a bipartite graph $R = \{(u, i)\}$ of links, where the link (u, i) indicates that item i was given a rating (observed or unobserved) by user u. For each $(u, i) \in R$, the rating r_{ui} belongs to some finite set S such as $\{1, 2, 3, 4, 5\}$. Given a set R^O of observed ratings, our goal is to classify the users and the items and to predict the rating r_{ui} of a link $(u, i) \in R$ for which the rating is not yet known.

Our generative model for the ratings is as follows. There are K groups of users and L groups of items. For each pair of groups k, ℓ, there is a probability distribution $p_{k\ell}(r)$ over S of the rating r that u gives i, assuming that u belongs entirely to group k and i belongs entirely to group ℓ.

To model mixed group memberships, each user u has a vector $\theta_u \in \mathbb{R}^K$, where θ_{uk} denotes the extent to which user u belongs to group k. Similarly, each item i has a vector $\eta_i \in \mathbb{R}^L$. These vectors are normalized; i.e., $\sum_k \theta_{uk} = \sum_\ell \eta_{i\ell} = 1$. The probability distribution of the rating r_{ui} is then a convex combination,

$$Pr[r_{ui} = r] = \sum_{k,\ell} \theta_{uk} \eta_{i\ell} p_{k\ell}(r). \qquad \textbf{[1]}$$

Abbreviating all these parameters as $\theta, \eta, \mathbf{p}$, the likelihood of the observed ratings is then

$$P(R^O|\theta, \eta, \mathbf{p}) = \prod_{(u,i)\in R^O} \sum_{k,\ell} \theta_{uk} \eta_{i\ell} p_{k\ell}(r_{ui}). \qquad \textbf{[2]}$$

As we discuss below, we infer the values of the parameters $\hat{\theta}, \hat{\eta}, \hat{\mathbf{p}}$ that maximize this likelihood using an efficient expectation-maximization algorithm. We can then use the inferred model to predict unobserved ratings r_{ui} for pairs $(u, i) \notin R^O$.

Our work differs from previous work on collaborative filtering in several ways. First, unlike matrix factorization approaches such as ref. 12 or their probabilistic counterparts (13–15), we do not think of the ratings $r_{ui} \in \{1, 2, 3, 4, 5\}$ as integers or real values. As has been established in the literature (16), giving a movie a rating of 5 instead of 1 does not mean the user likes it five times as much. Our results suggest that it is better to think of different ratings simply as different labels that appear on the links of the network. Moreover, our method yields a distribution over the possible ratings directly, rather than a distribution over integers or real numbers that must be somehow mapped to the space of possible ratings (13–15). From this point of view, our model is a bipartite MMSBM with metadata (or labels) on the edges; a similar model based on the stochastic block model (SBM), where each user and item belong to only one group, is given in ref. 2. An alternative approach would be to consider a multilayer representation of the data as in ref. 4.

Second, we do not assume that the matrices $\mathbf{p}$ have any particular structure. In particular, we do not assume homophily, where groups of users correspond to groups of items, and users prefer items that belong to their own group: That is, we do not assume that $\mathbf{p}(r)$ is larger on the diagonal for higher ratings r. Thus, our model can have arbitrary couplings between groups of users and items that are independent for each possible rating.

Third, unlike some approaches that use inference methods similar to ours (17), as stated above, our goal is not to predict the existence of links. In particular, we do not assume that users see only movies (say) that they like, and we do not treat missing links as zeros or low ratings. To put this differently, we are not trying to complete R to a full matrix of ratings, but only to predict the unobserved ratings in $R \setminus R^O$. Thus, the only terms in the likelihood of our model correspond to observed ratings.

As we describe below, our model also has the advantage of being mathematically tractable. It yields a highly efficient expectation-maximization algorithm for fitting the parameters:

