# Beyond Explicit Reports

## Comparing Data-Driven Approaches to Studying Underlying Dimensions of Music Preference

Kim, Jaehun; Manolios, Sandy; Demetriou, Andrew; Liem, Cynthia

**DOI**
10.1145/3320435.3320462

**Publication date**
2019

**Document Version**
Final published version

**Published in**
ACM UMAP 2019 - Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization

**Citation (APA)**
Kim, J., Manolios, S., Demetriou, A., & Liem, C. (2019). Beyond Explicit Reports: Comparing Data-Driven Approaches to Studying Underlying Dimensions of Music Preference. In *ACM UMAP 2019 - Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization* (pp. 285-293). ACM. https://doi.org/10.1145/3320435.3320462

# Beyond Explicit Reports: Comparing Data-Driven Approaches to Studying Underlying Dimensions of Music Preference

Jaehun Kim*
j.h.kim@tudelft.nl
Delft University of Technology
Delft, The Netherlands

Andrew M. Demetriou*
a.m.demetriou@tudelft.nl
Delft University of Technology
Delft, The Netherlands

Sandy Manolios*
s.manolios@tudelft.nl
Delft University of Technology
Delft, The Netherlands

Cynthia C. S. Liem
c.c.s.liem@tudelft.nl
Delft University of Technology
Delft, The Netherlands

## ABSTRACT

Prior research from the field of music psychology has suggested that there are factors common to music preference beyond individual genres. Specifically, research has shown that self-reported ratings of preference for individual musical genres can be reduced to 4 or 5 dimensions, which in turn have been shown to correlate to relevant psychological constructs, such as personality. However, the number of dimensions emerging from multiple studies has varied despite the care taken in conducting such research. Data-driven approaches offer opportunities to further this line of research with actual listening data, at a scale and scope surpassing that of traditional psychological studies. Although listening data can be considered more direct and comprehensive evidence of listening preference, transforming this data into meaningful measurements is non-trivial. In the current paper, we report on investigations seeking to find interpretable underlying dimensions of music taste, using implicit large-scale listening data. Offering a critical reflection on potential researchers' degrees of freedom, we adopt an explicit systematic approach, investigating the impact of varying different parameters, analysis, and normalization techniques. More precisely, we consider various ways to extract listening preference information from two large, openly available datasets of music listening behavior, making use of principal component analysis and variational autoencoders to extract potential underlying dimensions. Results and implications are discussed in light of prior psychological theory, and the potential of user listening data to further research on music preference.

## KEYWORDS

Multidisciplinary Approaches, Music Preferences, Listening Behavior, Latent Factor Models

*Authors contributed equally to the work


*UMAP '19, June 9–12, 2019, Larnaca, Cyprus*

ACM ISBN 978-1-4503-6021-0/19/06…$15.00
https://doi.org/10.1145/3320435.3320462

**ACM Reference Format:**
Jaehun Kim, Andrew M. Demetriou, Sandy Manolios, and Cynthia C. S. Liem. 2019. Beyond Explicit Reports: Comparing Data-Driven Approaches to Studying Underlying Dimensions of Music Preference. In *27th Conference on User Modeling, Adaptation and Personalization (UMAP '19), June 9–12, 2019, Larnaca, Cyprus.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3320435.3320462

## 1 INTRODUCTION

Despite interest and research contributions from multiple academic disciplines, the drivers of musical preference are incompletely understood. In music psychology, the general assumption is that measurement of music preference should show a finite number of interpretable underlying dimensions that cross genres. One such example has shown that self-reported preference for metal, alternative, and rock genres formed an "intense and rebellious" dimension, while preference for soul/funk, electronica/dance, and hip-hop/rap formed an "energetic and rhythmic" dimension [21].

While self-reports of music preference are useful in gaining insight into the minds of listeners, a further demonstration of the validity of such a dimensional scheme might be to examine implicit data, rather than explicit ratings of preference. For example, [21] examined the number of songs from certain genres that appeared in online music libraries (study 4). Given the prevalence of online streaming services, a reasonable approach might be to examine more detailed actual music listening behavior, beyond what songs simply appear in one's music library. One recent study used such data to examined diurnal and seasonal listening patterns [19].

### 1.1 Research Question

In light of previous studies that have shown evidence for underlying dimensions of music preference based on self-elicitation methods, in this work, we aim to assess whether a data-driven approach will also result in similarly interpretable 4 or 5 dimensional models as those found in prior research [21], when examining music listening behavior as opposed to explicit ratings of preference for music genres. To this aim, we will examine two large, publicly available datasets of music consumption behavior. As in prior music psychology research, we will extract principal components with factor rotations. In addition, we will apply a more recent non-linear approach from the data sciences: variational auto-encoding.

### 1.2 Methodological challenges

The current work is inspired by findings in music psychology, yet seeks to what extent data-driven approaches may be employed to replicate these. At present, our data-driven studies rely on existing, publicly available research datasets on music listening behavior.

Moving from traditional self-report based methodologies to data-driven methodologies allows for research on larger samples and larger-scale listening interactions. However, finding similarities between information obtained through self reports and information that can be distilled from listening behavior is a non-trivial task. Generally, for the analysis of music listening behavior, a number of decisions must be made by the researcher(s) to determine how to appropriately collect, pre-process, and analyze the data in each step of the research pipeline. Such decisions, previously referred to as 'researcher degrees of freedom' [30], often only have ambiguous criteria as guidelines and cause flexibility that may lead to falsely reporting positive results. In this paper, we seek to explicitly discuss and reflect on the researcher degrees of freedom encountered in the analysis of music listening behavior, in particular with regards to the various ways data may be pre-processed, and the various ways in which models might be selected for interpretation.

At a higher level, it also should be noted that the social and data-driven sciences have fundamental differences in methodological focus points. In the social sciences, theory-driven approaches are employed, in which the prime concern is demonstrating whether and to what degree independent and dependent variables are related, and interpreting the nature of these relationships. In contrast, data-driven (machine learning) methodology is particularly concerned with optimizing the fit of data to corresponding ground truth, while human interpretability of the variables and found relations is less of an explicit concern [17].

