# Are Representation Disentanglement and Interpretability Linked in Recommendation Models?

A Critical Review and Reproducibility Study\*

Ervin Dervishaj<br/>1[0000-0002-4192-1055], Tuukka Ruotsalo<br/>1,2[0000-0002-2203-4928], Maria Maistro<br/>1[0000-0002-7001-4817], and Christina Lioma<br/>1[0000-0003-2600-2701]

1 University of Copenhagen
2 LUT University
{erde,tr,mm,c.lioma}@di.ku.dk

Abstract. Unsupervised learning of disentangled representations has been closely tied to enhancing the representation interretability of Recommender Systems (RSs). This has been achieved by making the representation of individual features more distinctly separated, so that it is easier to attribute the contribution of features to the model's predictions. However, such advantages in interpretability and feature attribution have mainly been explored qualitatively. Moreover, the effect of disentanglement on the model's recommendation performance has been largely overlooked. In this work, we reproduce the recommendation performance, representation disentanglement and representation interpretability of five well-known recommendation models on four RS datasets. We quantify disentanglement and investigate the link of disentanglement with recommendation effectiveness and representation interpretability. While several existing work in RSs have proposed disentangled representations as a gateway to improved effectiveness and interpretability, our findings show that disentanglement is not necessarily related to effectiveness but is closely related to representation interpretability. Our code and results are publicly available at https://github. com/edervishaj/disentanglement-interpretability-recsys.

**Keywords:** Representation learning  $\cdot$  Disentanglement  $\cdot$  Feature attribution.

# 1 Introduction

In order to provide users with a personalised experience, recommender systems (RSs) build representations that encode user preferences from behavioural data. One way of extracting informative and more interpretable representations is by

<sup>\*</sup> This preprint has not undergone peer review (when applicable) or any post-submission improvements or corrections.

learning disentangled representations [2,7,26,30,40,48]. Real-world RSs data contain complex combinations of user preference factors (e.g., in a movie RS: movie genre, movie actors, mood of the user, time of the day, etc.). Given that one factor is often invariant to changes in other factors, unsupervised disentangled representation learning aims to encode each factor onto one of the dimensions of the latent space. Existing work claim that such representations improve interpretability because each latent dimension captures one semantically-meaningful explanatory factor of the data [2,5-7,7,11,21,24,30,48] (also called factor of variation). Disentangled representations have been recently studied – in RSs, and more generally – under the class of deep generative models, namely variational autoencoders (VAEs) [25] and generative adversarial networks (GANs) [16]. In order to force the deep neural network to disentangle the factors of variation of the data in the latent space, previous work modify the model's vanilla objective function with a regularization term [30], thereby acting as an "interpretability constraint on the latent dimensions" [39]. This change in the objective function has been empirically linked to reduced reconstruction capabilities in VAEs [21], pointing to a trade-off between disentanglement and downstream task performance [5, 24, 27].

Recently, the RS community has explored unsupervised disentangled representations for more accurate modeling of user preferences. However, previous work provides only a qualitative perspective on the model disentanglement, usually in the form of a visual inspection of the learned representations reduced to 2 dimensions [9, 18, 32, 46, 53]. The lack of an objective evaluation of disentanglement makes it unclear whether the findings of disentanglement, model effectiveness and improved representation interpretability generalize over different models and datasets.

In this work, we seek to address the following research questions:

- 1. Can we reproduce the recommendation effectiveness and disentanglement of existing RSs models aimed at learning disentangled representations?
- 2. What is the effect of disentanglement on recommendation effectiveness?
- 3. What is the effect of disentanglement on RSs representation interpretability?

First, we focus on reproducing recommendation effectiveness and disentanglement of state-of-the-art unsupervised disentangled RSs. Our results show that the reproducibility of the reported effectiveness of these models is dataset dependent, with differences up to 43% across datasets. To our best knowledge, only [35] has utilized disentanglement metrics to quantify disentanglement in RSs. However, we failed to reproduce their results which prompted us to conduct an empirical investigation that uses existing disentanglement metrics (disentanglement and completeness [12]). Specifically, we measure the disentanglement of user representations of five recommendation models on four datasets. Second, we provide the first study that quantifies representation interpretability with feature attribution methods and links it to disentanglement in RSs. We adapt two existing feature attribution approaches: LIME [38] and Shapley Value (SHAP) [31]. We use the feature-level scores produced by LIME and SHAP to define two measures, called LIME-global and SHAP-global, which quantify the degree of inter-

pretability of a model's representations. Finally, through a correlation analysis, we study the link between disentanglement, representation interpretability and effectiveness in RSs. Our findings do not support the alleged trade-off between disentanglement and effectiveness [21,24,27,39], due to no consistent statistically significant relation between the two. However, in line with prior qualitative work in RSs on interpretability and disentanglement [32,35,46], we find representation interpretability to be positively related to disentanglement.

# 2 Background

# 2.1 Disentangled Representations

Disentangled representations aim to separate the factors of variation in the data, i.e., changing one aspect of a data point should only affect the latent dimension responsible for the changed aspect, while keeping all the other dimensions unchanged. Since the data generating function is usually governed by few explanatory factors, separating these factors in a latent space makes for a more interpretable representation. Recently proposed models for disentangled representations are largely based on VAEs [21,30].  $\beta$ -VAE [21] was first to apply VAEs to learn a factorized representation of the independent explanatory factors from the data by enforcing an information bottleneck through a new hyperparameter  $\beta > 1$  which penalizes the KL-divergence term in the VAEs' evidence lower bound (ELBO) objective.