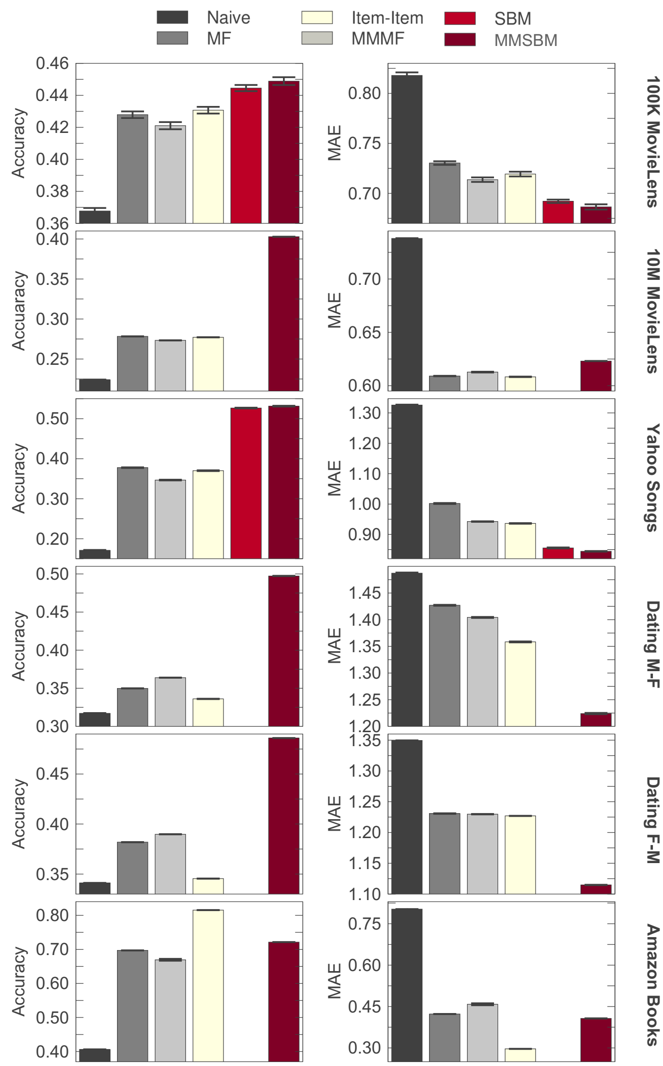

Fig. 1. Algorithm comparison. The performance of different approaches for predicting user–item ratings. From *Top* to *Bottom*, the datasets are MovieLens 100K, Movielens 10M, Yahoo! Songs, males rating females (M-F) in the LibimSeTi dataset, females rating males (F-M) in the LibimSeTi dataset, and Amazon Books. *Left* column displays the accuracy of the algorithms in each dataset, i.e., the fraction of ratings that are exactly predicted by each algorithm. *Right* column displays the MAE in the predicted vs. actual rating, treated as an integer or half-integer. In all cases, the bars are the average of a fivefold cross-validation and the error bars correspond to the SE of the mean. The SBM algorithm does not scale to the larger datasets, but achieves similar accuracy to the MMSBM on the datasets it can handle. The MMSBM achieves the best (highest) accuracy in five of six datasets and the best (lowest) MAE in four of six datasets.

The time each iteration takes is linear on the number of users, items, and observed links. As a result, we are able to handle large datasets and achieve a higher accuracy than standard methods. We also show that the probabilistic predictions made by our model are well calibrated in the frequentist sense (5), producing distributions of ratings statistically similar to real data.

Scalable Inference of Model Parameters

In most practical situations, marginalizing exactly over the group membership vectors θ and η and the probability matrices $\mathbf{p}$ (similar to ref. 2) is too computationally expensive. As an alternative we propose to obtain the model parameters that maximize the likelihood (2), using an expectation-maximization (EM) algorithm.

In particular, we use a classic variational approach (*Materials and Methods*) to obtain the following equations for the model parameters that maximize the likelihood:

$$\theta_{uk} = \frac{\sum_{i\in\partial u}\sum_{\ell}\omega_{ui}(k,\ell)}{d_u}, \qquad [3]$$

$$\eta_{i\ell} = \frac{\sum_{u\in\partial i}\sum_{k}\omega_{ui}(k,\ell)}{d_i}, \qquad [4]$$

$$p_{k\ell}(r) = \frac{\sum_{(u,i)\in R^O|r_{ui}=r}\omega_{ui}(k,\ell)}{\sum_{(u,i)\in R^O}\omega_{ui}(k,\ell)}. \qquad [5]$$

Here $\partial u = \{i|(u,i)\in R^O\}$ and $\partial i = \{u|(u,i)\in R^O\}$ denote the neighborhoods of u and i, respectively; $d_u = |\partial u|$ and $d_i = |\partial i|$ are the node degrees, i.e., the number of observed ratings for user u and item i, respectively; and

$$\omega_{ui}(k,\ell) = \frac{\theta_{uk}\eta_{i\ell}p_{k\ell}(r_{ui})}{\sum_{k',\ell'}\theta_{uk'}\eta_{i\ell'}p_{k'\ell'}(r_{ui})} \qquad [6]$$

is the variational method's estimate of the probability that the rating r_{ui} is due to u and i belonging to groups k and ℓ, respectively.

These equations can be solved with an EM algorithm. Starting with an estimate of θ, η, and $\mathbf{p}$, we repeat the following steps until the parameters converge to a fixed point: (*i*) (expectation step) use Eq. **6** to compute $\omega_{ui}(k,\ell)$ for $(u,i)\in R^O$; (*ii*) (maximization step) use Eqs. **3**–**5** to compute θ, η, and $\mathbf{p}$.

The number of parameters and terms in the sums in Eqs. **3**–**6** is $NK + ML + |R^O|KL$. Assuming that K and L are constant, each EM step is $O(N + M + |R^O|)$ and hence linear in the size of the dataset (Fig. S1*A*). As the set of observed ratings R^O is typically very sparse because only a small fraction of all possible user–item pairs have observed ratings, our algorithm is feasible even for very large datasets.

