Multi-style explainable matrix factorization techniques for recommender systems.

Olurotimi Nugbepo Seton
University of Louisville

Follow this and additional works at: https://ir.library.louisville.edu/etd
Part of the Other Computer Engineering Commons

Recommended Citation
https://doi.org/10.18297/etd/3661

This Doctoral Dissertation is brought to you for free and open access by ThinkIR: The University of Louisville's Institutional Repository. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of ThinkIR: The University of Louisville's Institutional Repository. This title appears here courtesy of the author, who has retained all other copyrights. For more information, please contact thinkir@louisville.edu.
MULTI-STYLE EXPLAINABLE MATRIX FACTORIZATION TECHNIQUES
FOR RECOMMENDER SYSTEMS

By

Olurotimi Nugbepo Seton
M.Sc., Computer Science,
Babcock University Ilishan-Remo, Nigeria

A Dissertation
Submitted to the Faculty of the
J.B. Speed School of Engineering of the University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy in Computer Science and Engineering

Department of Computer Science and Engineering
University of Louisville
Louisville, Kentucky

May 2021
MULTI-STYLE EXPLAINABLE MATRIX FACTORIZATION TECHNIQUES
FOR RECOMMENDER SYSTEMS

By

Olurotimi Nugbepo Seton
M.Sc., Computer Science,
Babcock University Ilishan-Remo, Nigeria

A Dissertation Approved On

April 13, 2021

by the following Dissertation Committee:

Dr. Olfa Nasraoui, Dissertation Advisor

Dr. Nihat Altiparmak

Dr. Hichem Frigui

Dr. Cara Cashon

Dr. Antonio Badia
ACKNOWLEDGEMENTS

Dr. Olfa Nasraoui, for her guidance, patience, encouragement and mentoring throughout my Ph.D program. I could not have done this without her unwavering support.

My dissertation committee for their support and guidance.

My lab mates, who have also become my friends, Pegah, Aneseh, Cagla, Khalid, Sami and Meriem as well as previous lab members Wenlong, Behnoush and Mohammed. I will like to also thank my supervisor at REACH, Dennis Keibler, for his support and flexibility, without which I could not have made progress in my dissertation.

My brothers and sister, parents and friends who have become family.

My wife, Tundun, for her support and understanding and unwavering support and love.

Finally, I will like to thank God for provision and inspiration.
ABSTRACT

MULTI-STYLE EXPLAINABLE MATRIX FACTORIZATION TECHNIQUES FOR RECOMMENDER SYSTEMS

Olurotimi Nugbepo Seton

April 13, 2021

Black-box recommender system models are machine learning models that generate personalized recommendations without explaining how the recommendations were generated to the user or giving them a way to correct wrong assumptions made about them by the model. However, compared to white-box models, which are transparent and scrutable, black-box models are generally more accurate. Recent research has shown that accuracy alone is not sufficient for user satisfaction. One such black-box model is Matrix Factorization, a State of the Art recommendation technique that is widely used due to its ability to deal with sparse data sets and to produce accurate recommendations. Recent work has proposed new Matrix Factorization models that are explainable by incorporating explanations derived from semantic knowledge graphs, user neighborhood, or item neighborhood graphs into the model learning process. These Explainable Matrix Factorization (EMF) methods have the benefit of providing explanations without sacrificing accuracy. However, their explanations tend to be limited to only one explanation style. In this dissertation, we propose a framework comprising new machine learning methods to build explainable models that can make recommendations with multiple explanation-styles, by hybridizing multiple EMF models and by proposing new EMF models that explain recommendations using tags.
The various pre-calculated explainability scores, leveraged in our proposed methods, have all been validated in prior work that conducted user studies to evaluate users’ satisfaction with each style individually. Unlike most existing work that generates explanations post-hoc, i.e., after the predictions have already been made, our framework is based on calculating explainability scores directly from available data, before the model is learned, and then using them as part of a regularization mechanism, to guide the model learning. Unlike post-hoc methods, our framework makes it possible to learn machine learning models that take into account the explanation scores, therefore ensuring higher transparency.

Our evaluation experiments show that our proposed methods provide accurate recommendations while also providing users with multiple styles of explanations about how data was used to generate each recommendation. Each explanation style also provides additional decision-making information that empowers the user to either trust or scrutinize the recommendations. Although, rooted in the hybrid recommendation framework, our proposed methods make a significant step forward in explainable AI and beyond existing hybrid frameworks, because the proposed hybridization mechanisms make an intentional effort to take into account the individual models’ explanations and not only their output predicted ratings.
TABLE OF CONTENTS

ACKNOWLEDGEMENTS iii
ABSTRACT iv
LIST OF TABLES ix
LIST OF FIGURES xiii
LIST OF ALGORITHMS xviii

CHAPTER

1 INTRODUCTION ......................................................... 1
  1.1 Motivation ....................................................... 1
    1.1.1 Examples of Explanations from the Movie Domain ......... 2
  1.2 Problem Statement ............................................. 11
  1.3 Rationale ....................................................... 12
  1.4 Research Contributions ....................................... 12
  1.5 Document Organization ........................................ 14

2 BACKGROUND AND LITERATURE REVIEW ......................... 16
  2.1 Collaborative Filtering (CF) .................................... 17
    2.1.1 Memory-based Methods .................................... 17
    2.1.2 Model-based Methods ..................................... 19
    2.1.3 Learning Algorithms for Matrix Factorization .......... 24
  2.2 Explanations in Recommender Systems ....................... 25
    2.2.1 Motivation ................................................ 25
    2.2.2 Explanations Styles in Recommender Systems .......... 27
    2.2.3 Using Tags in Collaborative Filtering Recommender Systems . 32
    2.2.4 Explainable Latent Factor Methods .................... 33
3 MULTIPLE EXPLANATION STYLES METHODS FOR MATRIX FACTORIZATION-BASED RECOMMENDER SYSTEMS

3.1 Rationale for Multi-style Explainable Matrix Factorization
   3.1.1 Rationale for Multiple Explanation Styles
   3.1.2 Rationale for Pre-calculated Explainability Scores

3.2 Hybridization Techniques for Multi-style Explainable Matrix Factorization
   3.2.1 Weighted Multi-style Explainable Matrix Factorization (W-MEMF)
   3.2.2 Regression-based Multi-style Explainable Matrix Factorization (R-MEMF)
   3.2.3 Switched Multi-style Explainable Matrix Factorization (S-MEMF)
   3.2.4 Asymmetric Multi-style Explainable Matrix Factorization (A-MEMF)

3.3 Tag-Assisted Explainable Matrix Factorization Methods
   3.3.1 Tag-based Explainability Graphs
   3.3.2 Tag-Assisted Explainable Matrix Factorization (TA-EMF)

3.4 Tag-boosted Multi-style Explainable Matrix Factorization
   3.4.1 Preferred Tag-boosted Multi-style EMF
   3.4.2 Relevant Tag-boosted Multi-style EMF

4 EXPERIMENTAL EVALUATION

4.1 Evaluation Metrics
   4.1.1 Accuracy Metrics
   4.1.2 Explainability Metrics
LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Top-3 Movies rated by Sample User A. Movies are ranked in descending order of the ratings.</td>
</tr>
<tr>
<td>2</td>
<td>Top-3 Movies Recommended for Sample User A using Matrix Factorization Model. Movies are ranked in descending order of the predicted ratings.</td>
</tr>
<tr>
<td>3</td>
<td>Top-3 Recommended Movies for Sample User A using $EMF_{UB}$. Movies are ranked in descending order of the predicted ratings.</td>
</tr>
<tr>
<td>4</td>
<td>Top-3 Recommended Movies for Sample User A using $EMF_{IB}$. Movies are ranked in descending order of the predicted ratings.</td>
</tr>
<tr>
<td>5</td>
<td>Top-3 Recommended Movies for Sample User B using $EMF_{UB}$. Movies are ranked in descending order of the predicted ratings.</td>
</tr>
<tr>
<td>6</td>
<td>Top-3 Recommended Movies for Sample User B using $EMF_{IB}$. Movies are ranked in descending order of the predicted ratings.</td>
</tr>
<tr>
<td>7</td>
<td>Top-3 Movies rated by Sample User B (Ranked in descending order of ratings)</td>
</tr>
<tr>
<td>8</td>
<td>Top-3 Movies Recommended for Sample User B using Matrix Factorization Model (Ranked in descending order of predicted ratings)</td>
</tr>
<tr>
<td>9</td>
<td>Summary of Research Contributions</td>
</tr>
<tr>
<td>10</td>
<td>Summary of notation used in this Chapter</td>
</tr>
<tr>
<td>11</td>
<td>Summary of Research Goals and Contributions</td>
</tr>
<tr>
<td>12</td>
<td>Summary of Data Used for Evaluating our Proposed Hybrid Methods</td>
</tr>
<tr>
<td>13</td>
<td>Explainability Coverage in Book and Movie Domains</td>
</tr>
<tr>
<td>14</td>
<td>Results of Hybrid Models on Test Data (Movie Domain). Best results are in bold</td>
</tr>
</tbody>
</table>
15 RMSE significance test results in the movie domain. Best results are in bold font. ................................................................. 80
16 NDCG@10 significance test results in the movie domain. Best results are in bold font. ................................................................. 81
17 Results of Hybrid Models on Test Data (Book Domain). Best results are in bold font. ................................................................. 81
18 MAP@10 significance test results in the book domain. Best results are in bold font. ................................................................. 82
19 MEP@10 significance test results in the movie domain (NSE-explainability) — $\theta^n = 0.05$. The winning method is in bold ........................................ 84
20 MEP@10 significance test results in the movie domain (ISE-explainability) — $\theta^i = 0.01$. The winning method is in bold ........................................ 85
21 MER@10 significance test results in the movie domain (NSE-explainability) — $\theta^n = 0.05$. The winning method is in bold ........................................ 85
22 MER@10 significance test results in the movie domain (ISE-explainability) — $\theta^i = 0.01$. The winning method is in bold ........................................ 85
23 Top-3 Movies rated by Sample User A. Movies are ranked in descending order of ratings. ................................................................. 89
24 Top-3 Movies Recommended for Sample User A using Matrix Factorization (MF) Model. Movies are ranked in descending order of the predicted ratings. 89
25 Top-3 Movies Recommended for Sample User A using User-based Explainable Matrix Factorization ($EMF_{UB}$) Model. Movies are ranked in descending order of the predicted ratings. ........................................ 89
26 Top-3 Movies Recommended for Sample User A using Item-based Explainable Matrix Factorization ($EMF_{IB}$) Model. Movies are ranked in descending order of the predicted ratings. ........................................ 91
Top-3 Movies Recommended for Sample User A using S-MEMF (Switched Multi-style Explainable Matrix Factorization using Smooth Maximum Switching Criterion) with $\alpha = 10.0$. Movies are ranked in descending order of the predicted ratings. .................................................. 93

Top-3 Movies Recommended for Sample User A using W-MEMF (Global) with $weight = 0.6$ (Slightly more weight attributed to $EMF_{UB}$ than $EMF_{IB}$). Movies are ranked in descending order of the predicted ratings. ............. 95

Top-3 Movies Recommended for User 13 using R-MEMF (Lasso) with $\beta_{NSE} = 0.4993$ and $\beta_{ISE} = 1.4354$. Movies are ranked in descending order of the predicted ratings. .................................................. 97

Top-3 Movies rated by Sample User B. Movies are ranked in descending order of the ratings. ................................................................. 100

Top-3 Movies recommended for Sample User B using Matrix Factorization (MF) Model. Movies are ranked in descending order of the predicted ratings. 101

Top-3 Movies Recommended for Sample User B using User-based Explainable Matrix Factorization ($EMF_{UB}$) Model. Movies are ranked in descending order of the predicted ratings. .................................................. 101

Top-3 Movies Recommended for Sample User B using Item-based Explainable Matrix Factorization ($EMF_{IB}$) Model. Movies are ranked in descending order of the predicted ratings. .................................................. 103

Top-3 Movies Recommended for Sample User B using S-MEMF (Switched Multi-style Explainable Matrix Factorization using Smooth Maximum Switching Criterion with $\alpha = 10.0$). Movies are ranked in descending order of predicted ratings. .................................................. 105

Top-3 Movies Recommended for Sample User B using W-MEMF (Global) with $weight = 0.6$: This means that slightly more weight was attributed to $EMF_{UB}$ than $EMF_{IB}$. Movies are ranked in descending order of ratings. . 107
Top-3 Movies Recommended for Sample User B using R-MEMF (Lasso) with $\beta_{NSE} = 0.4993$ and $\beta_{ISE} = 1.4354$. Movies are ranked in descending order of the predicted ratings.

RMSE vs Latent Factors (K) in the movie domain. $EMF_{TA}$, PrefTag, RelTag denote our proposed methods. Bold denote the best results.

MAP@10 vs Latent Factors (K) in the movie domain. $EMF_{TA}$, PrefTag, RelTag denote our proposed methods.

NDCG@10 vs Latent Factors (K) in the movie domain. $EMF_{TA}$, PrefTag, RelTag denote our proposed methods.

RMSE significance test results (K = 50, Common Relevant Tags = 5).

MAP@10 significance test results (K = 20, Common Preferred Tags = 10). Bold denotes the winner.

NDCG@10 significance test results (K = 50, Common Relevant Tags = 3). Bold denotes the winner.

MEP@10 significance test results (NSE-explainablity) — $\theta^n = 0.1$. Bold denotes the winner.

MER@10 significance test results (NSE-explainablity) — $\theta^n = 0.1$. Bold denotes the winner.

MEP@10 significance test results (ISE-explainablity) — $\theta^i = 0.1$. Bold denotes the winner.

MER@10 significance test results (ISE-explainablity) — $\theta^i = 0.1$. Bold denotes the winner.

Top-3 rated movies for Sample User.

Output of Relevant Tag-boosted Multi-style EMF for Sample User.

Output of Preferred Tag-boosted Multi-style EMF for Sample User.
<table>
<thead>
<tr>
<th>FIGURE</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Neighborhood-rating Style Explanations for top-3 movies recommended to Sample User A using $EMF_{UB}$</td>
</tr>
<tr>
<td>2</td>
<td>Influence Style Explanations for top-3 movies recommended to Sample User A using $EMF_{IB}$</td>
</tr>
<tr>
<td>3</td>
<td>Neighborhood-rating Style Explanations for top-3 movies recommended to Sample User B using $EMF_{UB}$</td>
</tr>
<tr>
<td>4</td>
<td>Influence Style Explanations for top-3 movies recommended to Sample User B using $EMF_{IB}$</td>
</tr>
<tr>
<td>5</td>
<td>Contributions of this work shown within the greater taxonomy of Matrix Factorization methods for Recommender Systems (RS). New techniques are shown with a bold outline.</td>
</tr>
<tr>
<td>6</td>
<td>Intermediate entities commonly used in explaining recommendation to users. Source: [1]</td>
</tr>
<tr>
<td>7</td>
<td>The Keyword Explanation Style. Source: [2]</td>
</tr>
<tr>
<td>8</td>
<td>Left: Example of user-based neighbors explanation style. Right: Example of item-based neighbors explanation style. Source: [3]</td>
</tr>
<tr>
<td>9</td>
<td>(a) Typical Explanation Style, (b) Explainable Matrix Factorization. Source: [4]</td>
</tr>
<tr>
<td>10</td>
<td>Contributions of this work shown within the greater taxonomy of Matrix Factorization methods. New techniques are shown with a bold outline.</td>
</tr>
<tr>
<td>11</td>
<td>Responses showing the distribution of explanation style preferences of 19 users for three movies from three different genres</td>
</tr>
</tbody>
</table>
12 Distribution of the number of distinct explanation styles selected by each respondent for the three recommended movies ............................................. 42
13 Tag-based Explanation Graph for Explaining the Recommended Movie, Pulp Fiction, to a User, obtained by combining the tag preferences (orange) and the tag relevance (blue) ................................................................. 58
14 Using Tag-based Explanation Graphs to explain the recommended movie, Coming to America to a sample user ................................................................. 59
15 Flowchart for evaluating our proposed Hybrid Multi-style EMF methods . 72
16 Tuning the Weighted Multi-style EMF (W-MEMF) Methods using the Validation Set (Movie Domain) (a) RMSE vs weight (b) MAP@10 vs weight (c) NDCG@10 vs weight ................................................................. 73
17 Tuning the Weighted Multi-style Methods using the Validation Set (Book Domain) (a) RMSE vs weight (b) MAP@10 vs weight (c) NDCG@10 vs weight 74
18 Tuning the Regression-based Multi-style EMF (R-MEMF) Methods using the Validation Set (Movie Domain) (a) RMSE vs λ (b) MAP@10 vs λ (c) NDCG@10 vs λ .................................................................................. 75
19 Tuning the Regression-based Multi-style EMF (R-MEMF) Methods using the Validation Set (Book Domain) (a) RMSE vs λ (b) MAP@10 vs λ (c) NDCG@10 vs λ .................................................................................. 76
20 Tuning the Switched Multi-style EMF (S-MEMF) Methods using the Validation Set (Movie Domain) (a) RMSE vs α (b) MAP@10 vs α (c) NDCG@10 vs α .................................................................................. 76
21 Tuning the Switched Multi-style EMF (S-MEMF) Methods using the Validation Set (Book Domain) (a) RMSE vs α (b) MAP@10 vs α (c) NDCG@10 vs α .................................................................................. 77
22 Tuning the Asymmetric Multi-style Methods using the Validation Set (Movie Domain) (a) RMSE vs iterations (b) MAP@10 vs iterations (c) NDCG@10 vs iterations ................................................................. 78
Tuning the Asymmetric Multi-style EMF (A-MEMF) Methods using the Validation Set (Movie Domain) (a) RMSE vs iterations (b) MAP@10 vs iterations (c) NDCG@10 vs iterations

Figure (a): $MPE@10$ vs explainability threshold $\theta^n$ using NSE Explainability Graph. Figure (b): $MER@10$ vs explainability threshold $\theta^n$. (Movie Domain)

Figure (a): $MPE@10$ vs explainability threshold $\theta^i$ using ISE Explainability Graph. Figure (b): $MER@10$ vs explainability threshold $\theta^i$. (Movie Domain)

Figure (a): $MPE@10$ vs explainability threshold $\theta^n$ using NSE Explainability Graph. Figure (b): $MER@10$ vs explainability threshold $\theta^n$. (Book Domain)

Figure (a): $MPE@10$ vs explainability threshold $\theta^i$ using ISE Explainability Graph. Figure (b): $MER@10$ vs explainability threshold $\theta^i$. (Book Domain)

Corresponding Explanations for movies recommended to Sample User A in Table 25 using the NSE-style generated by the $EMF_{UB}$ model. Top: First recommended movie. Middle: Second recommended movie. Bottom: Third recommended movie

Corresponding Explanations for movies recommended to Sample User A in Table 26 using the ISE-style generated by the $EMF_{IB}$ model. Top: First recommended movie. Middle: Second recommended movie. Bottom: Third recommended movie

Corresponding Explanations for movies recommended to Sample User A in Table 27 using the Multi-style Explanations generated by the S-MEMF (Switched Multi-style Explainable Matrix Factorization using Smooth Maximum Switching Criterion) model with $\alpha = 10.0$. Top and Middle: ISE style explanations for top 2 recommended movies. Bottom: NSE style explanation for third recommended movie
Corresponding Explanations (Left: NSE style, Right: ISE style) for movies recommended to Sample User A in Table 28 using the Multi-style Explanation generated using the W-MEMF (Global) model with weight $= 0.6$ (Slightly more weight attributed to $EMF_{UB}$ than $EMF_{IB}$). Top, Middle, Bottom: Explanations for first, second, and third movie, respectively . . . . . . . . . 96

Corresponding Explanations for movies recommended to Sample User A in Table 29 using the Multi-style Explanation generated with the R-MEMF (lasso) model with $\beta_{NSE} = 0.4993$ and $\beta_{ISE} = 1.4354$ (Only ISE Explanation Style presented to sample user). Top: Explanations for first movie. Middle: Explanations for second movie. Bottom: Explanation for third movie . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 98

Corresponding Explanations for the movies recommended to Sample User B in Table 32 using the NSE-style generated by the $EMF_{UB}$ model. Top: First recommended movie. Middle: Second recommended movie. Bottom: Third recommended movie . . . . . . . . . . . . . . . . 102

Corresponding Explanations recommended to Sample User B in Table 33 using the ISE-style generated by the $EMF_{IB}$ model. Top: First recommended movie. Middle: Second recommended movie. Bottom: Third recommended movie . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 104

Corresponding Explanations for movies recommended to Sample User B in Table 34 using the Multi-style Explanations generated by the S-MEMF (Switched Multi-style Explainable Matrix Factorization using Smooth Maximum Switching Criterion) model with $\alpha = 10.0$. Top: NSE style explanation for top recommended movie. Middle: ISE style explanations for second movie. Bottom: NSE style explanation for third recommended movie . . . . . 106
Explanation Styles (Left: NSE, Right: ISE) for recommended movies in Table 35 using the Multi-style Explanations generated using the W-MEMF (Global) model with weight = 0.6 (Slightly more weight attributed to $EMF_{UB}$ than $EMF_{IB}$). Top, Middle, Bottom: Explanations for first, second, and third movie, respectively.

Corresponding Explanations for recommended movies in Table 36 using the Multi-style Explanation generated using the R-MEMF (lasso) model with $\beta_{NSE} = 0.4993$ and $\beta_{ISE} = 1.4354$ (Hence, only the ISE Explanation Style is presented to the user). Top: Explanations for first movie. Middle: Explanations for second movie. Bottom: Explanation for third movie.

MAP@10 vs Number of Common Tags: Common Relevant Tags for RelTag and Common Preferred Tags for PrefTag.

NDCG@10 vs Number of Common Tags: Common Relevant Tags for RelTag and Common Preferred Tags for PrefTag.

MEP@10 vs Explainability Threshold ($\theta$). Our proposed methods are denoted as $EMF_{TA}$, PrefTags and RelTags.

MER@10 vs Explainability Threshold ($\theta$). Our proposed methods are $EMF_{TA}$, PrefTags and RelTags.
# LIST OF ALGORITHMS

<table>
<thead>
<tr>
<th>ALGORITHM</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Hybridized Multi-style EMF: used in the methods discussed in Sections 3.2.1 and 3.2.3</td>
<td>46</td>
</tr>
<tr>
<td>2 Weighted-Multi-style EMF (W-MEMF) using Global Weights</td>
<td>47</td>
</tr>
<tr>
<td>3 Weighted-Multi-style EMF (W-MEMF) using Personalized Weights</td>
<td>49</td>
</tr>
<tr>
<td>4 Switched-Multi-style EMF (S-MEMF) using Smooth Maximum Criterion</td>
<td>52</td>
</tr>
<tr>
<td>5 Preferred Tag-boosted Multi-style Explainable Matrix Factorization (PrefTag)</td>
<td>63</td>
</tr>
<tr>
<td>6 Relevant Tag-boosted Multi-style Explainable Matrix Factorization (RelTag)</td>
<td>65</td>
</tr>
</tbody>
</table>
CHAPTER 1

INTRODUCTION

1.1 Motivation

Recommender Systems (RSs) have become an increasingly crucial part of the online experience; they help users organize and understand their choices in a space with almost endless combinations of choices. The backbone of the modern RS are Machine Learning (ML) algorithms that have become increasingly accurate at predicting users’ preferences from data. As they become more and more accurate, however ML models have become increasingly difficult to explain. ML models whose path to making a decision cannot be explained are called Black-box models. These models are highly accurate but fail to explain their predictions to the users. Explainable Artificial Intelligence (XAI) can be defined as a model that produces details or reasons that make its functioning or reasoning clear or easy to understand for its target audience [5]. We use this definition in our work with the target users being the end users who require explanations to understand how their choices were selected for them by the recommendation engine. This research work focuses on the state of the art RSs that are based on Matrix Factorization (MF) [6] which are a family of black-box models that are highly accurate. The input to MF is a set of ratings ($R$) which holds the ratings given by a set of users $U$ to a set of items $I$, such that $r_{ui} \in R$ is the rating given by a user $u \in U$ to item $i \in I$, and is a value within a specified range. The rating matrix is very sparse because users do not rate every item in the set of items $I$. The set of items could be movies, destinations, music or even other users.