#### 2.2 Disentangled Representations in RSs

Disentangled representation learning has shown good empirical results by modeling the implicit biases in the data [5,6,21,24,33,41]. Inspired by this, the RS community has recently explored unsupervised disentangled representations: based on user intents [32,37,46,50], and supervised disentangled representations: based on item topics [18], user conformity to group (i.e., popularity bias) [49, 51, 53], long and short term user interests [52] and causal components [14,47]. Supervised disentangling models select some semantically meaningful item attributes, such as item topic [18], item attribute [35], etc., which are then used during training such that one dimension of the representation space encapsulates only one such item attribute. Unsupervised disentangling models, on the other hand, have the advantage that they do not need user/item side information, often not available in RSs datasets. The state of the art in RSs is MacridVAE [32], which assumes that user interactions are based on several user intent so it disentangles the user preferences into macro concepts (e.g., item categories). Ma et al. [32] represent each macro concept with a d-dimension vector to allow for a finer granularity of the user preferences, which they further disentangle by penalizing the KL-divergence in their VAE objective with  $\beta \gg 1$  (similar to [21]).

# 2.3 Disentanglement, Interpretability of Representations and Effectiveness

Disentangled representations have been linked to more interpretable representations [7,21], but at the expense of downstream task performance [5,24,27]. This is due to the regularization effect of disentanglement [30,39] during the learning process. In RSs, several work show qualitatively that disentanglement helps in learning representations that are more interpretable; [32] inspects the ability of disentangled representations in capturing the true item categories, [28,46] investigate how closely users' item reviews relate to the disentangled user intents and [35] compares the correlation in latent space between users and items on supervised disentangled dimensions and unsupervised ones, separately. While prior work on disentanglement in RSs proposes models that achieve state-of-the-art performance, the regularization aspect of disentanglement, and a possible trade-off with effectiveness, remains largely unexplored and only briefly mentioned by Nema et al. (see section 5.3 in [35]).

# 3 Experiments

In this section we describe the experimental setup of our reproducibility study on unsupervised disentangled RSs and the relation between disentanglement, recommendation effectiveness and representation interpretability. We describe the effectiveness, disentanglement and our proposed representation interpretability measures, the datasets and the recommendations models that we reproduce.

#### 3.1 Measures

Recommendation effectiveness measures We use Normalized Discounted Cumulative Gain (NDCG) [22], recall, Mean Reciprocal Rank (MRR) [43] and coverage [15], at cutoff 10, 50 and 100.

Disentanglement measures Several measures of disentanglement have been proposed, e.g., BetaVAE [21], FactorVAE [24], or the Disentanglement-Completeness-Informativeness (DCI) framework [12]. All of these measures require ground truth factors of variation (see section 3.2 on how we recover these ground truth factors). BetaVAE and FactorVAE also require a ground truth simulation function that, given factors of variation, can produce data samples. Since in RSs we lack such a simulation function, we use two measures from the DCI framework: disentanglement and completeness<sup>3</sup>. Both of these measures use simple estimators to predict the ground truth factors of variation of the data from a learned disentangled space. We describe these two measures next.

We assume that a model can learn an M-dimensional latent representation  $z \in \mathbb{R}^M$ . Given K binary ground truth factors of variation, we train K binary

<sup>&</sup>lt;sup>3</sup> We use the code release by [12] and the disentanglement\_lib framework [30].

classifiers  $f_j : \mathbb{R}^M \to \{0,1\}$  for  $j \in 1 ... K$ , that, given the latent representation z, predict the presence of each of the K factors. We collect in a matrix  $F \in \mathbb{R}^{M \times K}$  the importance<sup>4</sup> of dimension i of the latent space in predicting factor j. Then disentanglement (**D**) and completeness (**C**) are computed as:

$$\mathbf{D} = \sum_{i=1}^{M} \alpha_{i} D_{i}, \ D_{i} = 1 - H_{K}(P_{i}) \qquad \mathbf{C} = \sum_{j=1}^{K} \beta_{j} C_{j}, \ C_{j} = 1 - H_{M}(\tilde{P}_{j})$$

$$\alpha_{i} = \sum_{j=1}^{K} F_{ij} / \sum_{j=1}^{K} \sum_{i=1}^{M} F_{ij} \qquad \beta_{j} = \sum_{i=1}^{M} F_{ij} / \sum_{j=1}^{K} \sum_{i=1}^{M} F_{ij}$$

$$H_{K}(P_{i}) = -\sum_{j=1}^{K} P_{ij} \log_{K} P_{ij} \qquad H_{M}(\tilde{P}_{j}) = -\sum_{i=1}^{M} \tilde{P}_{ij} \log_{M} \tilde{P}_{ij}$$

$$P_{ij} = F_{ij} / \sum_{j=1}^{K} F_{ij} \qquad \tilde{P}_{ij} = F_{ij} / \sum_{i=1}^{M} F_{ij}$$

$$(1)$$

 ${f D}$  and  ${f C}$  are expressed as weighted sums of the disentanglement  $D_i$  of each dimension in the learned representation, and of the completeness  $C_j$  of each factor of variation, respectively. The entropy H specifies how feature importance probabilities  $P_j$  and  $\tilde{P}_j$  are distributed across the factors of variation and the dimensions of the representations, respectively. A higher entropy means that the classifier's feature importance is more uniformly distributed across factors/dimensions resulting in lower disentanglement/completeness, whereas a lower entropy means placing higher feature importance on a smaller subset of the dimensions of the learned representations, thus achieving better overall disentanglement. Both  ${f D}$  and  ${f C}$  range in [0,1], and the higher, the more disentangled/complete the representations. In this work, we use gradient boosting decision trees as binary classifiers and tune them on the representations of users in the validation set (section 3.2).

