Recommendation Independence*

Toshihiro Kamishima MAIL@KAMISHIMA.NET
Shotaro Akaho S.AKAHO@AIST.GO.JP
Hideki Asoh H.ASOH@AIST.GO.JP
National Institute of Advanced Industrial Science and Technology (AIST),
AIST Tsukuba Central 2, Umezono 1–1–1, Tsukuba, Ibaraki, Japan 305–8568

Jun Sakuma JUN@CS.TSUKUBA.AC.JP
University of Tsukuba, 1–1–1 Tennodai, Tsukuba, Ibaraki, Japan 305–8577; and
RIKEN Center for Advanced Intelligence Project, 1–4–1 Nihonbashi, Chuo-ku, Tokyo, Japan 103-0027

Editors: Sorelle A. Friedler and Christo Wilson

Abstract

This paper studies a recommendation algorithm whose outcomes are not influenced by specified information. It is useful in contexts potentially unfair decision should be avoided, such as job-applicant recommendations that are not influenced by socially sensitive information. An algorithm that could exclude the influence of sensitive information would thus be useful for job-matching with fairness. We call the condition between a recommendation outcome and a sensitive feature *Recommendation Independence*, which is formally defined as statistical independence between the outcome and the feature. Our previous independence-enhanced algorithms simply matched the means of predictions between sub-datasets consisting of the same sensitive value. However, this approach could not remove the sensitive information represented by the second or higher moments of distributions. In this paper, we develop new methods that can deal with the second moment, i.e., variance, of recommendation outcomes without increasing the computational complexity. These methods can more strictly remove the sensitive information, and experimental results demonstrate that our new algorithms can more effectively eliminate the factors that undermine fairness. Additionally, we explore potential applications for independence-enhanced recommendation, and discuss its relation to other concepts, such as recommendation diversity.

Keywords: Recommender System, Fairness, Independence

1. Introduction

A recommender system searches for items or information predicted to be useful to users, and its influence on users' decision-making has been growing. For example, online-retail-store customers check recommendation lists, and are more likely to decide to buy highly-rated items. Recommender systems have thus become an indispensable tool in support of decision-making.

Such decision-making support tools must be fair and unbiased, because users can make poor decisions if recommendations are influenced by specific information that does not match their needs. Hence, a recommendation algorithm that can exclude the influence of such information from its outcome would be very valuable. There are several representative scenarios in which the exclusion of specific information would be necessary. First, there are contexts in which recommendation services must be managed in adherence to laws and regulations. Sweeney presented an example of dubious advertisement placement that appeared to exhibit racial discrimination (Sweeney, 2013). In this case, the advertisers needed to generate personalized advertisements that were independent of racial information. Another concern is the fair treatment of information providers. An example in this context is the Federal Trade Commissions investigation of Google to determine whether the search engine ranks its own services higher than those of competitors (Forden, 2012). Algorithms that

[*] Our experimental codes are available at http://www.kamishima.net/iers/

can explicitly exclude information, whether or not content providers are competitors, would be helpful for alleviating competitors' doubts that their services are being unfairly underrated. Finally, a user sometimes needs to exclude the influence of unwanted information. Popularity bias, which is the tendency for popular items to be recommended more frequently (Celma and Cano, 2008), is a well-known drawback of recommendation algorithms. If information on the popularity of items could be excluded, users could acquire information free from unwanted popularity bias.

To fulfill the need for techniques to exclude the influence of specific information, several methods for fairness-aware data mining have been developed (for review, see Hajian et al., 2016). In these approaches, a classifier is designed to predict labels so that they are independent from specified sensitive information. By introducing this idea, we proposed the concept of *recommendation independence*. This is formally defined as unconditional statistical independence between a recommendation outcome and specified information. We call a recommendation that maintains the recommendation independence *independence-enhanced recommendation*. We developed two types of approaches to these recommendations. The first is a regularization approach, which adopts an objective function with a constraint term for imposing independence (Kamishima et al., 2012a, 2013; Kamishima and Akaho, 2017). The second is a model-based approach, which adopts a generative model in which an outcome and a sensitive feature are independent (Kamishima et al., 2016).

In this paper, we propose new methods for making independence-enhanced recommendations. Our previous model (Kamishima et al., 2013) took a regularization approach and combined probabilistic matrix factorization and a constraint term. However, because the constraint term was heuristically designed so that it matched means by shifting predicted ratings, it could not remove the sensitive information represented by the second or higher moments of distributions. Further, the approach could not control the range of predicted ratings, and thus would skew the rating distribution. For example, if all predicted ratings were shifted toward +1, the lowest ratings would not appear in the predictions. To remove these drawbacks without sacrificing computational efficiency, we developed two new types of constraint terms exploiting statistical measures: Bhattacharyya distance and mutual information.

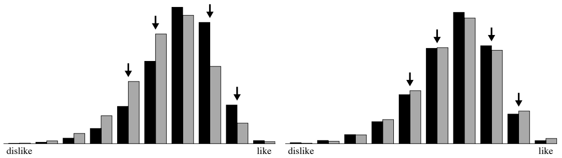
We performed more extensive experiments than in our previous studies in order to achieve more reliable verification. Here, we examine algorithms on three datasets and six types of sensitive features to confirm the effects of independence-enhancement. To verify the improvements that derive from considering the second moments, we quantitatively compared the quality of rating distributions using an independence measure.

Moreover, we explore scenarios in which independence-enhanced recommendation would be useful, and clarify the relation between recommendation independence and other recommendation research topics. We provide more examples of three types of scenarios in which enhancement of recommendation independence would be useful. As in the discussion in the RecSys 2011 panel (Resnick et al., 2011), rich recommendation diversity has been considered beneficial for making recommendations fair. We discuss the differences in the definitions of recommendation diversity and independence. We also note the relation to transparency and privacy in a recommendation context.

Our contributions are as follows.

- We develop new independence-enhanced recommendation models that can deal with the second moment of distributions without sacrificing computational efficiency.
- Our more extensive experiments reveal the effectiveness of enhancing recommendation independence and of considering the second moments.
- We explore applications in which recommendation independence would be useful, and reveal the relation of independence to the other concepts in recommendation research.

This paper is organized as follows. In section 2, we present the concept of recommendation independence, and discuss how the concept would be useful for solving real-world problems. Methods for independence-enhanced recommendation are proposed in section 3, and the experimental results are presented in section 4. Section 5 contains a discussion about recommendation independence and related recommendation issues, and section 6 concludes our paper.



(a) standard (b) independence-enhanced

Figure 1: Distributions of the predicted ratings for each sensitive value

2. Recommendation Independence

This section provides a formal definition of recommendation independence, and we show applications of this concept for solving real problems.

2.1. Formal Definition

Before formalizing the concept of recommendation independence, we will give a more intuitive description of this concept. Recommendation independence is defined as the condition that the generation of a recommendation outcome is statistically independent of a specified sensitive feature. This implies that information about the feature has no impact on the recommendation outcomes. We will take the example of a movie recommendation. In this example, we select the movie-release year as a sensitive feature, in order to prevent the release year from influencing the recommendation of individual movies. Therefore, assuming that there are two movies whose features are all the same except for their release year, this independence-enhanced recommender makes identical predictions for these two movies.

To illustrate the effect of enhancing recommendation independence, Figure 1 shows the distributions of predicted ratings for each sensitive value for the ML1M-Year dataset. The details will be shown in section 4; here, we briefly note that ratings for movies are predicted and the sensitive feature represents whether or not movies are released before 1990. Black and gray bars show the distributions of ratings for older and newer movies, respectively. In Figure 1(a), ratings are predicted by a standard algorithm, and older movies are highly rated (note the large gaps between the two bars indicated by arrowheads). When recommendation independence is enhanced (η=100) as in Figure 1(b), the distributions of ratings for older and newer movies become much closer (the large gaps are lessened); that is to say, the predicted ratings are less affected by the sensitive feature. It follows from this figure that the enhancement of recommendation independence reduces the influence of a specified sensitive feature on the outcome.

We can now formalize the above intuitive definition of the recommendation independence. Consider an event in which all the information required to make a recommendation, such as the specifications of a user and item and all features related to them, is provided and a recommendation result is inferred from this information. This event is represented by a triplet of three random variables: R, S, and F. R represents a recommendation outcome, which is typically a rating value or an indicator of whether or not a specified item is relevant. We call S a sensitive feature, using the terminology in the fairness-aware data mining literature. Finally, F represents all ordinary features (or features) related to this event other than those represented by R and S. The recommendation is made by inferring the value of R given the values of S and F based on a probabilistic recommendation model, $\Pr[R|S, F]$.

Based on information theory, the statement "generation of a recommendation outcome is statistically independent of a specified sensitive feature" describes the condition in which the mutual information between R and S is zero, $\mathrm{I}(R; S) = 0$. This means that we know nothing about R even if we obtain information about S, because information on S is excluded from R. This condition is equivalent to statistical independence between R and S, i.e., $\Pr[R] = \Pr[R|S]$, as denoted by $R \perp\!\!\!\perp S$.

