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.

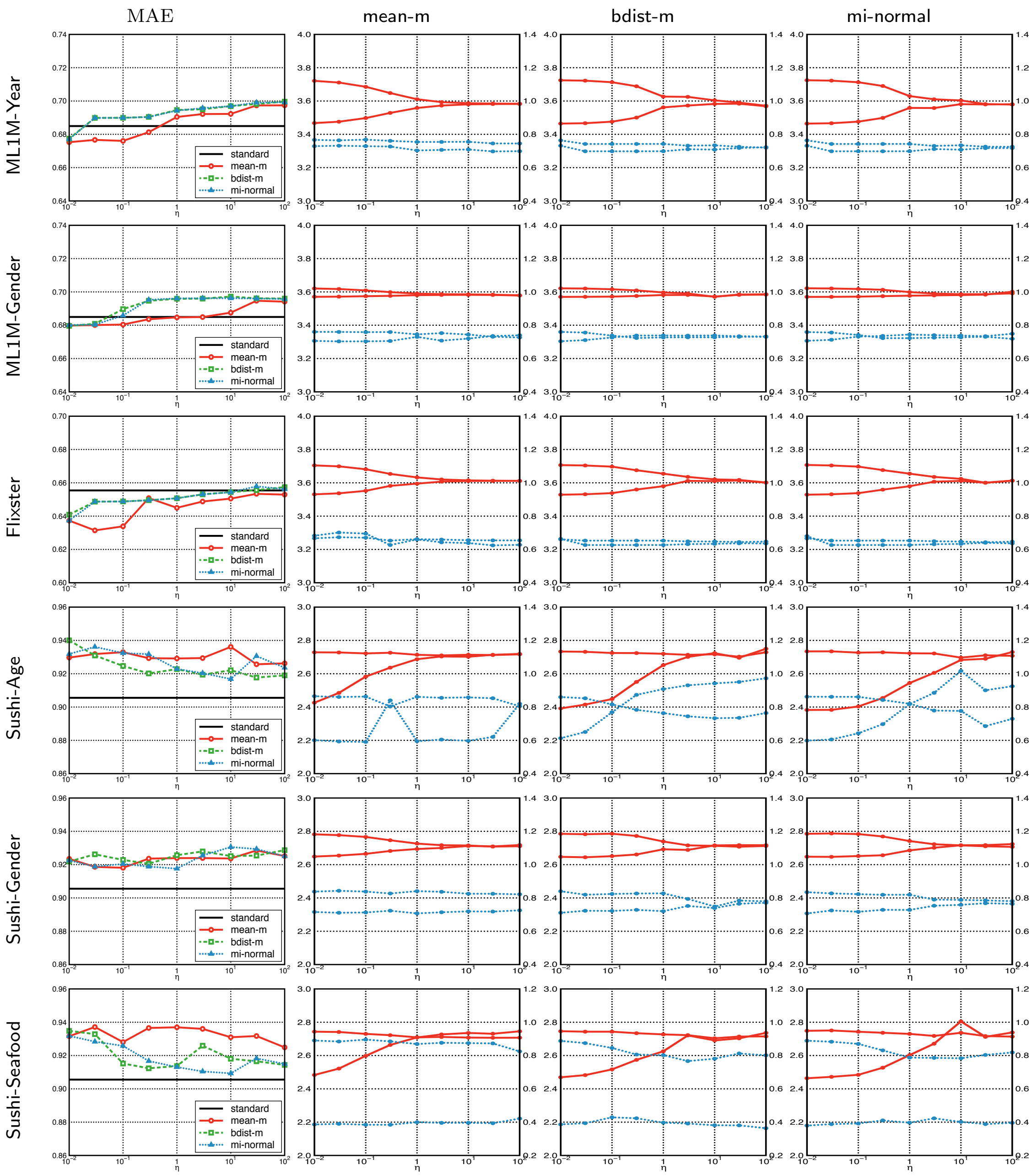
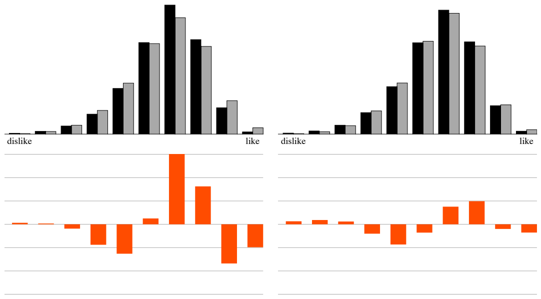


Figure 6: Changes of MAEs, means and standard deviations of predicted ratings

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	1.039	0.919	1.152	0.908	1.156	0.903
Sushi-Gender	0.881	0.965	0.872	0.973	0.878	0.974
Sushi-Seafood	0.909	1.038	0.895	1.059	0.894	1.058



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

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

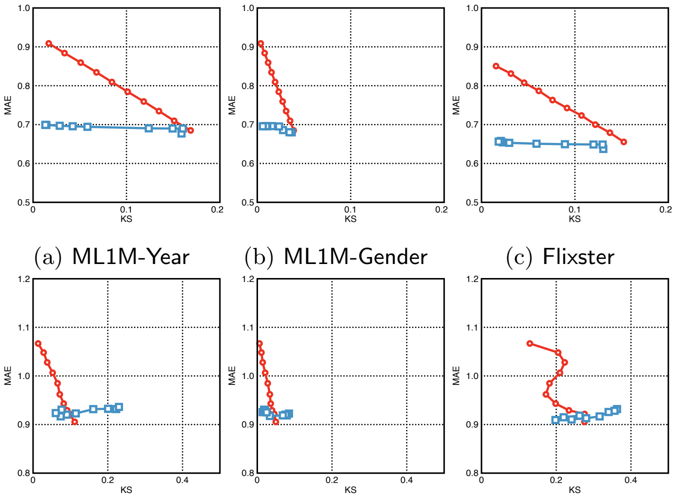
Appendix C. Comparison of Rating Distributions Predicted by mean-m and bdist-m Methods

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.