Results

The MMSBM Predicts Ratings Accurately. We test the performance of our algorithm in six datasets: the MovieLens 100K and 10M datasets, respectively; Yahoo! Songs; Amazon books (18, 19); and the LibimSeTi.cz dating agency (20), which (because it is primarily heterosexual) we split into two datasets, consisting of males rating females and vice versa. These datasets are diverse in the types of items, the sizes $|S|$ of the sets of possible ratings, and the density of observed ratings (Table S1). For each dataset we perform a fivefold cross-validation.

We compare our algorithm to four benchmark algorithms (see *Supporting Information, Benchmark Algorithms*): a baseline naive algorithm that assigns to each test rating r_{ui} the average of the observed ratings for item i; the item–item algorithm (21), which predicts r_{ui} based on the observed ratings of user u for items that are the most similar to i; "classical" matrix factorization (12); and mixed-membership matrix factorization (MMMF) (22). For all these benchmark algorithms except MMMF we use the implementation in the LensKit package (16), which is fast, highly optimized, and makes our results easily reproducible. For MMMF, we use the Matlab implementation provided by the authors (https://code.google.com/archive/p/m3f/). Additionally, for the smallest datasets, we also use the (unmixed) stochastic block model of ref. 2; however, that algorithm does not scale well to larger datasets (Fig. S1*B*).

For our algorithm, we set $K = L = 10$; i.e., we assume that there are 10 groups of users and 10 groups of items (recall that we do not assume any correspondence between these groups). We considered some other choices of K and L, but we found no differences in performance for $K, L \geq 10$ (Fig. S2). Because iterating the EM algorithm of Eqs. **3**–**6** can lead to different fixed points depending on its initial conditions, we perform 500 independent runs. We average the predicted probability distribution of ratings over the resulting fixed points, because we find they typically have comparable likelihood values (Fig. S3).

We can translate the resulting probability distribution of ratings into a single predicted rating by choosing an estimator; which one is optimal depends on the loss function or equivalently the measure of accuracy. We focus on two measures. For each algorithm, we define the accuracy as the fraction of ratings that are predicted exactly, and we also measure the mean absolute error (MAE). For these two, the optimal estimator is the mode and the median, respectively.

We find that in most datasets our approach outperforms the item–item algorithm, matrix factorization (MF), and MMMF (Fig. 1). Indeed, the accuracy, i.e., the fraction of exactly correct predictions, of the MMSBM is significantly higher than that of MF and MMMF for all of the datasets we tested and higher than the item–item algorithm in five of six datasets, the only exception

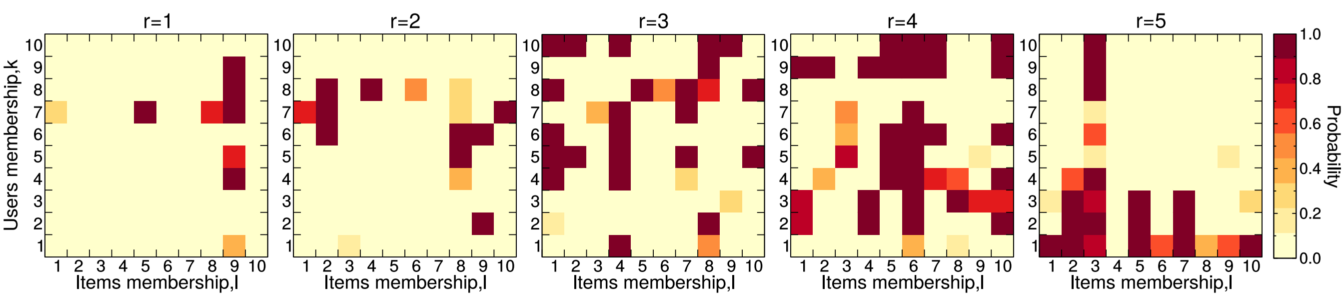

Fig. 2. Probability matrices in MMSBM. We show the inferred values for the probability matrices $\mathbf{p}$ from the MovieLens 100K dataset. *Left* to *Right*, the five matrices correspond to the ratings $r = 1, 2, 3, 4, 5$. For each one of them, the rows and columns correspond to the user's and item's groups; here $K = L = 10$. Each element, shown as a heat map, gives the probability $p_{k\ell}(r)$ that a user in group k gives a rating r to an item in group ℓ. The matrices are normalized such that $\sum_{r\in S} p_{k\ell}(r) = 1,\ \forall k,\ell$. Note that there is no ordering of the probability matrices that would make them diagonal.