2.2. Applications

We here consider recommendation applications for which independence should be enhanced.

2.2.1. Adherence to Laws and Regulations

Recommendation services must be managed while adhering to laws and regulations. We will consider the example of a suspicious advertisement placement based on keyword-matching (Sweeney, 2013). In this case, users whose names are more popular among individuals of African descent than European descent were more frequently shown advertisements implying arrest records. According to an investigation, however, no deliberate manipulation was responsible; rather, the bias arose simply as a side-effect of algorithms to optimize the click-through rate. Because similar algorithms of Web content optimization are used for online recommendations, such as for online news recommendations, similar discriminative recommendations can be provided in these contexts. For example, independence-enhanced recommendation would be helpful for matching an employer and a job applicant based not on gender or race, but on other factors, such as the applicant's skill level at the tasks required for the job.

Recommendation independence is also helpful for avoiding the use of information that is restricted by law or regulation. For example, privacy policies prohibit the use of certain types of information for the purpose of making recommendations. In such cases, by treating the prohibited information as a sensitive feature, the information can be successfully excluded from the prediction process of recommendation outcomes.

2.2.2. Fair Treatment of Content Providers

Recommendation independence can be used to ensure the fair treatment of content providers or product suppliers. The Federal Trade Commission has been investigating Google to determine whether the search engine ranks its own services higher than those of competitors (Forden, 2012). The removal of deliberate manipulation is currently considered to ensure the fair treatment of content providers. However, algorithms that can explicitly exclude information whether or not content providers are competitors would be helpful for dismissing the competitors' doubts that their services may be unfairly underrated.

Though this case is about information retrieval, the treatment of content providers in the course of generating recommendations can also be problematic. Consider the example of an online retail store that directly sells items in addition to renting a portion of its Web sites to tenants. On the retail Web site, if directly sold items are overrated in comparison to items sold by tenants, then the trade conducted between the site owner and the tenants is considered unfair. To carry on a fair trade, the information on whether an item is sold by the owner or the tenants should be ignored. An independence-enhanced recommender would be helpful for this purpose. Note that enhancing this type of recommendation independence is not disadvantageous to retail customers, because items are equally rated if they are equivalent and sold under equivalent conditions.

2.2.3. Exclusion of Unwanted Information

Users may want recommenders to exclude the influence of specific information. We give several examples. Enhancing independence is useful for correcting a popularity bias, which is the tendency for popular items to be recommended more frequently (Celma and Cano, 2008). If users are already familiar with the popular items and are seeking minor and long-tail items that are novel to them, this popularity bias will be unfavorable to them. In this case, the users can specify the volume of consumption of items as a sensitive feature, and the algorithm will provide recommendations that are independent of information about the popularity of items.

The deviations of preference data can be adjusted. As is well known, preference data are affected by their elicitation interface. For example, the response of users can be changed depending on whether or not predicted ratings are displayed (Cosley et al., 2003). By making a recommendation independent of the distinction of preference-elicitation interfaces, such unwanted deviations can be canceled.

When users explicitly wish to ignore specific information, such information can be excluded by enhancing the recommendation independence. Pariser recently introduced the concept of the filter bubble problem, which is the concern that personalization technologies narrow and bias the topics of interest provided to technology consumers, who do not notice this phenomenon (Pariser, 2011). If a user of a social network service wishes to converse with people having a wide variety of political opinions, a friend recommendation that is independent of the friends' political conviction will provide an opportunity to meet people with a wide range of views.

3. Independence-Enhanced Recommendation

In this section, after formalizing a task of independence-enhanced recommendation, we show an independence-enhanced variant of a probabilistic matrix factorization model. Finally, we show the previous and our new types of penalty terms required for enhancing the recommendation independence.

3.1. Task Formalization

We here formalize a task of *independence-enhanced recommendation*. Recommendation tasks can be classified into three types: finding good items that meet a user's interest, optimizing the utility of users, and predicting ratings of items for a user (Gunawardana and Shani, 2009). We here focus on the following predicting-ratings task. $X \in \{1, \ldots, n\}$ and $Y \in \{1, \ldots, m\}$ denote random variables for the user and item, respectively. R denotes a random variable for the recommendation outcome in the previous section, but this R is now restricted to the rating of Y given by X, and we hereafter refer to this R as a rating variable. The instance of X, Y, and R are denoted by x, y, and r, respectively.

As described in the previous section, we additionally introduced a random sensitive variable, S, which indicates the sensitive feature with respect to which the independence is enhanced. This variable is specified by a user or manager of recommenders, and its value depends on various aspects of an event as in the examples in section 2.2. In this paper, we restrict the domain of a sensitive feature to a binary type, $\{0, 1\}$. A training datum consists of an event, (x, y), a sensitive value for the event, s, and a rating value for the event, r. A training dataset is a set of N training data, $\mathcal{D} = \{(x_i, y_i, s_i, r_i)\}$, $i = 1, \ldots, N$. We define $\mathcal{D}^{(s)}$ as a subset consisting of all data in $\mathcal{D}$ whose sensitive value is s.

Given a new event, (x, y), and its corresponding sensitive value, s, a rating prediction function, $\hat{r}(x, y, s)$, predicts a rating of the item y by the user x, and satisfies $\hat{r}(x, y, s) = \mathrm{E}_{\Pr[R|x,y,s]}[R]$. This rating prediction function is estimated by using a regularization approach developed in the context of fairness-aware data mining (Kamishima et al., 2012b). This approach is to optimize an objective function having three components: a loss function, $\mathrm{loss}(r^*, \hat{r})$, an independence term, $\mathrm{ind}(R, S)$, and a regularization term, reg. The loss function represents the dissimilarity between a true rating value, r^*, and a predicted rating value, $\hat{r}$. The independence term quantifies the expected degree of independence between the predicted rating values and sensitive values, and a larger value of this term indicates the higher level of independence. The aim of the regularization term is to avoid overfitting. Given a training dataset, $\mathcal{D}$, the goal of the independence-enhanced recommendation is to acquire a rating prediction function, $\hat{r}(x, y, s)$, so that the expected value of the loss function is as small as possible and the independence term is as large as possible. The goal can be accomplished by finding a rating prediction function, $\hat{r}$, so as to minimize the following objective function:

$$\textstyle\sum_{(x_i,y_i,s_i,r_i)\in\mathcal{D}} \mathrm{loss}(r_i, \hat{r}(x_i, y_i, s_i)) \\ - \eta\, \mathrm{ind}(R, S) + \lambda\, \mathrm{reg}(\mathbf{\Theta}), \quad (1)$$

where $\eta > 0$ is an independence parameter to balance between the loss and the independence, $\lambda > 0$ is a regularization parameter, and $\mathbf{\Theta}$ is a set of model parameters.

3.2. An Independence-Enhanced Recommendation Model

We adopt a probabilistic matrix factorization (PMF) model (Salakhutdinov and Mnih, 2008) to predict ratings. Though there are several minor variants of this model, we here use the following model defined as equation (3) in (Koren, 2008):

$$\hat{r}(x, y) = \mu + b_x + c_y + \mathbf{p}_x^\top \mathbf{q}_y, \quad (2)$$

where μ, b_x, and c_y are global, per-user, and per-item bias parameters, respectively, and $\mathbf{p}_x$ and $\mathbf{q}_y$ are K-dimensional parameter vectors, which represent the cross effects between users and items. The parameters of the model are estimated by minimizing the squared loss function with a L_2 regularizer term. This model is proved to be equivalent to assuming that true rating values are generated from a normal distribution whose mean is equation (2). Unfortunately, in the case that not all entries of a rating matrix are observed, the objective function of this model is non-convex, and merely local optima can be found. However, it is empirically known that a simple gradient method succeeds in finding a good solution in most cases (Koren, 2008).

The PMF model was then extended to enhance the recommendation independence. First, the prediction function (2) was modified so that it is dependent on the sensitive value, s. For each value of s, 0 and 1, parameter sets, $\mu^{(s)}$, $b_x^{(s)}$, $c_y^{(s)}$, $\mathbf{p}_x^{(s)}$, and $\mathbf{q}_y^{(s)}$, are prepared. One of the parameter sets is chosen according to the sensitive value, and the rating prediction function becomes:

$$\hat{r}(x,y,s) = \mu^{(s)} + b_x^{(s)} + c_y^{(s)} + \mathbf{p}_x^{(s)\top} \mathbf{q}_y^{(s)}. \quad (3)$$

By using this prediction function (3), an objective function of an independence-enhanced recommendation model becomes:

$$\begin{aligned}\textstyle\sum_{(x_i,y_i,r_i,s_i)\in\mathcal{D}}(r_i - \hat{r}(x_i,y_i,s_i))^2 \\ - \eta\,\mathrm{ind}(R,S) + \lambda\,\mathrm{reg}(\mathbf{\Theta}),\end{aligned} \quad (4)$$

where the regularization term is a sum of L_2 regularizers of parameter sets for each value of s except for global biases, $\mu^{(s)}$. Model parameters, $\mathbf{\Theta}^{(s)} = \{\mu^{(s)}, b_x^{(s)}, c_y^{(s)}, \mathbf{p}_x^{(s)}, \mathbf{q}_y^{(s)}\}$, for $s \in \{0,1\}$, are estimated so as to minimize this objective. Once we learn the parameters of the rating prediction function, we can predict a rating value for any event by applying equation (3).