Thus, in theory-driven social sciences, insights are explicitly conditioned against relevant human theory and hypotheses, while in data-driven science, insights are expected to emerge from the data. With an increasing need for interpretability in many of today's artificial intelligence applications, an advantage of the former methodological line is that its research designs and outputs are designed with interpretability of results as a prime focus. However, the second methodological line operates on a larger search space, and thus may give a more comprehensive reading of what information may (or may not be) reflected by the data. In this work, while primarily performing data-driven analyses, we seek to balance these two viewpoints and to relate to prior psychological theory.

## 2 BACKGROUND

### 2.1 Music preference dimensions

The field of music psychology has shown evidence that preferences across music genres are not independent or random. Specifically, [21] suggest that personality characteristics are manifest in one's music listening behavior. For example, an individual who scores high in extraversion may have a preference for music that is typically present in social situations, such as genres one can dance to. A series of studies has shown evidence in favor of the presence of such interpretable underlying dimensions to music preferences, which have further been shown to correlate with psychological constructs, such as personality [3–5, 8, 10, 16, 21, 25].

**Table 1: Previous findings on music dimensions**

| Papers | No. of Dimensions | No. of Genres |
|---|---|---|
| Rentfrow & Gosling, 2003 [21] | 4 | 14 |
| George et al., 2007 [8] | 8 | 30 |
| Delsing et al., 2008 [4] | 4 | 11 |
| Schäfer & Sedlmeier, 2009 [26] | 6 | 25 |
| Dunn, Ruyter & Bouwhuis, 2011 [5] | 6 | 14 |
| Rentfrow, Goldberg & Levitin, 2011 [20] | 5 | 26 |
| Langmeyer et al., 2012 [16] | 4 | 14 |
| Brown, 2012 [3] | 4 | 12 |
| Greb, Schlotz & Steffens, 2017 [10] | 6 | 19 |

A key example is [21], in which Texas students (n > 1k) completed the Short Test of Music Preferences (STOMP). The researchers then performed a principal components analysis (PCA) with both orthogonal (e.g. varimax) and oblique (e.g. oblimin) rotations, and found 4 dimensions of music preferences. Those dimensions were: 1) Reflective and Complex, comprised of Blues, Jazz, Classical and Folk, 2) Intense and Rebellious, comprised of Rock, Alternative, and Heavy Metal, 3) Upbeat and Conventional, comprised of Country, Sound Tracks, Religious, and Pop, and 4) Energetic and Rhythmic, comprised of Rap/Hip-Hop, Soul/Funk, and Electronica/Dance. [20] later updated this model by dividing the Upbeat and Conventional dimension into Mellow, comprised of Pop, Soft-Rock, Soul and R&B, and Urban, comprised of Rap, Electronica, and Dance Music. This five factor model is called the MUSIC model and has now been used extensively in many studies about the diverse aspects of music preference, especially with regards to personality [25]. Subsequent studies showed a number of underlying dimensions, and while 4 dimensions were often found, overall the number varies from 4 [21] to 8 [8]. Reasons for the inconsistency may be differences in populations, as well as decisions made by researchers such as the procedure used to elicit preferences, and the number of genres rated which range from 11 [4] to 30 [8].

### 2.2 Self-reports vs. listening behavior

Most of the studies on the underlying structure of music preference are based on self reports, such as Likert-scale responses to assessments like the STOMP, or ratings of musical extracts. One notable exception is study 4 in [21], which consists of a confirmatory factor analysis (CFA) of their 4-dimensional model, performed on users' online music collections. The dataset was compiled by randomly selecting 20 songs per library, which were hand coded by judges into one of the 14 music genres present in the STOMP. They then assessed user's preference for each genre by the number of songs of that genre in the sample taken from their library.

One can wonder if self-reported preferences indeed match a user's music consumption. [5] collected listening time per genre, as well as responses to the STOMP for the same users, seeking to validate the model of [21]. [5] conclude that there are positive correlations, which however remain weak to moderate. This suggests that self-reported preferences do not perfectly match the listening behavior of the respondents. Several reasons for this difference can be evoked: genre definitions and their boundaries can be quite subjective; self-reported data can suffer from a social desirability

| Dataset | # Users | # Genres | Genre set |
|---|---|---|---|
| LFM1b[27] | 120,173 | 18 | Allmuic |
| MSD[2] | 109,959 | 27 | MASD[29] |

**Table 2: Description of datasets used in the experiments**

bias; and people may not listen to songs from genres proportionally to their liking of those genres. At the same time, listening data reflects a comprehensive, objective view on what people actually listened to, and therefore is of interest to us.

# 3 DATASETS

## 3.1 Acquisition

We studied diverse users' genre consumption profiles, which had previously been collected and processed into 2 publicly available datasets: LFM-1b [27], and the Million Song Dataset (MSD) [2]. Datasets were primarily chosen because they included the number of times a song was played by a user, thus resembling listening behavior. Secondary considerations included their overall size and the ease with which we could map appropriate genre classifications to the datasets. Our selection was not exhaustive, and was at least partially guided by convenience.

For purposes of comparison, we aimed to establish a user genre profile for both datasets such that the genre classifications of LFM-1b and MSD were as similar as possible to each other, and to prior research. The LFM-1b dataset provides user profiles separately from genre profiles (LFM1b-UGP) [28]. These genre profiles contain two types of classifications obtained by mapping the consumption data with data from two separate music catalogues: Allmusic and Freebase. The genre mappings differ significantly, with the first containing 20 and the second containing approximately 2000 genres. For the purposes of our study, we chose the former mapping. The profile is then established by aggregating the number of listening events of individual tracks by artist, and then remapping the artist to the corresponding genre [28]. For the MSD, we employed the Echonest user listening profile subset, which has the link to the MSD track id, which contains the music listening data of the users. For the experiments, we employed LFM1b-UGP with Allmusic genre mappings, and MSD with the MSD Allmusic Guide Style Dataset (MASD) genre mappings, where the vocabulary set is relatively clearer, and more similar to previous literature. Additionally we dropped the two most unpopular genres (`children's` and `spoken word`) from LFM1b-Allmusic, to establish even better resemblance with previous research. While we did not explore other potential genre mappings, some of which may have been significantly richer, it is likely that the specific mapping could have a significant effect on the results. For the purposes of this study we made a deliberate decision to seek out genre classifications that most closely resembled the number present in prior research.