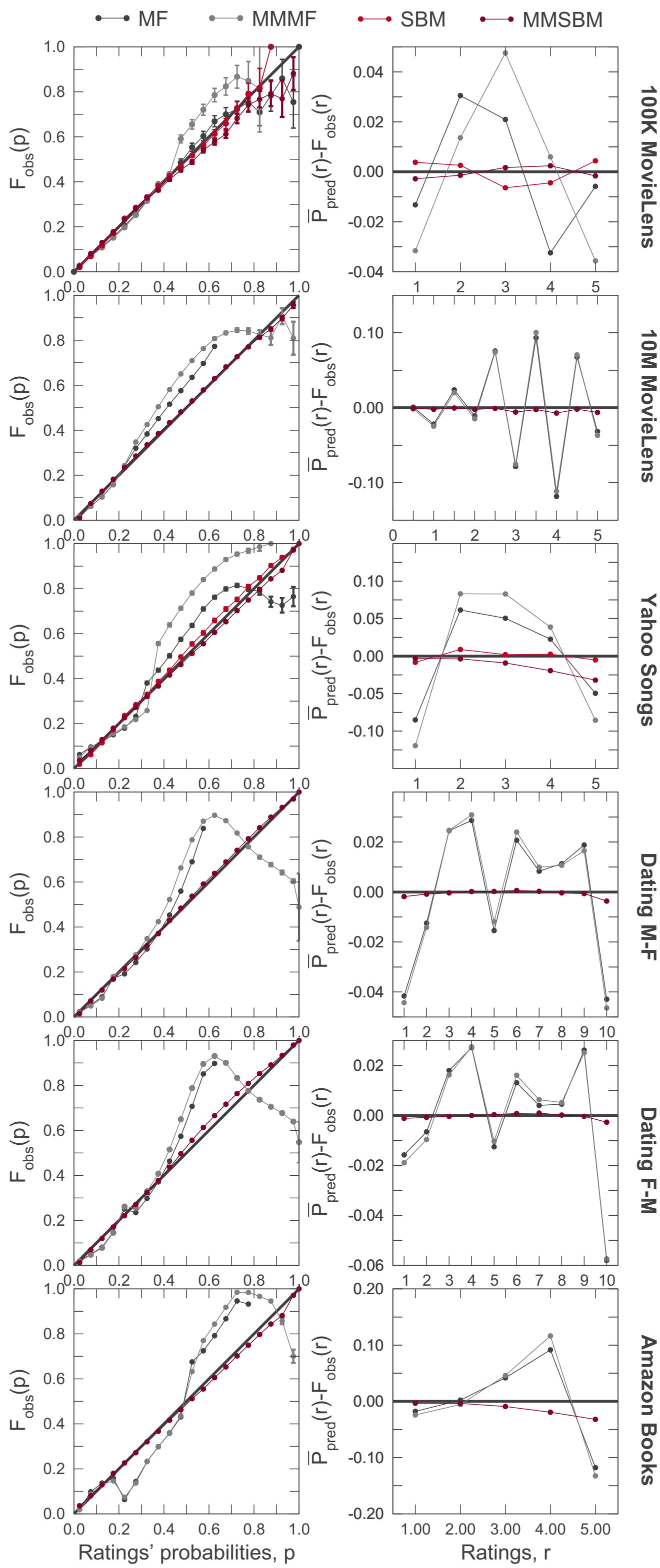

Fig. 3. Calibrating probabilistic predictions. *Left* column: Probabilistic calibration. For each $p \in (0, 1)$, we consider the set of held-out pairs for which each rating was predicted with probability p and show the fraction $F_{obs}(p)$ of those pairs for which this rating is correct. Points on the diagonal $F_{obs}(p) = p$ indicate that the algorithm captures the stochasticity in the ratings. *Right* column: Marginal calibration. For each possible rating r, we show the difference between the average probability $\overline{P}_{pred}(r)$ that r is predicted and the actual fraction of held-out pairs with rating r. Points on the line $\overline{P}_{pred}(r) - F_{obs}(r) = 0$ indicate that the algorithm matches the empirical distribution of ratings. All cases use fivefold cross-validation. The MMSBM and (on small datasets) the SBM are significantly better calibrated than other algorithms, producing rating distributions statistically similar to real data. (See *Supporting Information* for the probabilistic definition of MF and MMMF.)

being Amazon Books. The MAE of the MMSBM is the best (lowest) in four of the six datasets; item–item produces smaller MAE in Amazon Books; and item–item, MF, and MMMF produce smaller MAE in MovieLens 10M.

Interestingly, our approach produces results that are almost identical to those of the unmixed, fully marginalized SBM (2) for the two examples for which inference with the SBM is feasible. In particular, we achieve the same accuracy with $K = L = 10$ in the mixed-membership model as with 50 groups in the unmixed SBM. This result suggests that many of the groups observed in ref. 2 are in fact mixtures of a smaller number of groups and that the additional expressiveness of the MMSBM allows us to succeed with a lower-dimensional model. Moreover, the fact that the maximum-likelihood estimator of the MMSBM gives results that are as accurate as those obtained by sampling over SBMs suggests that the mixing of memberships is an appropriate substitute for sampling over partitions of users and items.

The MMSBM Generalizes Matrix Factorization. MF is one of the most successful and popular approaches to collaborative filtering, both in its classical (12) and its probabilistic form (13–17). However, as discussed, our MMSBM gives more accurate ratings, often by a large margin. Here, we propose an explanation for this improvement.