The accuracy of predictions is no longer considered sufficient as the sole evaluation metric for RSs [7] since explanations can be essential for some users to understand why certain items were recommended to them. In fact, explanations can help humans gain
insight into how results were generated and can be a beneficial way to identify biases and
detect errors in the model [8]. Post-hoc explanations [9] [10], among others, have been used
to explain the predictions of Machine Learning classifiers to users by explaining the results
of the model using feature importance and contributions to the prediction score. In contrast,
explainable Matrix Factorization methods such as [4] [3] [11] [12] have focused on explaining
the black-box Machine Learning model from within rather than after building the model,
by learning an explainable predictive model from the start. To this day, most explanation
methods assume one specific explanation style for all users and all items. However not only
is this not intuitively motivated - because different explanation styles may be more suitable
for different users, items, and/or domains - , but it is also likely that not all user and item
pairs will have an explanation source - for instance certain items (users) may not share
many similarities with other items (users), and may therefore lack neighborhood ratings
that are essential to neighborhood based explanation styles. In the following section, we
illustrate different explanations from real data and models, and show the importance of
explanations, while also showing the limitations of single-style explanations.

1.1.1 Examples of Explanations from the Movie Domain

In the following examples we refer to anonymous users from the MovieLens 1 data
set using the letter A, B, etc. Table 1 shows sample user A’s top-3 rated movies from the
real MovieLens movie dataset that consists of user ratings on movies. Table 2 shows the
top-3 recommended movies for sample user A, generated using a Matrix Factorization-based
(MF) recommender system model.

1https://grouplens.org/datasets/movielens/100k/
TABLE 1

Top-3 Movies rated by Sample User A. Movies are ranked in descending order of the ratings.

<table>
<thead>
<tr>
<th>Top-3 rated movies</th>
<th>Genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Wizard of Oz (1939)</td>
<td>Fantasy/Musical</td>
</tr>
<tr>
<td>The Shining (1980)</td>
<td>Horror/Mystery</td>
</tr>
<tr>
<td>One Flew Over the Cuckoo’s Nest (1975)</td>
<td>Drama/Comedy</td>
</tr>
</tbody>
</table>

TABLE 2

Top-3 Movies Recommended for Sample User A using Matrix Factorization Model. Movies are ranked in descending order of the predicted ratings.

<table>
<thead>
<tr>
<th>Top-3 recommended movies</th>
<th>Genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>They Made Me a Criminal (1939)</td>
<td>Noir/Crime</td>
</tr>
<tr>
<td>Shadow of Angels (1976)</td>
<td>Drama</td>
</tr>
<tr>
<td>Daens (1992)</td>
<td>Drama</td>
</tr>
</tbody>
</table>

Table 8 shows the output of MF which consists of the recommendations but without explanations. In this case, an explanation can take the form of information that is presented to the user to serve different goals such as exposing the reasoning behind a recommendation [13].

The lack of an explanation, which would help a user understand why a certain movie was recommended to them by the MF-based model, has motivated recent efforts to add explainability to MF. These methods based on Explainable Matrix Factorization (EMF) [4] [14] [3] [11] [15] [12], introduced explainability to the MF model learning process by adding an explainability term to MF’s objective function. The explainability term introduces a constraint which tries to bring the corresponding latent feature vectors of users and items, that are considered explainable to such users, closer in the latent space in order to push
them to the top of the top recommendations. This means that the final (explainable) recommendations made by EMF models can be different from the ones recommended by the classical MF technique. This is specifically because items that are explainable are favored in the final output ranking of EMF.

Tables 3 and 4 show the top-3 recommended movies using User-based EMF and Item-based EMF [4] [14] models, denoted as $EMF_{UB}$ and $EMF_{IB}$, respectively, for sample user A.

<table>
<thead>
<tr>
<th>TABLE 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Top-3 Recommended Movies for Sample User A using $EMF_{UB}$.</strong> Movies are ranked in descending order of the predicted ratings.</td>
</tr>
<tr>
<td><strong>Top-3 recommended movies</strong></td>
</tr>
<tr>
<td>Fire On the Mountain (1996)</td>
</tr>
<tr>
<td>Beyond Bedlam (1993)</td>
</tr>
<tr>
<td>All Things Fair (1996)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE 4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Top-3 Recommended Movies for Sample User A using $EMF_{IB}$.</strong> Movies are ranked in descending order of the predicted ratings.</td>
</tr>
<tr>
<td><strong>Top-3 recommended movies</strong></td>
</tr>
<tr>
<td>Two Friends (1986)</td>
</tr>
<tr>
<td>Mamma Roma (1962)</td>
</tr>
</tbody>
</table>

A key advantage of the EMF methods [4] [14] [3] [11] [15] [12] is the use of pre-calculated explainability scores which measure the strength of an explanation for an item to a user using the selected EMF method. These explainability scores can be presented to users as explanations for the recommended items using validated explanation styles. Figures
1 and 2 show the explanations presented to user A for the recommended movies using the $EMF_{UB}$ and $EMF_{IB}$ models, respectively. The recommendations from $EMF_{UB}$ are presented using Neighborhood-rating Style Explanation (NSE) [16] while the recommendations from $EMF_{IB}$ are presented using the Influence Style Explanation (ISE) [2].

Figure 1: Neighborhood-rating Style Explanations for top-3 movies recommended to Sample User A using $EMF_{UB}$
Figure 2: Influence Style Explanations for top-3 movies recommended to Sample User A using $EMF_{IB}$

The explanations from each individual explanation style provide useful additional information that can allow a user to make a decision about whether to follow each recommendation. Figure 1 (Top) shows that the first predicted movie, using $EMF_{UB}$, was rated highly by four similar users, but it was rated lowly by one similar user. The similar users used in this neighborhood style explanation (NSE) are users with similar preferences or ratings to the sample user. Figure 2 (Top) shows that the first movie recommended by
*EMF*$_{IB}$, is similar to other movies that the sample user had previously rated. The similar movies used in this influence style explanation (ISE) are movies which have received many similar ratings in common by many users. Figure 2 (Top) also shows, as the explanation for each recommended movie, the ratings given by the sample user to these similar movies. Note that a few of the other similar movies have not been previously rated by the sample user in question. Thus they are shown with no rating or a rating of zero.

Tables 7 and 8 shows the top-3 rated movies and recommended movies from the Matrix Factorization-based (MF) recommender system model, respectively for sample user B.

### TABLE 5

Top-3 Recommended Movies for Sample User B using *EMF*$_{UB}$. Movies are ranked in descending order of the predicted ratings.

<table>
<thead>
<tr>
<th>Top-3 recommended movies</th>
<th>Genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hedd Wyn (1992)</td>
<td>Drama/Romance</td>
</tr>
<tr>
<td>The Innocent Sleep (1995)</td>
<td>Thriller/Drama</td>
</tr>
<tr>
<td>Germinal (1993)</td>
<td>Drama/Romance</td>
</tr>
</tbody>
</table>

### TABLE 6

Top-3 Recommended Movies for Sample User B using *EMF*$_{IB}$. Movies are ranked in descending order of the predicted ratings.

<table>
<thead>
<tr>
<th>Top-3 recommended movies</th>
<th>Genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>Squeeze (1996)</td>
<td>Drama/Crime</td>
</tr>
<tr>
<td>Everest (1998)</td>
<td>Documentary/Short</td>
</tr>
</tbody>
</table>

Tables 5 and 6 show the top-3 recommended movies, using the *EMF*$_{UB}$ and *EMF*$_{IB}$ models respectively, for sample user B.
TABLE 7

Top-3 Movies rated by Sample User B (Ranked in descending order of ratings)

<table>
<thead>
<tr>
<th>Top-3 rated movies</th>
<th>Genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>French Twist (1995)</td>
<td>Comedy/Romance</td>
</tr>
<tr>
<td>Four Weddings and a Funeral (1994)</td>
<td>Romance/Comedy</td>
</tr>
</tbody>
</table>

TABLE 8

Top-3 Movies Recommended for Sample User B using Matrix Factorization Model (Ranked in descending order of predicted ratings)

<table>
<thead>
<tr>
<th>Top-3 recommended movies</th>
<th>Genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homage (1995)</td>
<td>Drama/Thriller</td>
</tr>
<tr>
<td>Sleepover (1995)</td>
<td>Drama</td>
</tr>
<tr>
<td>The Eighth Day (1996)</td>
<td>Drama/Comedy</td>
</tr>
</tbody>
</table>

Figures 3 and 4 show the corresponding NSE and ISE explanations for the top-3 recommended movies by $EMF_{UB}$ and $EMF_{IB}$ models for sample user B, respectively.
Figure 3: Neighborhood-rating Style Explanations for top-3 movies recommended to Sample User B using $EMF_{UB}$
In the NSE Explanation style, Figure 3 (top row) tells user B that one similar user has previously rated (lowly) the first recommended movie, using $EMF_{UB}$, while the third movie recommended by $EMF_{UB}$, had not been previously rated by any other similar users. Interestingly, none of the top-3 recommended movies have recorded a high rating by any of the users who are similar to user B. Hence, NSE style explanations do not justify the top-3 recommendations made by $EMF_{UB}$. In the ISE Explanation style, Figure 4 shows
how sample user B has previously rated several movies that are considered similar to the top movie, recommended using $EMF_{IB}$.

For the two sample users A and B, we observe that the EMF models succeeded to provide both recommendations and explanations for the recommended movies. However, a major limitation of these methods is that recommendations for each method can only be explained using one explanation style. This limitation becomes more obvious when the one available explanation style is not sufficient for explaining the recommended movie to the user, as shown in Figure 3 (Top Row) and Figure 4 (Bottom Row), where some of the explanations were not sufficient for explaining some of the recommended movies to the sample users. Recall that the user will eventually see only one list of top recommendations and each recommended movie in this list will be accompanied by at most one explanation style, specifically, NSE for $EMF_{UB}$ or ISE only for $EMF_{IB}$. Furthermore, some recommended movies will not have high explanation scores, and hence cannot be explained by a single style.

For all these reasons, this work introduces techniques for generating item recommendations that can be explained using either multiple explanation styles or by selecting the explanation style that reflects the target user’s explanation style preference, or some other criteria to be designed in Chapter 3. This work significantly expands the explainability proposed within latent factor models in [4] by designing new algorithms that can leverage not one, but multiple explanation sources and styles. The advantages of the proposed multiple-style explanation methods include an expanded explanation coverage (i.e., how many recommendations can be explained) without sacrificing model accuracy.

1.2 Problem Statement

The research questions that we attempt to answer are as follows: (1) Can we build a recommender system based on Matrix Factorization (MF) [6], that in addition to maintaining high accuracy, is able to explain the recommended items to users using multiple explanation styles? and (2) Will the proposed Multiple style explainability method provide
more explainability coverage than previously proposed single explainability style methods.

1.3 Rationale

This dissertation focuses on Matrix Factorization methods and on leveraging explainability graphs that can be generated from different Explainable Matrix Factorization (EMF) methods [4] [14] [3] [11] [15] [12] which can be used to explain the recommended items to users. We are motivated by the fact that given the same recommended item, users may have different preferences for different explanation styles that help them best understand why an item was recommended to them by an automated recommender system. To verify this variety in user preferences, we surveyed users to choose the best out of four explanation styles for each of three movies chosen from the MovieLens data set. 19 respondents provided anonymous feedback on their preferred explanation style. All users were shown the same three movies’ poster images and asked to select, from a list of four explanation styles, which one they preferred to examine if they wished to understand why a movie was recommended to them, regardless of how much they liked the movie. Each of the four explanation styles, shown to the users, have been previously validated in prior user studies. The results of this survey, presented in Chapter 3, showed that (1) user preferences varied among both users and items; and (2) no single explanation style emerged as a clear winner.

1.4 Research Contributions

Our research is motivated by (1) the need for explainability without sacrificing the accuracy of the state of the art latent factor-based recommendation methods, and (2) the need for accommodating different styles of explanations, motivated by our preliminary survey. Within this context, our research contributes to the recommender system and explainable Artificial Intelligence (AI) fields by proposing the following methods to expand the range of explainability styles within latent factor based recommendation methods:

1. A weighted hybridization technique for combining multiple EMF models to improve
the performance of the recommender system and explaining recommendations using multiple styles. This approach will be called Weighted Multi-style Explainable Matrix Factorization (W-MEMF).

2. A regression-based approach for hybridizing multiple EMF models to improve the performance of the recommender system and explaining recommendations using multiple styles. This approach will be called Regression-based EMF (R-MEMF).

3. A selection-based approach for hybridizing multiple EMF models to improve the performance of the recommender system and explaining recommendations using multiple styles. This approach will be called Regression-based EMF (R-MEMF).


5. A Tag-Aware Explainable Matrix Factorization (TA-EMF) method that utilizes user-generated tags to improve the performance of the MF model.

6. A tag-boosted Explainable Matrix Factorization method that can be used to improve the performance of EMF models and provide tag-based explanations and other explanation styles to users.

In addition, we evaluate the proposed algorithms with respect to both recommendation accuracy or relevance and explainability using standard metrics and public benchmarking data sets from two different domains, namely the movie domain and the book domain. Figure 5 situates our research contributions, relative to several relevant methods.

Our research findings shed new light on the performance of new mechanisms that we have investigated to design various hybridizations of explainable models. Although, rooted in the hybrid recommendation framework, our proposed methods make a significant step forward both relative to explainable AI and in relation to hybrid recommender systems, because the proposed hybridization mechanisms make an intentional effort to take into
account the individual models’ explanations and not only their output predicted ratings. Table 9 shows the summary of our research contributions.

Figure 5: Contributions of this work shown within the greater taxonomy of Matrix Factorization methods for Recommender Systems (RS). New techniques are shown with a bold outline.

1.5 Document Organization

The rest of this dissertation is organized as follows: Chapter 2 reviews the related work. Chapter 3 presents our proposed Multi-Style Explainable methods. Chapter 4 presents our evaluation experiments. Finally, Section 5 presents our conclusions.
### TABLE 9

Summary of Research Contributions

<table>
<thead>
<tr>
<th>Framework Component</th>
<th>Goal</th>
<th>Methodology</th>
<th>Validation Metrics</th>
<th>Evaluation Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Style</td>
<td>Generate personalized tag-based explanation</td>
<td>Tag-Assisted Explainable Matrix Factorization (Sec. 3.3)</td>
<td>Accuracy and Explainability Metrics e.g. RMSE, NDCG, MAP, MEP, MER, coverage</td>
<td>Sec. 4.3</td>
</tr>
<tr>
<td>Multi-style</td>
<td>Generate personalized multi-style explanations</td>
<td>Hybrid Multi-style Methods (Sec. 3.2)</td>
<td>Tag-boosting Multi-style Methods (Sec 3.4)</td>
<td>Sec 4.2</td>
</tr>
</tbody>
</table>

15
CHAPTER 2

BACKGROUND AND LITERATURE REVIEW

A surge in the usage of the internet by both individuals and corporations has given rise to a situation where users have access to more information than they can ever use or need. This in itself is a problem for marketers and e-commerce companies who have items they want users to be able to see. Using traditional database query methods based on popularity of the product ensures that only popular items will be seen and this comes with its own disadvantage; the marketer will miss out on customers who are not interested in popular items. There is a need for a way for customers to be able to get personalized results based on their interests and peculiar needs. There is also a need for e-commerce websites to be able to market items to the right users who will most likely purchase the items advertised. With this mutual need for a system that is both beneficial to the customers and the marketers, recommender systems have provided a way for users not to be overwhelmed by the vast amount of content they see online, while ensuring that the personalization experience is also an enjoyable one for users. [17] state that “Collaborative Filtering recommender system methods produce user specific recommendations of items based on patterns of ratings or usage without the need for exogenous information about either items or users”. Initial interest in recommender systems was in the domains of e-commerce and entertainment. However recommender systems have recently spread to almost every domain of human interaction online.

In this chapter, we review Collaborative Filtering recommendation techniques and recent efforts to add explanations to their outputs. We also review hybrid recommender systems.
2.1 Collaborative Filtering (CF)

Collaborative Filtering (CF) focuses more on the recorded interaction between the user and item than the description of the item [18]. The recorded interaction between users and items can be done explicitly or implicitly. Explicit interactions can be recorded in terms of ratings or likes given by the user to the item. CF is widely implemented and considered to be the most popular and powerful family of techniques used in recommender systems [19].

There are two main methods commonly used when using CF: memory-based/neighborhood methods and model-based methods [20]. Recommendations from memory-based methods are easy to interpret but might not be the most accurate; while recommendations from model-based methods are difficult to interpret but are highly accurate [20]

2.1.1 Memory-based Methods

Memory-based/neighborhood-based methods are a category of CF which use known user ratings for items directly when predicting users’ ratings for previously unseen items [20]. Memory-based methods work like word-of-mouth by making item recommendation for users based on the opinions of users similar to the target user. A user’s neighbors are very similar to the user in terms of how they rate items and their ratings tend to be highly correlated to that of the target user. Memory-based methods strengths include [20]:

- **Simplicity:** they are relatively simple to implement
- **Justifiability:** their recommendations can be easily understood and interpreted
- **Efficiency:** they do not require a training phase and pre-computing the nearest neighbors can be done offline and updated regularly. This is especially useful for large applications with many users and items
- **Stability:** only newly added data needs to be added to previously known data to update information about neighbors and recommendations. Therefore, this saves time because the whole process is not repeated for all user-item pairs when new data is added.
A limitation to automating the word-of-mouth process is that users will only ever get to see items that their neighbors have rated; therefore, many of items will go unseen. This problem of items going unseen is referred to as the problem of limited coverage which is a limitation of memory-based recommender systems. Another limitation of these methods is the cold start problem which occurs when a new user or item is added to the recommender system. In order for an item to get recommended, it must have been rated by at least one user and in order to discover the neighbors of a user, we have to know the user’s preference; but new users and items have none of this information and thus cannot receive recommendations or get recommended, respectively.

We explore the different neighborhood-styles in the following sections:

### 2.1.1.1 User-based Rating Prediction

We predict the rating of $r_{ui}$ for user $u$ for item $i$ by considering the k-nearest neighbors [21] of $u$ represented as $N(u)$. The k-nearest neighbors are the top-k users with the highest similarity to the user $u$ or the users who rate items in a very similar way to the user. We then select a subset of $N(u)$ who have rated $i$ and this subset of users is denoted as $N_i(u)$. Therefore $r_{ui}$ can be calculated as the average rating given to $i$ by these neighbors [20] as shown below:

$$
\hat{r}_{ui} = \frac{1}{|N_i(u)|} \sum_{v \in N_i(u)} r_{vi}
$$

Equation (1) has a major limitation which is the assumption that all users in $N_i(u)$ should be given the same level of similarity. We can use normalized weights to give higher importance to users more similar to the user $u$ and less weights to give lower importance to users less similar to the users $u$ as shown:

$$
\hat{r}_{ui} = \frac{\sum_{v \in N_i(u)} w_{uv}r_{vi}}{\sum_{v \in N_i(u)} |w_{uv}|}
$$

where $w_{uv}$ is the normalized weight for a neighbor $v$ and the absolute value of this weight is used in the denominator to ensure that negative weights do not produce ratings outside
the accepted range.

2.1.1.2 Item-based Recommendations

Unlike user-based methods that use information about similar users, item-based methods use rating information given to similar items and quite similar to user-based predictions, item-based recommendations for a user $u$ can be calculated as follows:

$$\hat{r}_{ui} = \frac{\sum_{j \in N_u(i)} w_{ij} r_{uj}}{\sum_{j \in N_u(i)} |w_{ij}|}$$

(3)

Where $N_u(i)$ are the items rated by user $u$ which are most similar to $i$ and $w_{ij}$ is the normalized weight of item $j$ which is considered to be similar to item $i$.

2.1.2 Model-based Methods

Model-based methods use the rating data to estimate or learn a model to make predictions [22] for users. These methods are usually more robust and better equipped to handle the sparsity problem which is a major limitation of the neighborhood method. They also tend to be better predictors than memory-based methods.

Examples of these methods include Bayesian belief nets CF [23] which are used for classification tasks provided the input data is categorical or Latent Semantic CF [24] uses a statistical modeling that discovers similar users or groups or communities by introducing hidden class variables in a mixture model setting [22]. In this research, we focus on CF model methods based on Matrix Factorization techniques which also ensure dimensionality reduction. Dimensionality reduction helps to tackle the problem of sparsity caused by sparse information in a high dimensional space by projecting the high-dimensional space into a lower dimensional space [25]. “Matrix Factorization techniques are a subset of model-based CF methods that assume that the similarity between users and items is induced by some factors hidden in the data” [26].
2.1.2.1 Principal Component Analysis (PCA)

PCA is a statistical dimensionality reduction method that can be used with large datasets that can be used to ensure minimal data loss [27]. PCA performs dimensionality reduction by constructing new features out of the original features of the dataset. The newly constructed features that retain the most information are considered to be the principal components of the dataset. The features that retain the most information are usually the features that have larger variability. The features with the least variability are discarded. These new features are usually uncorrelated with each other. The component that captures the most information will have a larger amount of variance than the next most informative component and so on. Dimensionality reduction is therefore achieved by neglecting the components with negligible contributions to the variance. These new features are called the Principal Components. PCA is usually applicable when the dataset is mostly Gaussian [14].

2.1.2.2 Singular Value Decomposition (SVD)

SVD is also a powerful technique for dimensionality reduction. The main problem it addresses is finding a feature space of lower dimensionality where the new features represent concepts for which we can also compute the strength of these concepts in the context of the collection [25]. SVD’s ability to derive semantic concepts in a space of low dimensionality made it an easy tool for Latent Semantic Analysis (LSA), which is a popular technique for text classification in the field of Information Retrieval.

The SVD algorithm is based on the theorem that it is always possible to decompose a given matrix $A^{m \times n}$ into $A = U \lambda V^T$ where $n$ is number of items and $m$ is the number of features, $U = U^{n \times r}$, $V = V^{m \times r}$ and $\lambda = \lambda^{r \times r}$. $\lambda$ is a diagonal-matrix that represents the strength of each concept that contains the singular values or eigenvalues, which are positive values sorted in decreasing order of magnitude. $U$ is called an item to concept matrix and $V$ is called a feature to concept matrix.

The get an approximation of the original matrix $A$, we truncate the singular values matrix
λ at row $i$ resulting in $A_i = U_i\lambda_i V_i^T$ which is the closest rank-$i$ matrix to $A$. $U$ is equivalent to the eigenvectors of $AA^T$ where $A^T A = V\lambda U^T U\lambda V^T = V\lambda^2 V^T$

2.1.2.3 Matrix Factorization (MF)

Matrix Factorization (MF) is a group of techniques, including SVD, which try to represent users and items using vectors of features which were derived from the ratings given by users to items they have interacted with in such a way that a strong association between user and item latent factors will result in a recommendation [26]. Let’s assume we have a set of users $U$ and a set of items $I$. The rating matrix is represented as $R$ which is a matrix that holds ratings given by users to items. Matrix factorization aims to factorize the rating matrix $R$ into 2 matrices $P$ and $Q$ such that the product of $P$ and $Q$ is approximately equal to the original rating matrix $R$ as shown below:

$$R \approx P \times Q^T$$

(4)

Where each row of $P$ represents the strength of the association between a corresponding user and features and each column of $Q$ represents the strength of the association between the corresponding item and features. The goal of MF is to find the least number of latent features we can use when factorizing our rating matrix $R$ such that the product of the resultant matrices $P$ and $Q^T$ will be approximately equal to the original rating matrix. The prediction score for user $u$ for item $i$ is calculated as follows:

$$\hat{r}_{ui} = p_u q_i^T$$

(5)

where $p_u$ is a vector representing the strength of the association between the user $u$ and the selected latent features and $q_i$ is a vector representing the strength of the association between the item $i$ and the selected latent vectors.

