HAL
open science

Co-clustering for fair recommendation

Gabriel Frisch, Jean-Benoist Leger, Yves Grandvalet

To cite this version:

Gabriel Frisch, Jean-Benoist Leger, Yves Grandvalet. Co-clustering for fair recommendation. 21st European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2021), Sep 2021, Bilbao, Spain. pp.607-630, 10.1007/978-3-030-93736-2_44 . hal-03239856v1

HAL Id: hal-03239856

https://hal.science/hal-03239856v1

Submitted on 27 May 2021 (v1), last revised 6 Jul 2021 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Co-clustering for fair recommendation

Gabriel Frisch[1], Jean-Benoist Leger[1], and Yves Grandvalet[1]

[1] *Université de Technologie de Compiègne,, CNRS, Heudiasyc UMR 7253, Compiègne France.*

2021, May

Abstract

Collaborative filtering relies on a sparse rating matrix, where each user rates a few products, to propose recommendations. The approach consists to approximate the sparse rating matrix with a simple model whose regularities allow to fill in the missing entries. The latent block model is a generative co-clustering model that can provide such an approximation. In this paper, we show that exogenous sensitive attributes can be incorporated in this model to ensure fair recommendations. Since users are only characterized by their ratings and their sensitive attribute, fairness is measured here by a parity criterion. Introducing the sensitive attribute in the latent block model leads to a classification of users that is independent from the sensitive attribute. We propose a definition of fairness for the recommender system that expresses that the ranking of items should be independent of the sensitive attribute. We show that our model ensures approximately fair recommendations provided that the classification of users approximately respects statistical parity.

1 Introduction

In simple terms, fairness is often loosely defined as the quality of treating people equally, with impartiality and rightfulness. Although imprecise, this definition stipulates that equal treatment refers to certain sensitive attributes shared by groups of people, such as gender, age, ethnicity, socioeconomic group, etc. In recent years, intensive research has highlighted the lack of fairness in decisions made by machine learning algorithms (6).

There are several stakeholders in a recommendation scenario. In the terminology of Burke *et al.* (8), we target consumer-fairness, where the

objective is to provide the same treatment to users of the recommender system, regardless of their sensitive attribute. We target recommender systems relying on collaborative filtering, which aims at building recommendations from the history of user ratings. These observed ratings are the basis for making automatic predictions about non-rated items, under the assumption that users can be clustered according to their past opinion behavior. Sensitive attributes are not used to fit the models, but some disparate treatments may nevertheless exist, possibly due to some societal or cultural effects that bias the sampling of data (11). In situations where the sensitive attribute can be collected, it therefore seems preferable to design algorithms that process sensitive attributes to remove their influence, rather than simply ignore them.

Many proposals have already been made on how fairness should be formally defined in collaborative filtering (12; 32). One common approach is the recommendation independence (22) or statistical parity, that requires the unconditional statistical independence between recommendations and a specified sensitive attribute. This equal treatment does not ensure equal impact (also called "equal opportunity"), which argues for equal recommendation quality between sensitive groups. Although some works (35) have argued that statistical parity may be overly restrictive, resulting in a poor quality of recommendations, we use here this definition to propose a fair collaborative filtering algorithm.

In this paper, we aim at producing fair recommendations using a co-clustering of users and items that respects statistical parity of users with respect to some sensitive attributes. For this purpose, we introduce a co-clustering model based on the Latent Block Model (LBM) that relies on an ordinal regression model that takes as inputs the sensitive attributes. We demonstrate that our model ensures approximately fair recommendations provided that the clustering of users approximately respects statistical parity. Finally, we conduct experiments on a real-world dataset to show that the proposed approach can help alleviate unfairness.

Related works

Several recent works have raised the issue of fairness in recommender systems. Kamishima *et al.* (22) have proposed methods for improving fairness, formalized as the independence of the predicted ratings with the sensitive attribute. Their methods are based on matrix factorization regularized by criteria that favor independence by controlling the moments of the distributions of rating among sensitive groups. Using the same definition of fairness,

Zhu *et al.* (36) proposed a tensor method that isolates sensitive attributes in sub-dimensions of the latent factor matrix. Unlike many other methods, this solution is capable of handling multiple and non-binary sensitive attributes. Yao and Huang (35) proposed four new metrics that deal with different types of unfairness and used them as penalty functions in augmented matrix factorization objectives.

All of the above methods are based on the fairness of predicted ratings, but an approximate fairness of ratings may not entail an approximate fairness of the recommender system that provides users with a short list of relevant items. With this in mind, Beutel *et al.* (4) provided new metrics based on pairwise comparisons and proposed a novel pairwise regularization approach to improve the fairness of the recommender system during training. Finally, further from recommender systems but still related to the model we use, the notion of statistical parity is often considered for fairness in clustering methods (1; 14; 3).

2 Model

The data used to build recommender systems can be aggregated in a matrix where rows are users, columns are items and entries the feedbacks. The model we propose is based on the Latent Block Model that considers a data matrix to group users and items based on their opinions.

2.1 The Latent block models

The Latent block models (LBM), also known as bipartite stochastic block models and introduced in (15), are generative probabilistic models enabling to cluster jointly the rows and the columns of a data matrix denoted $\boldsymbol{R}$. These co-clustering models assume a homogeneous block structure of the whole data matrix. This structure is unveiled by the reordering of rows and columns according to their respective cluster index; for k_1 row clusters and k_2 column clusters, the reordering reveals $k_1 \times k_2$ homogeneous blocks in the data matrix being possibly binary (15) categorical (23), or quantitative (26; 16).

The partitions of rows and columns are governed by the latent variables $\boldsymbol{U}$ and $\boldsymbol{V}$, $\boldsymbol{U}$ being the $n_1 \times k_1$ indicator matrix of row classes, and $\boldsymbol{V}$ being the $n_2 \times k_2$ indicator matrix of the column classes. The class indicator of row i is denoted $\boldsymbol{U}_i$, and similarly, the class indicator of column j is denoted $\boldsymbol{V}_j$. The LBM makes several assumptions on the dependency and on the form of the distributions:

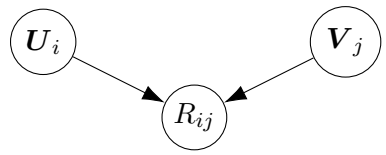


Figure 1: Graphical view of the Latent Block Model. Entries R_{ij} of the data matrix are independently generated according to the group membership $\boldsymbol{U}_i$ of row i and the group membership $\boldsymbol{V}_j$ of column j.

- The latent group memberships of rows and columns are assumed to be mutually independent and identically distributed, with respectively multinomial distributions $\mathcal{M}(1;\boldsymbol{\alpha})$ and $\mathcal{M}(1;\boldsymbol{\beta})$, where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{k_1})$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{k_2})$ are the mixing proportions of rows and columns:

$$p(\boldsymbol{U}, \boldsymbol{V}) = p(\boldsymbol{U})\, p(\boldsymbol{V}) = \prod_i p(\boldsymbol{U}_i;\boldsymbol{\alpha}) \prod_j p(\boldsymbol{V}_j;\boldsymbol{\beta}) \ .$$

- Conditionally to rows and columns assignments $(\boldsymbol{U}, \boldsymbol{V})$, the entries of the data matrix $\boldsymbol{R}$ are independent and identically distributed:

$$\begin{aligned} p(\boldsymbol{R}|\boldsymbol{U}, \boldsymbol{V};\boldsymbol{\theta}) &= \prod_{ij} p(R_{ij}|\boldsymbol{U}_i, \boldsymbol{V}_j) \ , \\ p(R_{ij}|U_{iq}V_{jl} = 1) &= \phi_{ql}(R_{ij}) \ , \end{aligned} \tag{1}$$

  with $\phi_{ql}(R_{ij})$ the density of the conditional distribution of R_{ij} depending on the group memberships of row i and column j.

2.2 Model proposed

The user feedback used for collaborative filtering can be implicit (history, browsing history, clicks...) or explicit. In the case of explicit evaluation data, users most often express their interest in items using a discrete rating scale. This rating scale suppose an order between levels, for example from 1 to 5 expressing the worst opinion to the best one. Models handling this type of data can assume that these scales are a discretization of the opinion of a user that may be better handled by a continuous variable. The method we propose to model ratings is based on a statistical co-clustering using ordered probit regression to model ordinal responses. Covariates encoding a sensitive user attribute can easily be included in the probit regression framework.

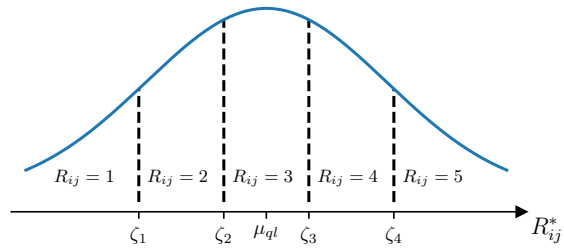


Figure 2: The conditional density function of R^*_{ij} and its relationship to R_{ij}. Fixed thresholds ζ_k, defines the discretization of R^*_{ij}.