Interpretability measures LIME [38], SHAP [31] and Integrated Gradients (IG) [42] are well-known local interpretability methods, i.e., they measure feature importance on the model's prediction per data sample. However, to our best knowledge, there exists no method that quantifies representation interpretability. To quantify the interpretability of the disentangled representations, we utilise the K binary classifiers used for disentanglement and completeness and adapt LIME and SHAP<sup>5</sup> into global measures, called LIME-global and SHAP-global. For each classifier  $f_j$ , we collect the mean absolute LIME/SHAP latent dimension importance across all user representations into a column vector  $\mathbf{s}_j \in \mathbb{R}^M$ . We concatenate all vectors  $\mathbf{s}_j$  to build matrix  $S \in \mathbb{R}^{M \times K}$ . As a final value for LIME-/SHAP-global, we normalize the columns of S into [0,1] and take the mean of the Jensen-Shannon (JS) divergence computed between every pair of columns of S. The intuition behind LIME-/SHAP-global and the application of the JS divergence is to penalise redundancy in the dimensions of the representation space

<sup>&</sup>lt;sup>4</sup> We use GradientBoostingClassifier from scikit-learn package as our binary classifier which provides impurity-based feature importances.

<sup>&</sup>lt;sup>5</sup> We do not use IG because the gradient boosting trees classifiers are non-differentiable.

#### Dervishaj et al.

6

and to promote sparsity in feature importance. If two binary classifiers place similar feature importance on the same subset of dimensions of the learned representation space, then the divergence between the classifiers' feature importance distribution will be 0. In this way, LIME-/SHAP-global penalises redundancy in the learned representations. Moreover, for the JS divergence (i.e., interpretability) to increase, the feature importance distribution should be concentrated only on some features (i.e., promoting sparsity in the features, and resulting in a simpler and more interpretable model [4,10,17,39]). LIME-/SHAP-global range is [0,1], with higher values denoting more interpretable representations.

Table 1: Dataset statistics after preprocessing. Min. IPU/IPI are the *minimum interactions per user/item*.

1							
	Amazon-CD	ML1M	Yelp	GR-Children			
Interactions	570 747	1000209	2762098	5067546			
Users	23024	6040	99011	117293			
Items	19444	3706	56441	42119			
Min. IPU	10	20	10	10			
Min. IPI	10	1	10	10			
Sparsity	99.877%	95.532%	99.953%	99.897%			

Table 2: Model summary. NP = non-personalised, MF = matrix factorisation, DAE = denoising autoencoder, VAE = variational autoencoder, DIS = explicitly disentangling model.

Model	Model Type						
wiodei	NP	MF	DAE	VAE	DIS		
Top-Popular	1						
PureSVD		1					
MultiDAE			1				
MultiVAE				1			
$\beta$ -VAE				1	1		
${\bf MacridVAE}$				1	1		

#### 3.2 Datasets

Table 1 shows the statistics of the datasets. We reproduce results for MovieLens 1M<sup>6</sup> [19] and GoodReads-Children<sup>7</sup> [44, 45], both used by unsupervised disentangled models in RSs. In addition, we evaluate our reproduced models also on

<sup>&</sup>lt;sup>6</sup> https://grouplens.org/datasets/movielens

<sup>&</sup>lt;sup>7</sup> https://mengtingwan.github.io/data/goodreads.html#datasets

two other datasets – Amazon-CD<sup>8</sup> [36] and Yelp<sup>9</sup> – that were not in the original papers, in order to investigate the generalizability of our results. We focus on user-item interactions, so we binarize the ratings by setting to 1 all ratings  $\geq 1$  and everything else to 0.

Sampling and splits We use 10-core [20,35] sampling <sup>10</sup> for all datasets except for ML1M, which includes only users with at least 20 interactions. Since we evaluate models that build user representations, a timestamp split might exclude some of the users during training. For this reason, we use random per-user trainvalidation-test splits with a ratio of 3:1:1.

Ground truth factors Quantifying disentanglement requires access to the ground truth factors of variation of the data [12,21,24,30]. To our best knowledge, no publicly available RSs dataset provides such factors. In this reproducibility study we follow closely [35] and utilize item content information to compose ground truth factors. Amazon-CD, ML1M, and Yelp provide tags/categories for each item. We keep the 100 most popular tags/categories, which we then group in 20 clusters using k-means clustering, with each tag/category represented as a vector over the items. Each item is part of a subset of clusters based on its corresponding tags/categories<sup>11</sup>. Finally, a user is assigned to a cluster if at least 50% of the items that the user has interacted with are part of that cluster. GoodReads-Children provides bookshelves for each book. We first order the bookshelves in decreasing order according to the number of books they contain and then manually merge bookshelves with similar names (e.g., 'picturebooks' and 'picture-book') and drop bookshelves whose name is not semantically meaningful (e.g., 'to-read', 'books-i-own')<sup>12</sup> to get a short final list of bookshelves. Finally, we assign users to shelves if at least 50% of their rated books fall within a shelf. We consider the clusters of categories (Amazon-CD, ML1M, Yelp) and the bookshelves (GoodReads-Children) as each dataset's respective set of ground truth factors of variation.

#### 3.3 Recommendation Models

We focus on unsupervised disentangling models given the lack of ground truth factors of variation in RSs. Table 2 lists our models. We reproduce two unsu-

<sup>&</sup>lt;sup>8</sup> https://cseweb.ucsd.edu/~jmcauley/datasets/amazon v2

 $<sup>^9</sup>$  https://www.yelp.com/dataset

We use the recpack [34] Python package to handle filtering, sampling and splitting of the datasets.

<sup>&</sup>lt;sup>11</sup> In ML1M, each tag is represented as a relevance score vector over the items. An item is assigned a cluster if the average relevance score of its tags that fall in the cluster is greater than M. We set M = 0.4 as per [35].