We start by giving an interpretation of MF as a special case of the MMSBM. In its simplest form, MF assumes that the expected rating that user u gives item i is $\bar{r}_{ui} = \tilde{\theta}_u \cdot \tilde{\eta}_i$, where $\tilde{\theta}_u$ and $\tilde{\eta}_i$ are K-dimensional vectors representing the user and the item, respectively. [One can apply a variety of noise models or loss functions, as well as regularization terms for the model parameters (12), but this does not significantly alter our discussion.] This can be interpreted as a mixed-membership model as follows: Assume that there are $K = L$ groups of users and items, that θ_{uk} is the probability that user u belongs to group k, and that η_{ik} is the probability that item i belongs to group k. Finally, assume that users in group k like only items in group k; in particular, users in k assign a baseline rating of 1 to items in group k and a rating of 0 to items in all other groups. Finally, let $s_u \geq 0$ and $s_i \geq 0$ be user and item "intensities" that correct for the fact that some users rate on average higher than others and that some items are generally more popular than others. Then the expected ratings are given by

$$\bar{r}_{ui} = \sum_k s_u \theta_{uk} s_i \eta_{ik}. \qquad [7]$$

If we set $\tilde{\theta}_{uk} = s_u \theta_{uk}$ and $\tilde{\eta}_{ik} = s_i \theta_{ik}$, this becomes the MF model $\bar{r}_{ui} = \tilde{\theta}_u \cdot \tilde{\eta}_i$. Thus, MF corresponds to a model where there is a one-to-one correspondence between groups of users and groups of items, and users in a given group like only items in the corresponding group. If these assumptions do not hold, in general MF will not be able to properly model user–item ratings.

Our MMSBM relaxes these assumptions by allowing the distribution of ratings to be given by arbitrary matrices **p**. MF is roughly equivalent to assuming that $p_{k\ell}$ is diagonal, at least for high ratings. We suggest that the improved performance of the MMSBM over MF is due to this greater expressive power. Indeed, Fig. 2 shows that the matrices **p** inferred by our model are far from diagonal (see also Fig. S4).

Moreover, the generality of the MMSBM allows it to account for many of the features of real ratings. For instance, different groups of users have different distributions of ratings: Users in group $k = 1$ rate most movies with $r = 5$, whereas those in $k = 7$ often give ratings $r = 1$. Similarly, movies in group $\ell = 3$ are consistently rated $r = 5$ by most users, whereas movies in $\ell = 9$ are rated $r = 1$ quite often. Interestingly, some groups of users agree on some movies but disagree on others: For example, users in groups $k = 9, 10$ agree that most movies in group $\ell = 3$ should be rated $r = 5$, but they disagree on movies in $\ell = 9$, rating them $r = 1$ and $r = 3$, respectively.

One can also compare our MMSBM to MMMF, because both attempt to take the mixed-membership nature of users and items into account. However, the analogy is not perfect: MMMF models ratings as the sum of a MF term and a correction that uses mixed group memberships that are unrelated to the feature vectors (22). Although this is an improvement over MF, it does not fundamentally remove the assumption that each group of users has a corresponding group of items that it prefers. Indeed, our numerical results show that the performance of MMMF is fairly close to that of MF in the datasets we considered.

The MMSBM Makes Well-Calibrated Probabilistic Predictions. Finally, our approach directly yields probabilistic predictions of the ratings, i.e., probability distributions on the discrete set S, and we can use the technique of frequentist calibration to see whether these predictions accurately capture the stochasticity of the data. Following ref. 5, we perform two types of calibration experiments. Probabilistic calibration means that, for each $r \in S$ and $p \in [0, 1]$, of the held-out pairs to which our approach assigns a rating of r with probability p, this is indeed the correct rating of a fraction p of them. (This differs slightly from ref. 5, where a probabilistic forecaster predicts the cumulative distribution of a continuous variable, but it seems to be a reasonable definition for discrete values.) Marginal calibration means that for each rating $r \in S$, the average probability we assign to r coincides with its actual frequency among the held-out pairs.

As we show in Fig. 3, the predictions of the MMSBM are indeed probabilistically and marginally well calibrated. Thus, in addition to giving accurate ratings in the sense of the MAE and the probability the rating is exactly correct, the MMSBM generates predictions that are statistically similar to real data, indicating that it captures the stochastic nature of the rating process.

Because MF and MMMF produce Gaussian distributions of real-valued ratings, to perform analogous calibration experiments we transform their predictions into a discrete probability distribution by integrating over the real numbers closest to each $r \in S$ (*Supporting Information*). For instance, if $S = \{1, 2, 3, 4, 5\}$, we define the probability that $r = 2$ as the integral of this continuous distribution over the interval $[1.5, 2.5)$. Fig. 3 shows that the resulting probabilistic predictions are not well calibrated, neither probabilistically nor marginally. One stark example of this is the MovieLens 10M dataset, where users use integer ratings much more often than half-integer ones. MF and MMMF cannot recognize this pattern and thus systematically underestimate and overestimate the probability of integer and half-integer ratings respectively. Similar, although less obvious, patterns cause MF and MMMF to be poorly calibrated in other datasets as well. Of course, one could attempt to infer a nonlinear mapping from continuous ratings to discrete ones, but this would increase the complexity of these models considerably. By treating each rating as a different label, the MMSBM adapts easily to the empirical distribution of ratings in each dataset.