2.2.1 Ordered probit in Latent Block Model

The ordered probit model (10) assumes the existence of a continuous, Gaussian distributed latent random variable, denoted $\boldsymbol{R}^*$. In a collaborative filtering context, this latent variable represents the underlying value, assumed to be continuous, assigned to an item by the user. The assumption of a single underlying continuous variable leading to ordinal ratings may be appropriate when ratings are not the result of a sequential process (9). The discrete observed ratings $\boldsymbol{R}$ are the result of the partition of the continuous space of $\boldsymbol{R}^*$ by a set of thresholds $\boldsymbol{\zeta}$ such that: $R_{ij} = 1$ if $-\infty < R^*_{ij} < \zeta_1$, $R_{ij} = 2$ if $\zeta_1 < R^*_{ij} < \zeta_2$, $\ldots$, $R_{ij} = K$ if $\zeta_{K-1} < R^*_{ij} < +\infty$ (see Figure 2).

We use the ordered probit model within a Latent Block Model (see Section 2.1), assuming that conditionally to row and column group assignments, the entries of $\boldsymbol{R}^*$ are independent and identically distributed with Gaussian distribution:

$$p(\, R^*_{ij} \Big| U_{iq}V_{jl} = 1; \mu_{ql}, \sigma\,) = \phi\Big(R^*_{ij}; \mu_{ql}, \sigma^2\Big) \;\; , \;\; \mu_{ql} \in \mathbb{R} \text{ and } \sigma \in \mathbb{R}^*_+ \qquad (2)$$

with $\phi\big(\cdot; \mu_{ql}, \sigma^2\big)$ the probability density function of the Gaussian distribution with mean μ_{ql} and variance σ^2. The conditional probability that a user i gives to the item j the rating with value k is then:

$$\begin{aligned} p(\, R_{ij} = k | U_{iq}V_{jl} = 1; \mu_{ql}\,) &= p(\, \zeta_{k-1} < R^*_{ij} < \zeta_k \Big| U_{iq}V_{jl} = 1; \mu_{ql}\,) \\ &= \Phi\Big(\zeta_k; \mu_{ql}, \sigma^2\Big) - \Phi\Big(\zeta_{k-1}; \mu_{ql}, \sigma^2\Big) \;\; , \end{aligned}$$

with $\Phi\big(\cdot; \mu_{ql}, \sigma^2\big)$ being the normal cumulative distribution function. To ensure model identifiability, the thresholds $\boldsymbol{\zeta}$ are fixed to equidistant predefined values.

2.2.2 Individual row and column effects

The Latent Block Model is well suited to collaborative filtering, in that it searches for users and items that share the same opinion patterns. However, a model that assumes that users in a given cluster share exactly the same opinion patterns is very restrictive. Instead, we assume here that opinions may be slightly different within a cluster, using a richer model than Equation (2) for the conditional distribution of R^*_{ij}. In addition to the cluster effect μ_{ql} derived solely from the group memberships of users and items, one deviation is induced by the user i and another by the item j :

$$p(\, R^*_{ij} \Big| U_{iq} V_{jl} = 1, A_i, B_j; \mu_{ql}\,) = \phi\Big(R^*_{ij}; \mu_{ql} + A_i + B_j, \sigma^2\Big) \ , \tag{3}$$

with latent variables $\boldsymbol{A}$ and $\boldsymbol{B}$ independently and identically distributed with:

$$A_i \overset{\text{iid}}{\sim} \mathcal{N}\Big(0, \sigma_A^2\Big), \qquad \sigma_A^2 \in \mathbb{R}_+^*$$
$$B_i \overset{\text{iid}}{\sim} \mathcal{N}\Big(0, \sigma_B^2\Big), \qquad \sigma_B^2 \in \mathbb{R}_+^*$$

These two variables encode different rating patterns for users and items such as systematic over- or under-rating relative to the user or item populations.

2.2.3 Sensitive attribute

We assume that, in addition to the matrix of ratings, we have access to a sensitive attribute s_i, describing here a binary feature of user i that should not intervene in the recommendation of items (more general sensitive attributes are considered in Appendix D). We introduce a latent variable C_j for each object j assuming that they interact with different strengths with the sensitive attribute. This interaction between the object j and the sensitive attribute s_i is added to the conditional distribution of R^*_{ij} (Equation 3):

$$p(\, R^*_{ij} \Big| U_{iq} V_{jl} = 1, A_i, B_j, s_i, C_j; \mu_{ql}\,) = \phi\Big(R^*_{ij}; \mu_{ql} + A_i + B_j + s_i C_j, \sigma^2\Big) \ ,$$

with

$$C_j \overset{\text{iid}}{\sim} \mathcal{N}\Big(0, \sigma_C^2\Big), \ \ \sigma_G^2 \in \mathbb{R}_+^* \ .$$

This model explains the ratings by $\mu_{ql} + A_i + B_j + s_i C_j$ and σ^2; the co-clustering is driven by μ_{ql}, and provided the effects of the sensitive attribute are well captured by $s_i C_j$, we expect the co-clustering to be independent of the sensitive attribute, which ensures fair recommendations as shown in Section 3.2. A summary of the model we propose is presented in Figure 3.

2.2.4 Modelling missingness

The datasets extracted from recommender systems are usually extremely sparse, with a high proportion of missing ratings, that is, ratings that were not provided by the users. The model we proposed so far does not accommodate missing observations, and suppose a fully observed data matrix $\boldsymbol{R}$.

The study of missing data identifies three main type of missingness (34): Missing Completely At Random (MCAR) and Missing At Random (MAR) referring to the mechanisms in which the probability of being missing does not depend on the variable of interest (here $\boldsymbol{R}^*$); and finally Missing Not At Random (NMAR) referring to the mechanisms in which the probability of being missing depends on the actual value of the missing data. A common implicit assumption in collaborative filtering is that ratings are MAR or CMAR: the presence/absence of ratings is assumed to convey no information whatsoever about the value of these ratings. For simplicity of statistical modelling we take the same assumption, although previous studies (27; 28) have shown a potential dependence between the presence of ratings and the underlying opinion. We introduce a simple Bernoulli missingness model generating $\boldsymbol{M} \in \{0,1\}^{n_1 \times n_2}$, a mask matrix where each entry M_{ij} is one with probability p and indicates whether the rating is observed: $M_{ij} = 1$ if R_{ij} is observed and 0 otherwise. Given the complete data matrix $\boldsymbol{R}^*$ and the mask matrix $\boldsymbol{M}$, the elements of the observed ratings $\boldsymbol{R}$ are generated as follows:

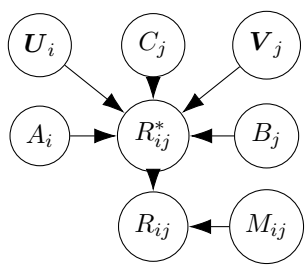
$$\left(R_{ij} \middle| R^*_{ij}, M_{ij}\right) = \begin{cases} \sum_{k=1}^{K} k \, \mathbb{1}_{]\zeta_{k-1};\zeta_k]}(R^*_{ij}) & \text{if} \quad M_{ij} = 1 \\ \text{NA} & \text{if} \quad M_{ij} = 0 \end{cases}$$

Any generative model under a MCAR or MAR process can be fitted separately from the missingness model as the overall likelihood can be factorized between the observed and non observed data. Under such assumptions, we show in AppendixA that ignoring non-observed ratings results in a proper fitting.