<sup>&</sup>lt;sup>12</sup> We follow [35], where uninformative bookshelves are dropped if marked as such by all authors.

pervised disentangling models: MacridVAE [32] and  $\beta$ -VAE<sup>13</sup> [21] and four non-disentangling models, as additional baselines: Top-Popular recommends only the most popular items, PureSVD [8] is a simple matrix factorization (MF) model, MultiDAE [29] and MultiVAE [29] are AE-based RSs. For each model, we use the official implementation released by the authors. We use the hyperopt [3] Python package to tune hyperparameters through 50 runs of Bayesian search optimization. All models are tuned for NDCG@100. For a fair comparison, we set common hyperparameter ranges – log-uniform distribution in [exp(-10), exp(-2)] for learning rate, integer uniform distribution in [2, 20] for latent dimensionality, {128, 256, 512, 1024} for batch size – and constrain training to a maximum of 500 epochs with early stopping. In selecting the hyperparameter tuning ranges<sup>14</sup>, we make sure to include all the extremes of the ranges reported by the authors of the models that we reproduce. We use five different randomization seeds for both model initialization and dataset splitting, and report mean scores of the models from the five seeds on the unseen test set.

Table 3: Effectiveness and disentanglement reproducibility results for [35] and [32]. Their results are shown as reported in the papers. Our results represent the mean of 5 runs tuned according to section 3.3.

9								
Model	ML1M				GR-CHILDREN			
Model	N@100	R@50	D	C	N@100	R@50	D	С
MultiDAE [32]	0.4045	0.4678	-	-	-	-	-	-
MultiDAE [35]	0.4040	0.4670	0.3810	0.3120	0.4250	0.5910	0.2870	0.2430
MultiDAE (ours)	0.4366	0.4500	0.1777	0.2605	0.3452	0.3807	0.1309	0.1849
Max rel. change	↑ 8.0%	↓ 3.8%	$\downarrow 53.4\%$	$\downarrow 16.5\%$	↓ 18.8%	↓ 35.6%	$\downarrow 54.4\%$	↓ 23.9%
MultiVAE [32]	0.4056	0.4583	-	-	-	-	-	-
MultiVAE [35]	0.4050	0.4580	0.3610	0.2940	0.4040	0.5770	0.3080	0.2630
MultiVAE (ours)	0.4360	0.4456	0.1859	0.2894	0.3432	0.3766	0.1377	0.1805
Max rel. change	↑ 7.7%	$\downarrow 2.8\%$	↓ 48.5%	↓ 1.6%	↓ 15.0%	↓ 34.7%	↓ 54.4%	↓ 31.4%
β-VAE [32]	0.4056	0.4582	-	-	-	-	-	-
$\beta$ -VAE [35]	0.4540	0.4110	0.7450	0.4730	0.4150	0.5860	0.4840	0.3030
$\beta$ -VAE (ours)	0.4070	0.4164	0.1879	0.2059	0.2999	0.3335	0.1053	0.2010
Max rel. change	$\downarrow 10.4\%$	$\downarrow 9.1\%$	$\downarrow 74.8\%$	$\downarrow 56.5\%$	$\downarrow 27.7\%$	↓ 43.1%	$\downarrow 78.2\%$	↓ 33.7%
MacridVAE [32]	0.4274	0.4904	-	-	-	-	-	-
MacridVAE (ours)	0.4580	0.4774	0.1593	0.2527	0.3764	0.4082	0.1593	0.2527
Max rel. change	$\uparrow 7.2\%$	$\downarrow 2.7\%$	-	-	-	-	-	-

# 4 Results and Discussion

#### 4.1 Reproducibility Results (RQ1)

In our literature review on disentanglement for RSs, we found [35] as the only work that explicitly measures disentanglement with existing metrics, so we fo-

<sup>&</sup>lt;sup>13</sup> We adapt  $\beta$ -VAE to use the multinomial distribution, similar to MultiVAE, and we set  $\beta > 1$  during tuning to enforce disentanglement [21].

<sup>&</sup>lt;sup>14</sup> We provide the optimal hyperparameters with our codes.

cus on reproducing their reported results and the results of MacridVAE [32]. In table 3 we give the original and reproduced results 15.

Regarding recommendation effectiveness, we observe some small discrepancies in ML1M; our NDCG@100 scores are generally higher (up to 8%) than the reported ones, with the exception of  $\beta$ -VAE which is 10% lower. Our recall@50 scores show a smaller difference from the original work results with  $\beta$ -VAE up to 9% lower score. In GoodReads-Children, we observe much larger differences with NDCG@100 and recall@50 up to 28% and 43% lower than the reported scores, respectively. We attribute the discrepancies to i) how we binarize explicit ratings; both [35] and [32] set ratings > 4 to 1, whereas we set to 1 all ratings > 1, and ii) how we tune the hyperparameters, especially the latent space dimensionality; both [35] and [32] use a fixed latent dimensionality, meanwhile we tune it according to section 3.3.

Regarding disentanglement, we observe significant discrepancies in our reproduced results. Even though Nema et al. [35] clearly describe the formulations of the metrics, we were not able to reproduce their scores, despite trying to implement their formulations as faithfully as possible. Moreover, retrieving the ground truth factors of variation – used to compute disentanglement scores – involves some hyperparameters; the threshold M of average relevance score for assigning an item to a cluster of its tags in ML1M, which we set according to [35], and the merging/dropping of bookshelves in GoodReads-Children. As a final attempt, we tried using logistic regression and random forests as classifiers, but the results were still very different. After these efforts, we reached out to the authors, but the code was not made available to us.

To investigate the generalizability of these results, we evaluate the reproduced models on two additional datasets (Amazon-CD and Yelp) and include PureSVD, as a MF baseline model that learns representations from interactions, and Top-Popular. For a fair comparison, we tune all the models according to section 3.3 and train them on implicit ratings (all ratings > 1 set to 1). We present these results in table 4. Table 4 shows MacridVAE as the best model across all effectiveness measures and all datasets (in Amazon-CD twice, and in GoodReads-Children 6-8% better than the next best model), which is consistent with [32]. Overall, neural models – MultiVAE and MultiDAE – perform much better than the MF model, as reported also by [13].