Comparing our data to that of the most similar prior psychological study to ours [21], we did not sample a few songs per user, but used complete listening behavior. We worked with the original genre associations of the datasets, avoiding hand-coded human genre mappings (although this results in different taxonomies than the original study). Details regarding the datasets are described in Table 2.

| Id | TF | IDF | Description |
|---|---|---|---|
| 0 | $c_{u,g}$ | 1 | raw count |
| 1 | $\frac{c_{u,g}}{\sum_{g' \in \mathcal{G}} c_{u,g'}}$ | 1 | user-normalized |
| 2 | $1 + \log c_{u,g}$ | 1 | sub-linear |
| 3 | $1 + \log c_{u,g}$ | $\log \frac{\lvert\mathcal{U}\rvert}{\lvert u \in \mathcal{U} : g \in u\rvert}$ | TF-IDF |
| 4 | N/A | N/A | Likert Scale [1..7] |

**Table 3: Normalization techniques investigated in this work. $c_{u,g}$ refers to the raw listening count of user $u$ to the genre $g$, and the $\mathcal{G}$ and $\mathcal{U}$ are the set of genres and users, respectively.**

## 3.2 Normalization

In previous work, data was collected using ratings on either 7 or 9 point Likert scales, resulting in all the declared genre preferences falling within the given range (either $[1, 7]$ or $[1, 9]$). However, many publicly accessible datasets have only an implicit proxy of such preferences, as they often contain the number of times a user has listened to specific tracks. Since there might be a substantial gap between such implicit feedback and "true" preference, models derived using raw listening counts and those established with answers from questionnaires on musical preferences are not directly comparable. A further concern is the large number of observations with 0 play counts in a number of genres. Whether or not the data should be examined in its raw form, or transformed to better conform to prior research is an open question.

One relatively simple and intuitive method to estimate the preference out of such implicit feedback can be to apply an appropriate normalization technique. As with the genre mappings, the choice of normalization schemes may affect results. In an attempt to be as objective as possible, we analyzed the raw results in addition to four normalization schemes.

In Information Retrieval (IR) and Recommender Systems (RS) literature, it is common practice to penalize repeated interactions using normalization. TF-IDF [24] is a common technique developed in the Information Retrieval (IR) and Natural Language Processing (NLP) domain, and is also commonly used in the RS field due to its simplicity and sound rationale. Specifically, we chose 3 different variations of TF-IDF strategies, illustrated in Table 3. We also included another normalization strategy that transforms the distribution of the given raw count per genre to $\mathcal{U}(1, 8)$, discretized to integers such that the resulted values proportionately resemble Likert-scale data, ranged 1-7, collected from participants in previous literature.

# 4 ANALYTIC STRATEGY

## 4.1 Principal Components Analysis

Following prior research e.g. [5, 21], we conducted a series of principal components analyses (PCA). Principal axis factoring techniques are sometimes also used to determine a number of underlying factors in data e.g. [20]. However, we opted not to explore the multitude of principal axis factor techniques available, so that our results would be as comparable as possible to most prior research. To examine the suitability of the datasets for the analysis of underlying structures, we computed measures of sampling adequacy (MSA) using the Kaiser-Meyer-Olkin procedure for the overall datasets

and each individual genre within each dataset. This procedure estimates the proportion of variance in a dataset and within individual variables that is common or shared variance. Scores of less than .5 are generally considered to indicate that the dataset is not suitable for factor analysis. In order to determine the number of factors to extract, we conducted a parallel analysis, which has been shown to be a satisfactory method for determining the number components to extract [35]. Specifically, parallel analysis compares the eigenvalues of the test dataset with those derived from a Monte Carlo simulation of a correlation matrix of uncorrelated variables, comprised of the same number of simulation observations and variables. Each component extracted from the test dataset is then compared to each component extracted from the simulated matrix, until the test components value falls below that of the simulated matrix [35]. The number of components falling above this threshold is then reported. However, other methods for estimating the number of factors to extract could have been employed, including visually inspecting the scree plots, or loadings from a range of extracted components to determine which model is the most interpretable.

One technique used to assist in determining and interpreting underlying dimensions commonly used in psychology research is factor rotation. As latent factors may be oriented in any way in multidimensional space, several rotations are often carried out until a simple structure is observed: specifically, that variables have relatively large loadings on one factor, and relatively small or negligible loadings on all others [7]. The output from the simple structure is then evaluated in light of what the components or factors could represent. In our case, we extracted loadings after using the orthogonal rotation most commonly reported in prior research, varimax, a commonly used oblique rotation, direct oblimin, as well as with no rotation.

### 4.2 $\beta$-VAE

Variational AutoEncoder (VAE) [15] is an unsupervised learning model that aims for learning distribution $p_\theta(x)$, from which one can draw a sample $x \in \mathcal{R}^d$. In many real world setups, often the data domain $x$ depends on certain underlying variables. This allows one to learn a joint distribution of $x$ and $z$, where $z \in \mathcal{R}^k$ is the latent variable that generates $x$. Here, instead of finding task-specific distribution of $z$, VAE attempts to find it by given task, with a function $f(z;\theta)$ that is complex enough to generate an object $x$ and random variable $z$ from simple distribution such as the normal distribution $\mathcal{N}(0, I)$.

However, to maximize the likelihood $\mathbb{E}_{p(z)}[p_\theta(x|z)]$, one should evaluate the integration of the likelihood over $z$, which is not efficient since most points of $z$ will not contribute much since particular $p(x)$ of one's interest would rely only on very small subspace of $z$, which means most of the case drawn points will not be helpful for finding $p$. To alleviate this problem, one can introduce an inference model $q_\phi(z|x)$ which estimates $z$ from the observation $x$, hopefully shares some commonalities with $p_\theta(x|z)$ since both of them related to same latent variable $z$ and data domain of $x$. As a neural network $f(z;\theta)$ typically embody $p_\theta(x|z)$, another neural network $g(x;\phi)$ can be chosen for the instance for $q_\phi(z|x)$. In this case, by introducing already highly probable data point $x$, learning process can be relaxed. Thus, the main objective of original VAE is formulated as follows:

$$\mathcal{L}(\theta, \phi; x, z) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x) \,||\, p(z)) \quad (1)$$

The first term of the objective ensures the model has minimal reconstruction error, and the second term makes sure the inference model $f_\theta$ resembles $p(z)$. By finding parameters that maximize this objective, one can find both an inference model $g_\phi$ and a generative model $f_\theta$.