The MMSBM Provides a Principled Method to Deal with the Cold Start Problem. Because the parameters of the MMSBM have a precise probabilistic interpretation, it can naturally deal with situations that are challenging for other algorithms. An example of this is the "cold start" problem, where we need to predict ratings for users or items for which we do not have training data (14, 23, 24).

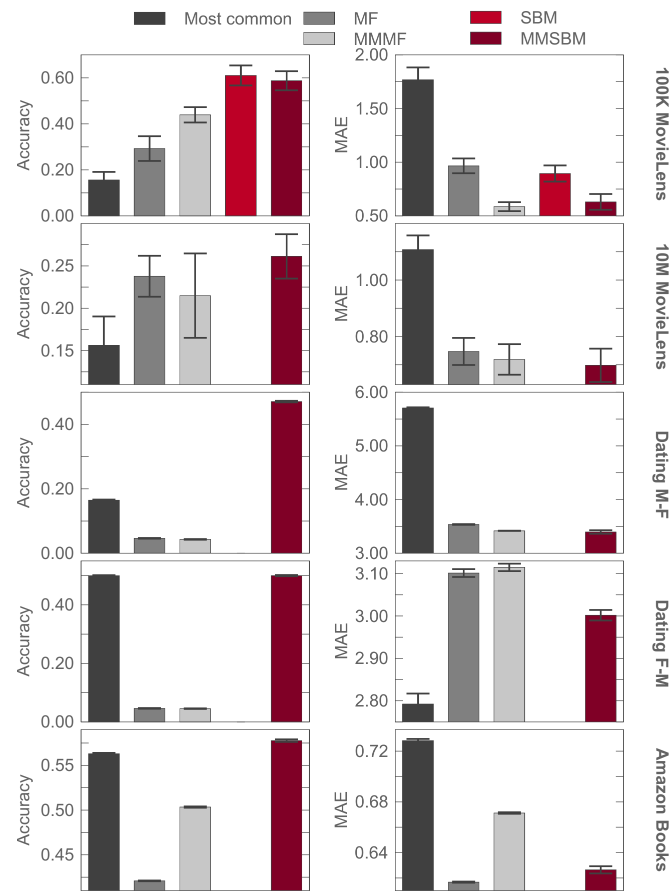

Fig. 4. Algorithm performance for the cold start problem. From *Top* to *Bottom*, the MovieLens 100K dataset with 0.17% of cold start cases on average, Movielens 10M (0.0015%), males rating females (M-F) in LibimSeTi (0.625%), females rating males (F-M) in LibimSeTi (0.31%), and Amazon Books (6.7%). We did not encounter any cold start cases in the cross-validation experiments with Yahoo! Songs; this is to be expected because Yahoo! Songs requires that users and songs have at least 20 ratings. *Left* column displays the accuracy for each dataset and *Right* column the mean absolute error. The bars show the average of fivefold cross-validation and the error bars show the SE.

In the MMSBM, the $\mathbf{p}$ matrices are the same for all users and items; in this sense, new users or items pose no particular difficulty. However, we have no information about their group membership vectors. In the absence of information about a new user n we can assume, a priori, that he or she belongs to each group to the same extent that a random existing user does. In practice, this means that we initially set his or her group membership vector to the average of the vectors of the observed users, $\theta_{nk} = \frac{1}{N}\sum_u \theta_{uk}$. We can treat $\eta_{i\ell}$ similarly for a new item i. This provides a principled method to deal with the cold start problem without additional elements (14).

In Fig. 4 we show that, in cold start situations, the MMSBM outperforms the other algorithms in most cases. MMSBM is always more accurate than MF and MMMF (although in one case the difference is not significant). In all but one case, the MMSBM is also more accurate than an algorithm that assigns the most common rating to an item. In terms of mean absolute error, our approach is more accurate than MF and MMMF in four of five datasets (in one, not significantly) and more accurate than using the most common rating in four of five cases.

Note that none of these approaches takes metadata on users or items into account, which is a standard approach to the cold start problem. For instance, one could assume that a new user will behave similarly to others of the same age, gender, etc. (Fig. S5), and compute the average membership vector over these users. We performed experiments restricting the average to users with same gender and/or age, but we found it did not significantly improve the performance (Fig. S6).

Discussion

We have shown that the MMSBM with its associated expectation-maximization algorithm is an accurate and scalable method to predict user–item ratings in a variety of contexts. It significantly outperforms other algorithms, including MF and MMMF, in most of the datasets we considered, both maximizing the probability that the predicted rating is exactly correct and minimizing the mean absolute error.