In table 5 we show the disentanglement and representation interpretability results of the considered models. We observe that PureSVD has the highest completeness across all datasets and the highest disentanglement in 2/4 datasets. MacridVAE, despite explicitly aimed at learning disentangled representations in RSs, shows the best disentanglement only in GoodReads-Children dataset. In the other datasets, its disentaglement and completeness scores are much lower than PureSVD, and MultiVAE in ML1M. In [32], the authors provide only a qualitative inspection of the representation space learned by MacridVAE (figure 2 in [32]). While their results indicate that MacridVAE can encode the items'

Note that the authors of MacridVAE do not report disentanglement scores in their experiments.

Table 4: Mean and standard deviation of recommendation effectiveness over 5 runs. Models tuned for NDCG@100. Best effectiveness results are given in bold.

Dotoset	Madal	EFFECTIVENESS					
Dataset	Model	NDCG@10	RECALL@10	MRR@10	COVERAGE@10		
Anseon-CD	Top-Popular	$0.0059 \pm 0.0002$	$0.0084 \pm 0.0002$	$0.0098 \pm 0.0005$	$0.0000 \pm 0.0000$		
	PureSVD	$0.0288 \pm 0.0003$	$0.0386 \pm 0.0007$	$0.0462 \pm 0.0006$	$0.0326 \pm 0.0006$		
Ŕ	MultiDAE	$0.0806 \pm 0.0016$	$0.1032 \pm 0.0018$	$0.1216 \pm 0.0030$	$0.4761 \pm 0.1953$		
20	MultiVAE	$0.0853 \pm 0.0019$	$0.1085 \pm 0.0028$	$0.1295 \pm 0.0028$	$0.4664 \pm 0.1194$		
<u> </u>	$\beta$ -VAE	$0.0570 \pm 0.0090$	$0.0762 \pm 0.0108$	$0.0864 \pm 0.0140$	$0.4067 \pm 0.1980$		
Α,	MacridVAE	$0.1621 \pm 0.0039$	$0.1824 \pm 0.0029$	$0.2560 \pm 0.0076$	$0.6624 \pm 0.0181$		
	Top-Popular	$0.0533 \pm 0.0030$	$0.0556 \pm 0.0038$	$0.1244 \pm 0.0049$	$0.0000 \pm 0.0000$		
₩	PureSVD	$0.3889 \pm 0.0008$	$0.3669 \pm 0.0011$	$0.6201 \pm 0.0006$	$0.1792 \pm 0.0016$		
MEIN	MultiDAE	$0.3806 \pm 0.0056$	$0.3632 \pm 0.0042$	$0.6142 \pm 0.0116$	$0.4420 \pm 0.0348$		
\$	MultiVAE	$0.3853 \pm 0.0047$	$0.3656 \pm 0.0042$	$0.6213 \pm 0.0056$	$0.4611 \pm 0.0275$		
$\overline{}$	$\beta$ -VAE	$0.3553 \pm 0.0222$	$0.3381 \pm 0.0183$	$0.5822 \pm 0.0287$	$0.4032 \pm 0.0145$		
	MacridVAE	$0.3963 \pm 0.0020$	$0.3807 \pm 0.0022$	$0.6269 \pm 0.0012$	$0.4793 \pm 0.0188$		
	Top-Popular	$0.0029 \pm 0.0003$	$0.0042 \pm 0.0004$	$0.0051 \pm 0.0007$	$0.0002 \pm 0.0000$		
	PureSVD	$0.0304 \pm 0.0002$	$0.0388 \pm 0.0003$	$0.0544 \pm 0.0002$	$0.0109 \pm 0.0001$		
790	MultiDAE	$0.0586 \pm 0.0010$	$0.0756 \pm 0.0010$	$0.1000 \pm 0.0018$	$0.2567 \pm 0.0514$		
46	MultiVAE	$0.0569 \pm 0.0015$	$0.0733 \pm 0.0016$	$0.0973 \pm 0.0031$	$0.2835 \pm 0.0544$		
	$\beta$ -VAE	$0.0458 \pm 0.0051$	$0.0597 \pm 0.0053$	$0.0788 \pm 0.0096$	$0.2212 \pm 0.0441$		
	MacridVAE	$0.0650 \pm 0.0013$	$0.0833 \pm 0.0016$	$0.1102 \pm 0.0021$	$0.3387 \pm 0.0150$		
GR. Children	Top-Popular	$0.0349 \pm 0.0095$	$0.0452 \pm 0.0100$	$0.0654 \pm 0.0206$	$0.0000 \pm 0.0000$		
	PureSVD	$0.2300 \pm 0.0006$	$0.2498 \pm 0.0008$	$0.3684 \pm 0.0008$	$0.0053 \pm 0.0001$		
	MultiDAE	$0.3452 \pm 0.0033$	$0.3807 \pm 0.0033$	$0.4989 \pm 0.0040$	$0.2233 \pm 0.0538$		
	MultiVAE	$0.3432 \pm 0.0022$	$0.3766 \pm 0.0034$	$0.4989 \pm 0.0027$	$0.2239 \pm 0.0429$		
بجيخ	$\beta$ -VAE	$0.2999 \pm 0.0096$	$0.3335 \pm 0.0081$	$0.4426 \pm 0.0147$	$0.3175 \pm 0.0912$		
G	MacridVAE	$0.3764 \pm 0.0038$	$0.4082 \pm 0.0036$	$0.5406 \pm 0.0046$	$0.3078 \pm 0.0189$		

Table 5: Mean and standard deviation of disentanglement and interpretability over 5 seeds. Models tuned for NDCG@100. Best disentanglement and interpretability results are given in bold.