$\beta$-VAE [11] introduces a more freedom to control the contribution of those two terms as follows:

$$\mathcal{L}(\theta, \phi; x, z, \beta) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \beta D_{KL}(q_\phi(z|x) \,||\, p(z)) \quad (2)$$

where $\beta$ is coefficient to control the contribution of the KL divergence term. Tuning $\beta$ properly allows the latent variable $z$ to have more "disentanglement" such that each dimension of the variable has better distinguishable meaning [11]. In our study, we examined the variability of the model fit with different setup of $\beta$ along with the number of dimensions of $z$, assuring that the resulting latent dimensions have a more clear, uncorrelated meaning.

As for the architecture of $f_\theta$ and $g_\phi$, we use a simple structure that has two hidden layers with $h$ units each. To infer the mean and log variance, we put two feed-forward layers that project the hidden activation to the space of $z$, such that one can draw a sample $z$ from $Q(z|x) = \mathcal{N}(\mu(x), \Sigma(x))$. Consequentially, we have neural networks as follows:

$$f_\theta(z) = \sigma(W_2 \cdot \sigma(W_1 \cdot z)), \quad z \sim Q(z|x) \quad (3)$$

$$g_\phi(x) = \sigma(W_4 \cdot \sigma(W_3 \cdot x) + b_4) \quad (4)$$

$$\mu(x) = W_\mu \cdot g_\theta(x) \quad (5)$$

$$\Sigma(x) = \exp(W_\Sigma \cdot g_\theta(x)) \quad (6)$$

where $W_1 \in \mathcal{R}^{h \times h}$ and $W_2 \in \mathcal{R}^{d \times h}$ belong to $\theta = \{W_1, W_2\}$ and $W_3 \in \mathcal{R}^{h \times k}$, $W_4 \in \mathcal{R}^{d \times h}$, $b_4 \in \mathcal{R}^d$ belong to $\phi = \{W_3, W_4, b_4\}$. $W_\mu \in \mathcal{R}^{k \times h}$ and $W_\Sigma \in \mathcal{R}^{k \times h}$ are the weights to transform the output of $g_\phi$ to estimate $\mu(x)$ and $\Sigma(x)$. $\sigma$ is a non-linear transformation for which functions such as Rectified Linear Unit (ReLU) [18] is chosen.

In particular, we set $h = 1024$ and $\sigma = ReLU$. Instead of a bias term, we applied Batch Normalization (BN) [12] for every layer in $f_\theta$ and $g_\phi$ except the output layer to stabilize the internal covariate, which is particularly problematic when the input data is not normalized and has high variance such as the case where raw playcount per genre is used. As for the optimization, we employed ADAM [14] and set the learning rate as 0.001 for all the models.

## 5 EVALUATION

### 5.1 Model Fit

To evaluate the model fit, we randomly split the samples into training and testing groups with same number of users in each. We fit our model on the training group, and evaluated the model fit on the testing group. Since unified measures are required to evaluate

between models, we used Root Mean Squared Error (RMSE) and the coefficient of determination ($R^2$). RMSE measures how accurate the reconstruction of the model is, and $R^2$ measures the percentage of variance explained by the reconstructions. We repeated this process for 5 trials within each normalization scheme, for both the PCA and $\beta$-VAE iterations.

## 5.2 Measuring the Latent Variable

*5.2.1 Disentanglement.* To measure the degree to which the obtained latent variable is informative, we employed the disentanglement that is proposed in [11]. The rationale of this measure is to determine to what extent each dimension of the latent variable $z$ independently affects each particular aspect or property of the generated data $x$.

*5.2.2 Saliency Map.* Considering $\beta$-VAE and its non-linearity, it is not as straight forward as linear models to interpret the relationship between the latent dimensions and input dimensions. To approximate such relationships, one of the most prevalent methods is to traverse from a latent point to its neighbor by changing values of a specific dimension of interest while fixing other dimension values in order to confirm which property of the generated data is affected by that latent dimension. However, since the data dimensions of our experimental setup hold relatively more independence than such complex input, one can produce similar estimates by computing the partial derivative of the generative model $f(z)$ with respect to input $z$ [1, 31]. The main motivation of this approach is that one can approximate the linear relationship between the input and the output of given model by taking its first-order Taylor expansion:

$$f_\theta(z) \approx w^T z + b \tag{7}$$

where $w \in \mathcal{R}^k$ is the partial derivative of the generator $f$ to the drawn point $z_0$, and $b$ is the residual. Specifically, $w$ is computed:

$$w = \frac{\partial f_\theta}{\partial z}\Big|_{z_0} \tag{8}$$

As one can see in (7), $w$ approximates the linear dependency of input data to the function output. More specifically, it suggests which input latent dimension should be changed to affect the generated data dimensions. Since $w$ is point-wise estimation, we employed $\mathbb{E}_z[w]$ as the estimator of the global linear approximation, which is drawn from $\mathcal{N}(0, I)$.