3.3. Independence Terms

Now, all that remains is to define the independence terms. Having previously introducing our independence term of mean matching (Kamishima et al., 2013), we here propose two new independence terms, distribution matching and mutual information.

3.3.1. Mean Matching

Our previous independence term in (Kamishima et al., 2013) was designed so as to match the means of two distributions $\Pr[R|S{=}0]$ and $\Pr[R|S{=}1]$, because these means match if R and S become statistically independent. The independence term is a squared norm between means of these distributions:

$$-\left(\frac{\mathbb{S}^{(0)}}{N^{(0)}} - \frac{\mathbb{S}^{(1)}}{N^{(1)}}\right)^2, \quad (5)$$

where $N^{(s)} = |\mathcal{D}^{(s)}|$ and $\mathbb{S}^{(s)}$ is the sum of predicted ratings over the set $\mathcal{D}^{(s)}$,

$$\mathbb{S}^{(s)} = \textstyle\sum_{(x_i,y_i,s_i)\in\mathcal{D}^{(s)}} \hat{r}(x_i,y_i,s_i). \quad (6)$$

We refer to this independence term as *mean matching*, which we abbreviate as mean-m.

3.3.2. Distribution Matching

To remedy the drawback that the mean-m term is designed to ignore the second moment, we propose a new independence term. Techniques for handling the independence between a continuous target variable and a sensitive feature have not been fully discussed. The method in (Calders et al., 2013) is basically the same as the approach of the mean-m. Pérez-Suay et al. (2017) proposed the use of the Hilbert-Schmidt independence criterion. However, this approach requires the computational complexity of $O(N^2)$ for computing a kernel matrix, and it is not scalable.

We therefore create our new independence term, *distribution matching with Bhattacharyya distance*, which we abbreviate as bdist-m. This bdist-m term can deal with the second moment of distributions in addition to the first moment. For this purpose, each of two distributions, $\Pr[R|S{=}0]$ and $\Pr[R|S{=}1]$, is modeled by a normal distribution, and the similarity between them are quantified by a negative Bhattacharyya distance:

$$\begin{aligned}&-\left(-\ln\int_{-\infty}^{\infty}\sqrt{\Pr[r|S{=}0]\Pr[r|S{=}1]}dr\right)\\ &= \frac{1}{2}\ln\left(\frac{2\sqrt{\mathbb{V}^{(0)}\mathbb{V}^{(1)}}}{\mathbb{V}^{(0)}+\mathbb{V}^{(1)}}\right) - \frac{\left(\frac{\mathbb{S}^{(0)}}{N^{(0)}} - \frac{\mathbb{S}^{(1)}}{N^{(1)}}\right)^2}{4(\mathbb{V}^{(0)}+\mathbb{V}^{(1)})},\end{aligned} \quad (7)$$

where $\mathbb{V}^{(s)}$ is the variance of predicted ratings over the training set $\mathcal{D}^{(s)}$. To estimate $\mathbb{V}^{(s)}$, we use the expectation of a posterior distribution of a variance parameter derived from equation (2.149) in (Bishop, 2006) to avoid the zero variance:

$$\mathbb{V}^{(s)} = \frac{2b_0 + \mathbb{Q}^{(s)} - (\mathbb{S}^{(s)})^2/N^{(s)}}{2a_0 + N^{(s)}}, \quad (8)$$

where $\mathbb{S}^{(s)}$ is equation (6), and $\mathbb{Q}^{(s)}$ is the squared sum of predicted ratings:

$$\mathbb{Q}^{(s)} = \textstyle\sum_{(x_i,y_i,s_i)\in\mathcal{D}^{(s)}} \hat{r}(x_i,y_i,s_i)^2. \quad (9)$$

a_0 and b_0 are hyper-parameters of a prior Gamma distribution. We use $2a_0{=}10^{-8}$ and $2b_0{=}10^{-24}$.

3.3.3. Mutual Information

We next propose our new independence term, *mutual information with normal distributions*, abbreviated as mi-normal. We employed mutual information for quantifying the degree of statistical independence. Distributions, $\Pr[R|S]$, are modeled by a normal distribution. Our new mi-normal term is defined as the negative mutual information between R and S:

$$\begin{aligned} &-\mathrm{I}(R;S) = -(\mathrm{H}(R) - \mathrm{H}(R|S)) \\ &= -(\mathrm{H}(R) - \textstyle\sum_s \Pr[S{=}s]\,\mathrm{H}(R|S{=}s)), \end{aligned} \tag{10}$$

where $\mathrm{H}(X)$ is a differential entropy function. We start with the second term of this equation. Because S is a binary variable, $\Pr[S{=}s]$ can be easily estimated by $N^{(s)}/N$. To estimate $\mathrm{H}(R|S{=}s)$, we model $\Pr[R|S{=}s]$ by a normal distribution. By using the formula of differential entropy of a normal distribution (e.g., see example 8.1.2 in Cover and Thomas, 2006), we get

$$\mathrm{H}(R|S{=}s) = \tfrac{1}{2} \ln 2\pi e \mathbb{V}^{(s)}. \tag{11}$$

We next turn to the first term of equation (10). $\Pr[R]$ is a mixture of two normal distributions:

$$\Pr[R] = \textstyle\sum_s \Pr[S{=}s] \Pr[R|S{=}s]. \tag{12}$$

We approximate $\Pr[R]$ by a single normal distribution with the same mean and variance of this mixture distribution. We consider this approximation proper, because it captures the first two moments of the mixture. Further, when R and S are completely independent, where two means of two normal distributions are equal, the mixture is precisely equivalent to a single normal distribution. This property is desirable, because we now try to let R and S be independent. By using the entropy of a normal distribution, we get

$$\mathrm{H}(R) = \tfrac{1}{2} \ln 2\pi e \mathbb{V}, \tag{13}$$

$$\mathbb{V} = \frac{4b_0 + (\mathbb{Q}^{(0)}{+}\mathbb{Q}^{(1)}) - \left(\mathbb{S}^{(0)}{+}\mathbb{S}^{(1)}\right)^2/N}{4a_0 + N}. \tag{14}$$

Finally, by substituting equations (11) and (13) into equation (10), we obtain a mi-normal independence term.

These new independence terms, bdist-m and mi-normal, are computationally efficient. Because these terms are smooth and analytically differentiable like a mean-m term, the objective function can be optimized efficiently. Their computational complexity is dominated by the sum and the squared sum of data, and it is $O(N)$. Because this complexity is on the same order as that of the loss function of the original PMF model, the total computational complexity of an independence-enhanced variant is the same as that of the original algorithm. Additionally, since these new terms take the first two moments of distributions into account, they give a much better approximation than the mean-m term. Finally, we note the difference between the bdist-m and mi-normal terms. A bdist-m term does not approximate $\Pr[R]$, unlike a mi-normal term. A mi-normal term can be straightforwardly extensible to the case that a sensitive feature is categorical type.

Table 1: Summary of experimental conditions

data	ML1M	Flixster	Sushi
# of users	$6,040$	$147,612$	$5,000$
# of items	$3,706$	$48,794$	100
# of ratings	$1,000,209$	$8,196,077$	$50,000$
rating scale	$1,2,\ldots,5$	$.5,1,\ldots,5$	$0,1,\ldots,4$
mean rating	3.58	3.61	1.27
# of latent factors, K	7	20	5
regularization param., λ	1	30	10
non-personalized MAE	0.934	0.871	1.081
standard MAE	0.685	0.655	0.906

4. Experiments

We implemented independence-enhanced recommenders and applied them to benchmark datasets. Below, we present the details regarding these datasets and experimental conditions, then compare three independence terms: mean-m, bdist-m, and mi-normal. These comparative experiments confirm the effectiveness of independence-enhanced recommenders and show the advantages gained by considering the second moments.