3 Inference and fair recommendations

3.1 A stochastic batch gradient descent of the variational criterion

The log-likelihood of the model is not tractable as it involves a sum that is combinatorially too large (7). We resort to a variational inference procedure (19) that introduces q_γ, a restricted parametric inference distribution defined



$$\boldsymbol{U}_i \overset{\text{iid}}{\sim} \mathcal{M}(1;\boldsymbol{\alpha}), \qquad \boldsymbol{\alpha} \in \mathbf{S}_{k_1-1}$$

$$\boldsymbol{V}_j \overset{\text{iid}}{\sim} \mathcal{M}(1;\boldsymbol{\beta}), \qquad \boldsymbol{\beta} \in \mathbf{S}_{k_2-1}$$

$$A_i \overset{\text{iid}}{\sim} \mathcal{N}\Big(0,\sigma_A^2\Big), \qquad \sigma_A^2 \in \mathbb{R}_+^*$$

$$B_j \overset{\text{iid}}{\sim} \mathcal{N}\Big(0,\sigma_B^2\Big), \qquad \sigma_B^2 \in \mathbb{R}_+^*$$

$$C_j \overset{\text{iid}}{\sim} \mathcal{N}\Big(0,\sigma_C^2\Big), \qquad \sigma_C^2 \in \mathbb{R}_+^*$$

$$\Big(R^*_{ij}\Big|U_{iq}=1,V_{jl}=1,A_i,B_j\Big) \overset{\text{ind}}{\sim} \mathcal{N}\Big(\mu_{ql}+A_i+B_j+s_iC_j,\sigma^2\Big)$$

$$\Big(R_{ij}\Big|R^*_{ij},M_{ij}\Big) = \begin{cases} \sum_{k=1}^{K} k\ \mathbb{1}_{]\zeta_{k-1};\zeta_k]}(R^*_{ij}) & \text{if} \quad M_{ij}=1 \\ \text{NA} & \text{if} \quad M_{ij}=0 \end{cases}$$

with $M_{ij} \overset{\text{iid}}{\sim} \mathcal{B}(p), \qquad p \in [0,1]$

and $\zeta_0 = -\infty < \zeta_1 < ... < \zeta_{K-1} < \zeta_K = \infty,$ fixed thresholds

Figure 3: Graphical view and summary of the ordered probit Latent Block Model with protected attribute $\boldsymbol{s}$. The discrete observed data R_{ij} is generated by the underlying continuous data R^*_{ij} and the mask entry M_{ij}.

on the latent variables of the model, to optimize the following lower bound on the log-likelihood:

$$\mathcal{J}(\gamma, \theta) = \log p(\boldsymbol{R}; \theta) - \mathrm{KL}\left(q_\gamma \,\|p(L|\boldsymbol{R})\right)$$

where KL stands for the Kullback-Leibler divergence, $\mathcal{H}$ for the differential entropy, $\theta = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \sigma^2, \sigma_A^2, \sigma_B^2, \sigma_C^2, p)$ is the concatenation of the model parameters, and $L = (\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ is the concatenation of the latent variables.

The variational distribution q_γ is chosen so that the computation of the criterion becomes easier:

$$\begin{array}{llll}
\forall i, & \boldsymbol{U}_i|\boldsymbol{R} \underset{q_\gamma}{\sim} \mathcal{M}\left(1; \boldsymbol{\tau}_i^{(U)}\right) & \forall j, & \boldsymbol{V}_j|\boldsymbol{R} \underset{q_\gamma}{\sim} \mathcal{M}\left(1; \boldsymbol{\tau}_j^{(V)}\right) \\
\forall i, & A_i|\boldsymbol{R} \underset{q_\gamma}{\sim} \mathcal{N}\left(\nu_i^{(A)}, \rho_i^{(A)}\right) & \forall j, & B_j|\boldsymbol{R} \underset{q_\gamma}{\sim} \mathcal{N}\left(\nu_j^{(B)}, \rho_j^{(B)}\right) \\
 & & \forall j, & C_j|\boldsymbol{R} \underset{q_\gamma}{\sim} \mathcal{N}\left(\nu_j^{(C)}, \rho_j^{(C)}\right) \ .
\end{array}$$

We also enforce the conditional independence of the latent variables, leading to the following fully factorized form:

$$\begin{aligned}
q_\gamma = & \prod_{i=1}^{n_1} \mathcal{M}\left(1; \boldsymbol{\tau}_i^{(U)}\right) \times \prod_{j=1}^{n_2} \mathcal{M}\left(1; \boldsymbol{\tau}_j^{(V)}\right) \\
& \times \prod_{i=1}^{n_1} \mathcal{N}\left(\nu_i^{(A)}, \rho_i^{(A)}\right) \times \prod_{j=1}^{n_2} \mathcal{N}\left(\nu_j^{(B)}, \rho_j^{(B)}\right) \\
& \times \prod_{j=1}^{n_2} \mathcal{N}\left(\nu_j^{(C)}, \rho_j^{(C)}\right) \ ,
\end{aligned} \tag{4}$$

where γ denotes the concatenation of all parameters of the variational distribution[1]. This conditional independence of the latent variables to $\boldsymbol{R}$ simplifies the criterion $\mathcal{J}(\gamma, \theta)$ to:

$$\mathcal{J}(\gamma, \theta) = \mathbb{E}_{q_\gamma}[\log p(\boldsymbol{R}|L)] - \mathrm{KL}\left(q_\gamma \,\|p(L; \theta)\right) \ . \tag{5}$$

As explained in Section 2.2.4, the optimization criterion relies only on the non-missing entries of $\boldsymbol{R}$ because the data is assumed to be missing at random. The full expansion of the criterion is given in Appendix A.

We resort to a batch stochastic optimization to maximize the variational criterion using noisy estimates of its gradient (31). Samples are drawn from the variational distribution (Equation 4) to estimate a noisy but unbiased gradient of the expectation of the conditional log-distribution of $\boldsymbol{R}$ (first term of Equation 5), which we then use to update our parameters as follows:

$$(\gamma, \theta)^{(t+1)} = (\gamma, \theta)^{(t)} + \eta \cdot \nabla_{(\gamma,\theta)} \mathcal{J}\left(\boldsymbol{R}_{(i:i+n),(j:j+n)}; \gamma, \theta\right) \ ,$$

[1] $\gamma = (\boldsymbol{\tau}^{(U)}, \boldsymbol{\tau}^{(V)}, \boldsymbol{\nu}^{(A)}, \boldsymbol{\rho}^{(A)}, \boldsymbol{\nu}^{(B)}, \boldsymbol{\rho}^{(B)}, \boldsymbol{\nu}^{(C)}, \boldsymbol{\rho}^{(C)})$

where n is the batch size and η is the adaptive learning rate based on the past gradients that were computed (Adam optimizer (24)).

Using a stochastic gradient algorithm instead of the usual EM algorithm alleviates the well-known initialization problems of the Latent Block Model, which result in unsatisfactory local maxima (5; 2). However, it requires the use of differentiable functions to back-propagate gradients through the automatic differentiation graph (30). For this purpose, the multinomial distributions are replaced by a differentiable Gumbel-Softmax distribution (20).

3.2 Fair recommendations

This section describes a theoretical result establishing a guarantee on the fairness of recommendations. This guarantee is subject to an assumption about the parity of the clustering of users that can be tested in practice, and that holds true for the experiments reported in Section 4 and Appendix D. We develop here the case of a binary sensitive attribute to simplify the exposition. The result is more general and applies to any discrete sensitive attribute. It is proven in this general sense in Appendix C.

Recommendations are partial orders between items. In collaborative filtering, the usual approach to producing recommendations is to estimate a relevance score for each item, which is then used to define a total order through numerical comparisons. With the parameters obtained by variational inference, we define the relevance score of item j for user i as:

$$\hat{R}_{ij} = \boldsymbol{\tau}_i^{(U)} \hat{\boldsymbol{\mu}} \boldsymbol{\tau}_j^{(V)^T} + \nu_i^{(A)} + \nu_j^{(B)} \quad . \tag{6}$$

This relevance score is computed from the maxima *a posteriori* of the latent variables encoding the user and item group memberships $(\boldsymbol{\tau}_i^{(U)}, \boldsymbol{\tau}_j^{(V)})$, that is, the trend related to the co-cluster to which (i, j) belongs, and the global effects related to user i and item j. It does not use the user's sensitive attribute s_i which is considered here as a nuisance parameter, properly taken into account during inference and then ignored when predicting a relevance score. It then becomes possible to compare items fairly with respect to the sensitive attribute.

Definition 3.1 (Fair comparison of items)**.** Given user i and any two items j and j', the comparison of items j and j' is said to be fair if it is freed from the evaluation bias regarding the sensitive attribute s: item j is fairly

preferred to item j' if $\hat{R}_{ij} > \hat{R}_{ij'}$, that is:

$$\boldsymbol{\tau}_i^{(U)} \hat{\boldsymbol{\mu}} \boldsymbol{\tau}_j^{(V)^T} + \nu_i^{(A)} + \nu_j^{(B)} > \boldsymbol{\tau}_i^{(U)} \hat{\boldsymbol{\mu}} \boldsymbol{\tau}_{j'}^{(V)^T} + \nu_i^{(A)} + \nu_{j'}^{(B)} \ .$$

The modelling of the observed data $\boldsymbol{R}$ incorporates the term $\nu_j^{(C)} s_i$, interpreted here as a spurious opinion bias related to the sensitive attribute. While it is important to ignore this term for a fair comparison of items, its inclusion into the model is important to allow the construction of clusters that are not affected by this spurious effect. These clusters can then be expected to be representative of all subpopulations defined by their sensitive attribute value, and thus to respect the statistical parity of users.

Definition 3.2 (Clustering ε-parity, binary sensitive attribute)**.** The clustering of users is said to respect ε-parity with respect to attribute s iff:

$$\forall q, \qquad \left| \frac{\#\{i | s_i = 1 \wedge u_{iq} = 1\}}{\#\{i | s_i = 1\}} - \frac{\#\{i | s_i = -1 \wedge u_{iq} = 1\}}{\#\{i | s_i = -1\}} \right| \leq \varepsilon \ , \tag{7}$$

where $\varepsilon \in \mathbb{R}_+$ measures the gap to exact parity, u_{iq} is the (hard) membership of user i to cluster q,and $\#\{i|\Omega\}$ is the number of users defined by the cardinality of the set Ω.