MF, when used in the context of recommender systems, can be seen as a minimization problem whose objective is to minimize the following objective function over known ratings:
\[ J = \sum_{u,i \in R} (r_{ui} - p_u q_i^T)^2 + \beta(\|p_u\|^2 + \|q_i\|^2) \] (6)

where a regularization term is used for both terms to avoid overfitting the training data while \(\beta\) is a coefficient for ensuring the smoothness of the newly added term. \(J\) is only convex with respect to \(p\) or \(q\) alone, but not to all other unknown parameters. Stochastic Gradient Descent (SGD) can be used to solve for the optimal parameters where the update rules for the user and item latent factor parameters \(p\) and \(q\) are given by:

\[ p_u^{(t+1)} \leftarrow p_u^{(t)} + \alpha(2(R_{u,i} - p_u^{(t)} q_i^{(t)}^T)q_i^{(t)} - \beta p_u^{(t)}) \] (7)

\[ q_i^{(t+1)} \leftarrow q_i^{(t)} + \alpha(2(R_{u,i} - p_u^{(t)} q_i^{(t)}^T)p_u^{(t)} - \beta q_i^{(t)}) \] (8)

An extension of Matrix Factorization is Non-Negative Matrix Factorization (NMF) [28] which can be used for learning parts of data or semantic features of textual information. Semantic features could be concepts or topics in the text domain; or in the movie domain, they could be similar genres, actors, producers, etc. This approach enforces a constrain that uses only non-negative values in the factorized matrices and this ensures that only additive combinations are allowed. This method was initially implemented in a Neural Network on a facial images database and it was found to perform better than PCA because this approach learnt features rather than whole images and was able to generalize better than the methods that learnt whole images.

2.1.2.4 Joint Matrix Factorization (JMF)

JMF is an extension of MF that is designed for merging data sources for matrix factorization. [29] used this approach to merge the rating matrix and information about users’ group behaviors to ensure the trustworthiness of the ratings provided by users. This method was used to ensure that malicious ratings were not used for recommendation generation. [30] incorporated both known ratings and information from other sources: information about the
user and information about the item. Information about the user included demographic information which was not always available due to privacy concerns while information about the item, which were movies in this instance, included the genre, actors, etc. Their approach was used to improve the recommendation accuracy of the MF-based recommender system. [31] proposed a novel movie similarity measure that was exploited by a JMF model for generating context-aware movie recommendations that took the movie mood into consideration.

### 2.1.2.5 Probabilistic Matrix Factorization (PMF)

PMF is an extension of MF that is designed to scale linearly with the number of observations and is especially suited for large, sparse and imbalanced data sets [32]. An extension to this model was made by adding a constraint that is based on the assumption that users who have rated the same set of items have similar preferences. The objective function for this method is given as follows:

\[
E = \frac{1}{2} \sum_{u=1}^{N} \sum_{i=1}^{M} I_{ui}(R_{ui} - U^T_u V_i)^2 + \frac{\lambda_U}{2} \sum_{i=1}^{N} \| U_u \|^2 + \frac{\lambda_V}{2} \| V_i \|^2
\]  

Where \( I_{ui} \) is the indicator function that is equal to 1 if the user \( u \) has rated item \( i \) and 0 otherwise, \( \lambda_U = \frac{\sigma^2}{\sigma_U^2} \), \( \lambda_V = \frac{\sigma^2}{\sigma_V^2} \) and \( \| . \| \) is the Frobenius norm. A local minimum can be found by using SGD on \( U \) and \( V \). In order to restrict the predicted value within the required range, a logistic function is used where \( g(x) = \frac{1}{1+e^{-x}} \).

The main advantage of this approach is its probabilistic approach to training the model by "finding only the point estimates of model parameters and hyper-parameters instead of inferring the full posterior distribution over them". Due to the use of only a subset of the data, this method is faster and computationally less expensive. [33] proposed the use of PMF for a factor analysis approach to solve the sparsity and low predictive accuracy limitations of basic MF by utilizing both users’ social information and ratings records.
2.1.2.6 Auto-encoders

Auto-encoders have also been used to mimic CF-based recommender systems [34] by exploiting information about the user’s preferences for items, expressed as ratings, to provide personalized recommendations. Each user and item is represented by a partially observed vector. [34] designed an item-based auto-encoder which accepts the partially observed vector for each user as input. These partially observed vectors are then projected into a low-dimensionality latent space, represented by the hidden units of the auto-encoder and then reconstructs the output in the output space. The output is a fully filled vector for each user representing the user’s expected ratings for each item. [34] proposed two methods: user-based auto-rec and item-based auto-rec and both methods performed better than the baselines they were compared to. Their experiments showed that the model’s prediction accuracy improved with increased numbers of hidden units with the best performance recorded when using 500 hidden units. However, a major limitation of this approach is that the recommendations are not explainable to the users. [35] proposed integrating Knowledge Graphs (KG) into the autoencoder model learning process in order to build models with explainable recommendations using KGs. [36] proposed integrating explainability graphs with pre-calculated explainability scores [4], into the model building process in order to generate explainable recommendations that can be presented using different visual explanation styles.

2.1.3 Learning Algorithms for Matrix Factorization

There are two main approaches used for minimizing the objective function given in Equation 6. These two approaches are briefly discussed as follows:

2.1.3.1 Stochastic Gradient Descent

This approach loops through all ratings in the training set and computes the prediction error for each user-item pair in the training data. The training prediction error for each
user-item pair \((e_{ui})\) is given as:

\[
e_{ui} = r_{ui} - p_u^T q_i^T
\]  \hspace{1cm} (10)

The algorithm then modifies the parameters in the opposite direction of the gradient by a quantity proportional to the learning rate \((\alpha)\). This approach’s popularity is due to its ease of implementation as well as it relatively fast running time \([6]\). The update methods for this approach are given by:

\[
q_i \leftarrow q_i + \alpha (e_{ui} p_u - \beta q_i)
\]  \hspace{1cm} (11)

\[
p_u \leftarrow p_u + \alpha (e_{ui} q_i - \beta p_u)
\]  \hspace{1cm} (12)

### 2.1.3.2 Alternating Least Squares

The fact that the user latent factor \(p_u\) and the item latent factor \(q_i\) are unknown makes the objective function of MF a non-convex function. However, the Alternating Least Squares approach seeks to change the problem into a quadratic one by fixing the values of \(p_u\)’s and recomputing the \(q_i\)’s by solving a least squares problem and vice-versa. The objective of this approach is to ensure that each step moves closer to the minimization of the objective function shown in Equation 6 until convergence is attained. This approach is especially useful in two main scenarios: parallelization and when the input data is implicit \([6]\).

### 2.2 Explanations in Recommender Systems

#### 2.2.1 Motivation

Although MF-based recommender systems have high predictive accuracy, they still work as black-box models which do not give the reasoning behind their suggestions to users. \([7]\) observed from their research that it was becoming apparent that accuracy measures alone are no longer sufficient for evaluating a recommender system. \([37]\) noted that good explanations could help inspire user trust and loyalty, among other desiderata which will
be discussed later on in this section. [2] observed that good explanations could also help users make more accurate decisions. Explanations help detect or estimate the likelihood of errors in recommendations. [8] identified two main sources of errors in recommender systems which are: model/process error and data error. Model/process errors usually arise when a recommender system uses a process to compute recommendations that do not match the user’s requirements e.g recommending a movie to the user without considering the season or time of the day. Data error on the other hand usually stems from inadequacies of the data used, and might be a result of insufficient data, bad data or high variance in the data.

In addition to trust building and error detection, explanations can also be used to justify recommendations and direct acceptance. Justification, in most cases is decoupled from the recommendation algorithm and can be descriptive [1]. Justification gives the user some insight into the reasoning behind the recommendation process, without fully exposing the underlying algorithm, which can help the user decide how much confidence to put in the output [14].

Some perceived benefits of explaining recommendations to users are summarized by [37] as follows:

- **Increased transparency and scrutability:** When users understand, in part or totally, what data was used or how their data was used to generate recommendations, the users are able to correct the recommender system if it is wrong and be more willing to accept the recommendations

- **Trust:** Trust was defined as the perceived confidence in a recommender system’s competence and user studies run by [38] showed that users returned to recommender systems they perceive to be trustworthy. However, trust may also be encouraged by a recommender system with higher accuracy and the interface may also be a factor that increases trust.

- **Persuasiveness:** Recommender systems that are trusted by users could also be able to persuade users to buy or try more items. In some systems where the metric
for success is time spent on the system, good explanations for recommending relevant content to the user, will persuade the user to spend more time.

- **Efficiency:** good explanations will help a user be more efficient on the system by helping the user to make faster and better decisions

### 2.2.2 Explanations Styles in Recommender Systems

Explanations have been used to justify recommendation by both earlier and more recent recommender systems. The choice of how to present these explanations to the users vary widely and depends on the domain, content, items, and the users among other factors. [1] stated that even though many implementations of explanations exist, their main objective still remains to show how a recommended item relates to a user’s preferences. A common technique for establishing the relationship between the user and the recommended item is the use of **intermediary entities** shown in Figure (6).

![Figure 6: Intermediate entities commonly used in explaining recommendation to users.](image)

Explanations of recommendations fall into one of three categories or Explanation Styles [2] described below:

- **Keyword Style Explanation (KSE):** This is an approach or explanation style used for explaining content-based recommendations. It seeks to answer the question “What is it about this item that speaks to my interest?” [1] called this explanation style feature-based explanation and described it as an approach that uses qualities or characteristics
of the recommended item as the intermediary entities. This explanation style helps users to make more accurate decisions [2]. An example of this explanation style is shown in Figure 7

<table>
<thead>
<tr>
<th>Slot</th>
<th>Word</th>
<th>Count</th>
<th>Strength</th>
<th>Explain</th>
</tr>
</thead>
<tbody>
<tr>
<td>DESCRIPTION</td>
<td>HEART</td>
<td>2</td>
<td>94.14</td>
<td>Explain</td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>BEAUTIFUL</td>
<td>1</td>
<td>17.07</td>
<td>Explain</td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>MOTHER</td>
<td>3</td>
<td>11.55</td>
<td>Explain</td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>READ</td>
<td>14</td>
<td>10.63</td>
<td>Explain</td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>STORY</td>
<td>16</td>
<td>9.12</td>
<td>Explain</td>
</tr>
</tbody>
</table>

Figure 7: The Keyword Explanation Style. Source: [2]

• Neighbor Style Explanation (NSE): This approach is used for CF-based RSs where users might be interested in knowing how users similar to themselves (neighbors) have rated the recommended item. Similar users’ ratings are grouped into good (4-5), neutral (3) or bad (1-2) groups. [1] further broke this explanation style into user-based explanations and item-based explanations. User-based explanations use other users as the intermediary entities. This approach was found to be more successful at persuading users to try an item but did not help them become efficient [8]. The item-based explanations used items as the intermediary entities. This approach helped users to make accurate decisions (efficient) and also improved the acceptance of the recommendations but often times users could not understand the relationship between the explaining items and the recommended item [39]. An example of this explanation style is shown in Figure 8.

• Influence Style Explanation (ISE): This approach presents a user with a list of the training items that had the most influence in deciding the recommendation of the item to the user. This explanation style allows users to understand how their data or previous inputs influence their recommendations.
Other explanation styles have been utilized across different domains using different representational styles. One of the most common ways to represent explanations is through the use of tags.

[1] introduced Tagsplanation which modified the KSE mentioned earlier and instead used tags to explain the recommended items. They used extra information about how the user has interacted with the tags (tag preference) and how the frequently the tag has been used to describe the item (tag relevance) to develop an explanation style based on tags.

[40] developed another tag-based explanation style which used tag clouds to explain recommendations to users. Their result contradicted [1] because they found that user’s recorded increased satisfaction using both personalized and non-personalized Tag Clouds. Researchers have used other entities such as context to explain recommendations as done by [41]. Their approach used contextual information such as “usage scenarios” and “accompanying persons” as explanations for recommendations. They proposed a new explanation style they coined as the context style explanation whose format was “Restaurant X is recommended to you because it is suitable for dates with your boyfriend/girlfriend”. The key
contextual factors they used were time, location, companion and purpose. A major limitation to their method is that their explanation style was only tested on a restaurant dataset and hence might not be effective for explaining recommendations in other domains.

[15] [11] proposed a new explanation style called the Inferred Fact Style Explanation. This explanation style uses knowledge graphs on users’ and semantic attributes. The uncertainty degree of the users’ preferences for semantic attributes is employed to justify a recommendation. Inference rules are used to derive new knowledge from known facts.

In addition to the explanations styles previously discussed, [42] defined some other explanation styles listed below:

- **Collaborative-Based Style Explanations**: These recommender engines have the user $u$’s ratings for items in $I$ as the assumed inputs. These are used with CF-based recommender systems where a user’s ratings for an item are used to find similar users called neighbors. The prediction for the recommended item is extrapolated from the neighbors’ ratings of the item. In most commercial systems, a popular implementation of this style is in the form "Customers who bought this item also bought...".

- **Content-Based Style Explanations**: These recommender engines have the user $u$’s ratings for some of the items in $I$ as the assumed inputs. Content-based explanation style is based on the items’ properties. This explanation style acts as a hybrid between CF and Content-based methods because the explanation style suggests that they compute the similarity between the features of highly rated items. An example of this style is Tagsplanation [1].

- **Case-Based Reasoning (CBR) Style Explanation**: This approach is quite similar to the Influence Style Explanation previously discussed where the explanation omits the features of the recommended item but instead focuses on similar items used to make the recommendation. The influence of an item is computed by looking at the difference in the recommendation with and without the item.

- **Knowledge and Utility-Based Style Explanation**: The assumed input to the recom-
mender engine are descriptions of user \( u \)'s needs or interests. The recommender engine then infers a match between the item \( i \) and the user \( u \)'s needs. This explanation style reflects how the user's needs are served by the item and uses these features that both meet the needs of the user and those that do not as the explanation for the recommendation.

- **Demographic Style Explanation:** The assumed input to the recommender engine is demographic information about user \( u \). Recommendations are therefore made by finding users with similar demographic information to the user \( u \). A recommended item \( i \) is extrapolated from how the similar users to user \( u \) rated item \( i \). Demographic explanations will be similar to ”Movie X was recommended to you because you are female and aged 20-40” but such an explanation might be considered too sensitive and not easily accepted by users.

Explanations have traditionally been implemented after the recommendation has been made. [43] proposed a personalized hybrid explanation for recommendations with is independent of the recommendation technique and combines basic explanation styles in order to provide the appropriate personalized explanation to users. They achieved this by building a black-box explanation system that uses the users’ preferences to generate content-based and collaborative-filtering-based style explanations which is independent of the recommender system technique used to generate the recommendations. Initially, all explanations are assigned the same weight but explanation weights get adjusted over time as the user’s explanation preferences are learnt. A limitation to this model is that it uses another black-box to provide explanations to the recommendations of the initial black-box which does not make it more transparent.

Framework ExpLOD [44] provides explanations by exploiting the information available in the Linked Open Data (LOD) cloud to generate natural language explanations for recommendations. Explanations are generated by combining features of movies a user has previously indicated as preferred and recommended items into a knowledge graph. The knowledge graph is constructed by using the most important features of the movies the
user has previously liked to explore the recommended movies to detect the occurrence of these features. Features that are present in more previously liked items are ranked higher than those that occur less. This ranked list of features are then presented to the user as the explanation for the movie. This approach was found to be accurate and also offered more interesting explanations. However, recommended items that have no information in the LOD cloud dataset will be at a disadvantage when using this method.

2.2.3 Using Tags in Collaborative Filtering Recommender Systems

[45] used tags to improve the performance of memory-based CF recommender system by proposing the integration of tags when building the user profile and estimating item similarity. To build a user’s profile, they considered the tags used by the user, items tagged by the user and the relationship between tags and the tagged items. Users neighbors are found by measuring the similarity between the tags used by a user and a prospective neighbor, the similarity between the users based on their ratings and the similarity between the user’s tag-item relationship. These terms are then summed up with each term having a weight. The weights all sum up to 1. They also proposed an item-based approach based on a similar weighted sum.

[46] proposed the use of tags in a recipe recommender system based on the user’s preferences for the ingredients in the recipe, represented by tags. They elicited the users for their preferred ingredients and their corresponding interest in the ingredient using a rating system. They created additional features vectors for the users and recipe tags which were then used in the prediction process.

[47] proposed an extended tag-induced MF technique for recommender systems which ”exploits correlations among tags derived the co-occurrence of tags to improve the performance of the RS even when the tag information is sparse”. They used the coupled similarity between tags to extend each item’s tags thus dealing with the tag sparsity problem. The item similarity, based on the extended tags, was then used as a regularization term for MF.
2.2.4 Explainable Latent Factor Methods

Latent Factor methods like Matrix Factorization (MF) are commonly used for recommender systems due to its high predictive accuracy. MF is also comparatively easy to extend and can easily be modified to use extra information in the recommendation process. Researchers have extended MF using tags [47] [45], content [48], contextual information [49]. These methods have mostly been aimed at improving the predictive accuracy of Matrix Factorization-based recommendation engines. Recently, user studies have found that higher accuracy will not improve the recommendation experience for users [7]. Explanations have been found to improve the experience by increasing the transparency of the black box recommender engine which improve trust, acceptance and effectiveness [39]. In this section, we discuss some approaches to improving the transparency of MF using explanations.

2.2.4.1 Explainable Matrix Factorization (EMF) Methods

[4] [3] proposed an extension to basic Matrix Factorization [6] that adds soft explainability constraints to bring users closer to items that considered explainable to such users in the latent space. The objective function for this method is shown in equation 13 where $\beta$ is the L2 regularization coefficient that controls the complexity of the Matrix Factorization model and $\lambda$ is the explainability term coefficient that controls the influence of the explainability constraint. $p_u$ is user $u$’s latent feature vector, $q_i$ is item $i$’s latent feature vector, $R^{TRAIN}$ is the set of user-item pairs with known ratings in the training set and $W_{u,i}$ is the explanation score for item $i$ for user $u$. The explanation score represents the strength of the explanation generated using the EMF method of item $i$ for user $u$. 

Figure 9: (a) Typical Explanation Style, (b) Explainable Matrix Factorization. Source: [4]
\[
\min J = \sum_{(u,i) \in R^{TRAIN}} (r_{ui} - p_u q_i^T)^2 + \frac{\beta}{2}(\|p_u\|^2 + \|q_i\|^2) + \frac{\lambda}{2}\|p_u - q_i\|^2 W_{ui}
\tag{13}
\]

[4] [3] proposed two methods for estimating the explanation score: **User-based EMF** and **Item-based EMF**. Both methods use a Neighborhood-rating technique for determining the fraction of a users similar to the target user or item. User-based EMF uses a **Neighborhood Style Explanation**, denoted as \(W\), which estimates the fraction of users most similar to user \(u\) who have rated the target item \(i\). Item-based EMF uses a **Influence Style Explanation**, denoted as \(E\), to estimate the fraction of items most similar to the target item, \(i\), the user had previously rated. These methods for calculating the explanation score proposed by [4] [3] are shown in Equation 14 where \(N_i\) is the set of items most similar to the target item \(i\), \(N_u\) is the set of similar users or neighbors of the user \(u\), \(I_u\) is the set of items rated by user \(u\) and \(U_i\) is the set of users who have rated item \(i\). The relationship between users and items based on the explanation score can be represented as a graph or an explainability graph. Therefore, different explanation styles will result in different explainability graphs. \(\theta\) is the threshold, for a selected explainability graph, beyond which an item is considered to be highly explainable to a user. \(\theta^n\) is the threshold using the NSE explainability graph while \(\theta^i\) is the threshold using the ISE explainability graph.

\[
W_{ui} = \begin{cases} 
\frac{|N_u \cap U_i|}{|N_u|} \cdot \frac{|N_u \cap U_i|}{|N_u|} \geq \theta^n \\
0, otherwise 
\end{cases}
\tag{14}
\]

\[
E_{ui} = \begin{cases} 
\frac{|N_i \cap I_u|}{|N_i|} \cdot \frac{|N_i \cap I_u|}{|N_i|} \geq \theta^i \\
0, otherwise 
\end{cases}
\]

[11] proposed semantic-aware EMF which utilizes semantic Knowledge Graphs (KGs) to provide explanations for recommended items to users. The objective function for this proposed method is shown in eq. 15.
\[
\min J = \sum_{(u, i) \in R_{TRAIN}} (r_{ui} - p_u q_i^T)^2 + \frac{\lambda}{2} \sum_{i, j \in S_{LDSD}} (S_{LDSD}^{i,j} - q_i q_j^T)^2 + \frac{\beta}{2} (\|p_u\|^2 + \|q_i\|^2) \quad (15)
\]

Where \(S_{LDSD}^{i,j}\) is an item-item similarity matrix based on the semantic distance between the items \(i, j\). The explanation scores for this model were based on semantic Knowledge Graphs as shown in eq. 16

\[
S_{UI}^{u,i} = \begin{cases} 
S_{Uf,u} S_{If,i} & \text{if } S_{Uf,u} S_{If,i} > \theta_s \\
0 & \text{otherwise} 
\end{cases} \quad (16)
\]

Where \(S_{Uf,u}\) is a user-based Knowledge Graph which stores the number of items with feature \(f\) liked by user \(u\). \(S_{If,i}\) is an item-based Knowledge Graph which with a value of 1 if item \(i\) possesses feature \(f\) and 0 otherwise. Finally, \(\theta_s\) is the threshold for items to be considered semantically explainable to users.

### 2.2.4.2 Other Explainable Latent Factor Methods

[50] proposed a method whose recommendations are considered to be simultaneously novel and explainable to the user. Their aim was to push items that haven’t been seen by the user but for which an explanation exists for the user into the top-N recommendation list with minimal loss in prediction accuracy. They achieve this by adding a novelty term to the objective function of EMF.

[51] proposed a method for explaining recommendation by integrating user sentiments from item reviews. They integrate two learning tasks, user preference modeling for recommendation and opinionated content modeling for explanation via a joint tensor factorization. The resultant method predicts not only as user’s preference over a list of items but also how the user will appreciate an item based on the features of the item.

### 2.2.5 Evaluating Explainability in Recommender Systems

Three metrics were proposed by [4] [14] [3] to evaluate recommendation explanations: MEP, MER, x-F score.
\[ MEP = \frac{1}{|U|} \sum_{u \in U} \frac{|R \cap W|}{|R|} \]  

(17)

\[ MER = \frac{1}{|U|} \sum_{u \in U} \frac{|R \cap W|}{|W|} \]  

(18)

\[ xF - score = 2 \times \frac{MEP \times MER}{MEP + MER} \]  

(19)

where \( U \) is the set of all users, \( R \) is the set of recommended items and \( W \) is the set of explainable items. MEP computes the ratio of simultaneously recommended and explainable items to the total number of recommended items across all users. MER calculates the ratio of the number of simultaneously recommended and explainable items across the total number of explainable items across all users.