4.1. Datasets and Experimental Conditions

We can now consider the experimental conditions in detail, including the datasets, methods, and evaluation indexes. To confirm the effectiveness of independence-enhancement more clearly, we tested our methods on the three datasets summarized in Tables 1 and 2. The first dataset was the

Table 2: Sizes, means, and variances of data subsets for each sensitive value

Datasets	size		mean		variance	
	$S{=}0$	$S{=}1$	$S{=}0$	$S{=}1$	$S{=}0$	$S{=}1$
ML1M-Year	456, 683	543, 526	3.72	3.46	1.15	1.30
ML1M-Gender	753, 769	246, 440	3.57	3.62	1.25	1.23
Flixster	3, 868, 842	4, 327, 235	3.71	3.53	1.18	1.18
Sushi-Age	3, 150	46, 850	2.39	2.75	1.75	1.60
Sushi-Gender	23, 730	26, 270	2.80	2.66	1.43	1.77
Sushi-Seafood	43, 855	6, 145	2.76	2.46	1.61	1.58

Movielens 1M dataset (ML1M) (Maxwell Harper and Konstan, 2015). We tested two types of sensitive features for this set. The first, Year, represented whether a movie's release year was later than 1990. We selected this feature because it has been proven to influence preference patterns, as described in section 2.1. The second feature, Gender, represented the user's gender. The movie rating depended on the user's gender, and our recommender enhanced the independence of this factor. The second dataset was the larger Flixster dataset (Jamali and Ester, 2010). Because neither the user nor the item features were available, we adopted the popularity of items as a sensitive feature. Candidate movies were first sorted by the number of users who rated the movie in a descending order, and a sensitive feature represented whether or not a movie was in the top 1% of this list. A total of 47.2% of ratings were assigned to this top 1% of items. The third dataset was the Sushi dataset[1] (Kamishima, 2003), for which we adopted three types of sensitive features: Age (a user was a teen or older), Gender (a user was male or female), and Seafood (whether or not a type of sushi was seafood). We selected these features because the means of rating between two subsets, $\mathcal{D}^{(0)}$ and $\mathcal{D}^{(1)}$, diverged.

We tested three independence terms in section 3.3: mean-m, bdist-m, and mi-normal. An objective function (4) was optimized by the conjugate gradient method. We used hyperparameters, the number of latent factors K, and a regularization parameter λ, in Table 1. For each dataset, $\mathcal{D}^{(0)}$ and $\mathcal{D}^{(1)}$, the parameters were initialized by minimizing an objective function of a standard PMF model without an independence term. For convenience, in experiments the loss term of an objective was rescaled by dividing it by the number of training examples. We performed a five-fold cross-validation procedure to obtain evaluation indexes of the prediction errors and independence measures.

We evaluated our experimental results in terms of prediction errors and the degree of independence. Prediction errors were measured by the mean absolute error (MAE) (Gunawardana and Shani, 2009). This index was defined as the mean of the absolute difference between the observed ratings and predicted ratings. A smaller value of this index indicates better prediction accuracy. To measure the degree of independence, we checked the equality between two distributions of ratings predicted on $\mathcal{D}^{(0)}$ and $\mathcal{D}^{(1)}$ datasets. As the measure of equality, we adopted the statistic of the two-sample Kolmogorov-Smirnov test (KS), which is a nonparametric test for the equality of two distributions. The KS statistic is defined as the area between two empirical cumulative distributions of predicted ratings for $\mathcal{D}^{(0)}$ and $\mathcal{D}^{(1)}$. A smaller KS indicates that R and S are more independent.

4.2. Experimental Results

In this section, we empirically examine the following two questions.

1. We consider whether independence-enhanced recommenders definitely enhance the recommendation independence. Such enhancement was not theoretically guaranteed for several reasons: the objective function (4) was not convex, the rating distributions were modeled by normal distributions, and we adopted an approximation in equation (13).
2. We compare the behaviors of the three independence terms. We specifically examine whether the bdist-m and mi-normal terms can take into account the second moment of distributions, which was ignored in the mean-m case.

To answer these questions, we generate two types of experimental results. First, we show the quantitative measures to evaluate the prediction accuracy and the degree of independence. Second, we qualitatively visualize the distributions of predicted ratings for the ML1M-Year dataset.

[1] http://www.kamishima.net/sushi/

4.2.1. Evaluation by Quantitative Measures

To examine the two questions, we quantitatively analyze the evaluation measures. We focus on the first question — that is, the effectiveness of independence-enhancement. Evaluation measures, MAE and KS, for six datasets are shown in Figure 2. We tested a standard PMF model and three independence-enhanced PMF models. We first investigate KSs, which measure the degree of independence, as shown in the right column of Figure 2. We can observe a decreasing trend of KSs along with the increase of an independence parameter, η. In addition, all independence-enhanced recommenders could achieve better levels of independence than those obtained by a standard PMF model, if the independence parameter is fully large. These trends verified that recommendation independence could be enhanced successfully. We then analyzed the MAEs, a measurement of prediction accuracy, from the left column of Figure 2. The MAEs derived from independence-enhanced PMF models were slightly increased when compared with those derived from a standard PMF model. Such losses in accuracy are inevitable, as we will discuss in section 5.1. However, the losses were slight, and these independent-enhancement models could accomplish good trade-offs between accuracy and independence. In summary, independence-enhanced recommenders could enhance recommendation independence successfully.

We then move on the second question. To examine the influence of considering the second moments, we checked the means and standard deviations of distributions for the ML1M-Year dataset. Test data were divided into two sub-datasets, $\mathcal{D}^{(0)}$ and $\mathcal{D}^{(1)}$, and means and standard deviations of predicted ratings were computed for each pair of sub-datasets. These pairs of means and standard deviations are depicted in Figure 3. For all types of independence terms, we found that the pairs of means approached each other as the independence parameter increased. Note that this trend was observed for all datasets. On the other hand, the pairs of standard deviations became closer as the value of η increased in the bdist-m and mi-normal cases, but not in the mean-m case. Similar trends were observed for all the

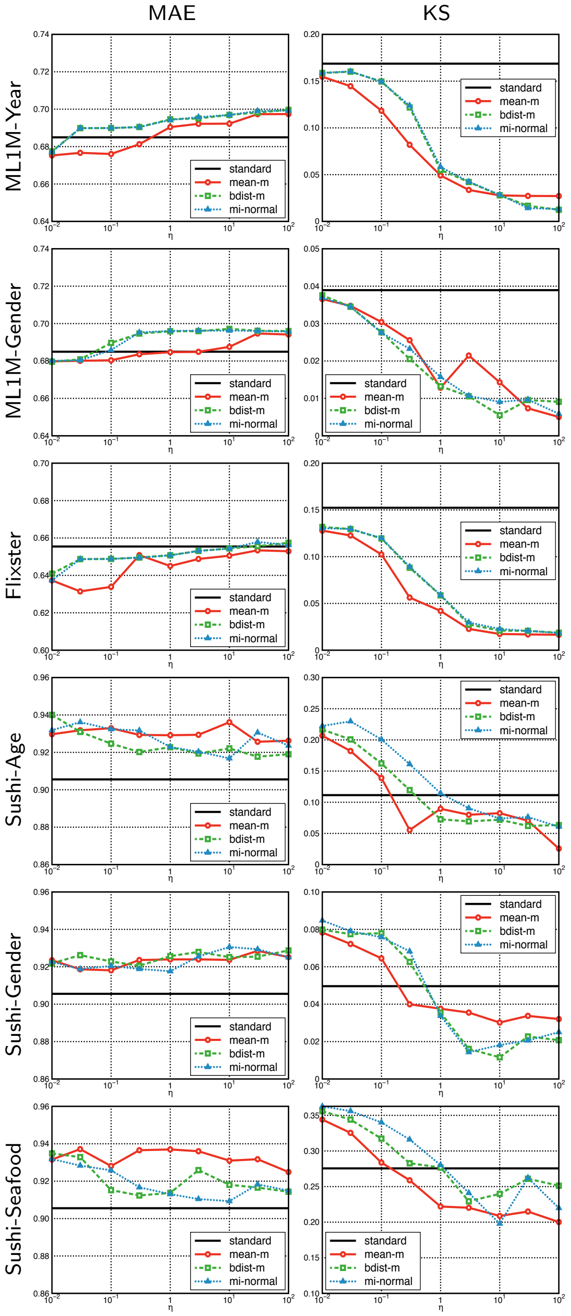
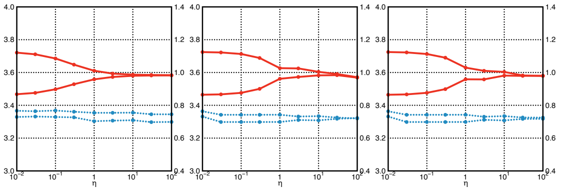


Figure 2: Changes in the MAE and KS measures

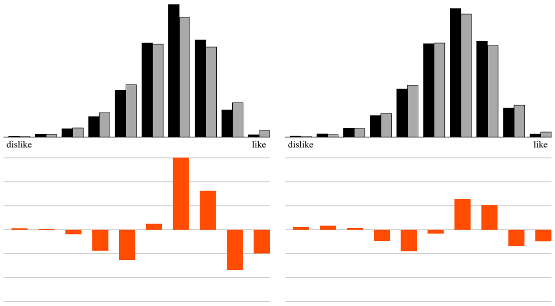
NOTE: The subfigure rows sequentially show the results for the ML1M-Year, ML1M-Gender, Flixster, Sushi-Age, Sushi-Gender, and ML1M-Seafood datasets, respectively. The X-axes of these subfigures represent the independence parameter, η, in a logarithmic scale. The Y-axes of subfigures in the first and the second columns represent MAEs in a linear scale and KSs in a linear scale, respectively. Note that the ranges are changed to emphasize the differences. The results for the bdist-m and mi-normal terms are overlapped in some subfigures.