In essence, clustering ε-parity requires that subpopulations of users defined by identical sensitive attributes be represented approximately equally in each user group. For the Latent Block Model, the hard membership u_{iq} of Definition 3.2 is given by the maximum *a posteriori* of the latent variable $\tau_{iq}^{(U)}$.

Our theoretical guarantee ensures that this approximate statistical parity in clusters is sufficient to get approximately fair recommendations from our model:

Definition 3.3 (ε-fair recommendation, binary sensitive attribute)**.** A recommender system is said to be ε-fair with respect to attribute s if for any two items j and j':

$$\left| \frac{\#\left\{i | s_i = 1 \wedge (\hat{R}_{ij} > \hat{R}_{ij'})\right\}}{\#\{i | s_i = 1\}} - \frac{\#\left\{i | s_i = -1 \wedge (\hat{R}_{ij} > \hat{R}_{ij'})\right\}}{\#\{i | s_i = -1\}} \right| \leq \varepsilon \ , \tag{8}$$

where $\varepsilon \in \mathbb{R}_+$ measures the gap to exact fairness

In essence, an ε-fair recommender system ensures that, for any two items, the proportion of users with the same preference is approximately identical in all the subpopulations of users defined by identical sensitive attributes.

Theorem 3.1 (Fair recommendation from clustering parity)**.** If the clustering of users in k_1 groups respects ε-parity (Definition 3.2 or Definition S1) then the recommender system relying on the relevance score defined in Equation (6) is $(k_1\varepsilon)$-fair (Definition 3.3 or Definition S2).

Proof: see Appendix C.

4 Experiment on MovieLens dataset

The final goal of a recommender system is to provide users with a shortlist of items that they might most enjoy. We choose here to directly assess the quality of the ranking rather than using proxy measures, such as root mean square error on ratings, that ignore relative rankings.

To measure the ranking performance of algorithms, we use the Normalized Discounted Cumulative Gain (21) (NDCG) that measures ranking quality by a penalized sum of the relevance scores of the ranking results:

$$NDCG@k = \frac{DCG@k}{IDCG@k} \text{ with } DCG@k = \sum_{i=1}^{k} \frac{rel_i}{\log(i+1)} \ ,$$

rel_i, the relevance of the results at each rank i before k and $IDCG@k$ being the $DCG@k$ computed with a perfect ranking.

We use the MovieLens 1M dataset (29) that contains one million ratings given by 6,040 users to 3,900 movies scaling from 1 to 5 (from least liked to most liked). The dataset also contains additional information about users: gender (binary), age category (seven levels) or occupation. We give here some experimental results where gender is the sensitive attribute, and additional results, in particular with age considered as the sensitive attribute, can be found in Appendix D.

4.1 Experimental Protocol

We estimate the average performances by predicting preferences on ratings that are concealed during training. These concealed ratings form our test set, with 20 ratings per user, which is about 10% of the available data. This process is repeated 5 times, with independent random draws, to produce stable average performances.

We compare our model (referred to as Parity LBM) with the baseline LBM that does not use the sensitive variable in the modelling (referred to as Standard LBM). We expect the latter model to create groups of users that do not respect clustering parity and to generate unfair recommendations.

We also compare to another co-clustering algorithm, weighted Bregman co-clustering (13) (referred to as Bregman co-clust) to compare the statistical parity of user groups inferred from another baseline. Finally, we compare with Singular Value Decomposition (SVD), a method popularized during the Netflix challenge (25) that still remains state of the art in collaborative filtering (33). All these baselines are implemented in the Python module `Surprise` (18).

The number of clusters in co-clustering and the number of factors in matrix factorization are both arbitrarily set to fifteen. Another comparison with more clusters, provided in Appendix D, produces qualitatively similar results.

We repeat the learning process 25 times from different random initializations to mitigate the initialization dependence that affects all optimization procedures. We select the best solution based on the optimization criteria, that is, the one with the highest likelihood for the LBM models and the lowest training reconstruction error for the other baselines.

4.2 Results and Discussion

4.2.1 Gender as sensitive attribute

User gender (binary in this dataset) is used as the sensitive attribute s_i. In the dataset, 27% of users self-identified as females, this proportion must be met in each group to respect clustering parity. To measure the dependence between gender and user group memberships, we compute the χ^2 statistic constructed from the contingency table of males and females counts in each group. Table 1 reports the p-value for testing the independence between groups and genders, with an asymptotical test. We recall that, under the null hypothesis of independence, the test statistic with k degrees of freedom has mean k and variance $2k$.

The fairness of recommendations resulting from this clustering parity is ascertained by computing the gap ε from exactly fair recommendations, as defined in Definition 3.3. Figure 4 displays these gaps, with lower values indicating a fairer recommendation; our model provides a significantly fairer recommendation compared to the standard Latent Block Model, which is itself much fairer than the two other baselines. The order observed in Table 1 is followed.

Figure 5 depicts the ranking performance of algorithms with the NDCG, averaged over all users, for a recommendation list of 10 items. SVD gets the best overall result, followed by the Latent Block Models that outper-

Table 1: Measures of statistical gender parity among user clusters. The number of user groups is $k_1 = 15$. The χ^2 statistic (with 14 degrees of freedom) is averaged over the five replicates of the experiment. A high value of the χ^2 statistic (or a low p-value) leads to the rejection of the clustering parity hypothesis.

Model	Parity LBM	Standard LBM	Bregman co-clust
χ^2 statistic	18.0	44.4	187
p-value	0.20	$5.1 \cdot 10^{-5}$	$< 10^{-15}$.

The results show that the methods that do not consider the sensitive variable in the modelling create groups that are dependent on gender. In contrast, our Parity-LBM model is consistent with the clustering parity hypothesis: the gender representation in groups is representative of the gender distribution in the overall dataset.

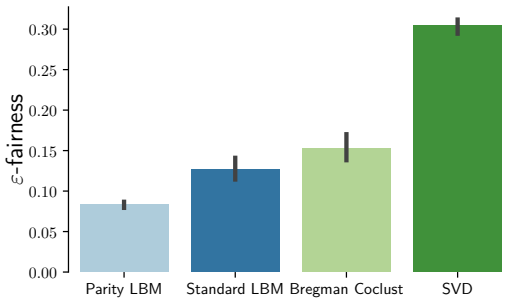


Figure 4: Gaps ε for the ε-fair recommendations (see Definition 3.3) provided by each model: a smaller ε-fairness indicates fairer recommendations.

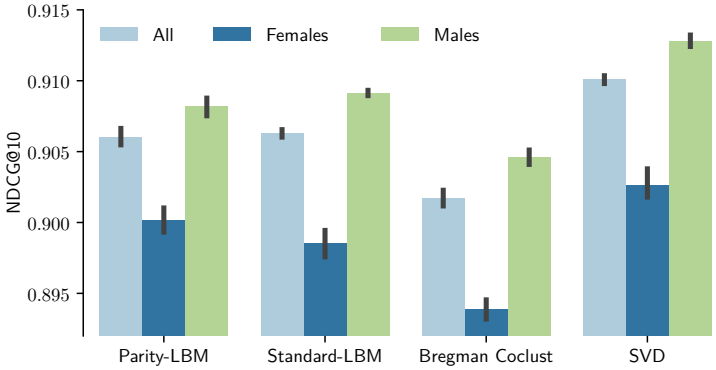


Figure 5: Normalized Discounted Cumulative Gain estimated on MovieLens-1M (the higher the better)

form Bregman co-clustering. The overall performances of our model and the standard LBM are not significantly different. Figure 5 also reports the average NDCG within each sensitive group. This performance measure shows that female users receive significantly less relevant recommendations than males with all algorithms. This disparate treatment can be related to equalized odds (17) in the classification framework, in that it focuses on truly relevant recommendations. The performance gap between the sensitive groups is reduced by our parity LBM compared to the standard LBM. Although the difference is the smallest among all comparisons, our model does not eliminate disparate treatment. As a cautionary note, although it is likely that the recommendations are less relevant to female users, under the assumption that the observed ratings are somewhat influenced by gender stereotypes, it is not possible to satisfactorily measure the performance of fair recommendations from the original rating matrix.