2.3 Hybrid Recommender System Techniques

[52] defined hybrid recommender systems as systems that combine two or more recommender systems in order to capitalize on the strengths of the individual recommender systems. [52] identified several hybridization techniques in use in traditional recommender systems. Weighted hybrid recommender systems evaluate the final predicted rating as a weighted aggregate of the individual recommendation models predictions. An example of a weighted hybrid recommender system is the Personalized-Tango [53], a newspaper article recommender system, which used a weighted average of the recommendations from a content-based and collaborative filtering recommender system as the hybrid model’s predicted rating. Hybrid recommender systems that use the switching technique select the model that meets a predetermined criterion. [54] proposed a switching hybrid recommender system that switched to the model that produced predictions that were closely correlated to the user’s ratings. A limitation to this approach is the added complexity of selecting the criterion for switching which could be considered as another parameter for tuning. Mixed hybrid recommender systems present the predictions from multiple recommendation models simultaneously to the user. Personalized TV (PTV) [55] used this approach to develop a programme listing service based on collaborative Filtering and Content-based recommender
systems. The Feature Combination approach uses the collaborative model as a source of feature information that can be used to improve the performance of the content-based recommender system. This allows the system to use only the necessary information from the collaborative recommender system without being totally dependent on it. This reduced dependence on the collaborative recommender system reduces the sensitivity of the hybrid on the number of users who have previously rated items. Cascaded hybrids use a staged process wherein an initial recommender system produces a list of candidates to be considered for recommendation. The second step refines the results from the first step by using the results as the model’s input data. These hybrids are usually more robust and less sensitive to noisy input data.

2.4 Chapter Summary

In this chapter, we reviewed CF methods and black-box recommender systems, in particular, those built using MF. We discussed the benefits of explanations, different explanation styles and when to use them. We also reviewed explainable latent factor models and the way they provide insights into how recommendations were made. In the next chapter, we will introduce our proposed methods and how they build on and improve some existing methods.
CHAPTER 3

MULTIPLE EXPLANATION STYLES METHODS FOR MATRIX FACTORIZATION-BASED RECOMMENDER SYSTEMS

In this chapter, we propose a methodology to build Multiple style Explainable Matrix Factorization models by leveraging a hybrid recommendation framework. In Section 3.1, we discuss why Multi-style EMF models are needed and the intuition behind our proposed methods. In Section 3.2, we introduce our proposed hybrid methods, inspired by the taxonomy of hybrid recommender systems proposed by [52]. In Section 3.3, we describe our novel content-boosted approach which leverages user-generated tags to generate tag-based explanations, an explanation style that was previously validated in [1] and [48]. Finally, in Section 3.4, we conclude the chapter.
Figure 10: Contributions of this work shown within the greater taxonomy of Matrix Factorization methods. New techniques are shown with a bold outline.
3.1 Rationale for Multi-style Explainable Matrix Factorization

The EMF methods, discussed in the previous chapter, were shown to maintain the predictive accuracy of the MF models, while also providing explanations that could be used to make the MF model building process more transparent. However, these methods are restricted to utilizing a single explanation style. This restriction becomes undesirable when the explanation style generated by the model may not be appropriate for the item being explained to the user or is not easily understood by the user. Moreover, some explanation styles may be unavailable or impossible to generate for certain user-item pairs. For instance although tag-based explanations have been validated in [1], an item that is missing sufficient tags cannot be explained. We therefore propose a framework for Multi-style Explainable Matrix Factorization that can employ multiple explanation styles when presenting recommended items to the users. We achieve these multiple-style explanations by leveraging a hybrid recommendation framework. One particular hybrid approach is content-boosting which we use to integrate user-generated tags into the EMF model building process. Other hybridization techniques combine multiple EMF models to produce a hybrid model with access to different explanation styles, generated by each individual model.

3.1.1 Rationale for Multiple Explanation Styles

As a precursor, we surveyed users to indicate the best explanation style they preferred from each of three movies, from different genres, chosen from the MovieLens data set. 19 respondents provided anonymous feedback on their preferred explanation style. All users were shown the same three movie poster images and asked to select, from a list of four explanation styles, which one they preferred to examine if they wished to understand why a movie was recommended to them, regardless of how much they liked the movie. Each of the four explanation styles shown to the users have been previously validated in prior user studies. The explanation styles presented to the users were:

1. **Explanation Style 1**: Inferred Facts Explanation Style (semantic Knowledge Graphs) [11]
2. **Explanation Style 2:** Tagsplanation (tag-based) [1]

3. **Explanation Style 3:** Item-based Neighborhood Style Explanation (Neighborhood ratings by the user for similar items) [4] [3]

4. **Explanation Style 4:** User-based Neighborhood Style Explanation (Neighborhood ratings by similar users on the recommended item) [4] [3]

We selected movies from three different genres, which were randomly selected from a list of movie genres. The three genres selected were comedy, horror and sci-fi. The users were shown each movie poster and asked "Regardless of whether you like the recommended movie below, please focus only on the "explanations" for why it was recommended. Which of the following explanation styles would you prefer to see?".

![Figure a: Responses for Sci-fi movie "Inception"](image1)

![Figure b: Responses for Comedy movie "Coming to America"](image2)

![Figure c: Responses for Horror movie "Annabelle"](image3)
Figure 11: Responses showing the distribution of explanation style preferences of 19 users for three movies from three different genres

Fig. 11 shows that users have different preferences in explanation styles which help them understand why an item was recommended to them. Furthermore, it also shows that users’ preferences might change based on the genre of the recommended movie. This shows that there is a need for EMF models that can present users with multiple or personalized explanation styles to help users understand recommendations generated by these models.

Figure 12: Distribution of the number of distinct explanation styles selected by each respondent for the three recommended movies

Fig. 12 shows that users have different preferences when selecting the explanation style that helped them understand why each movie was recommended to them. 42% of the respondents chose two distinct Explanation styles as the most useful for explaining the three recommended movies; while 26% chose a different preferred explanation style for explaining each of the three movies. Only 31% of the respondents found a single explanation style to be adequate for explaining all three recommended movies. These observations motivate
us to develop methods that can accommodate diverse explanation styles. In the following sections, we new methods for generating multiple style explanations for EMF models.

### 3.1.2 Rationale for Pre-calculated Explainability Scores

Our proposed methods incorporate explainability scores from different styles that have been published and validated in the literature to determine which item is considered to be explainable to a user. These explainability scores depend on the explanation style chosen, and our work specifically studies multiple explanation styles. For this reason we use the explainability scores calculated using three different styles, namely user-based neighborhood style (NSE) [3], item-based neighborhood style (ISE) [3], and in case tags are available for a given domain, tag-based style explanation (TSE), to be defined later in this chapter.

The pre-calculated explainability scores, leveraged in our proposed methods, that have been validated in prior work [14] that conducted user studies that found a higher subjective perception of transparency among user-item pairs for items that have higher objective user-based (NSE) and item-based (ISE) neighborhood style explanation scores. The tag-based style explanation scores, are justified by the user studies of [1] that validated the user’s satisfaction with preference-based and relevance-based tag-based explanations which are the main inspiration and basis for our tag-based explainability scores. Unlike prior work that computes these scores post-hoc, i.e., after the predictions have been made, our tag-based explainable methodology calculates these tag-based explainability scores directly from available data, before the model is learned, and then uses them as part of a regularization mechanism, to guide the model learning. For this reason, our proposed methods are inherently more transparent than post-hoc methods, because we learn models that directly depend on explanations, whereas post-hoc methods learn models that are completely dissociated from the explanations.
Summary of notation used in this Chapter

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( U )</td>
<td>set of all users</td>
</tr>
<tr>
<td>( I )</td>
<td>set of all items</td>
</tr>
<tr>
<td>( \mathcal{T} )</td>
<td>set of all user-generated tags</td>
</tr>
<tr>
<td>( \gamma )</td>
<td>hybrid model’s hyper-parameter</td>
</tr>
<tr>
<td>( P )</td>
<td>user latent factor matrix</td>
</tr>
<tr>
<td>( p_u )</td>
<td>latent factor vector for user ( u )</td>
</tr>
<tr>
<td>( Q )</td>
<td>item latent factor matrix</td>
</tr>
<tr>
<td>( q_i )</td>
<td>latent factor vector for item ( i )</td>
</tr>
<tr>
<td>( \varphi(\cdot) )</td>
<td>Pearson Correlation function</td>
</tr>
<tr>
<td>( \hat{r}_{ui} )</td>
<td>predicted rating for user-item pair ((u,i))</td>
</tr>
<tr>
<td>( \hat{r}_{ui}^{\text{hybrid}} )</td>
<td>predicted rating for user-item ((u,i)) pair from a hybrid model</td>
</tr>
<tr>
<td>( \hat{r}_{ui}^{(n)} )</td>
<td>predicted ratings for user-item pair from the (n)(^{th}) individual EMF model</td>
</tr>
<tr>
<td>( W_{ui}^{(n)} )</td>
<td>explanation score for user-item pair ((u,i)) from the (n)(^{th}) individual EMF model</td>
</tr>
<tr>
<td>( T_{ui}^{\text{UI}} )</td>
<td>Tag-based explanation score for user-item pair ((u,i))</td>
</tr>
<tr>
<td>( \vec{T}_u^{\text{pref}} )</td>
<td>user-tag preference vector for user ( u )</td>
</tr>
<tr>
<td>( \vec{T}_i^{\text{rel}} )</td>
<td>item-tag relevance vector for item ( i )</td>
</tr>
<tr>
<td>( \theta )</td>
<td>explainability threshold</td>
</tr>
</tbody>
</table>

3.2 Hybridization Techniques for Multi-style Explainable Matrix Factorization

Hybrid recommender systems combine two or more recommendation techniques to gain better performance with fewer of the drawbacks of any individual one [52]. We designed new Multi-style Explainable MF models using hybridization techniques which will result in recommender systems that combine two or more EMF models to gain better performance with fewer of the drawbacks of any individual EMF model while also providing a framework for multiple explanation styles. Table 10 lists all the notation used in this chapter.
The predicted ratings, $\hat{r}_{ui}$, of a user $u$ for an item $i$, from $N$ individual EMF models, are distinguished by an index $n$ as shown in Eq. 20 where $n$ denotes the $n^{th}$ EMF model.

$$\hat{r}_{EMF}^{ui} = \{\hat{r}^{(n)}_{ui} \mid n \in \{1, \ldots, N\} \}$$

Similarly, the explanation scores, $W_{ui}$, for an item $i$ to user $u$, from the individual EMF models, are distinguished using index $n$ shown in Eq. 21, as a set of scores for different explanation styles.

$$W_{EMF}^{ui} = \{W_{ui}^{(n)} \mid n \in \{1, \ldots, N\} \}$$

Therefore, our proposed Multi-style EMF model can be represented as shown in Eq. 22:

$$\hat{r}_{\gamma, \text{hybrid}}^{ui}, (W_{ui}^{(n)})_{n \in \{1, \ldots, N\}} = \text{hyb}(\hat{r}_{EMF}^{ui}, W_{EMF}^{ui}, \gamma)$$

where $\hat{r}_{\gamma, \text{hybrid}}^{ui}$ is the predicted rating of user $u$ for item $i$, by the hybrid model and $\gamma$ is the hybrid model’s hyper-parameter and $\text{hyb}(.)$ is a hybrid aggregation function that combines multiple recommendation model outputs such that $N$ is the number of explanation styles that are capable of explaining item $i$ to user $u$. The optimal value for $\gamma$ is learnt by using a validation set of user-item pairs denoted as $R_{val}$. The objective function for our proposed hybrid methods is given by:

$$\gamma_{opt} = \arg \min_{\gamma} J = \sum_{(u,i) \in R_{TRAIN}} (r_{ui} - \hat{r}_{ui}^{\gamma, \text{hybrid}})^2$$

Where $R_{TRAIN}$ is a user-item training set of ratings such that $R_{TRAIN} = R_{train} \cup R_{val}$ and $R_{train} \cap R_{val} = \emptyset$. Finally, $\gamma_{opt}$ are the optimal hybrid model hyper-parameter values chosen based on grid search and performance on the validation set $R_{val}$ or by directly optimizing an objective like Eq. 23.

Algorithm 1 describes the steps for calculating the hybrid predictions for the methods discussed in Sections 3.2.1 and 3.2.3. In Sections 3.2.1, 3.2.2, 3.2.3 and 3.2.4, we present four methods for hybrid multiple style explanations.
Algorithm 1 Hybridized Multi-style EMF: used in the methods discussed in Sections 3.2.1 and 3.2.3

**Input:** Data Matrix \( r_{ui} : (u, i) \in R \)  

**Output:** \( \hat{r}^{\gamma_{opt}, hybrid} \), hybridization function hyb(.)

1. Split ratings \( R \) into \( R = R^{TRAIN} \cup R^{test}, R^{TRAIN} = R^{train} \cup R^{val} \)
2. Build \( N \) distinct EMF models, \( \{P^n, Q^n\} \) using \( (u, i) \in R^{train} \) and compute \( \hat{r}^{EMF} \) and \( W^{EMF} \)
3. Find \( \gamma_{opt} \)
4. Compute Multi-style EMF predicted ratings \( \hat{r}_{ui}^{\gamma, hybrid} \) using Eq. 22

### 3.2.1 Weighted Multi-style Explainable Matrix Factorization (W-MEMF)

In this method, the objective is to learn the non-negative weights, \( w \), for the individual EMF model such that they minimize the objective function given in Eq. 23. The hybrid model’s prediction in Eq. 22 for a user-item pair is calculated as the weighted sum of each element in \( \hat{r}_{ui}^{EMF} \). The hybrid model hyper-parameter, \( \gamma \), are the non-negative weights in a vector \( w \).

#### 3.2.1.1 Hybrid Model Building

We investigate two approaches for minimizing the objective function in Eq. 23 using our proposed Weighted Multi-style Explainable Matrix Factorization (W-MEMF) method. These are described below:

1. **Global Weights:** In this approach, we assume the individual EMF models perform equally for all users such that a poor EMF model will be bad at predicting and explaining recommendations for all users and vice-versa. The objective, using this approach, is to estimate the weights for each individual EMF model that minimizes the hybrid model’s objective function shown in Eq. 23. In this method, \( \gamma = w_n \) and hybrid model’s predicted rating for a user-item pair, \( \hat{r}_{ui}^{global} \) is calculated as follows:
Algorithm 2 Weighted-Multi-style EMF (W-MEMF) using Global Weights

**Input:** Data Matrix $r_{ui} : (u, i) \in R, \Gamma$  
**Output:** $\hat{r}_{\gamma_{opt}.global}$

(a) Split into $R = R^{TRAIN} \cup R^{test}$, $R^{TRAIN} = R^{train} \cup R^{val}$

(b) Build $N$ distinct EMF models, $\{P^n, Q^n\}$ using $(u, i) \in R^{train}$ and compute $\hat{r}^{EMF}$ and $W^{EMF}$

(c) Find $\gamma_{opt}$:
   i. For $\gamma \in \Gamma$:
      A. Compute hybrid predictions, $\hat{r}^{\gamma.global}$ using Eq. 24 with $w_n = \gamma$
      B. Compute $RMSE^\gamma = \sum_{(u,i) \in R^{val}} (r_{ui} - \hat{r}^{\gamma.global}_{ui})^2$
   ii. Choose $\gamma_{opt} = \arg \min_{\gamma} RMSE^\gamma$

(d) Compute Multi-style EMF prediction $\hat{r}^{\gamma.hybrid}_{ui} = \hat{r}^{\gamma_{opt}.global}_{ui}$ using Eq. 24 with $w_n = \gamma_{opt}$

\[
\hat{r}^{\gamma.hybrid}_{ui} = \hat{r}^{w.hybrid}_{ui} = \hat{r}^{global}_{ui} = \sum_{n=1}^{N} w_n \hat{r}^{(n)}_{ui} \tag{24}
\]

Algorithm 2 shows the steps for calculating the predicted ratings using the Global W-MEMF method.

where $\sum_{n=1}^{N} w_n = 1$

2. **Personalized Weights:** In the personalized approach, we assume that individual EMF models may perform better for some users than others. Therefore, the objective is to learn the individual EMF models’ weights that are proportional to the accuracy of the model for each user $u$. We estimate the accuracy of the individual EMF models for each user, by calculating the Pearson Correlation coefficient between the user’s known ratings $r_{ui}$ in $R^{train}_u$, and the predicted ratings, $\hat{r}^{(n)}_{ui}$, for the user by the individual EMF models, as shown in Eq. 25. The intuition is that an EMF model that has a higher correlation with the known ratings of the user is a better predictor of the user’s
preferences and should also provide better explanations to the user. The steps for this method are shown in Algorithm 3.

\[
(w^\gamma_u)^{(n)} = \begin{cases} 
\varphi(r_u, \hat{r}_u^{(n)}) & \text{if } \varphi(r_u, \hat{r}_u^{(n)}) \geq \text{threshold} \\
0, & \text{otherwise}
\end{cases}
\]  

(25)

Where

\[
\varphi(r_u, \hat{r}_u^{(n)}) = \frac{\sum_{(u,i) \in R^{\text{train}}}(r_{ui} - \bar{r}_u)(\hat{r}_{ui} - \bar{\hat{r}}_u)}{\sqrt{\sum_{(u,i) \in R^{\text{train}}}(r_{ui} - \bar{r}_u)^2(\hat{r}_{ui} - \bar{\hat{r}}_u)^2}}
\]  

(26)

We then normalize the EMF models’ weights for each user as shown in Equation 27, where \(N\) is the number of individual models, to obtain the relative weights to be assigned to each individual EMF model’s predicted rating for user \(u\), as follows:

\[
w^u_n = \frac{(w^\gamma_u)^{(n)}}{\sum_{n=1}^{N}(w^\gamma_u)^{(n)}}
\]  

(27)

The predicted rating, \(\hat{r}_{ui}^{\text{personalized}}\), from the hybrid model for each user-item pair is calculated as the weighted sum of the predicted ratings which are the outputs of the individual EMF models, using the normalized weights as shown in Eq. 28.

\[
\hat{r}_{ui}^{\gamma,\text{hybrid}} = \hat{r}_{ui}^{w,\text{hybrid}} = \hat{r}_{ui}^{\text{personalized}} = \sum_{n=1}^{N} w^u_n \hat{r}_{ui}^{(n)}
\]  

(28)
Algorithm 3 Weighted-Multi-style EMF (W-MEMF) using Personalized Weights

**Input:** Data Matrix $r_{ui} : (u, i) \in R$

**Output:** $\hat{r}_{personized}$

(a) Split into $R = R^{TRAIN} \cup R^{test}$, $R^{TRAIN} = R^{train} \cup R^{val}$

(b) Build $N$ distinct EMF models, $\{P^n, Q^n\}$ using $(u, i) \in R^{train}$ and compute $\hat{r}^{EMF}$ and $W^{EMF}$

(c) For each user $u$:

   i. For each individual EMF model $n$:
      
      A. Compute personalized weight $w^u_n$ using Eq. 27

      ii. Compute Multi-style EMF predictions, $\hat{r}^{\gamma,hybrid}_{ui} = \hat{r}_{personized}$ using Eq. 28

3.2.2 Regression-based Multi-style Explainable Matrix Factorization (R-MEMF)

In this method, we consider the hybrid model’s prediction for a user-item pair as a regression problem. For the proposed hybrid model, the features of the regression model are the predicted ratings for the user-item pair from the individual EMF models, $\hat{r}^{EMF}_{ui}$. Finally, the objective of the regression model is to learn the coefficients for the features that minimize the hybrid model’s objective function in Eq. 23.

3.2.2.1 Hybrid Model Building

In this method, the hybridization hyper-parameters $\gamma$ are the feature coefficients vector, denoted as $\beta$. We propose three approaches for learning $\beta$ in the hybrid model building phase, where

$$\hat{r}^{hybrid}_{ui} = \beta_0 + \sum_{n=1}^{N} \beta_n \hat{r}^{(n)}_{ui}$$

1. **Linear Hybrid:** In this approach, we learn the feature coefficient vector, $\beta$, that minimizes the square error between the hybrid model predictions, $\hat{r}^{hybrid}_{ui}$ and the true ratings $r_{ui}$ in the training set $R^{train}$. The objective function for this method is given in Eq. 29.
\[
\beta = \arg \min_{\beta} \ J_{\text{linear}} = \sum_{(u,i) \in R^{\text{train}}} (r_{ui} - \hat{r}^{\text{hybrid}}_{ui})^2
\]  

(29)

2. **Lasso Hybrid:** In this approach, we learn the feature coefficient vector, \( \beta \), that minimizes the square error between the hybrid model predictions, \( \hat{r}^{\text{hybrid}}_{ui} \) and the true ratings \( r_{ui} \) in the training set, \( R^{\text{train}} \). To avoid over-fitting to the training data, we use a regularization term, weighted by coefficient \( \lambda \), that regulates the simplicity of the model to ensure generalization. For this approach, we use an L1-regularization term into the objective function as shown in Eq. 30. The L1-regularization selects the features with the most influence when minimizing the objective function.

\[
\beta = \arg \min_{\beta} \ J_{\text{lasso}} = \sum_{(u,i) \in R^{\text{train}}} (r_{ui} - \hat{r}^{\text{hybrid}}_{ui})^2 + \lambda \| \beta \|_1
\]  

(30)

3. **Ridge Hybrid:** In this approach, we learn the feature coefficient vector, \( \beta \), that minimizes the square error between the regression model predictions, \( \hat{r}^{\text{hybrid}}_{ui} \), and the true ratings \( r_{ui} \) in the training set, \( R^{\text{train}} \). We use an L2-regularization term with the objective function as shown in Eq. 31. The L2-regularization term helps reduce the model complexity by selecting the features that contribute towards minimizing the objective function. L2-regularization does not completely eliminate the influence of the other features but rather seeks to minimize their influence on the predictions.

\[
\beta = \arg \min_{\beta} \ J_{\text{ridge}} = \sum_{(u,i) \in R^{\text{train}}} (r_{ui} - \hat{r}^{\text{hybrid}}_{ui})^2 + \lambda \| \beta \|_2^2
\]  

(31)

3.2.3 **Switched Multi-style Explainable Matrix Factorization (S-MEMF)**

In this method, the hybrid model performs an intentional selection between the diverse individual recommendation models, at an item-level, by switching its output prediction between the different individual EMF models depending on the item being recommended, based on a switching criterion [52]. The switching criteria, in our case, are based on the
predicted ratings and explanation scores that are generated by the individual EMF models, as detailed below.

### 3.2.3.1 Switched Hybrid Model Building

In this method, the hybrid model’s hyper-parameter \( \gamma \) depends on the switching criteria selected. We investigate three switching criteria for building the hybrid model in the Switched Multi-style EMF hybridization phase:

1. **Explanation Score Switching Criterion:** Given predicted ratings \( \hat{r}_{ui}^{(n)} \) and explainability scores \( W_{ui}^{(n)} \), for each of the individual \( n^{th} \) models where \( n \in \{1, \ldots, n\} \), the final predicted rating from the hybrid model \( \hat{r}_{ui}^{\text{hybrid}} \), is selected using the switching condition shown in Eq. 32

\[
j = \arg \max_{n \in \{1,N\}} W_{ui}^{(n)}
\]

\[
\hat{r}_{ui}^{\text{hybrid}} = \hat{r}_{ui}^{\text{expl}} = \hat{r}_{ui}^{(j)}
\]  

(32)

The intuition is that the hybrid model will use the predicted rating from the individual EMF model with the highest explanation score for every user-item pair.

2. **Transparency Switching Criterion:** We define transparency for each user \( u \), as the degree of correlation between a user’s known ratings in \( R_{u}^{\text{train}} \) and the predicted ratings from an individual EMF model’s for the user, \( \hat{r}_{ui}^{(n)} \). Therefore, in this approach, the hybrid model selects all predicted ratings for a user from the individual EMF model with the highest correlation with the user’s known ratings, as shown in eq. 33.

\[
j = \arg \max_{n \in \{1,N\}} \varphi_{(u,i)} \in R_{u}^{\text{train}}(r_{u}, \hat{r}_{ui}^{(n)})
\]

\[
\hat{r}_{ui}^{\text{hybrid}} = \hat{r}_{ui}^{\text{transp}} = \hat{r}_{ui}^{(j)}
\]

(33)

Where \( \varphi(,) \) is the Pearson correlation coefficient calculated using Eq. 26. The intuition behind this switching criterion is that the individual model that is best at
predicting the preference of user \( u \) will also be the best at explaining the recommendations made to user \( u \).