# 6 RESULTS

## 6.1 Model Fit

*6.1.1 PCA.* Suitability for factor analysis was estimated with the KMO procedure. The LFM1B dataset showed better scores, with MSA scores above .9 for the overall dataset and for individual genres, within most normalization schemes. The MSD fared poorly, with scores around .6 for the overall dataset and for individual genres. This suggests that the genre mappings for the LFM1B dataset had more shared variance due to underlying factor structure than the MSD. To determine the number of factors to extract, we employed parallel analysis, the results of which can be found in Table 4. Other than that, we employed the standard pipeline; Singular Value Decomposition (SVD) is applied on the correlation matrix, from which we take the top-$k$ components. Although we applied different rotation techniques, there was no substantial difference among them.

| Normalization | LFM1b | MSD |
|---|---|---|
| raw count | 4 | 9 |
| user-normalized | 5 | 15 |
| sub-linear | 3 | 8 |
| TF-IDF | 3 | 8 |
| Likert [1..7] | 3 | 8 |

**Table 4: Optimal number of factors recommended by the parallel analysis. No variation is observed on within normalization; different splits does not affect the result.**

*6.1.2 $\beta$-VAE.* Since $\beta$ is one of the most influential factors that affects both the reconstruction and the disentanglement, we searched for a reasonable setup by sweeping a fixed range of candidate values $\beta' \in \{0.001, 0.01, 0.1, 1, 2.5, 5, 10, 25, 50, 100\}$ per all the normalizations and the datasets. From our experiments, both RMSE and disentanglement is improved with smaller values of $\beta'$. Since improvement of both measures plateaued in most of cases, we chose to pick $\beta = 0.001$ as the tentatively optimal value for our experimental setup for the rest of the paper. In most cases, disentanglement reached 1, and RMSE reached its best score. Note, that the case where the 'user-normalization' and 'TF-IDF' (on LFM1b) is applied implies that there might be room for improvement with lower setting of $\beta$.

*6.1.3 Between Models.* As Figure 1 illustrates, $\beta$-VAE in general shows substantially better fit in terms of the RMSE. The only exception was found when 'user-normalization' and 'TF-IDF' are applied on LFM1b, with latent dimensionality $\mathcal{R}^{15}$. Considering the aforementioned hyper-parameter search for $\beta$, this case might have been underfitted due to the insufficient search range.

However, in terms of $R^2$, PCA surpasses $\beta$-VAE's performance in most cases on the LFM1b data. This suggests that the PCA can cover more variance than $\beta$-VAE, while pointwise accuracy is better on $\beta$-VAE. On the other hand, $\beta$-VAE shows a significantly better performance than PCA on the MSD data. One explanation could be the different complexity of the genre mappings in the chosen datasets; the genre set from LFM1b could be relatively well separated linearly and less skewed compared to MSD, where, for instance, 6 out of 25 genres are sub-genres of 'rock'. It is also reflected in the result from parallel analysis, where on average more then 2 times the minimum number of factors are suggested to fit PCAs. Considering the result, the non-linearity and high capacity of $\beta$-VAE model can be more adaptable to such complexity.

## 6.2 Variability of Models

To investigate the variability introduced by different setups, we compute the correlation between normalization techniques between datasets and models, based on the average distance matrix, computed as follows:

$$corr_{n,m} = \rho(\tilde{C}^n_{i,j}, \tilde{C}^m_{i,j}), \quad i, j \in \mathcal{G} \backslash i == j \tag{9}$$

[1]For visualization purpose, we standardize the RMSE measure per dataset and normalization, which is denoted as $RMSE_z$.

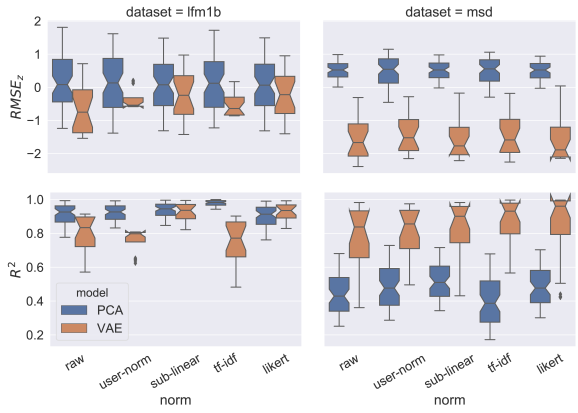



**Figure 1: Comparison between models in terms of model fit.**[1]

where $\tilde{C}^n$ refers to the approximated genre correlation matrix using nomalization $n$, and $\rho$ is Spearman's rank correlation. Specifically, $\tilde{C}$ is computed as follows:

$$\tilde{C} = \mathbb{E}[LL^\intercal] \tag{10}$$

where $L \in \mathcal{R}^{d \times k}$ is loadings in the case of PCA, and saliency map $w$ for $\beta$-VAE[2]. Figure 2 shows that there are substantial differences. For instance, for both MSD and LFM1b dataset, the PCA model shows a difference between 'user normalization' and other methods. However, no visible pattern is recognized in the case where $\beta$-VAE is used. We also examined the variability from the different rotation methods of PCA, but no substantial difference is observed, where all off-diagonal correlation in the matrix is higher than 0.85.

## 6.3 Does a Representative Model Exist?

Each permutation induced by different setup can be considered a different interpretation of the raw data, and as such it is not trivial to reject or accept each model when the goal of model selection rests mainly on reaching an understanding of the data. Thus, we investigated the representative model out of such variability for the sake of providing a sense of the middle-ground within the local model space that we studied. More specifically, the 'average' loadings are derived by applying the eigen decomposition on $\tilde{C}$, which is the averaged reconstructions of correlation matrix from each model. More formally:

$$\tilde{C} = V\Sigma^2 V^\intercal \tag{11}$$

where $V \in \mathcal{R}^{p \times k}$ is the eigenvectors and $\Sigma$ is eigenvalues of $\tilde{C}$. Averaged loadings can be calculated by $\tilde{L}^k = V^k\Sigma^k$, where the superscript $k$ indicates the loading is derived by using first $k$ components. Along with the marginalization over different normalizations, we also aggregated models drawn from different randomized splits.

Firstly, we compared the averaged model from 'user-normalized' cases and model averaged over other normalization techniques, which is indicated as visible clusters from Figure 2. Illustrated in Figure 5, it indicates that the 'non-user-normalized' model is in general one column shifted from the 'user-normalized' model, where the 1st to 4th components of the user-normalized model resemble the 2nd to 5th components of the 'other' models. This is confirmed by pairwise correlation. Considering that only the 'user normalization' reflects users normalized listening trend over genres, the first components of other models can be conjectured as 'listening intensity'. It is indeed shown that scores corresponding to the first principal component have very high correlations with the sum of normalized listening count ($> 0.95$), which implies that the listening intensity is most varying factor in the LFM1b dataset. As 2c suggests, a similar trend is observed in the MSD as well.