Dataset	Model	DISENTAN	GLEMENT	INTERPRETABILITY		
Dataset	Model	DISENTANG.	COMPLETE.	LIME-global	SHAP-global	
- 0	Top-Popular	-	-	-	-	
8	PureSVD	$0.2637 \pm 0.0082$	$0.3067 \pm 0.0077$	$0.4586 \pm 0.0179$	$0.5945 \pm 0.0201$	
ź	MultiDAE	$0.2255 \pm 0.0333$	$0.2568 \pm 0.0266$	$0.4285 \pm 0.0177$	$0.5207 \pm 0.0507$	
a di	MultiVAE	$0.2144 \pm 0.0079$	$0.2473 \pm 0.0145$	$0.4246 \pm 0.0124$	$0.5377 \pm 0.0353$	
Ameeon.Co	$\beta$ -VAE	$0.1301 \pm 0.0458$	$0.3054 \pm 0.1153$	$0.3379 \pm 0.0518$	$0.4284 \pm 0.0564$	
Α,	MacridVAE	$0.2024 \pm 0.0443$	$0.2611 \pm 0.0274$	$0.4669 \pm 0.0121$	$0.6267 \pm 0.0502$	
	Top-Popular	-	-	-	-	
←	PureSVD	$0.1565 \pm 0.0098$	$0.3586 \pm 0.0116$	$0.3035 \pm 0.0067$	$0.4390 \pm 0.0143$	
MIN	MultiDAE	$0.1777 \pm 0.0179$	$0.2605 \pm 0.0088$	$0.3351 \pm 0.0051$	$0.4784 \pm 0.0297$	
8	MultiVAE	$0.1859 \pm 0.0277$	$0.2894 \pm 0.0202$	$0.3168 \pm 0.0144$	$0.4694 \pm 0.0280$	
~	$\beta$ -VAE	$0.1305 \pm 0.0298$	$0.3251 \pm 0.0718$	$0.3013 \pm 0.0346$	$0.4109 \pm 0.0317$	
	MacridVAE	$0.1207 \pm 0.0091$	$0.2080 \pm 0.0381$	$0.3367 \pm 0.0204$	$0.4309 \pm 0.0332$	
	Top-Popular	-	-	-	-	
	PureSVD	$0.2987 \pm 0.0316$	$0.2687 \pm 0.0170$	$0.5200 \pm 0.0234$	$0.6601 \pm 0.0296$	
750	MultiDAE	$0.1909 \pm 0.0391$	$0.1917 \pm 0.0309$	$0.4744 \pm 0.0487$	$0.5330 \pm 0.0549$	
76	MultiVAE	$0.2190 \pm 0.0423$	$0.2323 \pm 0.0235$	$0.4884 \pm 0.0298$	$0.5712 \pm 0.0496$	
	$\beta$ -VAE	$0.1879 \pm 0.0148$	$0.2059 \pm 0.0199$	$0.4733 \pm 0.0179$	$0.5468 \pm 0.0133$	
	${\bf MacridVAE}$	$0.2208 \pm 0.0817$	$0.2366 \pm 0.0596$	$0.4473 \pm 0.0560$	$0.5571 \pm 0.1381$	
R. Children	Top-Popular	-	-	-	-	
	PureSVD	$0.1219 \pm 0.0059$	$0.2787 \pm 0.0097$	$0.4131 \pm 0.0081$	$0.4376 \pm 0.0104$	
	MultiDAE	$0.1309 \pm 0.0095$	$0.1849 \pm 0.0068$	$0.4379 \pm 0.0076$	$0.4916 \pm 0.0083$	
	MultiVAE	$0.1377 \pm 0.0190$	$0.1805 \pm 0.0319$	$0.4361 \pm 0.0099$	$0.4967 \pm 0.0259$	
چيخ	$\beta$ -VAE	$0.1053 \pm 0.0234$	$0.2010 \pm 0.0394$	$0.4287 \pm 0.0230$	$0.4645 \pm 0.0312$	
	MacridVAE	$0.1593 \pm 0.0219$	$0.2527 \pm 0.0230$	$0.4782 \pm 0.0162$	$0.5611 \pm 0.0341$	

ground truth category in the latent space, our quantitative evaluations show that its disentanglement can be surpassed by simpler models. Nevertheless, the interpretability of the representations of MacridVAE as measured by LIME/SHAP-global is the best among our reproduced models. This is in line with the original work findings (figure 3 in [32]) where the authors show that MacridVAE learns human-interpretable concepts in the dimensions of the latent space.

### 4.2 Effect of disentanglement on effectiveness (RQ2)

We study the relation between effectiveness, disentanglement and interpretability of learned representations through statistical correlation. Given that results within a model and/or dataset are correlated and violate the i.i.d assumption of common correlation coefficients, we use *repeated measurements correlation* [1] (RMCORR) that considers intra-group correlations. We present the intra-model RMCORR per dataset in fig. 1 and the intra-dataset RMCORR per model in fig. 2.

In table 4, MacridVAE shows the best recommendation effectiveness scores while being one of the models with the lowest disentanglement. While Ma et al. [32] attribute MacridVAE's state-of-the-art performance to its ability to disentangled user intentions, our reproducibility study shows a lack of consistent and statistically significant correlation between disentanglement and effectiveness measures; when accounting for the models (fig. 1) we observe a statistically significant correlation, however, no statistically significant correlation is observed when accounting for the datasets (fig. 2). This lack of correlation can be observed also for the other models and especially evident for PureSVD, which despite its highest disentanglement and completeness scores, falls behind the neural models in recommendation performance. On the other hand, the lack of correlation between disentanglement and effectiveness does not support the alleged trade-off between disentanglement and downstream task accuracy [5, 24, 27] in the RSs datasets and models that we study.