3. **Smooth Maximum Switching Criterion:** The hybrid prediction is computed using a smooth maximum of the individual model predictions \( \hat{r}_{ui}^{(n)} \), i.e., \( \hat{r}_{ui}^{\text{smooth}} = f_\alpha(\hat{r}_{ui}^{(n)}) \), where \( f_\alpha(\hat{r}_{ui}^{(n)}) \) is a smooth approximation of the maximum function, given in Eq. 34.

\[
f_\alpha(\hat{r}_{ui}^{EMF}) = \frac{\sum_{n=1}^{N} e^{\alpha \hat{r}_{ui}^{(n)}} \hat{r}_{ui}^{(n)}}{\sum_{m=1}^{N} e^{\alpha \hat{r}_{ui}^{(m)}}}
\]  

(34)

As \( \alpha \to \infty \), the function, \( f_\alpha \) is smooth and converges to the maximum function [56]. As \( \alpha \to -\infty \), \( f_\alpha \) converges to the minimum function. Finally, when \( \alpha = 0 \), the smooth maximum function returns the arithmetic mean of \( \hat{r}_{ui}^{(n)} \). Therefore, we consider the smooth function as an *adaptive aggregation* function and utilize it to find the right aggregation mechanism for \( \hat{r}_{ui}^{EMF} \). Algorithm 4 shows how the predicted ratings are calculated using this method.

**Algorithm 4** Switched-Multi-style EMF (S-MEMF) using Smooth Maximum Criterion

**Input:** Data Matrix \( r_{ui} : (u, i) \in R \), \( A \)

**Output:** \( \hat{r}_{ui}^{\alpha_{\text{opt}},\text{smooth}} \)

(a) Split into \( R = R^{\text{TRAIN}} \cup R^{\text{test}}, R^{\text{TRAIN}} = R^{\text{train}} \cup R^{\text{val}} \)

(b) Build \( N \) distinct EMF models, \( \{P^n, Q^n\} \) using \( (u, i) \in R^{\text{train}} \) and compute \( \hat{r}_{ui}^{EMF} \) and \( W^{EMF} \)

(c) Find \( \alpha_{\text{opt}} \):

i. For \( \alpha \in A \):

A. Compute hybrid predictions, \( \hat{r}_{ui}^{\alpha,\text{smooth}} = f_\alpha(\hat{r}_{ui}^{EMF}) \) using Eq. 34

B. Compute \( RMSE^\alpha = \sum_{(u,i) \in R^{\text{val}}} (r_{ui} - \hat{r}_{ui}^{\alpha,\text{smooth}})^2 \)

ii. Choose \( \alpha_{\text{opt}} = \arg \min_\alpha RMSE^\alpha \)

(d) Compute Multi-style EMF prediction \( \hat{r}_{ui}^{\alpha_{\text{opt}},\text{smooth}} = f_{\alpha_{\text{opt}}}(\hat{r}_{ui}^{EMF}) \) using Eq. 34
3.2.4 Asymmetric Multi-style Explainable Matrix Factorization (A-MEMF)

This method is inspired by [57] [11] where an Asymmetric Matrix Factorization algorithm exploits more than one domain in the process of building a heterogeneous MF model. The A-MEMF method extends this concept by building a hybrid model that is capable of handling more than one explanation style. This method considers the task of building a Hybrid Multi-style EMF model that combines two EMF models and is consequently able to explain recommendations using any of their individual explanation styles.

3.2.4.1 Hybrid Model Building

Building the hybrid model occurs in two phases which are discussed below.

1. **Anchoring Phase:** The individual EMF models consists of the user and item latent feature matrices, denoted as $P$ and $Q$ respectively. Therefore, given two EMF models, $EMF_t$ and $EMF_v$ and their predicted ratings denoted as $\hat{r}_{ui}^{(t)}$ and $\hat{r}_{ui}^{(v)}$ respectively, $\hat{r}^{(t)}$ can be factored as follows:

   $$\hat{r}_{ui}^{(t)} \approx P_u^{(t)}(Q_i^{(t)})^T$$  \hspace{1cm} (35)

   We consider the $EMF_t$ to be the anchor Explainable MF model and the latent space is generated using this EMF model.

2. **Transfer Phase:** In this phase, we transfer either the user or item latent feature matrix from $EMF_v$ and adapt it to learn the latent space of $EMF_t$. Therefore, we adapt the latent space basis from $EMF_v$, to match the previously obtained latent space built using $EMF_t$ while being anchored in either the user or item latent vectors. Since $\hat{R}^{(v)}$ and either $P^{(v)}$ or $Q^{(v)}$ are known, estimating $Q^{(t)}$ or $P^{(t)}$ respectively is a convex problem which can be easily solved. The result is a latent space originally constructed using $EMF_t$, but adapted by $EMF_v$. The adapted latent space, consequently is a hybrid of the individual EMF models and is also capable of explaining recommendations using their respective explanation styles. The hybrid model’s hyper-parameter
for this method is the number of iterations required for adapting to the new latent
space, beyond which there is no gain in performance.

We propose two approaches to building the hybrid model using this method:

(a) **User-Anchored Hybrid:** Update item latent features matrix, \( Q(t) \), in the
latent space while fixing the user latent features, \( P(v) \) which serves as an anchor.

The objective function for this approach is given by Eq. 36

\[
Q(t) = \arg\min_{Q(t)} J = \sum_{(u,i) \in R^{train}} (r_{ui} - \hat{r}_{ui}^{hybrid})^2 + \frac{\beta}{2} (\|P(v)\|_2^2 + \|Q(t)\|_2^2)
\]  

(36)

where

\[
\hat{r}_{ui}^{hybrid} = P(v) (Q^{(t)}_i)^T
\]  

(37)

Updating the values of the item latent matrix, \( Q(t) \) in the objective function
shown in Eq. 36 can be done using Stochastic Gradient Descent by selecting a
reasonable step size, \( \alpha \), which will converge to local minima in the latent space.

\[
q_{i}^{t+1} \leftarrow q_{i}^{t} + \alpha (2(r_{ui}^{train} - p_{u}^{v} (q_{i}^{t})^T p_{u}^{v} - \beta q_{i}^{t})
\]  

(38)

(b) **Item-Anchored Hybrid:** Update the user latent feature matrix \( P(t) \), in the
latent space while fixing the item latent features, \( Q^{(v)} \) which serves as the anchor.

The objective function for this approach is shown in Eq. 39

\[
P(t) = \arg\min_{P(t)} J = \sum_{(u,i) \in R^{train}} (r_{ui} - \hat{r}_{ui}^{hybrid})^2 + \frac{\beta}{2} (\|P(t)\|_2^2 + \|Q^{(v)}\|_2^2)
\]  

(39)

where

\[
\hat{r}_{ui}^{hybrid} = P^{(t)}_u (Q^{(v)}_i)^T
\]  

(40)
Updating the values of the item latent matrix, $P^{(t)}$ in the objective function shown in Eq. 36 can be done using Stochastic Gradient Descent by selecting a reasonable step size, $\alpha$, which will converge to local minima in the latent space.

$$p_{u_i}^{t+1} \leftarrow p_{u_i}^t + \alpha(2(r_{ui} - p_{u_i}^T q_{i}^u)q_{i}^v - \beta p_{u_i}^T)$$  \hspace{1cm} (41)
3.3 Tag-Assisted Explainable Matrix Factorization Methods

User-generated tags are a rich source of information and can be used to improve the recommendation model. Inspired by [1], we propose a novel tag-based explainability graph, that can be used in a tag-based explanation style method. Finally, we propose tag-boosted multi-style methods that integrate tagging information into explainable matrix factorization methods to provide multiple explanation-styles for each recommended item.

3.3.1 Tag-based Explainability Graphs

A tag-based explainability graph is a matrix that holds the explainability score of each user for every item available for the recommendation task. The explainability score is calculated by extracting users propensity for certain tags (tag preference) and relevance of tags to items (tag relevance). We used three types of relationships to construct three explainability graphs. The first graph is a user-based graph and it describes the relationship between all users and all available tags. The second graph is an item-based graph and it describes the relationship between all items and all available tags. The third graph is a combination of the aforementioned two graphs and it represents the relationship between users and items based on each user’s preference towards each item’s relevant tags. This combination is obtained using the product of the user-based and item-based graphs. In this work, we estimate the tag relevance and tag preference using the definitions given by [1] and use the Tagsplanation Explanation Style for all our proposed tag-based explanation methods. Therefore, three different tag-aware graphs are constructed using the user-generated tags, as follows, where $\mathcal{T} =$set of all tags, $U =$set of users, $I =$set of items.

1. Tag Preference graph ($T^{pref}$) is a bipartite graph, $T^{pref} = (U, \mathcal{T}, E^{pref})$.

2. Tag Relevance graph ($T^{rel}$) is a bipartite graph, $T^{rel} = (I, \mathcal{T}, E^{rel})$.

3. User-item tag-based explainability graph ($T^{UI}$) is a bipartite graph, $T^{UI} = (U, I, E^t)$.

The edge sets $E^{pref}$, $E^{rel}$ and $E^t$ are weighted edges with weights calculated as described in Sec 3.3.1.1, 3.3.1.2, and 3.3.1.3 respectively.
3.3.1.1 Tag Preference (User-tag Relationship)

A user’s tag preference is computed using a weighted average of the user’s rating of items tagged with that tag. Tag preference is denoted as \( \text{tagPref} \) and the tag preference of user \( u \) for tag \( t \) is calculated as follows:

\[
\text{tagPref}(u, t) = \frac{\left( \sum_{i \in I_u} r_{ui} \times \text{tagShare}(t, i) \right) + \bar{r}_u \times k}{\left( \sum_{i \in I_u} \text{tagShare}(t, i) \right) + k} \tag{42}
\]

Where \( \text{tagShare}(t, i) \) is defined as “as the number of times \( t \) has been applied to \( i \), divided by the number of times and tag has been applied to \( i \)” [1]. \( I_u \) is the set of items rated by user \( u \), \( \bar{r}_u \) is the average rating of user \( u \) across all items, \( r_{ui} \) is user \( u \)’s rating for item \( i \). Finally, \( k \) is a smoothing constant that accounts for users who have not rated any or too few items tagged with tag \( t \).

3.3.1.2 Tag Relevance (Item-tag Relationship)

A tag’s relevance to an item, denoted as \( \text{tagRel} \), can be calculated using correlation between users’ preferences for the tag and their preference for the recommended item. The correlation function used is the Pearson correlation:

\[
\text{tagRel}(t, i) = \begin{cases} \varphi(X, Y) & \text{if } t \text{ has been applied to } i \\ 0 & \text{otherwise} \end{cases} \tag{43}
\]

Where \( X \) is the set of ratings for item \( i \) by all users in \( U_t \) (set of users who have applied tag \( t \) to item \( i \)). This set of ratings is then adjusted by each user’s average rating to accommodate personal preferences. \( Y \) is defined as the set of inferred tag preference values toward tag \( t \) for all users in \( U_t \), adjusted by each user’s average rating. Therefore, \( X = \{r_{ui} : u \in U_t\} \) and \( Y = \{\text{tagPref}(u, t) : u \in U_t\} \).
3.3.1.3 User-Item Tag-based Explainability Graph

The tag-based explainability score for a user $u$ for item $i$ (Figure 13) is calculated as the dot product of the edge $V$ weight of tagPref graph and the tagRel graph, where $\vec{T}_{u}^{pref} = (tagPref(u, 1), ..., tagPref(u, |T|))$ and $\vec{T}_{i}^{rel} = (tagRel(i, 1), ..., tagRel(i, |T|))$.

$$T_{u,i}^{UI} = \begin{cases} 
\vec{T}_{u}^{pref} \cdot \vec{T}_{i}^{rel} & \text{if } \vec{T}_{u}^{pref} \cdot \vec{T}_{i}^{rel} > \theta_t \\
0 & \text{otherwise} 
\end{cases}$$

(44)

Where $\theta_t$ is a tag-based explainability threshold.

Figure 13: Tag-based Explanation Graph for Explaining the Recommended Movie, *Pulp Fiction*, to a User, obtained by combining the tag preferences (orange) and the tag relevance (blue).

3.3.1.4 Tag-based Explanation Style

The explanations generated by the above tag-based methods, will be presented to the user using the explanation style proposed by [1]. This explanation style presents user-generated
tags, sorted by the relevance of the tag to the recommended item, to the user as a keyword-style explanation. An example of a tagsplanation for a recommended movie is shown in Figure 14.

3.3.2 Tag-Assisted Explainable Matrix Factorization (TA-EMF)

In the previous chapter, we reviewed explainable matrix factorization algorithms which were implemented by [3] [11] which produce explanations for recommendations using the neighborhood technique and semantic information, respectively. The neighborhood technique is built on the premise that items liked by other users who are similar to the target user will most likely be liked by the target user as well. The semantic technique is based on estimating a user’s interest in an item by estimating the user’s interest in features of the item such as the actors and directors in a movie or author and publisher of a book. The explanation scores in these methods can be used to build an explainability matrix for each user-item pair. This matrix is then used in the process of learning the latent space vectors for both users and items. In this section, we propose a method that is driven by the User-Item tag-based ($T^{UI}$) explainability matrix presented in Section 3.3.1.
3.3.2.1 Model Building

The objective function minimized by our proposed method uses tag-based explainability scores instead of the neighborhood-based or semantic-based explainability scores:

\[
p_{u}, q_{i} = \arg \min_{p_{u, q_{i}}} J = \sum_{u, i \in R} (r_{ui} - p_{u}q_{i}^{T})^2 + \frac{\beta}{2} (\|p_{u}\|^2 + \|q_{i}\|^2) + \frac{\gamma}{2} \|p_{u} - q_{i}\|^2 T_{u, i}^{UI} \tag{45}
\]

The first two terms of Eq. 45 are from MF [6] represent the error after reconstruction using latent vectors and a regularization term to avoid over-fitting, respectively. \(\beta\) is a regularization coefficient that controls the smoothness of the regularization term. The third term adds the contribution of the explainability scores to the matrix factorization model as in [3] [11]. \(\gamma\) is a smoothing coefficient that controls the contribution of the explainability term to the learned parameters \(p_{u}\) and \(q_{i}\).

We utilize Stochastic Gradient Descent to update \(p\) and \(q\) iteratively until the convergence of \(J\).

The gradient of \(J\) with respect to \(p_{u}\):

\[
\frac{\partial J}{\partial p_{u}} = -2(r_{u,i} - p_{u}q_{i}^{T})q_{i} + \beta p_{u} - \gamma(p_{u} - q_{i})T_{u, i}^{UI} \tag{46}
\]

The gradient of \(J\) with respect to \(q_{i}\) is

\[
\frac{\partial J}{\partial q_{i}} = -2(r_{u,i} - p_{u}q_{i}^{T})p_{u} + \beta q_{i} + \gamma(p_{u} - q_{i})T_{u, i}^{UI} \tag{47}
\]

Using the gradients, the formulation of the update rules with learning parameters \(\alpha\) will be:

\[
p_{u}^{(t+1)} \leftarrow p_{u}^{(t)} + \alpha (2(r_{u,i} - p_{u}^{(t)}q_{i}^{(t)})^{T}q_{i}^{(t)} - \beta p_{u}^{(t)} - \gamma(p_{u}^{(t)} - q_{i}^{(t)})T_{u, i}^{UI})
\]

\[
q_{i}^{(t+1)} \leftarrow q_{i}^{(t)} + \alpha (2(r_{u,i} - p_{u}^{(t)}q_{i}^{(t)})^{T}p_{u}^{(t)} - \beta q_{i}^{(t)} + \gamma(p_{u}^{(t)} - q_{i}^{(t)})T_{u, i}^{UI}) \tag{48}
\]

Our proposed method differs from MF [6] because our method generates explanations simultaneously with recommendations. Our method differs from other explainable matrix

3.4 Tag-boosted Multi-style Explainable Matrix Factorization

We propose objective functions in this Section that are inspired by [6] [48] and [3]. We propose two new methods that incorporate EMF and tag-boosted methods in one approach. The intuition here is that since tags provide useful information, incorporating the user’s preference of a tag or the tag relevance of a tag for an item may lead to improved performance. This approach will allow recommendations to be presented to users using two widely accepted and previously validated [1] [4] [14] [3] explanation styles.

3.4.1 Preferred Tag-boosted Multi-style EMF

This approach integrates only the tags a user has shown some preference for into the matrix factorization model. This method is user-centered and only considers the contribution of the tags previously used by a user for an item.

3.4.1.1 Model Building

Our proposed method uses both the ISE-based EMF [3] explainability graph and the user preference square matrix ($S^{pref}$) in the process of building the user and item latent space vectors. The objective function, to be minimized, is given by:

$$p_u, q_i = \arg \min_{p_u, q_i} J = \sum_{u, i \in R} \left[ (r_{ui} - p_u q_i^T)^2 + \frac{\beta}{2}(\|p_u\|^2 + \|q_i\|^2) + \frac{\lambda}{2}\|p_u - q_i\|^2 E_{ui} + \frac{\gamma}{2} \sum_{v \in S^{pref}_{u,v}} (S^{pref}_{u,v} - p_u p_v^T)^2 \right] \quad (49)$$

where $S^{pref}(u, v) = \text{cosineSim}(u, v) = \frac{T^{pref}_u \cdot T^{pref}_v}{\|T^{pref}_u\| \|T^{pref}_v\|}$
The first three terms in Eq. 49 are similar to the Explainable Matrix Factorization (EMF) objective function [3]. \( r_{ui} \) represents the rating given for item \( i \) by user \( u \). \( p_u \) and \( q_i \) represent the low dimensional latent factor vectors of users and items, respectively. This version of EMF uses the ISE-based explainability graph \( (E_{u,i}) \) to represent the item-based explainability scores as given by Eq. 21. Our contribution is the addition of the fourth term to obtain a tag-boosted approach to integrate the information from the tags. \( S^{pref} \) is a user \( \times \) user similarity matrix that holds the similarity between every pair of users. For a target user \( u \), we get the subset of users \( S^{pref}_u \) such that \( v \in S^{pref}_u \) and \( u, v \) have used the same set of tags for any item \( i \). \( p_u \) and \( p_v \) are the latent factor vectors of users \( u \) and \( v \), respectively. \( \gamma \) is the tag-boosted term which weights the contribution of the new term. \( T^{pref}_u \) is the vector of preference weights given by user \( u \) to all tags. We use Stochastic Gradient Descent to optimize the objective function in 49.

The gradient of \( J \) with respect to \( p_u \) is calculated as follows:

\[
\frac{\partial J}{\partial p_u} = -2 (r_{ui} - p_u q_i^T) q_i + \beta p_u + \lambda (p_u - q_i) E_{ui} + \gamma (S^{pref}_{u,v} - p_u p_v^T) p_v
\]

The gradient of \( J \) with respect to \( q_i \) is calculated as follows:

\[
\frac{\partial J}{\partial q_i} = -2 (r_{ui} - p_u q_i^T) p_u + \beta q_i - \lambda (p_u - q_i) E_{ui}
\]

Using the gradients, the formulation of the update rules is:

\[
p_u^{(t+1)} \leftarrow p_u^{(t)} + \alpha \left( 2 (r_{ui} - p_u^{(t)} q_i^{(t)}^T) q_i^{(t)} - \beta p_u^{(t)} - \lambda (p_u^{(t)} - q_i^{(t)}) E_{ui} + \gamma (S^{pref}_{u,v} - p_u^{(t)} p_v^{(t)}^T) p_v^{(t)} \right)
\]

\[
q_i^{(t+1)} \leftarrow q_i^{(t)} + \alpha \left( 2 (r_{ui} - p_u^{(t)} q_i^{(t)}^T) p_u^{(t)} - \beta q_i^{(t)} + \lambda (p_u^{(t)} - q_i^{(t)}) E_{ui} \right)
\]

(50)

Algorithm 2 summarizes the steps of the Preferred Tag-boosted Multi-style EMF method.
Algorithm 5 Preferred Tag-boosted Multi-style Explainable Matrix Factorization (Pref-Tag)

**Input:** Train $R$, tag-Preference $T_u^{pref}$, number of factors $f$, number of common preferred tags $k$, $\alpha$, $\beta$, $\lambda$ and $\gamma$. **Output:** Latent factor matrices: $P$ and $Q$

1. **for** each item $i$:
   
   (a) calculate the set of neighborly items, using the Cosine similarity

2. **end for**

3. **for** each user-item pair $(u, i)$:
   
   (a) calculate $E_{u,i}$ using Eq. 14

4. **end for**

5. **for** each user pair $(u, v)$:
   
   (a) calculate $S^{pref}(u, v)$ using the Cosine similarity

6. **end for**

7. initialize latent factor matrices $P$ and $Q$

8. **for** each rating $r_{u,i}$ from the training set:
   
   (a) Calculate $p_{u}^{t+1}$ and $q_{i}^{t+1}$ using the update rule in Eq. 50
   (b) $t \leftarrow t + 1$

9. **end for**

### 3.4.2 Relevant Tag-boosted Multi-style EMF

This method utilizes the user-based explainability graph for EMF and item-centered tag similarity for the tag-boosted term. The tags we integrate into the model for this model are however obtained from the Tag-Relevance vector $\vec{T}_r^{rel}$. $S_{rel}$ is an $item \times item$ similarity matrix that holds the similarity between every item pair. For a target item $i$, we find the subset of items $S_i^{rel}$ such that $j \in S_i^{rel}$ and $i$ and $j$ have been tagged with the same tags. $q_i$ and $q_j$ are the latent factor vectors of the items $i$ and $j$ respectively. $\gamma$ is the tag-boosted term co-efficient which weights the contribution of the new term. TagRel is a $tags \times items$ matrix holding tag-item relevance values. The objective function to be minimized is given by:
\[ p_u, q_i = \arg \min_{p_u, q_i} J = \sum_{u,i \in R} \left[ (r_{u,i} - p_u q_i^T)^2 \right. \]

\[ \left. + \frac{\beta}{2} (\|p_u\|^2 + \|q_i\|^2) + \frac{\lambda}{2} \|p_u - q_i\|^2 W_{u,i} + \frac{\gamma}{2} \sum_{j \in S_{rel}^i} (S_{rel}^{i,j} - q_i q_j^T)^2 \right] \]

where \( S_{rel}^{i,j} = \cosineSim(i,j) = \frac{\mathbf{T}_{rel}^i \cdot \mathbf{T}_{rel}^j}{\|\mathbf{T}_{rel}^i\| \|\mathbf{T}_{rel}^j\|} \)

The gradient of \( J \) with respect to \( p_u \) is calculated as follows:

\[ \frac{\partial J}{\partial p_u} = -2(r_{u,i} - p_u q_i^T) q_i + \beta p_u + \lambda (p_u - q_i) W_{u,i} \]

The gradient of \( J \) with respect to \( q_i \) is calculated as follows:

\[ \frac{\partial J}{\partial q_i} = -2(r_{u,i} - p_u q_i^T) p_u + \beta q_i - \lambda (p_u - q_i) W_{u,i} + \gamma (S_{rel}^{i,j} - q_i q_j^T) q_j \]

Algorithm 3 summarizes the steps of the Relevant Tag-boosted Multi-style EMF method.
Using the gradients, the formulation of the update rules will be:

\[
\begin{align*}
    p^{(t+1)}_u & \leftarrow p^{(t)}_u + \alpha \left( 2(r_{u,i} - p^{(t)}_u q^{(t)}_i)^T q^{(t)}_i - \beta p^{(t)}_u - \lambda (p^{(t)}_u - q^{(t)}_i)W_{u,i} \right) \\
    q^{(t+1)}_i & \leftarrow q^{(t)}_i + \alpha \left( 2(r_{u,i} - p^{(t)}_u q^{(t)}_i)^T p^{(t)}_u - \beta q^{(t)}_i + \lambda (p^{(t)}_u - q^{(t)}_i)W_{u,i} + \gamma (S^{rel}_{i,j} - q^{(t)}_i q^{(t)}_j)^T q^{(t)}_j \right)
\end{align*}
\]  

(52)

**Algorithm 6** Relevant Tag-boosted Multi-style Explainable Matrix Factorization (RelTag)

**Input:** Train \( R \), tag-Relevance \( T^{rel} \), number of factors \( f \), number of common relevant tags \( k \), \( \alpha \), \( \beta \), \( \lambda \) and \( \gamma \). **Output:** Latent Factor Matrices: \( P \) and \( Q \)

1. **for** each user \( u \):
   (a) calculate the set of neighboring users using the Cosine similarity
2. **end for**
3. **for** each user-item pair \((u, i)\):
   (a) calculate \( W_{u,i} \) using Eq. 14
4. **end for**
5. **for** each item pair \((i, j)\) in tag-relevance matrix:
   (a) Calculate \( S^{rel}(i, j) \) using Cosine similarity
6. **end for**
7. initialize matrices \( P \) and \( Q \)
8. **for** each \( r_{u,i} \) from the training set:
   (a) Calculate \( p^{t+1}_u \) and \( q^{t+1}_i \) using the update rule in Eq. 52
   (b) \( t \leftarrow t + 1 \)
9. **end for**
CHAPTER 4

EXPERIMENTAL EVALUATION

In this chapter, we present the experimental evaluation of our proposed Multi-style Explainable Matrix Factorization methods by comparing them with baseline algorithms, in terms of the accuracy of the rating prediction and the top-n recommendation lists; as well as the explainability of the recommended items using appropriate metrics. Table 11 summarizes our evaluation plan.