Finally, we present some insights provided by our model on movies. We recall that the latent variable C_j, which is not used for fair prediction, captures the difference in opinion trends between female and male users on movie j. A high absolute value of C_j indicates a strongly gendered opinion for movie j. With our encoding of genres, negative C_j indicate a relative overrating by females and positive C_j indicate a relative overrating by males. We display the empirical cumulative distribution function (CDF) of C_j for movies conditionally on their genre (for some handpicked archetypal genres). The dominance of the CDF for a given genre expresses that, according to

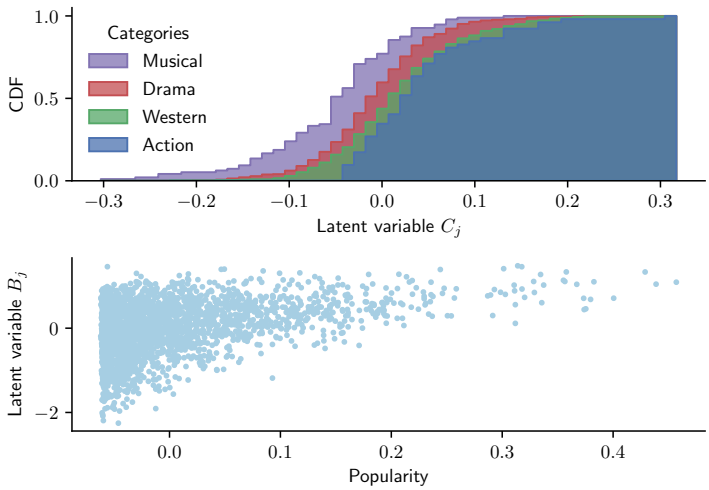


Figure 6: Top: cumulative distribution function of latent variable C_j conditionally on the genre of the movie. A dominating CDF indicates a genre for which females' opinions are more positive than males'. Bottom: scatter plot of the movie latent variable B_j versus popularity (ratio of ratings). High positive values of B_j (resp. popularity) correspond to movies that are the most liked (resp. popular).

our model, female users have a higher opinion than male users for the movies belonging to that genre. Figure 6 shows the results, which reflect stereotypes that women are more likely than men to positively evaluate musical films and dramas, while men are similarly inclined toward westerns and action films. These stereotypes are incorporated into our model to fit actual ratings, but ignored to deliver fair recommendations. The lists of extreme movies based on extreme (positive and negative) values of C_j is given in Appendix D.1.

The latent variable B_j encodes the overall opinion trend about movie j. Two interesting observations can be made from the scatter plot of B_j versus movie popularity (see bottom of Figure 6). First, unpopular movies are also the least appreciated according to our model; this supports the hypothesis that ratings are generated by a MNAR (Missing Not At Random) process, where a missing rating can be considered as weak negative feedback, assuming that users primarily rate items they like. This missingness process must still be taken into account in our model. Second, it shows that the most liked movies (according to our model) are not necessarily the most

popular (and will be recommended); the recommendations are not affected by popularity bias.

5 Conclusion

We proposed a new co-clustering method for fair recommendation. Our model combines the Gaussian Latent Block Model with an ordinal regression model. The sensitive attribute is adequately accounted for in the model, allowing the clustering of users to be unaffected by the effects of this attribute on ratings. This results in user clusters that approximately respect statistical parity. We base recommendation on a relevance score that ignores the sensitive attribute in order to compare items fairly. We provide theoretical guarantees ensuring approximately fair recommendations, for any known discrete sensitive attribute, provided that the clustering of users respects an approximate statistical parity that can be assessed in practice. Our analysis focuses on the fairness of preferences, as defined by the ranking of ratings, rather than on the predicted values themselves, which are less relevant for recommendation. Through experiments on real-world data, we show that our method significantly mitigates the unfairness of recommendations. Furthermore, the latent variables inferred by the model are also amenable to analyses that can help identify recommendation bias.

Our study supports that the absence of rating conveys some information that should be exploited. Previous works (27; 28) have already shown that the data used for collaborative filtering datasets can be strongly influenced by observational bias, which motivates dealing with missingness by a Missing Not At Random (MNAR) process. Societal biases may have a significant contribution to missingness, leading to an additional source of unfairness if missingness is not properly modeled. Studying fairness with MNAR processes is a highly relevant but extremely challenging direction for future research, as assessing the relevance of MNAR models in real situations requires data that are typically produced by online randomized experiments.

References

[1] Abbasi, M., Bhaskara, A., Venkatasubramanian, S.: Fair clustering via equitable group representations. In: Elish, M.C., Isaac, W., Zemel, R.S. (eds.) ACM Conference on Fairness, Accountability, and Transparency (FAccT). pp. 504–514 (2021), https://doi.org/10.1145/3442188.3445913

[2] Baudry, J.P., Celeux, G.: EM for mixtures. Statistics and Computing **25**(4), 713–726 (2015). https://doi.org/10.1007/s11222-015-9561-x

[3] Bera, S.K., Chakrabarty, D., Flores, N.J., Negahbani, M.: Fair algorithms for clustering (2019)

[4] Beutel, A., Chen, J., Doshi, T., Qian, H., Wei, L., Wu, Y., Heldt, L., Zhao, Z., Hong, L., Chi, E.H., Goodrow, C.: Fairness in Recommendation Ranking through Pairwise Comparisons, pp. 2212—-2220 (2019), https://doi.org/10.1145/3292500.3330745

[5] Biernacki, C., Celeux, G., Govaert, G.: Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. Computational Statistics & Data Analysis **41**, 561–575 (2003). https://doi.org/10.1016/S0167-9473(02)00163-9

[6] Binns, R.: Fairness in machine learning: Lessons from political philosophy. In: Friedler, S.A., Wilson, C. (eds.) Proceedings of the 1st Conference on Fairness, Accountability and Transparency. Proceedings of Machine Learning Research, vol. 81, pp. 149–159. PMLR, New York, NY, USA (23–24 Feb 2018), http://proceedings.mlr.press/v81/binns18a.html

[7] Brault, V., Mariadassou, M.: Co-clustering through latent bloc model: A review. Journal de la Société Française de Statistique **156**(3), 120–139 (2015), http://journal-sfds.fr/article/view/474/448

[8] Burke, R., Sonboli, N., Ordonez-Gauger, A.: Balanced neighborhoods for multi-sided fairness in recommendation. In: 1st Conference on Fairness, Accountability and Transparency. PMLR, vol. 81, pp. 202–214 (2018), http://proceedings.mlr.press/v81/burke18a.html

[9] Bürkner, P.C., Vuorre, M.: Ordinal regression models in psychology: A tutorial. Advances in Methods and Practices in Psychological Science **2**(1), 77–101 (2019)

[10] Daykin, A.R., Moffatt, P.G.: Analyzing ordered responses: A review of the ordered probit model. Understanding Statistics **1**(3), 157–166 (2002). https://doi.org/10.1207/S15328031US0103_02

[11] Daymonti, T.N., Andrisani, P.J.: Job preferences, college major, and the gender gap in earnings. Journal of Human Resources **19**(3), 408–428 (1984), https://EconPapers.repec.org/RePEc:uwp:jhriss:v:19:y:1984:i:3:p:408-428

[12] Gajane, P.: On formalizing fairness in prediction with machine learning. CoRR **abs/1710.03184** (2017), http://arxiv.org/abs/1710.03184

[13] George, T., Merugu, S.: A scalable collaborative filtering framework based on co-clustering. In: Fifth IEEE International Conference on Data Mining (ICDM) (2005)

[14] Ghadiri, M., Samadi, S., Vempala, S.: Socially fair k-means clustering. arXiv preprint arXiv:2006.10085 (2020)

[15] Govaert, G., Nadif, M.: Block clustering with Bernoulli mixture models: Comparison of different approaches. Computational Statistics & Data Analysis **52**(6), 3233–3245 (February 2008)

[16] Govaert, G., Nadif, M.: Latent block model for contingency table. Communications in Statistics - Theory and Methods **39**(3), 416–425 (2010). https://doi.org/10.1080/03610920903140197

[17] Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Advances in Neural Information Processing Systems 29. pp. 3315–3323 (2016), https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html

[18] Hug, N.: Surprise: A python library for recommender systems. Journal of Open Source Software **5**(52), 2174 (2020). https://doi.org/10.21105/joss.02174

[19] Jaakkola, T.S.: Tutorial on variational approximation methods. In: Advanced Mean Field Methods: Theory and Practice, pp. 129–159. MIT Press (2000)

[20] Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. arXiv preprint arXiv:1611.01144 (2016)

[21] Järvelin, K., Kekäläinen, J.: IR evaluation methods for retrieving highly relevant documents. In: ACM SIGIR Forum. vol. 51, pp. 243–250 (2017)

[22] Kamishima, T., Akaho, S., Asoh, H., Sakuma, J.: Recommendation independence. In: Conference on Fairness, Accountability and Transparency. pp. 187–201 (2018)

[23] Keribin, C., Brault, V., Celeux, G., Govaert, G.: Estimation and selection for the latent block model on categorical data. Statistics and Computing **25**(6), 1201–1216 (2015), https://hal.inria.fr/hal-01095957

[24] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

[25] Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. Computer **42**(8), 30–37 (Aug 2009)