(a) mean-m (b) bdist-m (c) mi-normal

Figure 3: Changes of means and standard deviations of predicted ratings according to the parameter, η

NOTE: The X-axes of these subfigures represent the independence parameter, η, in a logarithmic scale. These subfigures sequentially show the means and standard deviations of predicted ratings derived by the mean-m, bdist-m, and mi-normal methods for the ML1M-Year, respectively. Means and standard deviations for the two groups based on sensitive values are represented by the scales at the left and right side of these subfigures, respectively. Pairs of means and standard deviations are depicted by red solid and blue dotted lines, respectively.



(a) mean-m (b) mi-normal

Figure 4: Distributions of the ratings predicted by mean-m and mi-normal methods for each sensitive value

NOTE: In the upper charts, black and gray bars show the histograms of ratings for test data in $\mathcal{D}^{(0)}$ and $\mathcal{D}^{(1)}$, respectively. In the lower charts, we show the values for the black bins minus those for the gray bins.

datasets except for Sushi-Age and Sushi-Seafood. Note that, according to our analysis, the failure to control the second moments for these two datasets was due to the imbalanced distributions of sensitive values as in Table 2. We can conclude from these results that the methods using our new independence terms, bdist-m and mi-normal, can control the second moments of rating distributions, which our previous mean-m method could not control.

4.2.2. Qualitative Visualization of Predictions

To provide intuitive illustrations of independence-enhancement, we visualize the distributions of predicted ratings for the ML1M-Year dataset. In terms of the first research question, we show a comparison of the rating distributions predicted by the standard PMF model with those predicted by the mi-normal independence-enhanced PMF model in Figure 1. As described in section 2.1, the figure illustrates the effect of independence enhancement.

We here demonstrate the improvement in consideration of the second moments. By considering the second moments, our new models can remove the sensitive information more strictly, and can produce less skewed predicted ratings by adjusting their range. For this purpose, we visually compare the distributions of ratings predicted by ignoring the second moments, i.e., mean-m, with those predicted by considering them, i.e., mi-normal. Figure 4 shows the distributions of predicted ratings for each sensitive value for the ML1M-Year dataset, as in Figure 1. Figures 4(a) and 4(b) show the distributions of ratings predicted by the mean-m and mi-normal methods (η=100), respectively. First, the distributions for the datasets, $\mathcal{D}^{(0)}$ and $\mathcal{D}^{(1)}$, became much closer in the mi-normal case than in the mean-m case (see the smaller bars in the lower charts). This indicates that the mi-normal could more strictly remove sensitive information by considering the second moments of distributions. Second, we examine the skew of rating distributions. We concentrate on the rightmost bins in these histograms. The differences in these bins were larger than those obtained by a standard PMF model. This is because the distributions for $\mathcal{D}^{(1)}$ (gray) were shifted toward the plus side, and this bin contained all the test data whose ratings were predicted to be larger than the maximum of the rating range. The mi-normal method could achieve a smaller difference than the mean-m method by scaling the range of ratings, because the mi-normal method could control the second moments of distributions. However, such scaling was impossible in the mean-m case, because the mean-m could merely shift the distributions. This observation implies that the mi-normal method could produce a less skewed distribution of predicted ratings. Note that we observed similar trends in terms of the bdist-m method.

Finally, we comment on the difference between mi-normal and bdist-m. The differences in performance between these two methods in terms of accuracy and independence were very slight. Though the mi-normal method adopted an approximation, it could be straightforwardly extensible to the case that a sensitive feature is a categorical discrete variable. Therefore, it would be better to use the mi-normal method in general.

From the above, it may be concluded that all independence terms could successfully enhance recommendation independence, and that our new independence terms, mi-normal and bdist-m, can enhance independence more strictly than the previous mean-m term.

5. Discussion and Related Work

Finally, we will discuss the characteristics of recommendation independence, explore the relation between recommendation independence and diversity, and present related topics.

5.1. Discussion

In this section, we discuss three topics with respect to recommendation independence. First, let us consider why a sensitive feature must be specified in the definition of recommendation independence. In brief, a sensitive feature must be selected because it is intrinsically impossible to personalize recommendation outcomes if the outcomes are independent of any feature. This is due to the *ugly duckling theorem*, which is a theorem in pattern recognition literature that asserts the impossibility of classification without weighing certain features of objects as more important than others (Watanabe, 1969). Because recommendation is considered a task for classifying whether or not items are preferred, certain features inevitably must be weighed when making a recommendation. Consequently, it is impossible to treat all features equally. In the RecSys2011 panel (Resnick et al., 2011), a panelist also pointed out that no information is neutral, and thus individuals are always influenced by information that is biased in some manner.

Second, we will consider the indirect influences of sensitive features. In section 2.1, we incorporated a sensitive feature into a prediction model, $\Pr[R|S,F]$. It might appear that the model could be made independent by simply removing S, but this is not the case. By removing the sensitive information, the model satisfies the condition $\Pr[R|S,F] = \Pr[R|F]$. Using this equation, the probability distribution over (R,S,F) becomes

$$\begin{aligned}\Pr[R,S,F] &= \Pr[R|S,F]\Pr[S|F]\Pr[F]\\ &= \Pr[R|F]\Pr[S|F]\Pr[F]. \qquad (15)\end{aligned}$$

This is the conditional independence between R and S given F, i.e., $R \perp\!\!\!\perp S \mid F$, which is different from the unconditional independence between R and S, i.e., $R \perp\!\!\!\perp S$. Under a condition of conditional independence, if there are features in F that are not independent of S, the outcomes will be influenced by S through the correlated features. This phenomenon was observed in the example of online advertising in section 2.2.1. Even though no information on the individuals' races was explicitly exploited, such an incident could arise through the influence of other features that indirectly contain information about their races. Note that, within the fairness-aware data mining, such an indirect influence is called a red-lining effect (Calders and Verwer, 2010).

With respect to the fairness-aware data mining, Kamiran et al. (2013) discussed a more complicated situation, which they called conditional fairness. In this case, some of the features in F are explainable even if they are correlated with a sensitive feature in addition to being considered fair. For example, in the case of job-matching, if special skills are required for a target job, considering whether or not an applicant has these skills is socially fair even if it is correlated with the applicant's gender. Variables expressing such skills are treated as explainable variables, E, and the conditional independence between R and S given E, i.e., $R \perp\!\!\!\perp S \mid E$, is maintained. If E is a simple categorical variable, our method will be applicable by small modification. An independence term is computed for each dataset having the same explainable value, $e \in \mathrm{dom}(E)$, and the sum of these terms weighted by $\Pr[e]$ is used as a constraint term. However, it would not be as easy for the cases in which the domain of E is large or E is a continuous variable.

Finally, we will discuss the relation between accuracy and recommendation independence. Fundamentally, as recommendation independence is further enhanced, prediction accuracy tends to worsen. This is due to the decrease in available information for inferring recommendation outcomes. When information about S is not excluded, the available information is the mutual information between R and (S, F), i.e., $\mathrm{I}(R; S, F)$. The information becomes $\mathrm{I}(R; F)$ after excluding the information about S. Because

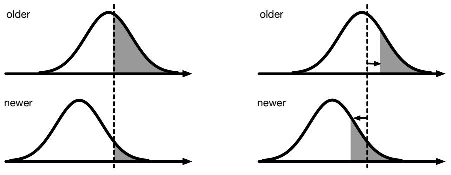
$$\mathrm{I}(R; S, F) - \mathrm{I}(R; F) = \mathrm{I}(R; S|F) \geq 0,$$

the available information is non-increasing when excluding the information on S. Hence, the trade-off for enhancing the independence generally worsens the prediction accuracy.

5.2. Relation to Recommendation Diversity

We will briefly discuss recommendation diversity (Kunaver and Požrl, 2017), which is an attempt to recommend a set of items that are mutually less similar. McNee et al. (2006) pointed out that recommendation diversity is important, because users become less satisfied with recommended items if similar items are repeatedly shown. To our knowledge, Ziegler et al. (2005) were the first to propose an algorithm to diversify recommendation lists by selecting items less similar to those already selected. Lathia et al. (2010) discussed the concept of temporal diversity that is not defined in a single recommendation list, but over temporally successive recommendations. Adomavicius and Kwon (2012) discussed the aggregate diversity and the individual diversity over the items that are recommended to a whole population of users and to a specific user, respectively.

We wish to emphasize that independence is distinct from recommendation diversity. There are three major differences between recommendation diversity and independence. First, while the diversity is the property of a set of recommendations, as described above, the independence is a relation between each recommendation and a specified sensitive feature. Hence, it is impossible to diversify a single recommendation, but a single recommendation can be independent if its prediction of ratings or its determination of whether to recommend or not is statistically independent from a given sensitive feature.