<table>
<thead>
<tr>
<th>Framework Component</th>
<th>Goal</th>
<th>Methodology</th>
<th>Validation Metrics</th>
<th>Evaluation Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Style</td>
<td>Generate personalized tag-based explanation</td>
<td>Tag-Assisted Explainable Matrix Factorization (Sec. 3.3)</td>
<td>Accuracy and Explainability Metrics such as RMSE, NDCG, MAP, MEP, MER, coverage</td>
<td>Sec. 4.3</td>
</tr>
<tr>
<td>Multi-style</td>
<td>Generate personalized multi-style explanations</td>
<td>Hybrid Multi-style Methods (Sec. 3.2)</td>
<td>Tag-boosting Multi-style Methods (Sec 3.4)</td>
<td>Sec 4.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Sec. 4.3</td>
</tr>
</tbody>
</table>

4.1 Evaluation Metrics

4.1.1 Accuracy Metrics

The Root Mean Squared Error (RMSE), shown in Eq. 53 is used to evaluate methods for predictive accuracy. $U$ is the set of all users, while $\hat{r}_{ui}$ is the predicted rating for item $i$ by
user $u$, and $r_{ui}$ is the actual rating of item $i$ by user $u$

$$RMSE = \sqrt{\frac{1}{|U|} \sum_{(u,i) \in R} (r_{ui} - \hat{r}_{ui})^2}. \quad (53)$$

However RMSE only measures reconstruction or rating estimation error which is not sufficient for recommendation quality. The latter places a higher emphasis on the ordering of recommended items, i.e., their ranking and further considers only the top K recommendations (which is indicated by the symbol @K in all the ranking based metrics below). Ranking is captured by the Normalized Discounted Cumulative Gain (NDCG) [58], shown in Eq. 54 which evaluates a method based on its ability to recommend relevant items in the correct order.

$$DCG@N = \sum_{i=1}^{N} \frac{rel_i}{\log_2(i+1)}$$

$$IDCG = \sum_{i=1}^{[RELP]} \frac{2^{rel_i} - 1}{\log_2(i+1)} \quad (54)$$

$$NDCG@N = \frac{DCG@N}{IDCG}$$

Where $rel_i$ is the graded relevance of result at position $i$ and $RELP$ is the list of relevant items, ordered by relevance, in the recommended list, up to position $p$.

Another classical metric that considers the order of recommendations, is the Mean Average Precision (MAP@N), which evaluates recommendation quality by placing an emphasis on the relevance of an item to the user.

$$MAP@N = \frac{1}{|U|} \sum_{u=1}^{U} \frac{1}{m} \sum_{n=1}^{N} P_u(n) \cdot rel_u(n). \quad (55)$$

$U$ denotes the set of all users, while $m$ is the number of relevant items, and $N$ is the number of desired recommendations. $P_u(n)$ is the precision for user $u$, which is the ratio of simultaneously recommended and relevant items to the total number of recommended items up to position $n$. $rel(n)$ is either 0 or 1, indicating whether the $n^{th}$ item is relevant in the context of implicit data (where rated us considered to be relevant).
While all three above metrics capture different aspects of recommendation quality, NDCG is considered as the most agreed upon standard for benchmarking recommendation quality because of its applicability to all contexts where relevance can be nuanced such as in the case of ratings. In contrast, MAP inherently assumes a binary relevance (0 or 1), and is thus more suitable in the context of Information retrieval and implicit data. That said, we report all metrics following the standard in the recommendation systems community.

### 4.1.2 Explainability Metrics

In addition to recommendation accuracy, we evaluated our models using explainability metrics, namely MEP, MER and explainability coverage given by [3]:

\[
MEP = \frac{1}{|U|} \sum_{u \in U} \frac{|R \cap E|}{|R|}
\]

\[
MER = \frac{1}{|U|} \sum_{u \in U} \frac{|R \cap E|}{|E|}
\]

\[
coverage = \frac{\text{number of explainable user – item pairs}}{\text{number of all user – item pairs}}
\]

\[
\text{number of explainable user – item pairs} = \sum_{u \in U} \sum_{i \in E} \text{explainable}(u, i)
\]

\[
\text{number of all user – item pairs} = \sum_{u \in U} \sum_{i \in I}
\]

\[
U \text{ represents the set of users, while } R \text{ is the set of recommended items, and } E \text{ denotes the set of explainable items. MEP computes the proportion of simultaneously recommended and explainable items to the total number of recommended items across all users. Similarly, MER calculates the proportion of simultaneously recommended and explainable items to the total number of explainable items, averaged across all the users. Coverage measures the fraction of all user-item pairs that are explainable using an explainable model, meaning pairs with non-zero explainability score. We computed the above explainability metrics using the neighborhood explainability graphs (user-based and item-based) as explained next, which capture what items are considered to be “explainable” in order to compute the explainability metrics.}

68
4.1.2.1 Taking into Account the Style in Explainability Metrics

Although the explainability metrics are generic in their definition (see Eqs. 56 and 57, they can incorporate the explainability scores from different styles to determine the set $E$ where the membership condition essentially considers an item to be explainable to a user if its explainability score exceeds a given threshold. However these explainability scores depend on the explanation style chosen, and our work specifically studies multiple explanation styles. For this reason when evaluating explainability metrics, we report the metrics obtained using the explainability scores calculated using all the different styles investigated in this work, namely user-based neighborhood style (NSE), item-based neighborhood style (ISE), and in case tags are available for a given domain, tag-based neighborhood style (TSE).

4.1.3 Hypothesis

We hypothesize that Matrix Factorization methods with at least one explanation style, can provide higher explainability coverage and higher explainability than Matrix Factorization methods with only one or no explanation style, without sacrificing accuracy significantly.

The null hypothesis $\mu_0$: our proposed hybrid multi-style explainable matrix factorization methods will have a mean performance less than or equal to the mean performance of the baseline methods using the appropriate metrics.

The Alternate hypothesis $\mu_A$: our proposed methods’ mean performance will be better than the mean performance of the baselines methods using the appropriate metrics.

4.1.3.1 Validity of Explainability Metrics

The explainability metrics are an aggregation of the explainability scores that have been validated in prior work [14] that conducted user studies that found a higher subjective
perception of transparency among user-item pairs for items that have higher objective NSE and ISE style explanation scores. In the case of tag-based neighborhood style explanation scores, we are justified by the user studies of [1] that validated the user’s satisfaction with preference-based and relevance-based tag-based explanations which are the main inspiration and basis for our tag-based explainability scores.

4.2 Evaluation of Hybrid Multi-style Explainable Matrix Factorization Methods

In this section, we describe our experiments for evaluating the performance of our proposed hybrid methods. We evaluate the performance of our methods using data from two different domains: movies (the 100k MovieLens\(^1\) benchmark dataset) and books (using the BookCrossing\(^2\) benchmark dataset). The MovieLens dataset includes 943 users, 1682 movies and 100,000 known user ratings while the Book-Crossing dataset includes 278,858 users, 271,379 books and 1,000,000 known user ratings. We filtered out the implicit ratings and users who have rated less than five movies from the BookCrossing data set. Therefore, the data set , after filtering, includes 3586 users, 7602 movies and 84981 ratings. Our Hybrid methods are built on top of the predicted rating matrices and explainability graphs obtained from the individual single-style Explainable Matrix Factorization methods [3]. We restrict the number of individual EMF models used to evaluate our proposed hybrid methods to the following two models: user-based EMF, \(EMF_{UB}\) and item-based EMF, \(EMF_{IB}\).

<table>
<thead>
<tr>
<th>TABLE 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summary of Data Used for Evaluating our Proposed Hybrid Methods</td>
</tr>
<tr>
<td>Domain</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>MovieLens(100k)</td>
</tr>
<tr>
<td>BookCrossing</td>
</tr>
</tbody>
</table>

\(^1\)https://grouplens.org/datasets/movielens/
\(^2\)http://www2.informatik.uni-freiburg.de/ cziegler/BX/
4.2.1 Experimental Setting

We split the rating data into three sets with 80% assigned to the training set, 10% assigned to the validation set, for tuning the hybrid methods hyper-parameters, and the remaining 10% assigned to the test set used to calculate all the reported metrics. After learning the hybrid methods hyper-parameters, we trained the hybrid model on training data that combines the training set and the validation set. The $EMF_{UB}$ and $EMF_{IB}$ models were tuned using 5-fold cross-validation and random search to find the model hyper-parameters $\alpha$, $\beta$ and $\lambda$ where $\alpha$ is the learning rate, $\beta$ is the regularization term and $\lambda$ is the explainability term coefficient. We optimized the models’ parameters using Stochastic Gradient Descent. Figure 15 shows the flowchart for evaluating our proposed hybrid Multi-style EMF methods.

We compare the performance of our proposed hybrid models to basic Matrix Factorization [6], user-based EMF and item-based EMF [4] [3].

4.2.2 Explainability Coverage

We evaluate the explainability coverage of all EMF methods in the different domains using Eq. 58.

<table>
<thead>
<tr>
<th>Explanation Style</th>
<th>Book Domain Coverage (%)</th>
<th>Movie Domain Coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSE only</td>
<td>14.10</td>
<td>51.18</td>
</tr>
<tr>
<td>ISE only</td>
<td>6.76</td>
<td>55.10</td>
</tr>
<tr>
<td>NSE and ISE</td>
<td>2.54</td>
<td>43.06</td>
</tr>
<tr>
<td>Multi-style</td>
<td><strong>18.31</strong></td>
<td><strong>63.22</strong></td>
</tr>
</tbody>
</table>

TABLE 13

Explainability Coverage in Book and Movie Domains
Table 13 shows that using multiple explanation styles increases the number of explainable user-item pairs. 43% of all user-item pairs can be explained using both NSE explanation style and ISE explanation style in the movie domain. However, 63% of all user-item pairs can be explained using either or both NSE and ISE explanation style. That means that our proposed hybrid methods increase the explainability coverage by leveraging the predicted ratings and explanation styles of all the individual EMF models. Therefore, the explanations generated by everyone of the individual EMF models become valid expla-
nations for the recommended items. In the book domain, using multiple explanation styles provides explanations for 18.31% of all user-item pairs. This is due to the higher sparsity of the BookCrossing data set used in our evaluation as shown in Table 12.

4.2.3 Evaluation of Weighted Multi-style Explainable Matrix Factorization (W-MEMF)

Figure 16: Tuning the Weighted Multi-style EMF (W-MEMF) Methods using the Validation Set (Movie Domain) (a) RMSE vs weight (b) MAP@10 vs weight (c) NDCG@10 vs weight

Figure 16 shows the result of tuning our proposed Weighted-MEMF methods, in the Movie domain, where weight is the value of the non-negative weights assigned to the individual EMF models. The values of weight were varied between 0 and 1 by increments of 0.1. In the Global W-MEMF method, when weight = 0, the predicted rating $\hat{r}_{global}$, is based on only the predicted ratings from the $EMF_{IB}$ model and when weight = 1, $\hat{r}_{global}$ is based on only the predicted rating from the $EMF_{UB}$ model. Figure 16 shows the performance of our proposed W-MEMF method, on the validation set while varying weight. Figure 16 (a) shows that W-MEMF methods have the least error when weight = 0.2. Figure 16 (b) shows the best performing model is the personalized W-MEMF method for MAP@10.
Finally, Figure 16 (c) shows that the Global W-MEMF method performs best, in terms of NDCG@10, when $weight = 0.1$.

Figure 17 shows the results of tuning our proposed Weighted-MEMF methods, in the Book domain. The hybrid method performs best, in terms of RMSE, when $weight = 0.3$ using the Global W-MEMF method. When evaluating MAP@10, the Global W-MEMF approach also performs best when $weight = 0.1$. Finally, when evaluating NDCG@10, the Global W-MEMF approach performs best when $weight = 0.1$ and 0.4.

![Figure 17: Tuning the Weighted Multi-style Methods using the Validation Set (Book Domain)](image)

(a) RMSE vs weight  (b) MAP@10 vs weight  (c) NDCG@10 vs weight
4.2.4 Regression-based Multi-style Explainable Matrix Factorization (R-MEMF)

Figure 18 shows the results of tuning the hybrid hyper-parameter for our proposed Regression-based Multi-style EMF (R-MEMF) methods in the movie domain. The hybrid model performs best using the Lasso regularization approach for all three accuracy metrics. For RMSE, the best $\lambda = 0.0001$, for MAP@10, the best $\lambda = 0.01$ and NDCG@10, the best $\lambda = 1e-5$. In the book domain, the Lasso regularization approach performs best for RMSE and NDCG@10 using $\lambda = 0.0001$ and $\lambda = 0.001$, respectively. For MAP@10, the linear approach performs best.
4.2.5 Switched Multi-style Explainable Matrix Factorization (S-MEMF)
Figure 20 shows the results of tuning our Switching hybrid models where $\alpha$ is the smoothing parameter used in the smoothing maximum function used by the mostrelevant switching condition. We used values between -10 and 10 to estimate the performance of the smoothing maximum function as $\alpha \in (\infty, \infty)$. Figure 20 (a) shows that the hybrid predicted rating matrix has the least squared error when $\alpha = 0$. Figure 20 (b) shows that the best MAP@10 was observed using the most-relevant selection criteria for $\alpha = 10$. However, using the most-explainable selection criteria works almost as well as the most-relevant selection criteria for this metric. Finally, Figure 20 (c) shows that the best performing selection criteria for the NDCG@10 metric was the most-relevant selection criteria when $\alpha = -0.5$.

Figure 21: Tuning the Switched Multi-style EMF (S-MEMF) Methods using the Validation Set (Book Domain) (a) RMSE vs $\alpha$ (b) MAP@10 vs $\alpha$ (c) NDCG@10 vs $\alpha$
4.2.6 Evaluation of Asymmetric Multi-style Explainable Matrix Factorization (A-MEMF)

Figure 22: Tuning the Asymmetric Multi-style Methods using the Validation Set (Movie Domain) (a) RMSE vs iterations (b) MAP@10 vs iterations (c) NDCG@10 vs iterations

Figure 22 shows the results of tuning the Asymmetric Multi-style EMF methods using the validation set. In the movie domain, the Item-Anchored approach performs best for RMSE after training the model for 20 iterations. For MAP@10 and NDCG@10, the User-Anchored approach performs best after training the model for 10 iterations. In the book domain, shown in Figure 23, the best performance for RMSE is observed at 5 iterations using the Item-anchored approach. The Item-anchored approach also performs best for NDCG@10 after 100 iterations. For NDCG@10, the User-anchored approach performed best after 5 iterations.
4.2.7 Evaluation on Test Data

4.2.7.1 Evaluation of Recommendation Accuracy

Table 14 shows the result of evaluating our proposed methods and the baseline methods using the test data. Our proposed method, Asymmetric Multi-style EMF, using the Item-Anchored approach (A-MEMF (Item-Anchored)), outperformed the baseline methods for RMSE. The improvement in performance is significant, as shown in Table 15, at p-value < .05 when the mean performance is compared to the mean performance of the baseline methods. Our proposed Regression-based Multi-style EMF method, using the Lasso regularization (R-MEMF(Lasso)), outperformed the baseline methods for NDCG@10 and the improvement is also significant when compared to MF and EMFUB. However, it is moderately significant when compared to EMFIB as shown in Table 16.

Table 17 shows the result of evaluating our proposed methods and the baseline methods using the test set in the book domain. The best performing algorithms in this domain were
Item-based EMF and basic Matrix Factorization, denoted as $EMF_{IB}$ and $MF$, respectively, closely followed in terms of NDCG by our proposed switched multi-style S-MEMF (transparent) approach.

### TABLE 14

Results of Hybrid Models on Test Data (Movie Domain). Best results are in bold.

<table>
<thead>
<tr>
<th>method</th>
<th>RMSE</th>
<th>MAP@10</th>
<th>NDCG@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF</td>
<td>0.19432 ± 0.00014</td>
<td>0.00073 ± 0.001</td>
<td>0.7961 ± 0.005</td>
</tr>
<tr>
<td>$EMF_{UB}$</td>
<td>0.19025 ± 0.00031</td>
<td>0.00049 ± 0.0002</td>
<td>0.7971 ± 0.006</td>
</tr>
<tr>
<td>$EMF_{IB}$</td>
<td>0.18518 ± 0.00054</td>
<td>0.00104 ± 0.0006</td>
<td>0.7988 ± 0.005</td>
</tr>
<tr>
<td>A-MEMF (Item-Anchored)</td>
<td><strong>0.18415 ± 0.0007</strong></td>
<td>7.4074e−5 ± 1e−5</td>
<td><strong>0.80118 ± 0.0052</strong></td>
</tr>
<tr>
<td>A-MEMF (User-Anchored)</td>
<td>0.18855 ± 0.0007</td>
<td>0.00013 ± 0.0</td>
<td>0.7997 ± 0.0054</td>
</tr>
<tr>
<td>S-MEMF(Smooth)</td>
<td>0.18569 ± 0.0004</td>
<td>8.5580e−5 ± 2e−5</td>
<td>0.7991 ± 0.004</td>
</tr>
<tr>
<td>R-MEMF (Lasso)</td>
<td>0.1890 ± 0.00069</td>
<td>3.1531e−5 ± 3e−5</td>
<td><strong>0.8016 ± 0.0047</strong></td>
</tr>
<tr>
<td>W-MEMF (Global)</td>
<td>0.18479 ± 0.00038</td>
<td>7.5621e−5 ± 1e−5</td>
<td>0.7994 ± 0.007</td>
</tr>
<tr>
<td>W-MEMF (Personalized)</td>
<td>0.18533 ± 0.0005</td>
<td>6.8417e−5 ± 2e−5</td>
<td>0.7992 ± 0.005</td>
</tr>
</tbody>
</table>

### TABLE 15

RMSE significance test results in the movie domain. Best results are in bold font.

<table>
<thead>
<tr>
<th>Model 1</th>
<th>Model 2</th>
<th>$P$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF</td>
<td>A-MEMF (Item-Anchored)</td>
<td>1.4532e-19</td>
</tr>
<tr>
<td>$EMF_{UB}$</td>
<td>A-MEMF (Item-Anchored)</td>
<td>3.7647e-15</td>
</tr>
<tr>
<td>$EMF_{IB}$</td>
<td>A-MEMF (Item-Anchored)</td>
<td>1.5673e-3</td>
</tr>
</tbody>
</table>
TABLE 16

NDCG@10 significance test results in the movie domain. Best results are in bold font.

<table>
<thead>
<tr>
<th>Model 1</th>
<th>Model 2</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF</td>
<td>R-MEMF (Lasso)</td>
<td>0.0133</td>
</tr>
<tr>
<td>$EMF_{UB}$</td>
<td>R-MEMF (Lasso)</td>
<td>0.0467</td>
</tr>
<tr>
<td>$EMF_{IB}$</td>
<td>R-MEMF (Lasso)</td>
<td>0.1239</td>
</tr>
</tbody>
</table>

TABLE 17

Results of Hybrid Models on Test Data (Book Domain). Best results are in bold font.

<table>
<thead>
<tr>
<th>method</th>
<th>RMSE</th>
<th>MAP@10</th>
<th>NDCG@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF</td>
<td>0.1668 ± 0.0004</td>
<td>0.00018 ± 0.0000</td>
<td>0.8599 ± 0.0031</td>
</tr>
<tr>
<td>$EMF_{UB}$</td>
<td>0.1651 ± 0.0003</td>
<td>0.00024 ± 0.0000</td>
<td>0.8422 ± 0.0019</td>
</tr>
<tr>
<td>$EMF_{IB}$</td>
<td>0.1646 ± 0.0002</td>
<td>0.00028 ± 0.0007</td>
<td>0.8451 ± 0.0049</td>
</tr>
<tr>
<td>A-MEMF (Item-Anchored)</td>
<td>0.1807 ± 0.0012</td>
<td>0.00036 ± 0</td>
<td>0.8324 ± 0.0090</td>
</tr>
<tr>
<td>A-MEMF (User-Anchored)</td>
<td>0.1694 ± 0.0002</td>
<td>0.00030 ± 0</td>
<td>0.8418 ± 0.0010</td>
</tr>
<tr>
<td>S-MEMF (Smooth)</td>
<td>0.1708 ± 0.0008</td>
<td>0.00019 ± 0</td>
<td>0.8486 ± 0.0070</td>
</tr>
<tr>
<td>S-MEMF (Transparent)</td>
<td>0.1804 ± 0.0011</td>
<td>0.00036 ± 0</td>
<td>0.8537 ± 0.0040</td>
</tr>
<tr>
<td>R-MEMF (Lasso)</td>
<td>0.2064 ± 0.0018</td>
<td>9.5841 ± 0</td>
<td>0.8477 ± 0.1040</td>
</tr>
<tr>
<td>R-MEMF (Linear)</td>
<td>0.2064 ± 0.00013</td>
<td>0.00018 ± 0</td>
<td>0.8437 ± 0.0050</td>
</tr>
<tr>
<td>W-MEMF (Global)</td>
<td>0.1716 ± 0.0008</td>
<td>9.33247 ± 0</td>
<td>0.8435 ± 0.003</td>
</tr>
<tr>
<td>W-MEMF (Personalized)</td>
<td>0.1663 ± 0.0012</td>
<td>8.2769 ± 0</td>
<td>0.8452 ± 0.005</td>
</tr>
</tbody>
</table>

Table 17 shows the result of evaluating our proposed methods and the baseline methods using the test set. In terms of RMSE, the best performing model was the User-based EMF model. In terms of MAP@10, our proposed method, Switched Multi-style EMF, using transparency as the selection criteria, was the best performing model. The improvement in performance, compared to the baseline methods, was significant and is shown in Table 18.
at p-value < .05 when the mean performance is compared with the mean performance of the baseline methods.