Consequentially, we also considered globally averaged models over all permutations within dataset and model whose loadings are illustrated in Figure 4 and 6, respectively. Although compared models show a large difference, resulted components resemble each other substantially. Figure 3 illustrates the agreement between components that are derived from two globally averaged models, and Table 5 lists 3 genres that scored the highest and lowest values in the averaged model components, which are again selected based on the agreement. As one can see in Figure 3a, with LFM1b, absolute correlation of each component is higher than 0.9 except the first component that still holds positive correlation 0.73[3].

However, as depicted in Figure 3b, the two models fit with MSD dataset do not agree with each other as much as in LFM1b's case. It suggests that only a few components agree, while the location of them varies. We assume it is caused by the complexity of the MSD dataset, due to its rather complicated vocabulary set. The $R^2$ measure also indicates that MSD dataset is a more difficult dataset to fit for PCA especially, where $k = \mathcal{R}^{15}$ solution still shows worse fit than $k = \mathcal{R}^2$ model on LFM1b.

# 7 DISCUSSION

## 7.1 Is Consensus Enough?

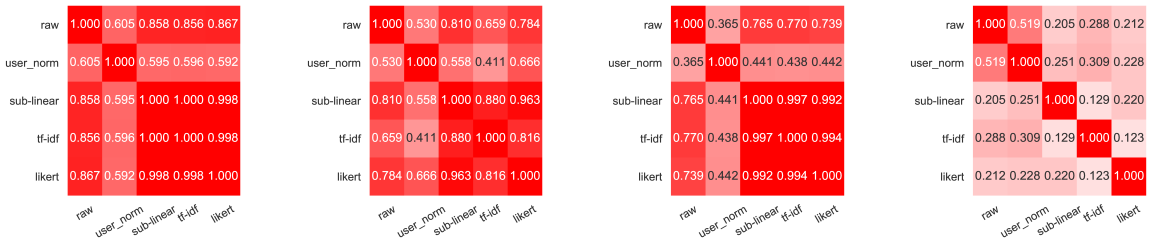
Results show that using various viewpoints on the same data, including permutations such as normalization, different models, different setups, bring different models. Out of this number of possibilities, choosing a model for further interpretation remains non-trivial since objectively measuring interpretability is not well established, to the best of our knowledge. Model fit based on reconstruction error might not necessarily reflect the degree to which the model is interpretable. Disentanglement, which at least assures that each dimension has less intertwined conceptual meaning between latent variables, helps in one aspect, but does not tackle the "quality" of such disentangled dimensions. As a result, although VAE in general performs better than PCA on RMSE and reaches similar level of disentanglement with our setup, this does not necessarily imply that the PCA approach is worse than VAE in terms of interpretability. Results indicate that individual PCA and VAE often encode incompatible dimensions from the same dataset.

One least-failing solution might be examining a model on the intersection of models, which we tried in 6.3. Results suggest that

[2]We normalized each rows of $w$ by the L2 norm to regulate the scale introduced by data normalization techniques.

[3]p-values for all the correlation indicate they are significant at $p < 0.01$ level.

[4]it is sorted by the loading values of each genres to the component. (i.e. the punk and `heavy metal` have the most and second most high value on PC1 of the representative model, and `vocal` has the lowest value, etc.)

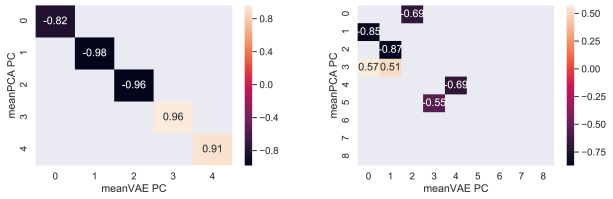



**(a) PCA-LFM1b** **(b) $\beta$-VAE-LFM1b** **(c) PCA-MSD** **(d) $\beta$-VAE-MSD**

**Figure 2: Variability introduced by the normalization techniques across datasets and models. Each plot visualizes a correlation matrix, derived from correlations between entries of the distance matrices corresponding to the different normalization techniques.**

| | LFM1b | | | | MSD | |
|---|---|---|---|---|---|---|
| Relatedness[4] | PC1 | PC2 | PC3 | PC4 | PC1 | PC2 |
| +++ | punk | rnb | country | heavy metal | Hip_Hop_Rap | Electronica |
| ++ | heavy metal | rap | blues | blues | Pop_Latin | Dance |
| + | rock | reggae | folk | reggae | RnB_Soul | Reggae |
| - | new age | heavy metal | new age | electronic | Punk | Pop_Contemporary |
| – | easy listening | classical | electronic | alternative | Rock_College | Country_Traditional |
| — | vocal | new age | rap | pop | Pop_Indie | Rock_Contemporary |

**Table 5: Summary of components that are strongly agreed between two models.**




**(a) LFM1b model agreement** **(b) MSD model agreement**

**Figure 3: Model agreement measured by component-wise rank correlation. Only significant values ($p < 0.01$) are displayed for clarity.**

two models averaged on normalizations reached a certain degree of agreement on LFM1b. However, on the MSD models results are rather divergent which is also suggested by substantially higher dimensionality required to encode the data sufficiently.

Considering the overall results, a formal, comprehensive definition of the interpretability is the key to solve a substantial amount of the questions we posed during the present work. If there is no human factor to determine which model is better, at least for machine learning approaches it is not trivial how to determine which solution is more humanly interpretable.

## 7.2 Limitations

One limitation of the present work is the lack of a sample elicited genre classification. As no fully objective classification exists, it is unclear what classification should be employed to answer such questions. We chose classifications in an effort to make our work comparable to prior research, yet it remains unclear how different classifications might change results.