Additionally, because the model and its parameters are readily interpretable, it can be extended to (and performs well in) situations that are challenging for other approaches, such as a cold start where no prior information is available about a new user or item; one could also consider extensions of the model that take into account metadata for users (e.g., age and gender) and/or items (e.g., genre), analogous to unmixed stochastic block models with node metadata (25).

Finally, because the MMSBM assigns a probability to each possible rating, it is amenable to frequentist calibration, and we found that its predictions are in fact statistically similar to real data as measured by probabilistic and marginal calibration (5). We believe that this performance is due to the fact that the MMSBM is a more expressive generalization of matrix factorization, allowing each pair of user and item groups to have an arbitrary probability distribution of ratings. Matrix factorization is a widely used tool with many applications beyond recommendation; given our findings, it may make sense to use the MMSBM in those other applications as well.

Materials and Methods

We maximize the likelihood **2** as a function of $\theta, \eta, \mathbf{p}$, using an EM algorithm. We start with a standard variational trick that changes the log of a sum into a sum of logs, writing

$$\log P(R^O|\theta,\eta,\mathbf{p}) = \sum_{(u,i)\in R^O} \log \sum_{k\ell} \theta_{uk}\eta_{i\ell}p_{k\ell}(r_{ui})$$

$$= \sum_{(u,i)\in R^O} \log \sum_{k\ell} \omega_{ui}(k,\ell)\,\frac{\theta_{uk}\eta_{i\ell}p_{k\ell}(r_{ui})}{\omega_{ui}(k,\ell)}$$

$$\geq \sum_{(u,i)\in R^O} \sum_{k\ell} \omega_{ui}(k,\ell)\log\frac{\theta_{uk}\eta_{i\ell}p_{k\ell}(r_{ui})}{\omega_{ui}(k,\ell)}. \qquad [8]$$

Here $\omega_{ui}(k,\ell)$ is the estimated probability that a given ranking r_{ui} is due to u and i belonging to groups k and ℓ, respectively, and the lower bound in the third line is Jensen's inequality $\log\bar{x} \geq \overline{\log x}$. This lower bound holds with equality when

$$\omega_{ui}(k,\ell) = \frac{\theta_{uk}\eta_{i\ell}p_{k\ell}(r_{ui})}{\sum_{k'\ell'}\theta_{uk'}\eta_{\ell' i}p_{k'\ell'}(r_{ui})}, \qquad [9]$$

giving us the update Eq. **6** for the expectation step.

For the maximization step, we derive update equations for the parameters $\theta, \eta, \mathbf{p}$ by taking derivatives of the log-likelihood Eq. **8**. Including Lagrange multipliers for the normalization constraints, we obtain

$$\theta_{uk} = \frac{\sum_{i\in\partial u}\sum_{l}\omega_{ui}(k,\ell)}{\sum_{i\in\partial u}\sum_{k\ell}\omega_{ui}(k,\ell)} = \frac{\sum_{i\in\partial u}\sum_{l}\omega_{ui}(k,\ell)}{d_u}, \qquad [10]$$

$$\eta_{i\ell} = \frac{\sum_{u\in\partial i}\sum_{k}\omega_{ui}(k,\ell)}{\sum_{u\in\partial i}\sum_{k\ell}\omega_{ui}(k,\ell)} = \frac{\sum_{u\in\partial i}\sum_{k}\omega_{ui}(k,\ell)}{d_i}, \qquad [11]$$

where d_u and d_i are the degrees of the user u and item i, respectively. Finally, including a Lagrange multiplier for the normalization constraints, we have

$$p_{k\ell}(r) = \frac{\sum_{(u,i)\in R^O|r_{ui}=r}\omega_{ui}(k,\ell)}{\sum_{(u,i)\in R^O}\omega_{ui}(k,\ell)}. \qquad [12]$$

ACKNOWLEDGMENTS. We thank C. Shalizi for helpful comments and suggestions. We also thank L. Mackey providing us his code and assistance for the MMMF algorithm. This work was supported by a James S. McDonnell Foundation Research Award (to R.G. and M.S.-P.), Spanish Ministerio de Economia y Comptetitividad Grants FIS2013-47532-C3 (to A.G.L., R.G., and M.S.-P.) and FIS2015-71563-ERC (to R.G.), European Union Future and Emerging Technologies (FET) Grant 317532 [multilevel complex networks and systems (MULTIPLEX), to R.G. and M.S.-P.], the John Templeton Foundation (C.M.), and the army research office (ARO) under Contract W911NF-12-R-0012 (to C.M.).