### TABLE 18

MAP@10 significance test results in the book domain. Best results are in bold font.

<table>
<thead>
<tr>
<th>Model 1</th>
<th>Model 2</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF</td>
<td>S-MEMF (Transparent)</td>
<td>2.3393e-35</td>
</tr>
<tr>
<td>$EMF_{UB}$</td>
<td>S-MEMF (Transparent)</td>
<td>7.4028e-37</td>
</tr>
<tr>
<td>$EMF_{IB}$</td>
<td>S-MEMF (Transparent)</td>
<td>5.9854e-28</td>
</tr>
</tbody>
</table>

#### 4.2.7.2 Explainability Evaluation

In this section, we compared the explainability performance of our proposed methods with that of the baseline methods, using objective explainability metrics presented in Section 4.1.2. These metrics are an aggregation of the explainability scores that have been validated in prior work [14] that conducted user studies that found a higher subjective perception of transparency among user-item pairs for items that have higher objective NSE and ISE style explanation scores.

In the following, we present our results for each of the two domains, first movies then books.

1. **Movie Domain**

   Figure 24 shows the MEP@10 metric results in the movie domain. The Asymmetric Multi-style EMF (A-MEMF) outperformed the other methods when evaluated with the user-based neighborhood explainability style and the item-based neighborhood explainability style. Tables 19 and 20 show the results of the significance tests for MEP@10 values, using the user-based neighborhood explainability and item-based neighborhood explainability, respectively. We evaluated the significance using the explainability thresholds $\theta^u = 0.1$ and $\theta^i = 0.1$ for the user-based and item-based neighborhood explainability styles, respectively. Tables 21 and 22 show the results
of the significance tests for MER@10 values using the explainability threshold values $\theta^n = 0.1$ and $\theta^i = 0.1$. The result shows that the R-MEMF method, using the Lasso regularization approach, performed significantly better than the baseline methods.

Figure 24: Figure (a): MEP@10 vs explainability threshold $\theta^n$ using NSE Explainability Graph. Figure (b): MER@10 vs explainability threshold $\theta^n$. (Movie Domain)
Figure 25: Figure (a): \textit{MEP@10 vs explainability threshold $\theta^i$ using ISE Explainability Graph}. Figure (b): \textit{MER@10 vs explainability threshold $\theta^i$.} (Movie Domain)

\begin{table}[h]
\centering
\caption{MEP@10 significance test results in the movie domain (NSE-explainablity) — $\theta^n = 0.05$.}
\begin{tabular}{|l|l|l|}
\hline
Model 1 & Model 2 & $P$-value \\
\hline
MF & \textbf{R-MEMF(Lasso)} & 4.7787e-4 \\
$EMF_{UB}$ & \textbf{R-MEMF(Lasso)} & 6.1312e-10 \\
$EMF_{IB}$ & \textbf{R-MEMF(Lasso)} & 2.6478e-12 \\
\hline
\end{tabular}
\end{table}

The winning method is in bold.
TABLE 20

MEP@10 significance test results in the movie domain (ISE-explainablity) — $\theta^i = 0.01$.

The winning method is in bold

<table>
<thead>
<tr>
<th>Model 1</th>
<th>Model 2</th>
<th>$P$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF</td>
<td>R-MEMF(Lasso)</td>
<td>0.0151</td>
</tr>
<tr>
<td>$EMF_{UB}$</td>
<td>R-MEMF(Lasso)</td>
<td>3.1564e-7</td>
</tr>
<tr>
<td>$EMF_{IB}$</td>
<td>R-MEMF(Lasso)</td>
<td>6.3748e-7</td>
</tr>
</tbody>
</table>

TABLE 21

MER@10 significance test results in the movie domain (NSE-explainablity) — $\theta^n = 0.05$.

The winning method is in bold

<table>
<thead>
<tr>
<th>Model 1</th>
<th>Model 2</th>
<th>$P$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF</td>
<td>R-MEMF(Lasso)</td>
<td>6.3028e-4</td>
</tr>
<tr>
<td>$EMF_{UB}$</td>
<td>R-MEMF(Lasso)</td>
<td>1.6928e-9</td>
</tr>
<tr>
<td>$EMF_{IB}$</td>
<td>R-MEMF(Lasso)</td>
<td>3.3550e-12</td>
</tr>
</tbody>
</table>

TABLE 22

MER@10 significance test results in the movie domain (ISE-explainablity) — $\theta^i = 0.01$.

The winning method is in bold

<table>
<thead>
<tr>
<th>Model 1</th>
<th>Model 2</th>
<th>$P$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF</td>
<td>R-MEMF(Lasso)</td>
<td>0.0280</td>
</tr>
<tr>
<td>$EMF_{UB}$</td>
<td>R-MEMF(Lasso)</td>
<td>1.7596e-6</td>
</tr>
<tr>
<td>$EMF_{IB}$</td>
<td>R-MEMF(Lasso)</td>
<td>2.2289e-6</td>
</tr>
</tbody>
</table>

2. Book Domain
In the book domain, basic MF outperformed our proposed hybrid methods, in terms of...
MER@10 using both the NSE explainability graph and the ISE explainability graph. In terms of MEP@10, using the NSE explainability graph, there is no clear winner as shown in Figure 26 (a). However, when $\theta^n > 0.04$, the basic MF model outperforms other models.

4.2.7.3 Examples

The following examples show the multi-style explanations for two anonymous sample users, denoted as Sample User A and Sample User B, in the movie domain. For each user, we show the sample user’s top-3 rated movies, from the training set, and the movies recommended to the user using a Matrix Factorization-based recommender system model. A user will be considered to “like” a movie if the rating is 3 or above, out of a maximum of 5. We also show the top-3 recommended movies using $EMF_{UB}$ and $EMF_{IB}$ and their corresponding explanations. Finally, we show the top-3 movies recommended by the hybrid multi-style EMF methods and their corresponding multi-style explanations using R-MEMF (lasso regularization), W-MEMF (global weights) and S-MEMF (Smoothing Maximum criterion). For both users, the R-MEMF (lasso) model’s recommendations were generated with $\beta_{NSE} = 0.4993$ and $\beta_{ISE} = 1.4354$ which implies that $EMF_{IB}$ was had more influence in calculating the users’ predicted ratings for the recommended movies. The W-MEMF (global) model’s recommendations were generated with weight = 0.6 which implies there was slightly more weight attributed to predicted ratings from $EMF_{UB}$ by the hybrid model. The S-MEMF (smooth) model’s recommendations were generated with $\alpha = 10$ which implies that for each user-item pair, the EMF method with the highest predicted rating was selected as the hybrid model’s predicted rating.

1. Sample User A

Table 23 shows the top-3 movies rated by Sample User A, from the training data, in the movie domain. Table 24 shows the top-3 movies recommended by the MF model. Being a black box model, MF’s recommendations have no accompanying
explanations. Table 25 shows the top-3 movies recommended by the user-based EMF model, $EMF_{UB}$. The accompanying explanations are shown in Figure 28, using the NSE explanation style. Table 26 shows the top-3 movies recommended by the item-based EMF model, $EMF_{IB}$ and the corresponding explanations are shown in Figure 29, using the Influence Style Explanation (ISE). Tables 27, 28 and 29 show the top-3 recommended movies using the S-MEMF, W-MEMF and R-MEMF hybrid methods, respectively.

Figure 30 shows the presentation of the multiple explanation styles where the recommended movies are explained to the user, calculated using S-MEMF with $\alpha = 10.0$, using the explanation style with the highest explanation score for the recommended movie. Figure 31 shows the presentation of the multiple explanation styles, calculated using W-MEMF with weight $= 0.6$ (hence attributing slightly more weight to $EMF_{UB}$), where the recommended movies are explained to the sample user using both NSE and ISE explanation styles. Figure 32 shows the presentation of the multiple explanation styles, calculated using R-MEMF with lasso regularization, which resulted in regression coefficients $\beta_{NSE} = 0.4993$ and $\beta_{ISE} = 1.4354$. Therefore, the ISE explanation style is presented to the sample user for explaining the recommended movies because of its higher feature coefficient.

Figures 30, 31 and 32 show the importance of the multi-style explanation methods which insight into the contributions of the individual EMF models and also how the presented explanation styles were selected. Therefore in addition to our proposed methods (S-MEMF, W-MEMF and R-MEMF) being accurate, and to improving explainability and coverage relative to the single style baseline EMF methods, they have the added advantage of being more transparent.
### TABLE 23

Top-3 Movies rated by Sample User A. Movies are ranked in descending order of ratings.

<table>
<thead>
<tr>
<th>Top-3 rated movies</th>
<th>Genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gone with the Wind (1939)</td>
<td>Drama/Romance</td>
</tr>
<tr>
<td>Dirty Dancing (1987)</td>
<td>Romance/Dance</td>
</tr>
<tr>
<td>Adventures of Priscilla, Queen of the Desert (1994)</td>
<td>Musical</td>
</tr>
</tbody>
</table>

### TABLE 24

Top-3 Movies Recommended for Sample User A using Matrix Factorization (MF) Model. Movies are ranked in descending order of the predicted ratings.

<table>
<thead>
<tr>
<th>Top-3 recommended movies</th>
<th>Genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>Niagara, Niagara (1997)</td>
<td>Drama/Romance</td>
</tr>
<tr>
<td>Sweet Nothing (1995)</td>
<td>Drama</td>
</tr>
<tr>
<td>The Eighth Day (1996)</td>
<td>Drama/Comedy</td>
</tr>
</tbody>
</table>

### TABLE 25

Top-3 Movies Recommended for Sample User A using User-based Explainable Matrix Factorization ($EMF_{UB}$) Model. Movies are ranked in descending order of the predicted ratings.

<table>
<thead>
<tr>
<th>Top-3 recommended movies</th>
<th>Genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beyond Bedlam (1993)</td>
<td>Drama</td>
</tr>
<tr>
<td>Bliss (1997)</td>
<td>Drama/Romance</td>
</tr>
<tr>
<td>Hearts and Minds (1996)</td>
<td>Drama/History</td>
</tr>
</tbody>
</table>
Figure 28: Corresponding Explanations for movies recommended to Sample User A in Table 25 using the NSE-style generated by the $EMF_{UB}$ model. Top: First recommended movie. Middle: Second recommended movie. Bottom: Third recommended movie.
TABLE 26

Top-3 Movies Recommended for Sample User A using Item-based Explainable Matrix Factorization ($EMF_{IB}$) Model. Movies are ranked in descending order of the predicted ratings.

<table>
<thead>
<tr>
<th>Top-3 recommended movies</th>
<th>Genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw Deal (1948)</td>
<td>Crime</td>
</tr>
<tr>
<td>Buddy (1997)</td>
<td>Family/Drama</td>
</tr>
</tbody>
</table>
Figure 29: Corresponding Explanations for movies recommended to Sample User A in Table 26 using the ISE-style generated by the $EMF_{IB}$ model. Top: First recommended movie. Middle: Second recommended movie. Bottom: Third recommended movie.
TABLE 27

Top-3 Movies Recommended for Sample User A using S-MEMF (Switched Multi-style Explainable Matrix Factorization using Smooth Maximum Switching Criterion) with $\alpha = 10.0$. Movies are ranked in descending order of the predicted ratings.

<table>
<thead>
<tr>
<th>Top-3 recommended movies</th>
<th>Genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beyond Bedlam (1993)</td>
<td>Drama</td>
</tr>
<tr>
<td>Raw Deal (1948)</td>
<td>Crime</td>
</tr>
</tbody>
</table>
Figure 30: Corresponding Explanations for movies recommended to Sample User A in Table 27 using the Multi-style Explanations generated by the S-MEMF (Switched Multistyle Explainable Matrix Factorization using Smooth Maximum Switching Criterion) model with $\alpha = 10.0$. Top and Middle: ISE style explanations for top 2 recommended movies. Bottom: NSE style explanation for third recommended movie.
TABLE 28

Top-3 Movies Recommended for Sample User A using W-MEMF (Global) with weight = 0.6 (Slightly more weight attributed to $EMF_{UB}$ than $EMF_{IB}$). Movies are ranked in descending order of the predicted ratings.

<table>
<thead>
<tr>
<th>Top-3 recommended movies</th>
<th>Genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Forbidden Christ (1950)</td>
<td>Drama/Mystery</td>
</tr>
<tr>
<td>Scream of Stone (1991)</td>
<td>Drama</td>
</tr>
<tr>
<td>Fire on the Mountain (1996)</td>
<td>Documentary</td>
</tr>
</tbody>
</table>
Figure 31: Corresponding Explanations (Left: NSE style, Right: ISE style) for movies recommended to Sample User A in Table 28 using the Multi-style Explanation generated using the W-MEMF (Global) model with weight = 0.6 (Slightly more weight attributed to $EMF_{UB}$ than $EMF_{IB}$). Top, Middle, Bottom: Explanations for first, second, and third movie, respectively
Top-3 Movies Recommended for User 13 using R-MEMF (Lasso) with $\beta_{NSE} = 0.4993$ and $\beta_{ISE} = 1.4354$. Movies are ranked in descending order of the predicted ratings.

<table>
<thead>
<tr>
<th>Top-3 recommended movies</th>
<th>Genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Koldum Klaka (1994)</td>
<td>Drama</td>
</tr>
<tr>
<td>Scream of Stone (1991)</td>
<td>Drama</td>
</tr>
<tr>
<td>The Forbidden Christ (1950)</td>
<td>Drama/Mystery</td>
</tr>
</tbody>
</table>
Figure 32: Corresponding Explanations for movies recommended to Sample User A in Table 29 using the Multi-style Explanation generated with the R-MEMF (lasso) model with $\beta_{NSE} = 0.4993$ and $\beta_{ISE} = 1.4354$ (Only ISE Explanation Style presented to sample user). Top: Explanations for first movie. Middle: Explanations for second movie. Bottom: Explanation for third movie.

Figure 28 shows the NSE explanations, which say that the recommended movies are liked by most of the users with similar interests to User A. The ISE explanations, shown in Figure 29, say that this user has previously liked most of the movies that
are similar to the recommended movies.

In the case, depicted in Figure 30, the multi-style approach successfully chose the right explanation style for each movie (namely, ISE for the top 2 movies and NSE for the third recommended movie). The ISE explanations say that this user liked several movies that are similar to the top 2 recommended movies. The NSE explanation say that the third recommended movie is liked by most of the users with similar interests to User A.

The NSE explanations on the left, in Figure 31, show no evidence that similar users liked any of the recommended movies. However, the ISE explanations on the right say that this user liked several movies that are similar to all three recommended movies. This example shows how different explanation styles give the user a chance to scrutinize the predicted recommendations and make a better informed decision, compared to either no explanations at all (the case of Black Box methods), or a single style explanation.

2. **Sample User B**

Table 30 shows the top-3 movies rated by Sample User B, from the training data, in the movie domain. Table 31 shows the top-3 movies recommended by the MF model. The recommendations have no accompanying explanations. Table 32 shows the top-3 movies recommended by the user-based EMF model, $EMF_{UB}$. The accompanying explanations are shown in Figure 33, using the NSE explanation style, where only the top-2 movies can be explained using this explanation style. The explanations basically say that there is no strong evidence that the recommended movies are liked by most of the users with similar interests to User B. The explanations empower the user with more information to scrutinize the model’s output predictions. Table 33 shows the top-3 movies recommended by the item-based EMF model, $EMF_{IB}$ and the corresponding explanations are shown in Figure 34, using the Influence Style Explanation (ISE). Tables 34 and 35 show the top-3 recommended movies using the S-MEMF and W-MEMF hybrid methods, respectively.
Figure 35 shows the presentation of the multiple explanation styles where the recommended movies are explained to the user, calculated using S-MEMF with $\alpha = 10.0$, using the explanation style with the highest explanation score for the recommended movie.

Figure 36 shows the presentation of the multiple explanation styles, calculated using W-MEMF with $weight = 0.6$ (hence attributing slightly more weight to $EMF_{UB}$), where the recommended movies are explained to the sample user using both NSE and ISE explanation styles.

Figure 37 shows the presentation of the multiple explanation styles, calculated using R-MEMF with lasso regularization, which resulted in regression coefficients $\beta_{NSE} = 0.4993$ and $\beta_{ISE} = 1.4354$. Therefore, the ISE explanation style is presented to the sample user for explaining the recommended movies because of its higher feature coefficient.

Figures 35, 36 and 37 show the importance of the multi-style explanation methods which insight into the contributions of the individual EMF models and also how the presented explanation styles were selected. Therefore in addition to our proposed methods (S-MEMF, W-MEMF and R-MEMF) being accurate, and to improving explainability and coverage relative to the single style baseline EMF methods, they have the added advantage of being more transparent.

**TABLE 30**

Top-3 Movies rated by Sample User B. Movies are ranked in descending order of the ratings.

<table>
<thead>
<tr>
<th>Top-3 rated movies</th>
<th>Genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>French Twist (1995)</td>
<td>Comedy/Romance</td>
</tr>
<tr>
<td>Four Weddings and a Funeral (1994)</td>
<td>Romance/Comedy</td>
</tr>
</tbody>
</table>
### TABLE 31

Top-3 Movies recommended for Sample User B using Matrix Factorization (MF) Model. Movies are ranked in descending order of the predicted ratings.

<table>
<thead>
<tr>
<th>Top-3 recommended movies</th>
<th>Genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homage (1995)</td>
<td>Drama/Thriller</td>
</tr>
<tr>
<td>Sleepover (1995)</td>
<td>Drama</td>
</tr>
<tr>
<td>The Eighth Day (1996)</td>
<td>Drama/Comedy</td>
</tr>
</tbody>
</table>

### TABLE 32

Top-3 Movies Recommended for Sample User B using User-based Explainable Matrix Factorization (EMF$_{UB}$) Model. Movies are ranked in descending order of the predicted ratings.

<table>
<thead>
<tr>
<th>Top-3 recommended movies</th>
<th>Genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hedd Wyn (1992)</td>
<td>Drama/Romance</td>
</tr>
<tr>
<td>The Innocent Sleep (1995)</td>
<td>Thriller/Drama</td>
</tr>
<tr>
<td>Germinal (1993)</td>
<td>Drama/Romance</td>
</tr>
</tbody>
</table>
Figure 33: Corresponding Explanations for the movies recommended to Sample User B in Table 32 using the NSE-style generated by the $EMF_{UB}$ model. Top: First recommended movie. Middle: Second recommended movie. Bottom: Third recommended movie.
TABLE 33

Top-3 Movies Recommended for Sample User B using Item-based Explainable Matrix Factorization ($EMF_{IB}$) Model. Movies are ranked in descending order of the predicted ratings.

<table>
<thead>
<tr>
<th>Top-3 recommended movies</th>
<th>Genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>Squeeze (1996)</td>
<td>Drama/Crime</td>
</tr>
<tr>
<td>Everest (1998)</td>
<td>Documentary/Short</td>
</tr>
</tbody>
</table>
Figure 34: Corresponding Explanations recommended to Sample User B in Table 33 using the ISE-style generated by the $EMF_{IB}$ model. Top: First recommended movie. Middle: Second recommended movie. Bottom: Third recommended movie.
TABLE 34

Top-3 Movies Recommended for Sample User B using S-MEMF (Switched Multi-style Explainable Matrix Factorization using Smooth Maximum Switching Criterion with $\alpha = 10.0$). Movies are ranked in descending order of predicted ratings.

<table>
<thead>
<tr>
<th>Top-3 recommended movies</th>
<th>Genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>Squeeze (1996)</td>
<td>Romance/Drama</td>
</tr>
<tr>
<td>Mad Dog Time (1996)</td>
<td>Adaptation/Drama</td>
</tr>
<tr>
<td>La Vie est Belle (1987)</td>
<td>Sci-fi/Action</td>
</tr>
</tbody>
</table>
Figure 35: Corresponding Explanations for movies recommended to Sample User B in Table 34 using the Multi-style Explanations generated by the S-MEMF (Switched Multi-style Explainable Matrix Factorization using Smooth Maximum Switching Criterion) model with $\alpha = 10.0$. Top: NSE style explanation for top recommended movie. Middle: ISE style explanations for second movie. Bottom: NSE style explanation for third recommended movie.
TABLE 35

Top-3 Movies Recommended for Sample User B using W-MEMF (Global) with $weight = 0.6$: This means that slightly more weight was attributed to $EMF_{UB}$ than $EMF_{IB}$. Movies are ranked in descending order of ratings.

<table>
<thead>
<tr>
<th>Top-3 recommended movies</th>
<th>Genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hedd Wyn (1992)</td>
<td>Drama/Romance</td>
</tr>
<tr>
<td>The Forbidden Christ (1950)</td>
<td>Drama/Mystery</td>
</tr>
<tr>
<td>Squeeze (1996)</td>
<td>Drama/Crime</td>
</tr>
</tbody>
</table>
Figure 36: Explanation Styles (Left: NSE, Right: ISE) for recommended movies in Table 35 using the Multi-style Explanations generated using the W-MEMF (Global) model with \( \text{weight} = 0.6 \) (Slightly more weight attributed to \( EMF_{UB} \) than \( EMF_{IB} \)). Top, Middle, Bottom: Explanations for first, second, and third movie, respectively
TABLE 36

Top-3 Movies Recommended for Sample User B using R-MEMF (Lasso) with $\beta_{NSE} = 0.4993$ and $\beta_{ISE} = 1.4354$. Movies are ranked in descending order of the predicted ratings.

<table>
<thead>
<tr>
<th>Top-3 recommended movies</th>
<th>Genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Forbidden Christ (1950)</td>
<td>Drama/Mystery</td>
</tr>
<tr>
<td>Squeeze (1996)</td>
<td>Drama/Crime</td>
</tr>
<tr>
<td>Everest (1998)</td>
<td>Documentary/Short</td>
</tr>
</tbody>
</table>
Figure 37: Corresponding Explanations for recommended movies in Table 36 using the Multi-style Explanation generated using the R-MEMF (lasso) model with $\beta_{NSE} = 0.4993$ and $\beta_{ISE} = 1.4354$ (Hence, only the ISE Explanation Style is presented to the user). Top: Explanations for first movie. Middle: Explanations for second movie. Bottom: Explanation for third movie.

The NSE explanation for the third movie, shown in Figure 35, basically says that the recommended movie is liked by most of the users with similar interests to User B. The NSE explanation for the top recommended movie says that this recommended
movie is liked by one user and disliked by another user, among users with similar interests to User B. The ISE explanation for the second recommended movie shows that User B liked two similar movies, while disliking another two similar movies. This example shows how different explanation styles give the user a chance to scrutinize the recommendations and make a better informed decision about whether to follow them, compared to either having no explanations at all (the case of Black Box methods such as MF), or single style explanation methods like EMF.

The NSE explanations, in Figure 36, on the left say that the recommended movies are not liked by the users with similar interests to User B. The ISE explanations on the right say that User B did not like movies that are similar to the recommended movies. This example shows how different explanation styles give the user a chance to scrutinize the recommendations and make a better informed decision about whether to follow them, compared to either having no explanations at all (the case of Black Box methods such as MF), or single style explanation methods like EMF.

The ISE explanations, in Figure 37, basically say that this user (User B) liked movies that are similar only to the third top recommended movie. However, the user did not like movies that are similar to the first or second top recommended movies, thus allowing the user to scrutinize the predictions and accept only one or none of the predictions.