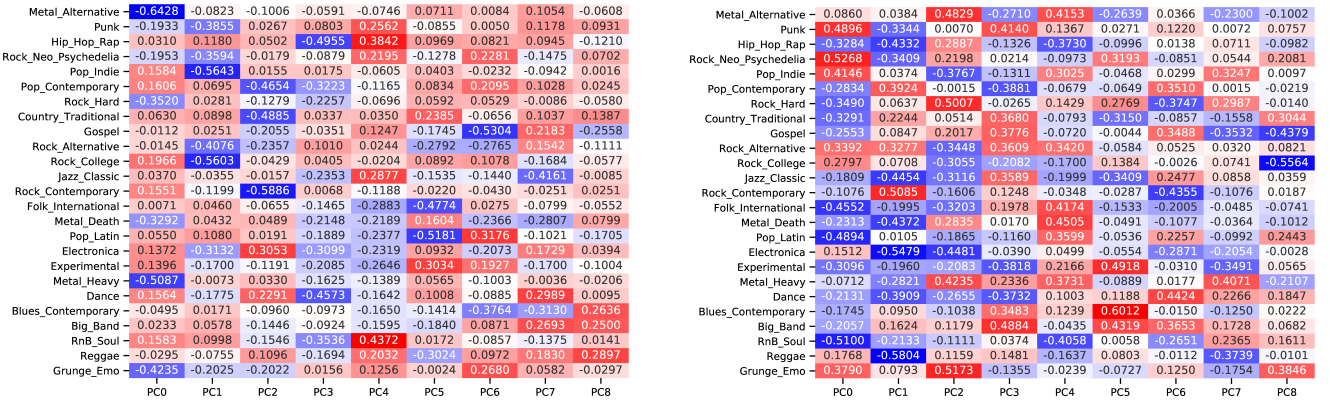
While the sample size in the present work is significantly larger than that observed in prior research, the population in question is not fully representative as it is limited to the users of the online services from which the data was released. Therefore, claims about generalizable dimensions of music listeners worldwide is not possible from our results. Further, explanatory variables such as socio-economic status, age, and geographic region were not available for both datasets, and therefore not employed in our analysis. Neither were data from personality assessments, which was a primary theoretical motivation for many of the psychology studies cited. Furthermore, the demographic data available for our LFM dataset was not equally distributed across regions, or age groupings.

While the intensity for one's preference for music overall has been shown to relate to how specifically one consumes music beyond the simple frequency of one's listening behavior [26], our only estimate of preference intensity is the overall number of music streams.

## 7.3 Future Work

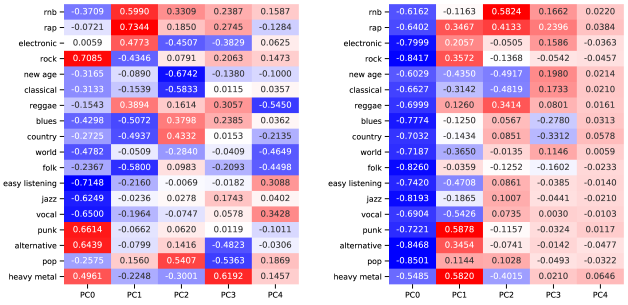
This research opens many leads for further studies. Firstly, an appropriately stratified sample across regions and age groups could confirm and extend findings. Indeed, only one dataset among those used included demographic data, which was not evenly distributed across regions, and was mainly composed of residents of the United States (20%) that were between 19 and 32 years old (70% of the total). Further, we were unable to compare the results by age and by country of users. This would have allowed us to compare the differences in taste among different populations.

A further extension could be to conduct an analysis more typical of music psychology research. This might entail conducting principal components and/or exploratory factor analyses on datasets similar to those used in this study. This could be followed up with



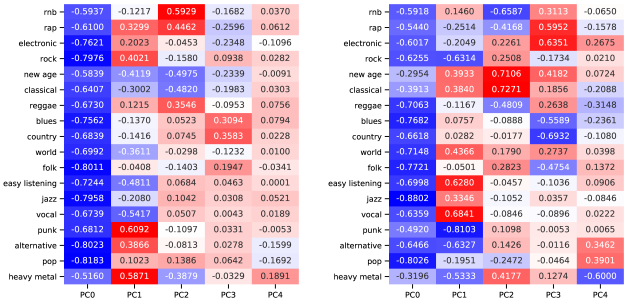

(a) PCA-MSD (b) $\beta$-VAE-MSD

**Figure 4: Representative loadings from possible models trained on MSD dataset with dimensions $k = 9$.**




(a) user normalized (b) other normalizations

**Figure 5: Comparison between average loadings from user-normalized PCA models and other models**




(a) PCA-LFM1b (b) $\beta$-VAE-LFM1b

**Figure 6: Representative loadings from possible models trained on LFM1b dataset with dimensions $k = 5$.**

a series of factor rotations, until a simple structure is observed. The components or factors could then be interpreted by the researchers, and examined for interpretability in light of the 4 or 5 dimensional models most commonly observed in music psychology research. As a final step, a confirmatory factor analysis can be conducted to assess model fit on a separate dataset comprised of similar genre mappings, or a randomly split half of the same dataset.

Another interesting step in the continuity of this work would be to ask users to complete additional questionnaires. Education level and other socio-demographic information about the users would allow us replicate, compare, and confirm findings on cultural omnivores and their characteristics in different countries over the last 30 years [6, 9, 13, 23, 32–34].

Many researchers have shown correlations between personality dimensions and music taste, although results are mixed [22, 25]. It would be useful to further those studies by correlating personality with dimensions derived from actual consumption data.

## 8 CONCLUSION

While it is clear that decisions made by researchers have the potential to impact results, we present here a data-driven process by which some of those effects may be accounted for. It is achieved by seeking the representative model out of a potential model space, determined, and potentially greatly inflated by, various decisions regarding data handling and analysis setups. Nevertheless, results still indicate variability with respect to the dataset. This suggests the need for further in-depth investigations on the effect of variability of dataset, genre-mappings, as well as consumption patterns on modeling the underlying dimensions of music preference.

If universality exists within music listening preferences as is suggested in prior research, taking a further step towards contributing to the development of the theoretical foundations investigated in the field of Music Psychology may be a potent next step from which more insight can emerge regarding the modelling of music listening behavior.

## ACKNOWLEDGMENTS

This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

## REFERENCES

[1] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. 2010. How to Explain Individual Classification Decisions. *Journal of Machine Learning Research* 11 (2010), 1803–1831.