(a) standard (b) diversified

Figure 5: Enhancement of recommendation diversity

Second, recommendation independence depends on the specification of a sensitive feature that is a function of an item and a user. On the other hand, recommendation diversity basically depends on the specification of how items are similar. Even though similarity metrics are not explicitly taken into account, similarities are implicitly considered in a form, for example, whether or not a pair of items are the same. Therefore, independence and diversity are applicable to different situations. Because a sensitive feature can represent the characteristics of users, independence can be applicable for coping with the factors of users, such as the ML1M-Gender or Sushi-Age feature. Inversely, the relative difference between item properties is hard to process by using independence. For example, when using a similarity for diversity, one can represent whether or not two items are the same color, but it is not easy to capture such a feature by using a sensitive feature. In this way, diversity and independence are directed at different aspects in a recommendation context.

Third, while diversity seeks to provide a wider range of topics, independence seeks to provide unbiased information. Consider a case like that in Figure 1, where there are two types of candidate movies, older and newer, and older movies tend to be highly rated. As shown in Figure 5(a), a standard recommender selects top-rated movies, which are illustrated by shading. Newer movies are less recommended than older movies, because newer ones are rated lower. To enhance diversity, newer movies are added to a recommendation list instead of removing the older movies, as shown in Figure 5(b). As a result, both older and newer movies are recommended, and a wider range of movie topics is provided to users. However, the predicted ratings are still affected by whether a movie is newer or older. In other words, this diversified recommendation list is biased in the sense that it is influenced by the release year of movies. This is highly contrasted with the case of recommendation independence shown in Figure 1(b). However, in this latter case, the inverse situation applies: even though the recommendation independence is enhanced, a recommender might select movies having highly skewed topics with respect to sensitive features other than the one specified. Therefore, as shown in this example, the purposes of recommendation diversity and independence are different.

5.3. Other Related Topics

In addition to the recommendation diversity, the concept of recommendation independence has connection with the following research topics.

We adopted techniques for fairness-aware data mining to enhance the independence. Fairness-aware data mining is a general term for mining techniques designed so that sensitive information does not influence the mining outcomes. Pedreschi et al. (2008) first advocated such mining techniques, which emphasized the unfairness in association rules whose consequents include serious determinations. Datta et al. (2016) quantified the influence of a sensitive future by a surrogate data test. Another technique of fairness-aware data mining focuses on classifications designed so that the influence of sensitive information on the predicted class is reduced (Kamishima et al., 2012b; Calders and Verwer, 2010; Kamiran et al., 2012). These techniques would be directly useful in the development of an independence-enhanced variant of content-based recommender systems, because content-based recommenders can be implemented by standard classifiers. Specifically, class labels indicate whether or not a target user prefers a target item, and the features of objects correspond to features of item contents.

The concept behind recommendation transparency is that it might be advantageous to explain the reasoning underlying individual recommendations. Indeed, such transparency has been proven to improve the satisfaction of users (Sinha and Swearingen, 2002), and different methods of explanation have been investigated (Herlocker et al., 2000). In the case of recommendation transparency, the system convinces users of its objectivity by demonstrating that the recommendations were not made with any malicious intention. On the other hand, in the case of independence, the objectivity is guaranteed based on mathematically defined principles. However, it would be useful to examine the interface to quantify the degree of recommendation independence to improve user experiences. Just as it can be helpful to show the confidence of the prediction accuracy, it would be helpful to display the measures of independence. Comparing the independence-enhanced recommendations with non-enhanced recommendations would also be beneficial.

Because independence-enhanced recommendation can be used to avoid the exposure of private information if the private information is represented by a sensitive feature, these techniques are related to privacy-preserving data mining. To protect the private information contained in rating information, dummy ratings are added (Weinsberg et al., 2012). In addition, privacy attack strategies have been used as a tool for detecting discrimination discovery (Ruggieri et al., 2014).

6. Conclusions

We proposed the concept of recommendation independence to exclude the influence of specified information from a recommendation outcome. We previously attempted to enhance recommendation independence, but these attempts merely shifted the predicted ratings. In this paper, we developed new algorithms that can deal with the second moments of rating distributions, and thus the sensitive information could be more strictly removed. The advantages of new algorithms were demonstrated by the experimental results. In addition, we explored applications of independence-enhancement, and clarified the relations between recommendation independence and other recommendation topics, such as diversity.

There are many capabilities required for independence-enhanced recommendation. While our current technique is mainly applicable to the task of predicting ratings, we plan to develop another algorithm for the find-good-items task.

We plan to explore other types of independence terms or approaches. Because sensitive features are currently restricted to binary types, we will also try to develop independence terms that can deal with a sensitive feature that is categorical or continuous. Methods for handling more complicated conditional fairness like that described in section 5.1 need to be developed.

Acknowledgments

We gratefully acknowledge the valuable comments and suggestions of Dr. Paul Resnick. We would like to thank the Grouplens research lab and Dr. Mohsen Jamali for providing datasets. This work is supported by MEXT/JSPS KAKENHI Grant Number JP24500194, JP15K00327, and JP16H02864.

References

G. Adomavicius and Y. Kwon. Improving aggregate recommendation diversity using ranking-based techniques. *IEEE Trans. on Knowledge and Data Engineering*, 24(5):896–911, 2012.

C. M. Bishop. *Pattern Recognition And Machine Learning*. Information Science and Statistics. Springer, 2006.

T. Calders and S. Verwer. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21: 277–292, 2010.

T. Calders, A. Karim, F. Kamiran, W. Ali, and X. Zhang. Controlling attribute effect in linear regression. In *Proc. of the 13th IEEE Int'l Conf. on Data Mining*, pages 71–80, 2013.

Ò. Celma and P. Cano. From hits to niches?: or how popular artists can bias music recommendation and discovery. In *Proc. of the 2nd KDD Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition*, 2008.

D. Cosley, S. K. Lam, I. Albert, J. A. Konstan, and J. Riedl. Is seeing believing? how recommender interfaces affect users' opnions. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, pages 585–592, 2003.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications and Signal Processing. Wiley, second edition, 2006.

A. Datta, S. Sen, and Y. Zick. Algorithmic transparency via quantitative input influence. In *IEEE Symposium on Security and Privacy*, 2016.

S. Forden. Google said to face ultimatum from FTC in antitrust talks. Bloomberg, Nov. 13 2012. ⟨http://bloom.bg/PPNEaS⟩.

A. Gunawardana and G. Shani. A survey of accuracy evaluation metrics of recommendation tasks. *Journal of Machine Learning Research*, 10:2935–2962, 2009.

S. Hajian, F. Bonchi, and C. Castillo. Algorithmic bias: from discrimination discovery to fairness-aware data mining. The 22nd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, Tutorial, 2016.

J. L. Herlocker, J. A. Konstan, and J. Riedl. Explaining collaborative filtering recommendations. In *Proc. of the Conf. on Computer Supported Cooperative Work*, pages 241–250, 2000.

M. Jamali and M. Ester. A matrix factorization technique with trust propagation for recommendation in social networks. In *Proc. of the 4th ACM Conf. on Recommender Systems*, pages 135–142, 2010.

F. Kamiran, A. Karim, and X. Zhang. Decision theory for discrimination-aware classification. In *Proc. of the 12th IEEE Int'l Conf. on Data Mining*, pages 924–929, 2012.

F. Kamiran, I. Žliobaitė, and T. Calders. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and Information Systems*, 35:613–644, 2013.

T. Kamishima. Nantonac collaborative filtering: Recommendation based on order responses. In *Proc. of The 9th Int'l Conf. on Knowledge Discovery and Data Mining*, pages 583–588, 2003.

T. Kamishima and S. Akaho. Considerations on recommendation independence for a find-good-items task. In *Workshop on Responsible Recommendation*, 2017.

T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Enhancement of the neutrality in recommendation. In *The 2nd Workshop on Human Decision Making in Recommender Systems*, 2012a.

T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Proc. of the ECML PKDD 2012, Part II*, pages 35–50, 2012b. [LNCS 7524].

T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Efficiency improvement of neutrality-enhanced recommendation. In *The 3rd Workshop on Human Decision Making in Recommender Systems*, 2013.

T. Kamishima, S. Akaho, H. Asoh, and I. Sato. Model-based approaches for independence-enhanced recommendation. In *Proc. of the IEEE 16th Int'l Conf. on Data Mining Workshops*, pages 860–867, 2016.

Y. Koren. Factorization meets the neighborhood: A multifaceted collaborative filtering model. In *Proc. of the 14th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pages 426–434, 2008.

M. Kunaver and T. Požrl. Diversity in recommender systems — a survey. *Knowledge-Based Systems*, 123:154–162, 2017.

N. Lathia, S. Hailes, L. Capra, and X. Amatriain. Temporal diversity in recommender systems. In *Proc. of the 33rd Annual ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 210–217, 2010.