4.2.7.4 Analysis of Results

In the movie domain, our proposed methods outperformed the baseline methods in terms of predicted accuracy, relevance of recommended movies to users, and explainability of the recommended movies using two explanation styles. The significance tests also show that our methods’ superior performance, compared to the baseline methods, was significant (p < .05). In the book domain, our proposed method, Switched Multi-style EMF (S-MEMF), using transparency as the selection criteria, performed better than the baseline methods in terms of MAP@10 and as well as the basic MF method in terms of NDCG@10.
Therefore, our proposed hybrid method produced more accurate recommendations with the added benefit of explaining the recommendations to users with the explanation style more closely aligned with their preferences.
4.3 Evaluation of Tag-boosted Multi-style Explainable Matrix Factorization

In this section, we present our offline experiments for evaluating the performance of our proposed Tag-boosted Multi-style EMF methods. We evaluate our model using the HetRec data set from the movie domain only since the book crossings data set did not have tags associated with both users and items, as in the case of the HetRec data.

4.3.1 Experimental Settings

The HetRec data consists of 2,113 users, 10,197 movies, 13,222 tags and 855,598 ratings. Similar to previous works on tag data in the literature, we applied some filters to the data to reduce the sparsity of the dataset and increase the strength of the tag-based relationships between users and movies. We selected users who had rated at least 50 unique movies and used at least 10 unique tags. We also selected movies that have been rated by at least 50 unique users and tagged with at least 10 unique tags. Our filtered data ended up with 264 users, 1239 movies, 5293 tags and 21,214 ratings. The filtered data was split into 2 sets: a training set consisting of 90% of each user’s known ratings and a test set consisting of 10% of each user’s known ratings. We normalized the ratings in the dataset, between 0 and 1, using the maximum rating value. The models hyper-parameters were learnt using 5-fold cross validation. We optimized all model parameters using Stochastic Gradient Descent. We compared our methods to the following baseline methods: Basic Matrix Factorization [6] and Explainable Matrix Factorization with two different explanation styles(user-based and item-based) [4] [3].

4.3.2 Evaluation on Test Data

4.3.2.1 Evaluation of Recommendation Accuracy

Tables 37, 38 and 39 show the results of comparing the accuracy of our proposed methods with the baseline methods while varying the number of latent factors, $K$. The results show that the proposed tag-boosted Multi-style EMF methods, denoted as RelTag

---

3https://grouplens.org/datasets/hetrec-2011/
and PrefTag, outperform the baseline methods. For RMSE, RelTag performed better than the baseline methods when $K = 10$ and $K = 50$. For MAP@10, which measures the mean relevance of the top-10 recommended movies for each user, the PrefTag method outperformed the baseline methods and thus recommends more relevant movies to the users. The PrefTag method performed best when $K = 20$ and $K = 50$. For NDCG@10, which considers the relative ranking of recommended movies compared to their true ranking for each user, the RelTag method outperformed the baseline methods and thus, recommended movies to users, in a ranked order, that closely aligned with their true movie rankings.

**TABLE 37**
RMSE vs Latent Factors (K) in the movie domain. $EMF_{TA}$, PrefTag, RelTag denote our proposed methods. Bold denote the best results

<table>
<thead>
<tr>
<th>K</th>
<th>MF</th>
<th>$EMF_{UB}$</th>
<th>$EMF_{IB}$</th>
<th>$EMF_{TA}$</th>
<th>PrefTag</th>
<th>RelTag</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>.1474 ± .0035</td>
<td><strong>.1423 ± .0014</strong></td>
<td>.1466 ± .009</td>
<td>.1469 ± .0023</td>
<td>.1543 ± .0026</td>
<td>.1520 ± .0009</td>
</tr>
<tr>
<td>10</td>
<td>.1347 ± .0042</td>
<td>.1326 ± .0016</td>
<td>.1331 ± .0019</td>
<td>.1332 ± .0024</td>
<td>.1724 ± .0017</td>
<td><strong>.1313 ± .0031</strong></td>
</tr>
<tr>
<td>20</td>
<td>.1541 ± .0019</td>
<td><strong>.1410 ± .0020</strong></td>
<td>.1491 ± .0021</td>
<td>.1566 ± .0010</td>
<td>.1655 ± .0009</td>
<td>.1411 ± .0018</td>
</tr>
<tr>
<td>50</td>
<td>.1641 ± .0025</td>
<td>.1707 ± .0015</td>
<td>.1685 ± .0038</td>
<td>.3037 ± .0004</td>
<td>.3238 ± .0003</td>
<td><strong>.1633 ± .0020</strong></td>
</tr>
</tbody>
</table>

**TABLE 38**
MAP@10 vs Latent Factors (K) in the movie domain. $EMF_{TA}$, PrefTag, RelTag denote our proposed methods

<table>
<thead>
<tr>
<th>K</th>
<th>MF</th>
<th>$EMF_{UB}$</th>
<th>$EMF_{IB}$</th>
<th>$EMF_{TA}$</th>
<th>PrefTag</th>
<th>RelTag</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>.0046 ± .0027</td>
<td><strong>.4713 ± .0000</strong></td>
<td>.3306 ± .2150</td>
<td>.0042 ± .0027</td>
<td>.0034 ± .0069</td>
<td>.0005 ± .0003</td>
</tr>
<tr>
<td>10</td>
<td><strong>.4245 ± .1407</strong></td>
<td>.2841 ± .2293</td>
<td>.0017 ± .0017</td>
<td>.0008 ± .0004</td>
<td>.0115 ± .0010</td>
<td>.0051 ± .0003</td>
</tr>
<tr>
<td>20</td>
<td>.0017 ± .0009</td>
<td>.0026 ± .0008</td>
<td>.0022 ± .0010</td>
<td>.0010 ± .0009</td>
<td><strong>.0157 ± .0006</strong></td>
<td>.0039 ± .0007</td>
</tr>
<tr>
<td>50</td>
<td>.0016 ± .0006</td>
<td>.0021 ± .0014</td>
<td>.0021 ± .0011</td>
<td>.0022 ± .0006</td>
<td><strong>.0152 ± .0006</strong></td>
<td>.0023 ± .0003</td>
</tr>
</tbody>
</table>

The significance tests on the results of our proposed methods resulted in p-values that are shown in Tables 40, 41 and 42. For all metrics, our proposed methods outperformed the baseline methods except when we compare the RMSE values of RelTag and basic MF.

Figure 38 shows the influence of the number of common relevant tags and preferred tags between similar items and users on our proposed methods. We observe that increasing
TABLE 39

NDCG@10 vs Latent Factors (K) in the movie domain. $EMF_{TA}$, PrefTag, RelTag denote our proposed methods

<table>
<thead>
<tr>
<th>K</th>
<th>MF</th>
<th>$EMF_{UB}$</th>
<th>$EMF_{IB}$</th>
<th>$EMF_{TA}$</th>
<th>PrefTag</th>
<th>RelTag</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>.8573 ± .0038</td>
<td>.8488 ± .0060</td>
<td>.8547 ± .0042</td>
<td>.8558 ± .0062</td>
<td>.8614 ± .0081</td>
<td>.8777 ± .0000</td>
</tr>
<tr>
<td>10</td>
<td>.8526 ± .0131</td>
<td>.8605 ± .0037</td>
<td>.8599 ± .0023</td>
<td>.8580 ± .0020</td>
<td>.8477 ± .0045</td>
<td>.8777 ± .0000</td>
</tr>
<tr>
<td>20</td>
<td>.8228 ± .0047</td>
<td>.8448 ± .0079</td>
<td>.8528 ± .0050</td>
<td>.8508 ± .0055</td>
<td>.8486 ± .0056</td>
<td>.8488 ± .0033</td>
</tr>
<tr>
<td>50</td>
<td>.8471 ± .0049</td>
<td>.8490 ± .0036</td>
<td>.8474 ± .0040</td>
<td>.8498 ± .0024</td>
<td>.8482 ± .0028</td>
<td>.8777 ± .0000</td>
</tr>
</tbody>
</table>

Figure 38: MAP@10 vs Number of Common Tags: Common Relevant Tags for RelTag and Common Preferred Tags for PrefTag

the number of common relevant tags between items improves the performance of the model, in terms of the mean number of relevant movies recommended users. However, the performance decreases beyond 9 common relevant tags between items. However, the impact is more noticeable using our proposed PrefTag method which performs best using 5 common preferred tags between users.

Figure 39 shows the influence of the number of common relevant tags and preferred
tags between similar items and users on our proposed methods. Our proposed RelTag method performs best using 3 common relevant tags between items.

**TABLE 40**

RMSE significance test results($K = 50$, Common Relevant Tags = 5).

<table>
<thead>
<tr>
<th>Model 1</th>
<th>Model 2</th>
<th>$P$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF</td>
<td>RelTag</td>
<td>0.2353</td>
</tr>
<tr>
<td>$EMF_{UB}$</td>
<td>RelTag</td>
<td>3.9481e-8</td>
</tr>
<tr>
<td>$EMF_{IB}$</td>
<td>RelTag</td>
<td>9.5791e-4</td>
</tr>
<tr>
<td>$EMF_{TA}$</td>
<td>RelTag</td>
<td>3.9526e-32</td>
</tr>
</tbody>
</table>
TABLE 41

MAP@10 significance test results (K = 20, Common Preferred Tags = 10). Bold denotes the winner

<table>
<thead>
<tr>
<th>Model 1</th>
<th>Model 2</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF</td>
<td>PrefTag</td>
<td>1.8827e-19</td>
</tr>
<tr>
<td>$EMF_{UB}$</td>
<td>PrefTag</td>
<td>2.0378e-19</td>
</tr>
<tr>
<td>$EMF_{IB}$</td>
<td>PrefTag</td>
<td>3.1445e-18</td>
</tr>
<tr>
<td>$EMF_{TA}$</td>
<td>PrefTag</td>
<td>8.1562e-20</td>
</tr>
</tbody>
</table>

TABLE 42

NDCG@10 significance test results (K = 50, Common Relevant Tags = 3). Bold denotes the winner

<table>
<thead>
<tr>
<th>Model 1</th>
<th>Model 2</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF</td>
<td>RelTag</td>
<td>5.7748e-15</td>
</tr>
<tr>
<td>$EMF_{UB}$</td>
<td>RelTag</td>
<td>1.8414e-13</td>
</tr>
<tr>
<td>$EMF_{IB}$</td>
<td>RelTag</td>
<td>1.8388e-14</td>
</tr>
<tr>
<td>$EMF_{TA}$</td>
<td>RelTag</td>
<td>2.5197e-16</td>
</tr>
</tbody>
</table>

4.3.2.2 Explainability Evaluation

We evaluate our proposed methods and baselines using three explainability styles: user-based neighborhood explainability, item-based neighborhood explainability and tag-based explainability styles. We denote the explainability thresholds for these styles using $\theta^n$, $\theta^i$ and $\theta^t$ respectively. Figure 40 shows the Mean Explainability Precision scores (MEP@K) for the top-10 recommended items for users using the above listed explainability styles. Our proposed methods outperform the baseline methods for the user-based and item-based neighborhood explainability styles. Tables 43 and 44 show the significance test p-values of our proposed method, RelTag, compared to the baseline methods, using the user-based
neighborhood explainability style for $\theta^n = 0.1$. Tables 45 and 46 show the significance test p-values of our proposed method, PrefTag, compared to the baseline methods, using the item-based neighborhood explainability style for $\theta^i = 0.1$

**TABLE 43**

MEP@10 significance test results (NSE-explainablity) — $\theta^n = 0.1$. Bold denotes the winner

<table>
<thead>
<tr>
<th>Model 1</th>
<th>Model 2</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF</td>
<td>RelTag</td>
<td>2.787e-27</td>
</tr>
<tr>
<td>$EMF_{UB}$</td>
<td>RelTag</td>
<td>3.283e-27</td>
</tr>
<tr>
<td>$EMF_{IB}$</td>
<td>RelTag</td>
<td>5.935e-27</td>
</tr>
<tr>
<td>$EMF_{TA}$</td>
<td>RelTag</td>
<td>1.565e-27</td>
</tr>
</tbody>
</table>

**TABLE 44**

MER@10 significance test results (NSE-explainablity) — $\theta^n = 0.1$. Bold denotes the winner

<table>
<thead>
<tr>
<th>Model 1</th>
<th>Model 2</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF</td>
<td>RelTag</td>
<td>8.412e-27</td>
</tr>
<tr>
<td>$EMF_{UB}$</td>
<td>RelTag</td>
<td>1.017e-26</td>
</tr>
<tr>
<td>$EMF_{IB}$</td>
<td>RelTag</td>
<td>1.529e-26</td>
</tr>
<tr>
<td>$EMF_{TA}$</td>
<td>RelTag</td>
<td>5.544e-27</td>
</tr>
</tbody>
</table>

**TABLE 45**

MEP@10 significance test results (ISE-explainablity) — $\theta^i = 0.1$. Bold denotes the winner

<table>
<thead>
<tr>
<th>Model 1</th>
<th>Model 2</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF</td>
<td>PrefTag</td>
<td>4.193e-25</td>
</tr>
<tr>
<td>$EMF_{UB}$</td>
<td>PrefTag</td>
<td>2.440e-24</td>
</tr>
<tr>
<td>$EMF_{IB}$</td>
<td>PrefTag</td>
<td>2.285e-24</td>
</tr>
<tr>
<td>$EMF_{TA}$</td>
<td>PrefTag</td>
<td>1.295e-21</td>
</tr>
</tbody>
</table>
Figure 40: MEP@10 vs Explainability Threshold ($\theta$). Our proposed methods are denoted as $EMF_{TA}$, PrefTags and RelTags.
Figure 41: MER@10 vs Explainability Threshold ($\theta$). Our proposed methods are $EMF_{TA}$, PrefTags and RelTags
TABLE 46

MER@10 significance test results (ISE-explainablity) — $\theta^i = 0.1$. Bold denotes the winner

<table>
<thead>
<tr>
<th>Model 1</th>
<th>Model 2</th>
<th>$P$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF</td>
<td>PrefTag</td>
<td>4.006e-20</td>
</tr>
<tr>
<td>$EMF_{UB}$</td>
<td>PrefTag</td>
<td>4.105e-20</td>
</tr>
<tr>
<td>$EMF_{IB}$</td>
<td>PrefTag</td>
<td>4.561e-20</td>
</tr>
<tr>
<td>$EMF_{TA}$</td>
<td>PrefTag</td>
<td>5.622e-20</td>
</tr>
</tbody>
</table>

4.3.2.3 Examples

Table 47 shows the top-3 rated movies for a sample user from the data. The results of the proposed Tag-boosted Multi-style EMF, methods RelTag and PrefTag, are shown in Tables 48 and 49 respectively.

TABLE 47

Top-3 rated movies for Sample User

<table>
<thead>
<tr>
<th>Top-3 rated movies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean and Sober (Drama)</td>
</tr>
<tr>
<td>Strangers on a train (Thriller)</td>
</tr>
<tr>
<td>Indiana Jones and the Temple of Doom (Adventure/Action)</td>
</tr>
</tbody>
</table>
### TABLE 48
Output of Relevant Tag-boosted Multi-style EMF for Sample User

<table>
<thead>
<tr>
<th>Top-3 Recommended Movies</th>
<th>Neighbor-rating Style Explanation</th>
<th>Tag-based Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rupan Sansei (Action)</td>
<td>2 similar users rated this movie as 5 stars</td>
<td>-</td>
</tr>
<tr>
<td>Forrest Gump (Drama/Romance)</td>
<td>1 similar user rated this movie as 5 stars</td>
<td>vietnam, oscar (best picture), classic</td>
</tr>
<tr>
<td>The Lion King (Family)</td>
<td>3 similar users rated this movie as 4 stars</td>
<td>-</td>
</tr>
</tbody>
</table>

### TABLE 49
Output of Preferred Tag-boosted Multi-style EMF for Sample User

<table>
<thead>
<tr>
<th>Top-3 Recommended Movies</th>
<th>Influence Style Explanation</th>
<th>Tag-based Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris (Drama)</td>
<td>You rated 2 similar movies as 4 stars</td>
<td>-</td>
</tr>
<tr>
<td>Le Fabuleux Destin Poulain (Romance/Comedy)</td>
<td>You rated 1 similar movie as 3 stars</td>
<td>-</td>
</tr>
<tr>
<td>Pulp Fiction (Crime/Drama)</td>
<td>You rated 3 similar movies as 4 stars</td>
<td>quentin tarantino, hit-men, comedy</td>
</tr>
</tbody>
</table>

Tables 48 and 49 show the advantage of using multiple explanation styles. Although the tag-based explanations are not available for every recommended item, when present, tag-based explanations provide useful information about the recommended movie. The ISE and NSE styles provide explanations to the user about why the movie was recommended.
but the tag-based explanation tells the user about the possible content of the recommended movie. For example, in Table 49, the ISE explanation tells the user about movies similar to the recommended movie “Pulp Fiction” but the tag-based explanation gives some insight into what the user might find interesting about the movie with a tag that describes the director, “quentin tarantino”, another that describes the genre of the movie, “comedy”, and finally, a tag that describes an important part of the plot of the movie “hitmen”.

4.3.2.4 Analysis of Results

The influence of the number of similar relevant or preferred tags between movies and users, respectively on our proposed tag-boosted methods seems to be restricted to the movie domain. This is due to the vast number of entities that can be represented using user-generated tags in the movie domain. However, many movies do not share the same number of entities (e.g. actors, directors, producers, genres among others) which are represented by factual tags or relevant tags. This is not the case for subjective tags, which can be used by users to describe emotions and hence broadly describe multiple movies. In other domains, where there might be fewer entities that can be represented with user-generated tags, we might obtain different results.

4.4 Chapter Summary

To summarize this chapter, we try to answer the hypotheses that we stated at the start in Sec. 4.1.3. Table 13 showed that the hybrid methods improved the explainability coverage by providing multiple explanation-styles for most user-item pairs. In the book domain, Table 17 showed that the hybrid methods improved the MF model performance, while providing multiple explanation styles for the recommendations. In the movie domain, our results, in Sec. 4.2.7, showed that different explanation styles can be leveraged simultaneously to produce explainable recommendations with multiple styles of explanations, and with significantly higher explainability metrics, without sacrificing the accuracy of the recommendations. We also showed, in Sections 4.2.7.1, 4.2.7.2, 4.3.2.1, and 4.3.2.2, that the
performance of the hybrid and tag-boosted methods significantly outperformed the baseline methods ($p < .05$) in terms of accuracy and explainability, while illustrating with real examples of the algorithms’ outputs how the diverse set of explanations complement one another and can appeal to diverse user preferences and item types in explanation styles on a case by case basis. Our research findings shed new light on the performance of the novel approaches that we proposed to design various hybridizations of explainable models. Although, rooted in the hybrid recommendation framework, our proposed methods make a significant step forward in explainable AI and beyond the confines of hybrids. This is because the proposed hybridization mechanisms make an intentional effort to take into account the individual models’ explanations and not only their output predicted ratings.
CHAPTER 5

CONCLUSION AND FUTURE WORK

Matrix Factorization (MF) is an accurate model-based Collaborative Filtering technique commonly used in recommender systems due to its accuracy and robustness in handling sparse data. However, MF is limited by the opaqueness of the recommendation process which makes it difficult to understand how the recommendations were generated. Explainable Matrix Factorization (EMF) adds an explainability constraint to the MF algorithm in the model building phase. The explainability constraint introduced a degree of transparency to the MF algorithm. EMFs is also able to provide explanations for recommendations to users thus making MF models explainable.

In this work, we focused on filling the gaps created by the Explainable Matrix Factorization (EMF) algorithms which were only capable of providing one explanation style to all users for all items. These EMF algorithms were insensitive to a user’s preferred explanation style. Secondly, EMF algorithms were only able to explain a fraction of the total number of user-item pairs available to the recommender system. Users had different explanation style preferences, even for the same recommended item, as verified by our preliminary questionnaire.

To tackle these limitations of explainability coverage and explanation preference-insensitivity, we proposed four hybridization techniques for combining multiple EMF models that improved the accuracy, explainability and coverage of the resultant hybrid model. We also proposed two tag-boosting techniques to provide users with tag-based explanations that complemented explanations generated by EMF models.

The first hybrid method we proposed was the Weighted Multi-style EMF (W-MEMF) algorithm which assigned weights to the individual EMF models to build an improved hy-
brid model capable of explaining recommendations with multiple explanation styles. We explored two approaches (global and personalized) for estimating the weights assigned to the individual EMF models when building the hybrid model. Secondly, we proposed the Regression-based Multi-style EMF (R-MEMF) algorithm which models the hybridization problem as a regression problem. We explored solving the regression problem using different regularization mechanisms to select the most important features for building a good predictor of the user’s true preference. Thirdly, we proposed the Switched Multi-style EMF (S-MEMF) hybridization algorithm which selected the best EMF model for a recommendation task based on an appropriate selection criteria such as transparency and model accuracy. Finally, we proposed the Asymmetric Multi-style EMF (A-MEMF) hybridization algorithm which realized the hybridization using a two-step approach. In the anchoring step, we built an EMF model which would serve as an anchor. In the transfer step, we learn the basis matrix from another EMF model that can adapt in the latent space built in the anchoring step. This newly adapted latent space is used to generate recommendations that can be explained using the explanation styles of the input EMF models.

The first Tag-boosted approach we proposed, Preferred-tags Boosted EMF, used the tag-preference latent space to find users who have the same preference for user-generated tags and used information about these similar users as additional information for improving the performance of EMF models. The second Tag-boosted approach we proposed, Relevant-tags Boosted EMF, used the tag-relevance latent space to find items that have been tagged with the same tags considered to be relevant to describing the item. This extra information was used to improve the performance of the EMF models. These proposed methods also had the extra benefit of providing tags that could be considered as explanations for the recommended items to users.

We evaluated our proposed methods by measuring the accuracy, recommendations, explainability coverage and explainability of the recommended items. We also evaluated the explainability of our proposed hybrid models using neighborhood-based and explainability measures in different domains of movies and books. Our research findings shed new light
on the performance of new mechanisms that we have investigated to design various ways to hybridize several explainable models. Although, rooted in a hybrid recommendation framework, our proposed methods make a significant step forward in explainable AI and beyond the confines of hybrid methods, because the proposed hybridization mechanisms make an intentional effort to take into account the individual models’ explanations and not only their output predicted ratings.

Our work opens several directions for expansion, including exploring different hybridization mechanisms or utilizing additional individual explainable latent factor recommendation algorithms in the hybrid model building process. Moreover, online user studies should be performed to evaluate the users’ satisfaction with different explanation styles and generation mechanisms, within different recommendation tasks and contexts.
REFERENCES


128


CURRICULUM VITAE

NAME: Olurotimi Nugbepo Seton

ADDRESS: Computer Science and Engineering Department
J.B Speed School of Engineering
University of Louisville
Louisville, KY 40292
United States of America.

EDUCATION: Ph.D., Computer Science & Engineering, May 2021
University of Louisville, Louisville, KY, USA.
M.Sc. in Computer Science, May 2014
Babcock University Ilishan-Remo, Ogun, Nigeria
B.Sc in Computer Science (Technology), May 2009
Babcock University Ilishan-Remo, Ogun, Nigeria

WORK EXPERIENCE: Graduate Assistant, REACH, University of Louisville, KY, USA, 2016
Software Development Intern, Hussmann Corporation, St. Louis, MO, USA, 2018
Data Science Intern, Brown-Forman Corporation, Louisville, KY, USA, 2019
AWARDS: CSE Doctoral Award, 2020
Dissertation Completion Award, 2020
Richard Tapia Celebration of Diversity in Computing Scholarship, 2020

PUBLICATIONS:

1. Olurotimi Seton, Olfa Nasraoui, Pegah Sagheb Haghighi, Mohammed Alshammari, and Khalil Damak. Tag-boosted Multi-style Explainable Matrix Factorization for Collaborative Filtering Recommendation. (to be submitted)

2. Olurotimi Seton and Olfa Nasraoui. Hybridization Techniques for Multiple style Explainable Matrix Factorization-based Recommendation. (to be submitted)