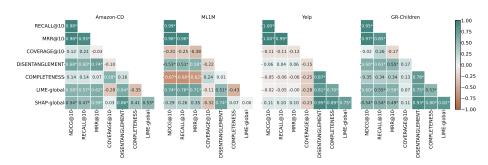


Fig. 1: Repeated measurements correlation of effectiveness, disentanglement and representation interpretability measures for each dataset. (\*) denotes statistical significance at p < 0.05.

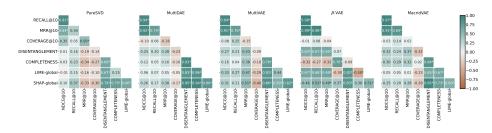


Fig. 2: Repeated measurements correlation of effectiveness, disentanglement and representation interpretability measures for each model. (\*) denotes statistical significance at  $\mathbf{p} < 0.05$ .

# 4.3 Effect of disentanglement on representation interpretability (RQ3)

In figs. 1 and 2 we observe a strong positive correlation (RMCORR  $\in$  [0.51, 0.95]) between disentanglement and LIME-/SHAP-global measures. This correlation holds across all the datasets and models that we run, especially for MacridVAE. This validates Ma et al. [32] claim that MacridVAE is able to encode user intents on the dimensions of the learned latent space. While one of the benefits of learning disentangled representations is better representation interpretability [2, 7, 30, 39, 48], to our best knowledge this is the first study to explicitly quantify this connection in RSs.

## 4.4 Limitations

In the context of RSs, the lack of ground truth factors of variation can prevent an objective evaluation of disentanglement. In this reproducibility study, similar to existing literature [35, 37, 46, 49–51, 53], we derive factors of variation from a combination of item content information and interaction data, which could have affected our ability to reproduce the disentanglement results of Nema et al. [35]. In this work, we investigate the link between disentanglement and interpretability with feature attribution methods like LIME and SHAP. However, both of these methods have their own limitations [23] which our proposed LIME/SHAP-global inherit. Finally, we note that we focus here in the interpretability of representations and how they relate to the ground truth factors of variation, which does not necessarily translate in the interpretability of the entire model and its downstream task output [39].

## 5 Conclusion

We presented the first reproducibility study of representation disentanglement in RSs and its association to recommendation effectiveness and representation interpretability. In our study, we found that it is non-trivial to reproduce disentanglement results without access to the ground truth factors of variation of existing work. The differences between our reproduced recommendation effectiveness scores and those reported in existing work are within 10% in ML1M and within 43% in GoodReads-Children. We also presented an adaptation of LIME and SHAP for quantifying representation interpretability. Our correlation analysis on the link between disentanglement and interpretability showed a strong positive correlation between the two, supporting qualitative evidence on their direct connection of prior work [9, 18, 32, 46, 53]. On the other hand, different from existing literature in representation disentanglement in other domains [21, 24, 27, 39], we did not find a consistent and statistically significant correlation between disentanglement and recommendation effectiveness in the datasets and models that we used.

# References

- Bakdash, J.Z., Marusich, L.R.: Repeated measures correlation. Frontiers in psychology 8, 456 (2017)
- Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence 35(8), 1798–1828 (2013)
- 3. Bergstra, J., Yamins, D., Cox, D.: Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In: International conference on machine learning. pp. 115–123. PMLR (2013)
- 4. Bohanec, M., Bratko, I.: Trading accuracy for simplicity in decision trees. Machine Learning 15, 223–250 (1994)
- 5. Burgess, C.P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., Lerchner, A.: Understanding disentangling in  $\beta$ -VAE. arXiv preprint arXiv:1804.03599 (2018)
- Chen, R.T., Li, X., Grosse, R.B., Duvenaud, D.K.: Isolating sources of disentanglement in variational autoencoders. Advances in neural information processing systems 31 (2018)
- Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: Infogan: Interpretable representation learning by information maximizing generative adversarial nets. Advances in neural information processing systems 29 (2016)
- 8. Cremonesi, P., Koren, Y., Turrin, R.: Performance of recommender algorithms on top-n recommendation tasks. In: Proceedings of the fourth ACM conference on Recommender systems. pp. 39–46 (2010)
- 9. De Divitiis, L., Becattini, F., Baecchi, C., Del Bimbo, A.: Disentangling features for fashion recommendation. ACM Transactions on Multimedia Computing, Communications and Applications 19(1s), 1–21 (2023)
- Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)
- 11. Dupont, E.: Learning disentangled joint continuous and discrete representations. Advances in neural information processing systems **31** (2018)
- 12. Eastwood, C., Williams, C.K.: A framework for the quantitative evaluation of disentangled representations. In: International conference on learning representations (2018)