F. Maxwell Harper and J. A. Konstan. The movielens datasets: History and context. *ACM Trans. on Interactive Intelligent Systems*, 5(4), 2015.

S. M. McNee, J. Riedl, and J. A. Konstan. Accurate is not always good: How accuracy metrics have hurt recommender systems. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, pages 1097–1101, 2006.

E. Pariser. *The Filter Bubble: What The Internet Is Hiding From You.* Viking, 2011.

D. Pedreschi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In *Proc. of the 14th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pages 560–568, 2008.

A. Pérez-Suay, V. Laparra, G. Mateo-García, J. Muños-Marí, L. Gómez-Chova, and G. Camps-Valls. Fair kernel learning. In *Proc. of the ECML PKDD 2017*, 2017.

P. Resnick, J. Konstan, and A. Jameson. Panel on the filter bubble. The 5th ACM Conf. on Recommender Systems, 2011. ⟨http://acmrecsys.wordpress.com/2011/10/25/panel-on-the-filter-bubble/⟩.

S. Ruggieri, S. Hajian, F. Kamiran, and X. Zhang. Anti-discrimination analysis using privacy attack strategies. In *Proc. of the ECML PKDD 2014, Part II*, pages 694–710, 2014. [LNCS 8725].

R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems 20*, pages 1257–1264, 2008.

R. Sinha and K. Swearingen. The role of transparency in recommender systems. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, pages 830–831, 2002.

L. Sweeney. Discrimination in online ad delivery. *Communications of the ACM*, 56(5):44–54, 2013.

S. Watanabe. *Knowing and Guessing – Quantitative Study of Inference and Information.* John Wiley & Sons, 1969.

U. Weinsberg, S. Bhagat, S. Ioannidis, and N. Taft. Blurme: Inferring and obfuscating user gender based on ratings. In *Proc. of the 6th ACM Conf. on Recommender Systems*, pages 195–202, 2012.

C. N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *Proc. of the 14th Int'l Conf. on World Wide Web*, pages 22–32, 2005.

Proceedings of Machine Learning Research 81:1–20, 2018 Conference on Fairness, Accountability, and Transparency

Recommendation Independence [Supplementary Materials]

Toshihiro Kamishima MAIL@KAMISHIMA.NET
Shotaro Akaho S.AKAHO@AIST.GO.JP
Hideki Asoh H.ASOH@AIST.GO.JP
National Institute of Advanced Industrial Science and Technology (AIST),
AIST Tsukuba Central 2, Umezono 1–1–1, Tsukuba, Ibaraki, Japan 305–8568

Jun Sakuma JUN@CS.TSUKUBA.AC.JP
University of Tsukuba, 1–1–1 Tennodai, Tsukuba, Ibaraki, Japan 305–8577; and
RIKEN Center for Advanced Intelligence Project, 1–4–1 Nihonbashi, Chuo-ku, Tokyo, Japan 103-0027

Editors: Sorelle A. Friedler and Christo Wilson

Appendix A. Changes of MAEs, means and standard deviations of predicted ratings

Figure 6 is the full version of Figure 3. This shows the changes of MAEs, means and standard deviations of predicted ratings according to the parameter, η, for all datasets. We focus on pairs of standard deviations depicted by blue dotted lines in the right three columns of the subfigures. While the mean-m term is designed to ignore the second moments of distributions, these moments can be taken into account by the bdist-m and mi-normal terms. Hence, the behaviors of standard deviations, which are the square roots of the second moments, disclose the distinctions of the three independence terms. The observations of the standard deviations may be summarized as:

- ML1M-Gender: all three independence terms could make pairs of standard deviations converge.
- ML1M-Year, Flixster, Sushi-Gender: bdist-m and mi-normal terms could make standard deviations converge, but a mean-m term could not.
- Sushi-Age, Sushi-Seafood: for all three independence terms, standard deviations did not converge.

In summary, there were no cases for which a mean-m term could make standard deviations converge, but a mi-normal term or a bdist-m term could not. From this fact, we can conclude that our new independence terms, bdist-m and mi-normal, enhanced recommendation independence more strictly than the mean-m term.

Appendix B. The Analysis of the Failure to Control the Second Moments for Some Datasets

In section 4.2.1, we briefly described that the failure to control the standard deviations for Sushi-Age and Sushi-Seafood was due to the instability of the predictions. We here show more evidences. Table 3 shows MAEs for two groups, $\mathcal{D}^{(0)}$ and $\mathcal{D}^{(1)}$, under the condition of recommendation independence being fully enhanced ($\eta = 100$). Errors for a Sushi-Age dataset such that S=0, and for a Sushi-Seafood dataset such that S=1 (underlined in the Table) were relatively large. From Table 2, it may be seen that the numbers of data for these two groups were very small, and as a result, predictions became unstable. This instability made it difficult to control the shapes of distributions, and thus standard deviations failed to converge. Despite these difficult conditions, we emphasize that all independence recommenders succeeded to in modifying the first moments of distributions appropriately.

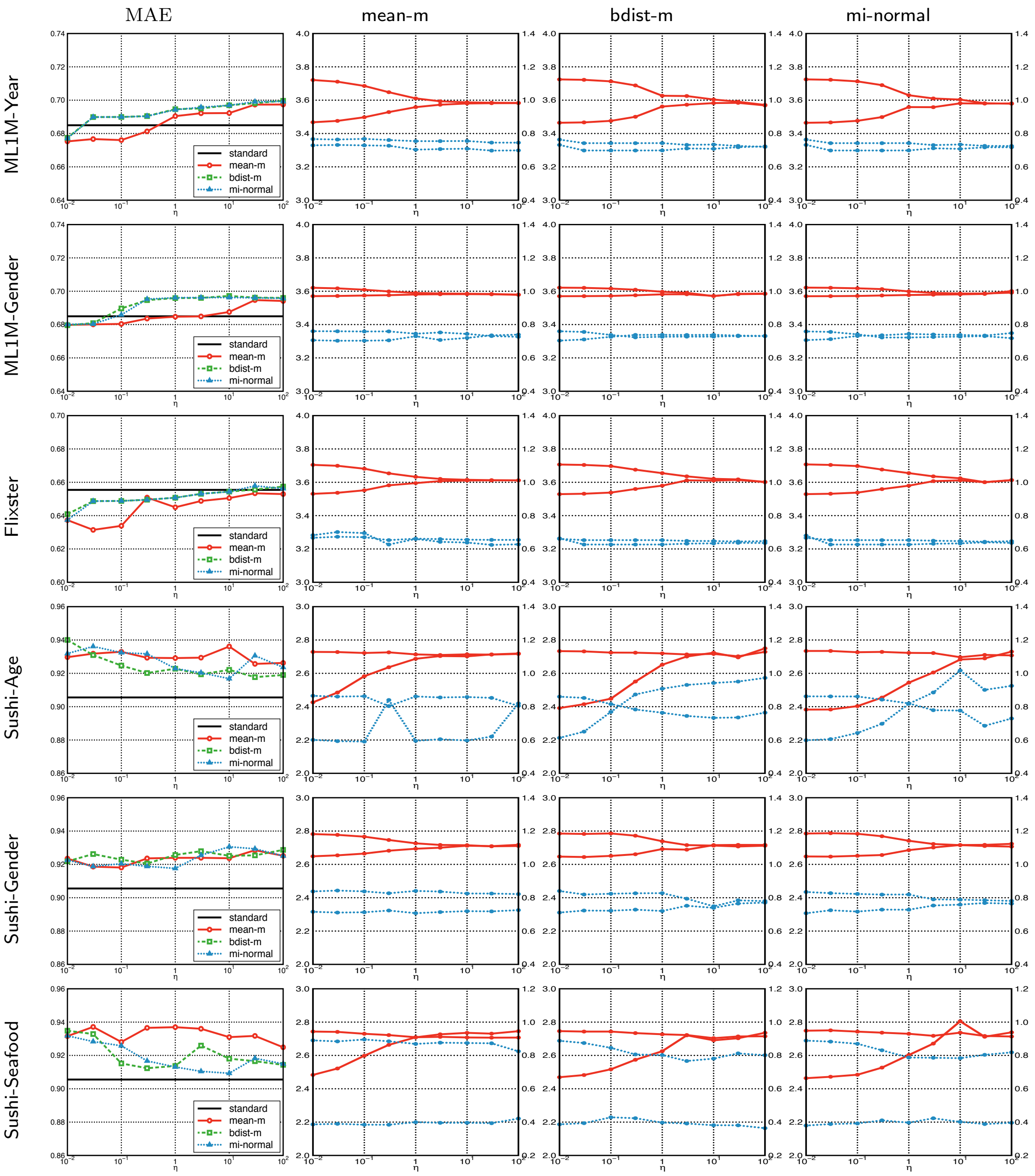
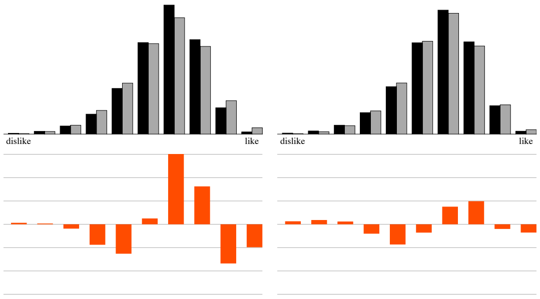