1. Su X, Khoshgoftaar TM (2009) A survey of collaborative filtering techniques. *Adv Artif Intell* 2009:421425.
2. Guimerà R, Llorente A, Moro E, Sales-Pardo M (2012) Predicting human preferences using the block structure of complex social networks. *PLoS One* 7(9):e44620.
3. Airoldi EM, Blei DM, Fienberg SE, Xing EP (2008) Mixed membership stochastic blockmodels. *J Mach Learn Res* 9:1981–2014.
4. Peixoto TP (2015) Model selection and hypothesis testing for large-scale network models with overlapping groups. *Phys Rev X* 5:011033.
5. Gneiting T, Balabdaoui F, Raftery AE (2007) Probabilistic forecasts, calibration and sharpness. *J R Stat Soc B* 69(2):243–268.
6. Cemgil AT (2009) Bayesian inference for nonnegative matrix factorisation models. *Comput Intell Neurosci* 2009:785152.
7. Berry MW, Browne M, Langville AN, Pauca VP, Plemmons RJ (2007) Algorithms and applications for approximate nonnegative matrix factorization. *Comput Stat Data Anal* 52(1):155–173.
8. Ding C, He X, Simon HD (2005) On the equivalence of nonnegative matrix factorization and spectral clustering. *SDM* 5:606–610.
9. Kim J, Park H (2008) Sparse nonnegative matrix factorization for clustering, Technical report (Georgia Institute of Technology, Atlanta).
10. Brunet JP, Tamayo P, Golub TR, Mesirov JP (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci USA* 101(12):4164–4169.
11. Ball B, Karrer B, Newman MEJ (2011) Efficient and principled method for detecting communities in networks. Phys Rev E Stat Nonlin Soft Matter Phys 84:036103.
12. Koren Y, Bell R, Volinsky C (2009) Matrix factorization techniques for recommender systems. *Computer* 42:30–37.
13. Meeds E, Ghahramani Z, Neal RM, Roweis ST (2006) Modeling dyadic data with binary latent factors. *Advances in Neural Information Processing Systems 19*, eds Schölkopf B, Platt J, Hoffman T (MIT Press, Cambridge, MA), pp 977–984.
14. Salakhutdinov R, Mnih A (2008) Probabilistic matrix factorization. *Advances in Neural Information Processing Systems 20*, eds Platt JC, Koller D, Singer Y, Roweis ST (MIT Press, Cambridge, MA), pp 1257–1264.
15. Shan H, Banerjee A (2010) Generalized probabilistic matrix factorizations for collaborative filtering. *Proceedings of the 2010 IEEE International Conference on Data Mining* (IEEE Computer Soc, Washington, DC), pp 1025–1030.
16. Ekstrand MD, Ludwig M, Konstan JA, Riedl JT (2011) Rethinking the recommender research ecosystem: Reproducibility, openness, and LensKit. *Proceedings of the fifth ACM Conference on Recommender Systems* [Association for Computing Machinery (ACM), New York], pp 133–140.
17. Gopalan P, Hofman JM, Blei DM (2013) Scalable recommendation with Poisson factorization. arXiv:1311.1704.
18. McAuley J, Targett C, Shi Q, van den Hengel A (2015) Image-based recommendations on styles and substitutes. *SIGIR'15* (ACM, New York), pp 43–52.
19. McAuley J, Pandey R, Leskovec J (2015) Inferring networks of substitutable and complementary products. *KDD'15* (ACM, New York), pp 785–794.
20. Brozovsky L, Petricek V (2007) *Recommender System for Online Dating Service* (VSB, Ostrava, Czech Republic).
21. Sarwar B, Karypis G, Konstan J, Riedl J (2001) Item-based collaborative filtering recommendation algorithms. *WWW'01* (ACM, New York), pp 285–295.
22. Mackey L, Weiss D, Jordan MI (2010) *Mixed Membership Matrix Factorization* (Omnipress, Haifa, Israel), pp 711–718.
23. Schein AI, Popescul A, Ungar LH, Pennock DM (2002) *Methods and Metrics for Cold-Start Recommendations* (ACM, New York), pp 253–260.
24. Park ST, Chu W (2009) *Pairwise Preference Regression for Cold-Start Recommendation* (ACM, New York), pp 21–28.
25. Newman MEJ, Clauset A (2015) Structure and inference in annotated networks. Nat Commun 7:11863.
26. Paterek A (2007) Improving regularized singular value decomposition for collaborative filtering. *Proceedings of the KDD Cup Workshop at SIGKDD'07, 13th ACM International Conference on Knowledge Discovery and Data Mining* (ACM, New York), pp 39–42.
27. Gardner W (2003) Learning characteristics of stochastic-gradient-descent algorithms: A general study, analysis, and critique. *Signal Process* 6(2):113–133.
28. Holland PW, Laskey KB, Leinhardt S (1983) Stochastic blockmodels: First steps. *Soc Network* 5:109–137.
29. Nowicki K, Snijders TAB (2001) Estimation and prediction for stochastic blockstructures. *J Am Stat Assoc* 96:1077–1087.
30. Guimerà R, Sales-Pardo M (2009) Missing and spurious interactions and the reconstruction of complex networks. *Proc Natl Acad Sci USA* 106(52):22073–22078.