- Ferrari Dacrema, M., Cremonesi, P., Jannach, D.: Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In: Proceedings of the 13th ACM conference on recommender systems. pp. 101–109 (2019)
- 14. Gao, C., Wang, S., Li, S., Chen, J., He, X., Lei, W., Li, B., Zhang, Y., Jiang, P.: Cirs: Bursting filter bubbles by counterfactual interactive recommender system. ACM Transactions on Information Systems 42(1), 1–27 (2023)
- Ge, M., Delgado-Battenfeld, C., Jannach, D.: Beyond accuracy: evaluating recommender systems by coverage and serendipity. In: Proceedings of the fourth ACM conference on Recommender systems. pp. 257–260 (2010)
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems 27 (2014)
- 17. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM computing surveys (CSUR) **51**(5), 1–42 (2018)
- Guo, Z., Li, G., Li, J., Chen, H.: TopicVAE: Topic-aware disentanglement representation learning for enhanced recommendation. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 511–520 (2022)
- 19. Harper, F.M., Konstan, J.A.: The movielens datasets: History and context. ACM transactions on interactive intelligent systems (tiis) 5(4), 1–19 (2015)
- He, X., Liao, L., Zhang, H., Nie, L., Hu, X., Chua, T.S.: Neural collaborative filtering. In: Proceedings of the 26th international conference on world wide web. pp. 173–182 (2017)
- 21. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.:  $\beta$ -VAE: Learning basic visual concepts with a constrained variational framework. In: International conference on learning representations (2016)
- 22. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. ACM Transactions on Information Systems (TOIS) **20**(4), 422–446 (2002)
- Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., Wortman Vaughan, J.: Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In: Proceedings of the 2020 CHI conference on human factors in computing systems. pp. 1–14 (2020)
- 24. Kim, H., Mnih, A.: Disentangling by factorising. In: International Conference on Machine Learning. pp. 2649–2658. PMLR (2018)
- 25. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: International Conference on Learning Representations (ICLR) (2014)
- 26. Lake, B.M., Ullman, T.D., Tenenbaum, J.B., Gershman, S.J.: Building machines that learn and think like people. Behavioral and brain sciences 40, e253 (2017)
- 27. Lezama, J.: Overcoming the disentanglement vs reconstruction trade-off via jacobian supervision. In: International Conference on Learning Representations (2019)
- Li, Y., Zhao, P., Wang, D., Xian, X., Liu, Y., Sheng, V.S.: Learning disentangled user representation based on controllable VAE for recommendation. In: International Conference on Database Systems for Advanced Applications. pp. 179–194. Springer (2021)
- 29. Liang, D., Krishnan, R.G., Hoffman, M.D., Jebara, T.: Variational autoencoders for collaborative filtering. In: Proceedings of the 2018 world wide web conference. pp. 689–698 (2018)
- 30. Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., Bachem, O.: Challenging common assumptions in the unsupervised learning of disentangled

- representations. In: international conference on machine learning. pp. 4114–4124. PMLR (2019)
- 31. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Advances in neural information processing systems **30** (2017)
- 32. Ma, J., Zhou, C., Cui, P., Yang, H., Zhu, W.: Learning disentangled representations for recommendation. Advances in neural information processing systems 32 (2019)
- 33. Meo, C., Mahon, L., Goyal, A., Dauwels, J.:  $\alpha$ -TCVAE: On the relationship between disentanglement and diversity. In: The Twelfth International Conference on Learning Representations (2024)
- 34. Michiels, L., Verachtert, R., Goethals, B.: Recpack: An (other) experimentation toolkit for top-n recommendation using implicit feedback data. In: Proceedings of the 16th ACM Conference on Recommender Systems. pp. 648–651 (2022)
- 35. Nema, P., Karatzoglou, A., Radlinski, F.: Disentangling preference representations for recommendation critiquing with  $\beta$ -VAE. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. pp. 1356–1365 (2021)
- 36. Ni, J., Li, J., McAuley, J.: Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). pp. 188–197 (2019)
- 37. Ren, X., Xia, L., Zhao, J., Yin, D., Huang, C.: Disentangled contrastive collaborative filtering. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 1137–1146. SIGIR '23 (2023)
- 38. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)
- Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., Zhong, C.: Interpretable machine learning: Fundamental principles and 10 grand challenges. Statistic Surveys 16, 1–85 (2022)
- Schmidhuber, J.: Learning factorial codes by predictability minimization. Neural computation 4(6), 863–879 (1992)
- 41. Shi, Y., Paige, B., Torr, P., et al.: Variational mixture-of-experts autoencoders for multi-modal deep generative models. Advances in neural information processing systems **32** (2019)
- 42. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International conference on machine learning. pp. 3319–3328. PMLR (2017)
- 43. Voorhees, E.: Proceedings of the 8th text retrieval conference. TREC-8 Question Answering Track Report pp. 77–82 (1999)
- 44. Wan, M., McAuley, J.: Item recommendation on monotonic behavior chains. In: Proceedings of the 12th ACM conference on recommender systems. pp. 86–94 (2018)
- 45. Wan, M., Misra, R., Nakashole, N., McAuley, J.: Fine-grained spoiler detection from large-scale review corpora. arXiv preprint arXiv:1905.13416 (2019)
- 46. Wang, X., Jin, H., Zhang, A., He, X., Xu, T., Chua, T.S.: Disentangled graph collaborative filtering. In: Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval. pp. 1001–1010 (2020)
- 47. Wang, X., Li, Q., Yu, D., Cui, P., Wang, Z., Xu, G.: Causal disentanglement for semantics-aware intent learning in recommendation. IEEE Transactions on Knowledge and Data Engineering (2022)

- 48. Wang, X., Chen, H., Tang, S., Wu, Z., Zhu, W.: Disentangled representation learning. arXiv preprint arXiv:2211.11695 (2022)
- 49. Yang, Y., Huang, C., Xia, L., Huang, C., Luo, D., Lin, K.: Debiased contrastive learning for sequential recommendation. In: Proceedings of the ACM Web Conference 2023. pp. 1063–1073 (2023)
- 50. Zhang, L., Liu, G., Liu, X., Wu, J.: Denoising item graph with disentangled learning for recommendation. IEEE Transactions on Knowledge and Data Engineering (2024)
- 51. Zhao, Z., Chen, J., Zhou, S., He, X., Cao, X., Zhang, F., Wu, W.: Popularity bias is not always evil: Disentangling benign and harmful bias for recommendation. IEEE Transactions on Knowledge and Data Engineering (2022)
- 52. Zheng, Y., Gao, C., Chang, J., Niu, Y., Song, Y., Jin, D., Li, Y.: Disentangling long and short-term interests for recommendation. In: Proceedings of the ACM Web Conference 2022. pp. 2256–2267 (2022)
- 53. Zheng, Y., Gao, C., Li, X., He, X., Li, Y., Jin, D.: Disentangling user interest and conformity for recommendation with causal embedding. In: Proceedings of the Web Conference 2021. pp. 2980–2991 (2021)