Figure 6: Changes of MAEs, means and standard deviations of predicted ratings

NOTE: The subfigure rows sequentially show the results for the ML1M-Year, ML1M-Gender, Flixster, Sushi-Age, Sushi-Gender, and ML1M-Seafood datasets, respectively. The X-axes of these subfigures represent the independence parameter, η, in a logarithmic scale. The Y-axes of subfigures in the first column represent MAEs in a linear scale. Red solid lines with points, green broken lines, and blue dotted lines show results by the mean-m, bdist-m, and mi-normal models, respectively. Black solid lines without points show non-personalized MAEs in Table 1, which are errors of an original probabilistic matrix factorization model. Note that results for bdist-m and mi-normal terms overlapped in some subfigures. The Y-axes of subfigures in the other three columns represent the means and standard deviations of predicted ratings in a linear scale. Means and standard deviations for two groups based on sensitive values are represented by the scales at the left and right side of these subfigures, respectively. Pairs of means and standard deviations are depicted by red solid and blue dotted lines, respectively.

Table 3: Absolute Mean Errors Per Sensitive Value

Methods	mean-m		mi-normal		bdist-m	
Datasets	S=0	S=1	S=0	S=1	S=0	S=1
ML1M-Year	0.683	0.709	0.684	0.712	0.685	0.712
ML1M-Gender	0.678	0.742	0.680	0.743	0.680	0.744
Flixster	0.681	0.628	0.684	0.631	0.687	0.631
Sushi-Age	<u>1.039</u>	0.919	<u>1.152</u>	0.908	<u>1.156</u>	0.903
Sushi-Gender	0.881	0.965	0.872	0.973	0.878	0.974
Sushi-Seafood	0.909	<u>1.038</u>	0.895	<u>1.059</u>	0.894	<u>1.058</u>



(a) mean-m (b) bdist-m

Figure 7: Distributions of the ratings predicted by mean-m and bdist-m methods for each sensitive value

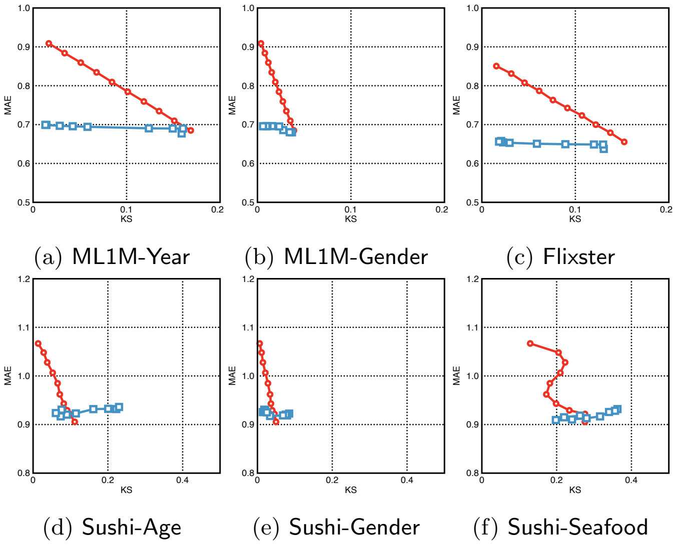
Appendix C. Comparison of Rating Distributions Predicted by mean-m and bdist-m Methods

Figure 7 is the same as the Figure 4 except for comparing the mean-m and bdist-m methods. Similar trends were observed as in the case of comparison with the mi-normal method.

Appendix D. Efficiency of an Accuracy-Independence Trade-Off

We next examined the efficiency of the trade-off between accuracy and independence. Before discussing this trade-off, we first consider the following two baseline errors. The first baseline is the *non-personalized MAE*, defined as the MAE when the mean ratings are always offered. This corresponds to the expected MAE when randomly recommending items. This type of non-personalized recommendation can be considered completely independent, because R is statistically independent from all the other variables, including S. The second baseline is the *standard MAE*, defined as the MAE when ratings are predicted by an original probabilistic matrix factorization model without an independence term. Due the above trade-off, the error in ratings predicted theoretically by an independence-enhanced recommender, would be larger or equal to the standard MAE. We show these two baselines for the three datasets in Table 1. Compared to the MAEs in Figure 2, MAEs produced by our independence-enhanced recommenders were substantially smaller than their corresponding non-personalized MAEs and were nearly equal to or slightly worse than their corresponding standard MAEs.

To analyze the accuracy-independence trade-off, we compared independence-enhanced recommenders with another baseline, a partially random recommender. This partially random recommender offers ratings by a standard recommender, but the ϕ% of items were exchanged with randomly selected items. This replacement could be simulated by replacing the ϕ% values of predicted ratings with the corresponding mean ratings, which were the ratings of a non-personalized recommender. We show the comparison of results obtained by a mi-normal term in Figure 8. Note that results of mean-m and bdist-m terms were similar to those of the mi-normal term. The accuracies of partially random recommenders were much worse than those of our mi-normal at the same level of independence, though results were rather unstable in Sushi-Age and Sushi-Seafood cases. Based on these re-



(a) ML1M-Year (b) ML1M-Gender (c) Flixster

(d) Sushi-Age (e) Sushi-Gender (f) Sushi-Seafood

Figure 8: Comparison of independence-enhanced and partially random recommendations

NOTE: The curves in the charts show the changes in the KS and MAE. The X-axes of these subfigures represent the Kolmogorov-Smirnov static (KS) in a linear scale. The Y-axes represent the prediction errors measured by the mean absolute error (MAE) in a linear scale. Because a smaller KS and MAE respectively indicate greater independence and better accuracy, the bottom left corner of each chart is preferable. Red lines with circles show the indices derived by a partially random recommender whose mixture ratio, ϕ, was changed from 0% to 90%. Blue lines with squares show the results of a mi-normal model when the independence parameter, η, is increased from 10^{-2} to 10^{2}.

sults, we conclude that independence-enhanced recommenders could increase independence with a smaller sacrifice in accuracy than random replacement.

In summary, these experimental results suggest that independence-enhanced recommenders efficiently exclude sensitive information.

Appendix E. Changes in Preference to Movie Genres

To show how the patterns of recommendations were changed, we show the genre-related differences of mean ratings in Table 4. The ML1M-Year and ML1M-Gender data were first divided according to the eighteen kinds of movie genres provided in the original data. Each genre-related data set further divided into two sets according to their sensitive values, the mean ratings were computed for each set, and we showed the differences of these mean ratings. We targeted three types of ratings: the original true ratings, and ratings predicted by mean-m, bdist-m, and mi-normal methods ($\eta = 100$). We selected the six genres for which the absolute differences between original mean ratings for two subsets were largest. In Table 4(a), the genres in which newer movies were highly rated are presented in the upper three rows, and the lower three rows show the genres in which older movies were favored. In Table 4(b), the genres preferred by females are listed in the upper three rows, and the lower three rows show the genres preferred by males.

In this table, because the mi-normal method showed very similar trend with the mean-m and bdist-m methods, we hereafter focus on the mi-normal method. It could be seen that the absolute differences in the mi-normal columns were generally reduced compared to those of the corresponding original differences if they were originally large. For example, in the case of ML1M-Year data, the large absolute difference in the *Fantasy* 0.593 was reduced to 0.308, but the small absolute difference in *Animation* 0.040 was widen to 0.305. This meant that the independence-recommenders did not merely shift the predicted ratings according to the sensitive values to en-

Table 4: Genre-related differences of mean ratings

(a) ML1M-Year data set: old - new

genre	original	mean-m	bdist-m	mi-normal
Animation	−0.040	−0.302	−0.301	−0.305
Documentary	0.113	−0.120	−0.117	−0.121
Film-Noir	0.238	−0.037	0.025	0.007
Western	0.524	0.254	0.231	0.232
Mystery	0.563	0.295	0.303	0.297
Fantasy	0.593	0.331	0.308	0.308

(b) ML1M-Gender data set: male - female

genre	original	mean-m	bdist-m	mi-normal
Children's	−0.214	−0.160	−0.160	−0.146
Musical	−0.213	−0.154	−0.154	−0.146
Romance	−0.100	−0.046	−0.047	−0.036
Crime	0.024	0.081	0.089	0.094
Film-Noir	0.074	0.125	0.145	0.134
Western	0.103	0.155	0.147	0.163

hance independence. By excluding the information about sensitive features, the differences between mean ratings were occasionally widened. This is because the balance between independence and accuracy is considered in equation (1), and thus the ratings for events that are more helpful for enhancing independence are drastically changed.