[26] Lomet, A., Govaert, G., Grandvalet, Y.: Model Selection for Gaussian Latent Block Clustering with the Integrated Classification Likelihood. Advances in Data Analysis and Classification **12**(3), 489–508 (2018), https://hal.archives-ouvertes.fr/hal-00913680

[27] Marlin, B.M., Zemel, R.S., Roweis, S.T., Slaney, M.: Collaborative filtering and the missing at random assumption. In: Twenty-Third Conference on Uncertainty in Artificial Intelligence (UAI). pp. 267–275 (2007)

[28] Marlin, B.M., Zemel, R.S., Roweis, S.T., Slaney, M.: Collaborative filtering and the missing at random assumption. CoRR **abs/1206.5267** (2012), http://arxiv.org/abs/1206.5267

[29] Movielens 1M datasets. https://grouplens.org/datasets/movielens/

[30] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc. (2019), http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[31] Ranganath, R., Gerrish, S., Blei, D.: Black Box Variational Inference. In: Kaski, S., Corander, J. (eds.) Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 33, pp. 814–822. PMLR, Reykjavik, Iceland (22–25 Apr 2014), http://proceedings.mlr.press/v33/ranganath14.html

[32] Räz, T.: Group fairness: Independence revisited. arXiv preprint arXiv:2101.02968 (2021)

[33] Rendle, S., Zhang, L., Koren, Y.: On the difficulty of evaluating baselines: A study on recommender systems. arXiv preprint arXiv:1905.01395 (2019)

[34] Rubin, D.B.: Inference and missing data. Biometrika **63**(3), 581–592 (1976), http://www.jstor.org/stable/2335739

[35] Yao, S., Huang, B.: Beyond parity: Fairness objectives for collaborative filtering. CoRR **abs/1705.08804** (2017), http://arxiv.org/abs/1705.08804

[36] Zhu, Z., Hu, X., Caverlee, J.: Fairness-aware tensor-based recommendation. In: 27th ACM International Conference on Information and Knowledge Management. pp. 1153—-1162 (2018). https://doi.org/10.1145/3269206.3271795

A Computation of the variational log-likelihood criterion

The criterion we want to optimize is:

$$\mathcal{J}(q_\gamma, \theta) = \mathcal{H}(q_\gamma) + \mathbb{E}_{q_\gamma}\left[\mathcal{L}(\boldsymbol{R}, \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}; \theta)\right] \quad . \tag{S9}$$

We chose to restrict the space of the variational distribution q_γ in order to get a fully factorized form:

$$\begin{aligned} q_\gamma = \textstyle\prod_{i=1}^{n_1} \mathcal{M}\Big(1; \tau_i^{(U)}\Big) \ \times \ \prod_{j=1}^{n_2} \mathcal{M}\Big(1; \tau_j^{(V)}\Big) \\ \times \textstyle\prod_{i=1}^{n_1} \mathcal{N}\Big(\nu_i^{(A)}, \rho_i^{(A)}\Big) \times \prod_{j=1}^{n_2} \mathcal{N}\Big(\nu_j^{(B)}, \rho_j^{(B)}\Big) \\ \times \textstyle\prod_{j=1}^{n_2} \mathcal{N}\Big(\nu_j^{(C)}, \rho_j^{(C)}\Big) \end{aligned} \tag{S10}$$

where γ denotes the parameters concatenation of the variational distribution[2] q_γ. The entropy is additive across independant variables so we get:

$$\mathcal{H}(q_\gamma) = \mathcal{H}(q_\gamma(\boldsymbol{U})) + \mathcal{H}(q_\gamma(\boldsymbol{V})) + \mathcal{H}(q_\gamma(\boldsymbol{A})) + \mathcal{H}(q_\gamma(\boldsymbol{B})) + \mathcal{H}(q_\gamma(\boldsymbol{C})) \ ,$$

with the following terms:

$$\begin{aligned} \mathcal{H}(q_\gamma(\boldsymbol{U})) &= -\sum_{iq} \tau_{iq}^{(U)} \log \tau_{iq}^{(U)} \\ \mathcal{H}(q_\gamma(\boldsymbol{V})) &= -\sum_{jl} \tau_{jl}^{(U)} \log \tau_{jl}^{(V)} \\ \mathcal{H}(q_\gamma(\boldsymbol{A})) &= \frac{1}{2} \sum_i \log \rho_i^{(A)} + \frac{n_1}{2}(\log 2\pi + 1) \\ \mathcal{H}(q_\gamma(\boldsymbol{B})) &= \frac{1}{2} \sum_j \log \rho_j^{(B)} + \frac{n_2}{2}(\log 2\pi + 1) \\ \mathcal{H}(q_\gamma(\boldsymbol{C})) &= \frac{1}{2} \sum_j \log \rho_j^{(C)} + \frac{n_2}{2}(\log 2\pi + 1) \end{aligned}$$

The independence of the latent variables allows to rewrite the expectation of the complete log-likelihood as:

$$\begin{aligned} \mathbb{E}_{q_\gamma}[\mathcal{L}(\boldsymbol{R}, \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})] = \mathbb{E}_{q_\gamma}[\mathcal{L}(\boldsymbol{U})] + \mathbb{E}_{q_\gamma}[\mathcal{L}(\boldsymbol{V})] \\ + \mathbb{E}_{q_\gamma}[\mathcal{L}(\boldsymbol{A})] + \mathbb{E}_{q_\gamma}[\mathcal{L}(\boldsymbol{B})] + \mathbb{E}_{q_\gamma}[\mathcal{L}(\boldsymbol{C})] \\ + \mathbb{E}_{q_\gamma}[\mathcal{L}(\boldsymbol{R}|\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})] \ , \end{aligned}$$

[2] $\gamma = (\boldsymbol{\tau}^{(U)}, \boldsymbol{\tau}^{(V)}, \boldsymbol{\nu}^{(A)}, \boldsymbol{\rho}^{(A)}, \boldsymbol{\nu}^{(B)}, \boldsymbol{\rho}^{(B)}, \boldsymbol{\nu}^{(C)}, \boldsymbol{\rho}^{(C)})$

with the following terms:

$$\mathbb{E}_{q_\gamma}\mathcal{L}(\boldsymbol{U}) = \mathbb{E}_{q_\gamma}\left[\sum_{iq} U_{iq}\log\alpha_q\right] = \sum_{iq}\tau_{iq}^{(U)}\log\alpha_q$$

$$\mathbb{E}_{q_\gamma}\mathcal{L}(\boldsymbol{V}) = \mathbb{E}_{q_\gamma}\left[\sum_{jl} V_{jl}\log\beta_l\right] = \sum_{jl}\tau_{jl}^{(V)}\log\beta_l$$

$$\begin{aligned}
\mathbb{E}_{q_\gamma}\mathcal{L}(\boldsymbol{A}) &= -\frac{n_1}{2}\log 2\pi - \frac{n_1}{2}\log\sigma_A^2 - \frac{1}{2\sigma_A^2}\sum_i \mathbb{E}_{q_\gamma}A_i^2 \\
&= -\frac{n_1}{2}\log 2\pi - \frac{n_1}{2}\log\sigma_A^2 - \frac{1}{2\sigma_A^2}\sum_i\left(\left(\nu_i^{(A)}\right)^2 + \rho_i^{(A)}\right) \\
\mathbb{E}_{q_\gamma}\mathcal{L}(\boldsymbol{B}) &= -\frac{n_2}{2}\log 2\pi - \frac{n_2}{2}\log\sigma_B^2 - \frac{1}{2\sigma_B^2}\sum_i\left(\left(\nu_i^{(B)}\right)^2 + \rho_i^{(B)}\right) \\
\mathbb{E}_{q_\gamma}\mathcal{L}(\boldsymbol{C}) &= -\frac{n_2}{2}\log 2\pi - \frac{n_2}{2}\log\sigma_C^2 - \frac{1}{2\sigma_C^2}\sum_j\left(\left(\nu_j^{(C)}\right)^2 + \rho_j^{(C)}\right)
\end{aligned}$$

and as the entries of the data matrix $\boldsymbol{R}$ are independent and identically distributed:

$$\mathbb{E}_{q_\gamma}\mathcal{L}(\boldsymbol{R}|\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{U},\boldsymbol{V}) = \mathbb{E}_{q_\gamma}\mathcal{L}\Big(\boldsymbol{R}^{(\mathrm{o})}\Big|\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{U},\boldsymbol{V}\Big) + \mathcal{L}\Big(\boldsymbol{R}^{(\neg o)}\Big) \quad \text{(S11)}$$

where $\boldsymbol{R}^{(\mathrm{o})}$ denotes the set of observed ratings and $\boldsymbol{R}^{(\neg o)}$, the set of non-observed ratings, where $R_{ij} = \mathrm{NA}$. From Equation S11, it becomes clear that maximizing $\mathbb{E}_{q_\gamma}\mathcal{L}(\boldsymbol{R}^{(\neg o)})$ is not necessary to infer the model parameters used for prediction and therefore ignoring the non-observed data is correct. The expectation of the conditional log-likelihood (first term of right side of Equation S11) is numerically estimated by sampling from q_γ.