[2] Thierry Bertin-Mahieux, Daniel P. W. Ellis, Brian Whitman, and Paul Lamere. 2011. The Million Song Dataset. In *Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR*. University of Miami, 591–596.

[3] Robert A Brown. 2012. Music preferences and personality among Japanese university students. *International Journal of Psychology* 47, 4 (2012), 259–268.

[4] Marc JMH Delsing, Tom FM Ter Bogt, Rutger CME Engels, and Wim HJ Meeus. 2008. Adolescents' music preferences and personality characteristics. *European Journal of Personality: Published for the European Association of Personality Psychology* 22, 2 (2008), 109–130.

[5] Peter Gregory Dunn, Boris de Ruyter, and Don G Bouwhuis. 2012. Toward a better understanding of the relation between music preference, listening behavior, and personality. *Psychology of Music* 40, 4 (2012), 411–428.

[6] Michael Emmison. 2003. Social class and cultural mobility: reconfiguring the cultural omnivore thesis. *Journal of sociology* 39, 3 (2003), 211–230.

[7] Leandre R Fabrigar, Duane T Wegener, Robert C MacCallum, and Erin J Strahan. 1999. Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods* 4, 3 (1999), 272.

[8] Darren George, Kelly Stickle, Faith Rachid, and Alayne Wopnford. 2007. The association between types of music enjoyed and cognitive, behavioral, and personality factors of those who listen. *Psychomusicology: A Journal of Research in Music Cognition* 19, 2 (2007), 32.

[9] Amir Goldberg. 2011. Mapping shared understandings using relational class analysis: The case of the cultural omnivore reexamined. *Amer. J. Sociology* 116, 5 (2011), 1397–1436.

[10] Fabian Greb, Wolff Schlotz, and Jochen Steffens. 2017. Personal and situational influences on the functions of music listening. *Psychology of Music* (2017), 0305735617724883.

[11] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. $\beta$-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *5th International Conference on Learning Representations, ICLR*.

[12] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on Machine Learning, ICML (JMLR Workshop and Conference Proceedings)*, Vol. 37. JMLR.org, 448–456.

[13] Mads Meier Jæger and Tally Katz-Gerro. 2008. The rise of the cultural omnivore 1964-2004. *Research department of social policy and welfare services Working Paper* 9 (2008).

[14] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3th International Conference on Learning Representations, ICLR*.

[15] Diederik P. Kingma and Welling Max. 2014. Auto-Encoding Variational Bayes. In *2th International Conference on Learning Representations, ICLR*.

[16] Alexandra Langmeyer, Angelika Guglhör-Rudan, and Christian Tarnai. 2012. What do music preferences reveal about personality? *Journal of individual differences* (2012).

[17] Cynthia CS Liem, Markus Langer, Andrew Demetriou, Annemarie MF Hiemstra, Achmadnoer Sukma Wicaksana, Marise Ph Born, and Cornelius J König. 2018. Psychology Meets Machine Learning: Interdisciplinary Perspectives on Algorithmic Job Candidate Screening. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Springer, 197–253.

[18] Vinod Nair and Geoffrey E. Hinton. 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on Machine Learning ICML*. Omnipress, 807–814.

[19] Minsu Park, Jennifer Thom, Sarah Mennicken, Henriette Cramer, and Michael Macy. 2019. Global music streaming data reveal diurnal and seasonal patterns of affective preference. *Nature Human Behaviour* (2019), 1.

[20] Peter J Rentfrow, Lewis R Goldberg, and Daniel J Levitin. 2011. The structure of musical preferences: a five-factor model. *Journal of personality and social psychology* 100, 6 (2011), 1139.

[21] Peter J Rentfrow and Samuel D Gosling. 2003. The do re mi's of everyday life: the structure and personality correlates of music preferences. *Journal of personality and social psychology* 84, 6 (2003), 1236.

[22] Peter J Rentfrow and Samuel D Gosling. 2006. Message in a ballad: The role of music preferences in interpersonal perception. *Psychological science* 17, 3 (2006), 236–242.

[23] Gabriel Rossman and Richard A Peterson. 2015. The instability of omnivorous cultural taste over time. *Poetics* 52 (2015), 139–153.

[24] Gerard Salton and Michael McGill. 1984. *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company.

[25] Thomas Schäfer and Claudia Mehlhorn. 2017. Can personality traits predict musical style preferences? A meta-analysis. *Personality and Individual Differences* 116 (2017), 265–273.

[26] Thomas Schäfer and Peter Sedlmeier. 2009. From the functions of music to music preference. *Psychology of Music* 37, 3 (2009), 279–300.

[27] Markus Schedl. 2016. The LFM-1b Dataset for Music Retrieval and Recommendation. In *ICMR*. ACM, 103–110.

[28] Markus Schedl and Bruce Ferwerda. 2017. Large-Scale Analysis of Group-Specific Music Genre Taste from Collaborative Tags. In *ISM*. IEEE Computer Society, 479–482.

[29] Alexander Schindler, Rudolf Mayer, and Andreas Rauber. 2012. Facilitating Comprehensive Benchmarking Experiments on the Million Song Dataset. In *Proceedings of the 13th International Society for Music Information Retrieval Conference, ISMIR*. FEUP Edições, 469–474.

[30] Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science* 22, 11 (2011), 1359–1366.

[31] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *CoRR* abs/1312.6034 (2013).

[32] Alexander Vander Stichele and Rudi Laermans. 2006. Cultural participation in Flanders: Testing the cultural omnivore thesis with population data. *Poetics* 34, 1 (2006), 45–64.

[33] Alan Warde and Modesto Gayo-Cal. 2009. The anatomy of cultural omnivorousness: The case of the United Kingdom. *Poetics* 37, 2 (2009), 119–145.

[34] Alan Warde, David Wright, and Modesto Gayo-Cal. 2007. Understanding cultural omnivorousness: Or, the myth of the cultural omnivore. *Cultural sociology* 1, 2 (2007), 143–164.

[35] William R Zwick and Wayne F Velicer. 1986. Comparison of five rules for determining the number of components to retain. *Psychological bulletin* 99, 3 (1986), 432.