Stochastic gradient optimization To optimize the criterion with stochastic gradient descent, we express the variational log-likelihood criterion on a

single rating:

$$\begin{aligned}\mathcal{J}(R_{ij};q_\gamma,\theta) = \mathbb{E}_{q_\gamma}\Big[\mathcal{L}\Big(R_{ij}^{(\mathrm{o})}\Big|\boldsymbol{U}_i,\boldsymbol{V}_j,A_i,B_j,C_j\Big)\Big] \\ + \frac{1}{n_2}(\mathcal{H}(q_\gamma(\boldsymbol{U}_i)) + \mathcal{H}(q_\gamma(A_i)) + \mathbb{E}_{q_\gamma}[\mathcal{L}(\boldsymbol{U}_i)] + \mathbb{E}_{q_\gamma}[\mathcal{L}(A_i)]) \\ + \frac{1}{n_2}(\mathcal{H}(q_\gamma(\boldsymbol{V}_j)) + \mathcal{H}(q_\gamma(B_j)) + \mathbb{E}_{q_\gamma}[\mathcal{L}(\boldsymbol{V}_j)] + \mathbb{E}_{q_\gamma}[\mathcal{L}(B_j)]) \\ + \frac{1}{n_2}(\mathcal{H}(q_\gamma(C_j)) + \mathbb{E}_{q_\gamma}[\mathcal{L}(C_j)])\end{aligned}$$

A batch of data, $\boldsymbol{R}_{(i:i+n),(j:j+n)}$, consists of a $(n \times n)$ sub-matrix randomly sampled from the original matrix $\boldsymbol{R}$.

B Clustering ε-parity and ε-fair recommendation for arbitrary discrete sensitive attribute

Definition S1 (Clustering ε-parity, arbitrary discrete sensitive attribute)**.** The clustering of users is said to respect ε-parity with respect to the discrete attribute $s \in \mathcal{S}$ iff:

$$\forall (t,t') \in \mathcal{S}^2,\ \forall q,\ \left|\frac{\#\{i|s_i = t \wedge u_{iq} = 1\}}{\#\{i|s_i = t\}} - \frac{\#\{i|s_i = t' \wedge u_{iq} = 1\}}{\#\{i|s_i = t'\}}\right| \leq \varepsilon\ , \tag{S12}$$

where $\varepsilon \in \mathbb{R}_+$ measures the gap to exact parity, u_{iq} is the (hard) membership of user i to cluster q,and $\#\{i|\Omega\}$ is the number of users defined by the cardinality of the set Ω.

Definition S2 (ε-fair recommendation, arbitrary discrete sensitive attribute)**.** A recommender system is said to be ε-fair with respect to the dicrete attribute $s \in \mathcal{S}$ if for any two items j and j':

$$\forall (t,t') \in \mathcal{S}^2,\ \left|\frac{\#\left\{i|s_i = t \wedge (\hat{R}_{ij} > \hat{R}_{ij'})\right\}}{\#\{i|s_i = t\}} - \frac{\#\left\{i|s_i = t' \wedge (\hat{R}_{ij} > \hat{R}_{ij'})\right\}}{\#\{i|s_i = t'\}}\right| \leq \varepsilon\ , \tag{S13}$$

where $\varepsilon \in \mathbb{R}_+$ measures the gap to exact fairness

C Proof of Theorem 3.1

Theorem C.1 (Fair recommendation from clustering parity)**.** If the clustering of users in k_1 groups respects ε-parity (Definition 3.2 or Definition S1)

then the recommender system relying on the relevance score defined in Equation (6) is $(k_1\varepsilon)$-fair (Definition 3.3 or Definition S2).

Proof. Suppose that $\boldsymbol{\tau}^{(U)}$, the maximum *a posteriori* of $\boldsymbol{U}$, is a binary matrix; $\boldsymbol{\tau}^{(U)}$ is thus a $n_1 \times k_1$ indicator matrix of row classes membership. Then, given user i, item j is said to be preferred to item j' if $\hat{R}_{ij} > \hat{R}_{ij'}$, that is:

$$\begin{aligned}
\hat{R}_{ij} > \hat{R}_{ij'} &\iff \boldsymbol{\tau}_i^{(U)}\hat{\boldsymbol{\mu}}\boldsymbol{\tau}_j^{(V)^T} + \nu_i^{(A)} + \nu_j^{(B)} > \boldsymbol{\tau}_i^{(U)}\hat{\boldsymbol{\mu}}\boldsymbol{\tau}_{j'}^{(V)^T} + \nu_i^{(A)} + \nu_{j'}^{(B)} \\
&\iff \boldsymbol{\tau}_i^{(U)}\hat{\boldsymbol{\mu}}\left(\boldsymbol{\tau}_j^{(V)} - \boldsymbol{\tau}_{j'}^{(V)}\right)^T > \nu_{j'}^{(B)} - \nu_j^{(B)} \\
&\iff \boldsymbol{\tau}_i^{(U)}\boldsymbol{a} > b \\
&\iff \boldsymbol{a}_{d_i} > b \ , \qquad \text{(S14)}
\end{aligned}$$

with $\boldsymbol{a} \in \mathbb{R}^{k_1}$ defined by $\boldsymbol{a} = \hat{\boldsymbol{\mu}}\left(\boldsymbol{\tau}_j^{(V)} - \boldsymbol{\tau}_{j'}^{(V)}\right)^T$, $b \in \mathbb{R}$ defined by $b = \nu_{j'}^{(B)} - \nu_j^{(B)}$ and $d_i \in \{1, \cdots, k_1\}$ being the group indicator of user i: $\tau_{i,d_i}^{(U)} = 1$.

Suppose ε-parity, from Definition S1 (Definition 3.2 is a particular case of Definition S1), we have

$$\forall (t,t'), \qquad \forall q, \qquad \left| \frac{\#\{i|s_i = t \wedge d_i = q\}}{\#\{i|s_i = t\}} - \frac{\#\{i|s_i = t' \wedge d_i = q\}}{\#\{i|s_i = t'\}} \right| \leq \varepsilon$$

therefore,

$$\forall (t,t'), \quad \forall q, \quad \left| \mathbb{1}_{\boldsymbol{a}_{d_i} > b} \frac{\#\{i|s_i = t \wedge d_i = q\}}{\#\{i|s_i = t\}} - \mathbb{1}_{\boldsymbol{a}_{d_i} > b} \frac{\#\{i|s_i = t' \wedge d_i = q\}}{\#\{i|s_i = t'\}} \right| \leq \varepsilon \mathbb{1}_{\boldsymbol{a}_{d_i} > b}$$

By summing over all groups, we get:

$$\forall (t,t'), \quad \sum_q \left| \frac{\mathbb{1}_{\boldsymbol{a}_{d_i} > b} \#\{i|s_i = t \wedge d_i = q\}}{\#\{i|s_i = t\}} - \frac{\mathbb{1}_{\boldsymbol{a}_{d_i} > b} \#\{i|s_i = t' \wedge d_i = q\}}{\#\{i|s_i = t'\}} \right| \leq \varepsilon \sum_q \mathbb{1}_{\boldsymbol{a}_{d_i} > b}$$

and from the triangular inequality,

$$\begin{aligned}
\forall (t,t'), \quad & \left| \frac{\sum_q \mathbb{1}_{\boldsymbol{a}_{d_i} > b} \#\{i|s_i = t \wedge d_i = q\}}{\#\{i|s_i = t\}} - \frac{\sum_q \mathbb{1}_{\boldsymbol{a}_{d_i} > b} \#\{i|s_i = t' \wedge d_i = q\}}{\#\{i|s_i = t'\}} \right| \leq \varepsilon \sum_q \mathbb{1}_{\boldsymbol{a}_{d_i} > b} \\
\forall (t,t'), \quad & \left| \frac{\#\{i|s_i = t \wedge \boldsymbol{a}_{d_i} > b\}}{\#\{i|s_i = t\}} - \frac{\#\{i|s_i = t' \wedge \boldsymbol{a}_{d_i} > b\}}{\#\{i|s_i = t'\}} \right| \leq \varepsilon k_1
\end{aligned}$$

And, applying (S14), the result is obtained:

$$\forall(t,t'),\quad \left|\frac{\#\left\{i|s_i=t\wedge(\hat{R}_{ij}>\hat{R}_{ij'})\right\}}{\#\left\{i|s_i=t\right\}}-\frac{\#\left\{i|s_i=t'\wedge(\hat{R}_{ij}>\hat{R}_{ij'})\right\}}{\#\left\{i|s_i=t'\right\}}\right|\leq\varepsilon k_1$$

□ □

D Supplemental results for MovieLens 1M

D.1 Gender as sensitive attribute

Supplemental analysis of the model

We list in Tables S2 and S3 the most extreme movies according to the inferred value of their latent variable C_j. Variable C_j encodes the difference in opinion between the sensitive groups, not the overall opinion. For example, a movie may well be liked by most people but liked even more by males. Table S2 lists movies for which females have a better opinion than males and Table S3 lists movies for which males have a better opinion than females.

Higher number of groups

We did not optimize the hyper-parameters of the compared models. We present here additional experiments to illustrate that the conclusions of Section 4 apply to different hyper-parameter settings. Using a substantially larger number of groups ($k_1 = 50$ user groups and $k_2 = 50$ item groups) or a larger dimension of latent factors for SVD (also 50), the statistical gender parity measures given in Table S4 and the recommendation performance given in Figure S7 are qualitatively similar to the ones given in Table 1 and Figure 5.

D.2 Age as sensitive attribute

The age range of the users is indicated within the following intervals: 'Under 18','18-24', '25-34', '35-44', '45-49', '50-55' and '56+'. The counts of users in each age category is displayed in Figure S8.

User age is treated as sensitive: we introduce seven binary sensitive attributes s_i encoding for the seven categories of user age. We use a one-hot encoding of the seven categories of user age and introduce for the purpose seven binary sensitive attributes $s_i^1, \cdots, s_i^7$ and their item associated latent

Table S2: List of movies with the largest gap in opinion between females and males for which females have a better opinion than males

Title	Year	Genders	Cj
Dirty Dancing	1987	Musical—Romance	0.31
Rocky Horror Picture Show, The	1975	Comedy—Horror—Musical—Sci-Fi	0.26
Sound of Music, The	1965	Musical	0.24
Grease	1978	Comedy—Musical—Romance	0.23
Jumpin' Jack Flash	1986	Action—Comedy—Romance—Thriller	0.23
Gone with the Wind	1939	Drama—Romance—War	0.22
Newsies	1992	Children's—Musical	0.21
Strictly Ballroom	1992	Comedy—Romance	0.21
Steel Magnolias	1989	Drama	0.20
Sense and Sensibility	1995	Drama—Romance	0.20
Full Monty, The	1997	Comedy	0.19
Much Ado About Nothing	1993	Comedy—Romance	0.18
Thelma & Louise	1991	Action—Drama	0.18
Swing Kids	1993	Drama—War	0.17
Fried Green Tomatoes	1991	Drama	0.17
Ever After: A Cinderella Story	1998	Drama—Romance	0.17
Anastasia	1997	Animation—Children's—Musical	0.17
Little Women	1994	Drama	0.17
Color Purple, The	1985	Drama	0.17
To Wong Foo, Thanks for Everything!	1995	Comedy	0.17

Table S3: List of movies with the largest gap in opinion between females and males for which males have a better opinion than females

Title	Year	Genders	Cj
Good, The Bad and The Ugly, The	1966	Action—Western	-0.32
Animal House	1978	Comedy	-0.30
Caddyshack	1980	Comedy	-0.27
Dumb & Dumber	1994	Comedy	-0.27
Exorcist, The	1973	Horror	-0.24
Clockwork Orange, A	1971	Sci-Fi	-0.24
Patton	1970	Drama—War	-0.23
Godfather: Part II, The	1974	Action—Crime—Drama	-0.22
Reservoir Dogs	1992	Crime—Thriller	-0.22
Saving Private Ryan	1998	Action—Drama—War	-0.22
Airplane!	1980	Comedy	-0.21
Eyes Wide Shut	1999	Drama	-0.21
Aliens	1986	Action—Sci-Fi—Thriller—War	-0.21
Predator	1987	Action—Sci-Fi—Thriller	-0.20
Apocalypse Now	1979	Drama—War	-0.20
Unforgiven	1992	Western	-0.20
Evil Dead II (Dead By Dawn)	1987	Action—Adventure—Comedy—Horror	-0.20
Big Trouble in Little China	1986	Action—Comedy	-0.20
Godfather, The	1972	Action—Crime—Drama	-0.20

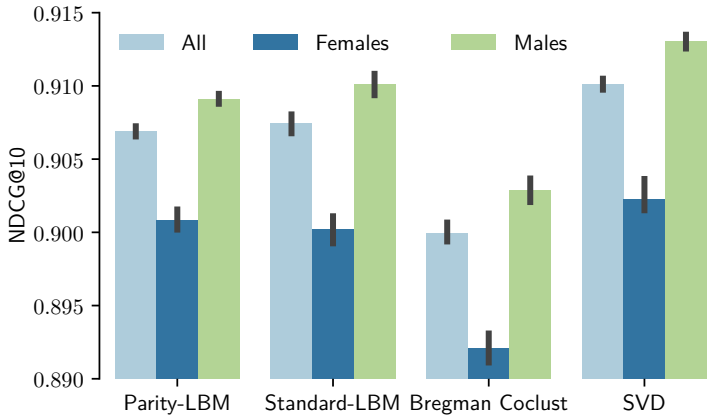


Figure S7: Normalized Discounted Cumulative Gain estimated on MovieLens-1M with $k_1 = k_2 = 50$ groups for clustering methods and 50 factors for the SVD.

Table S4: Measures of gender statistical parity. The number of user groups is $k_1 = 50$. The χ^2 statistic (with 49 degrees of freedom) is averaged over the five replicates of the experiment. A high value of the χ^2 statistic (or a low p-value) leads to the rejection of the statistical parity hypothesis.

Model	Parity LBM	Standard LBM	Bregman co-clust
χ^2 statistic	20	94	105
p-value	0.999	$1.1 \cdot 10^{-4}$	$5.8 \cdot 10^{-6}$

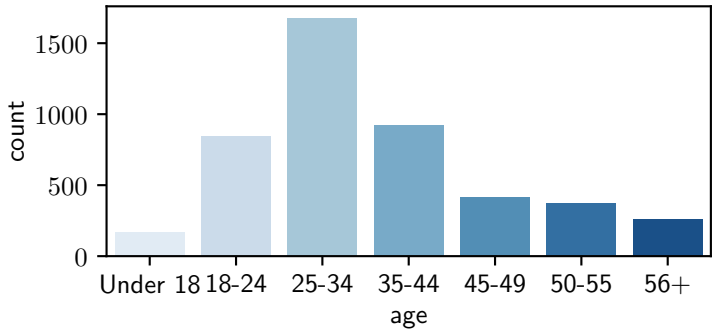


Figure S8: Count of users in each age category.

Table S5: Measures of statistical parity with respect to age category. The number of group of users is $k_1 = 15$. A high value of the χ^2 statistic (or a low p-value) leads to the rejection of the statistical parity hypothesis. The χ^2 statistic is averaged on the five folds of the cross-validation. Degrees of freedom is 14.

Model	Parity LBM	Standard LBM	Bregman co-clust
χ^2 statistic	99	144	577
p-value	0.12	$5.1 \cdot 10^{-5}$	$< 10^{-15}$

variables $C_j^1, \cdots, C_j^7$. We use the protocol described in Section 4 with the exception that our Parity-LBM is initialized from estimates obtained with the Standard-LBM. Table S5 presents results of the χ^2 statistics constructed from the contingency table of user age counts in each group. The methods that do not consider the sensitive variable in the modelling create groups that are dependent on the age and assuming the statistical parity with our Parity-LBM model is reasonable.

Finally, we illustrate the interpretability of the estimates of the latent variables $C_j^1, \cdots, C_j^7$ related to movies. For each age category k, we select the thirty movies with the largest value of the latent variables C_j^k. These movies have the largest positive opinion bias for users in the given age category. Figure S9 displays a boxplot of the release years of these films for all user age categories. The greater variability in the distribution for older users means that they have a comparatively higher opinion of older movies than younger users. If user age were the sensitive attribute, the recommendations would not account for these differences.

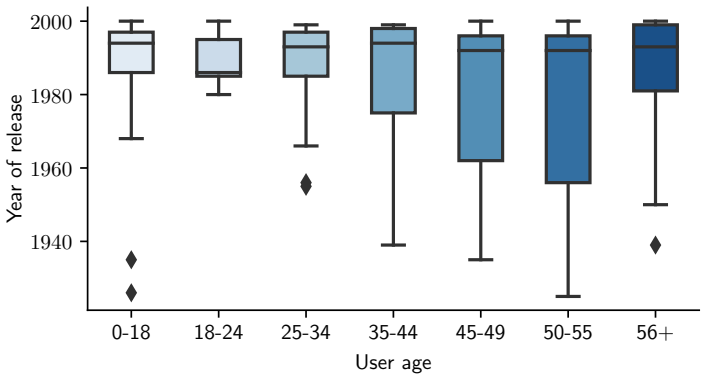


Figure S9: Release years of the thirty most extreme movies according to the inferred positive value of the latent variables $C_j^1, \cdots, C_j^7$. Each latent variable C_j^k is matched with its corresponding user age category.