CCA and a Multi-way Extension for Investigating Common Components between Audio, Lyrics and Tags.

Matt McVicar[1] and Tijl De Bie[2] *

Intelligent Systems Lab, University of Bristol
matt.mcvicar@bristol.ac.uk, tijl.debie@gmail.com

Abstract. In our previous work, we used canonical correlation analysis (CCA) to extract shared information between audio and lyrical features for a set of songs. There, we discovered that what audio and lyrics share can be largely captured by two components that coincide with the dimensions of the core affect space: valence and arousal. In the current paper, we extend this work significantly in three ways. Firstly, we exploit the availability of the Million Song Dataset with the MusiXmatch lyrics data to expand the data set size. Secondly, we now also include social tags from Last.fm in our analysis, using CCA also between the tag space and the lyrics representations as well as between the tag and the audio representations of a song. Thirdly, we demonstrate how a multi-way extension of CCA can be used to study these three datasets simultaneously in an incorporated experiment. We find that 2-way CCA generally (but not always) reveals certain mood aspects of the song, although the exact aspect varies depending on the pair of data types used. The 3-way CCA extension identifies components that are somewhere in between the 2-way results and, interestingly, appears to be less prone to overfitting.

Keywords: Canonical Correlation Analysis, Mood Detection, Million Song Dataset, MusiXmatch, Last.fm.

1 Introduction

In this paper we ask what is shared between the audio, lyrics and social tags of popular songs. We employ canonical correlation analysis (CCA) to find maximally correlated projections of these three feature domains in an attempt to discover commonalities and themes. In our previous work [16] we attempted to maximise the correlation between audio and lyrical features and discovered that the optimal correlations related strongly to the mood of the piece.

We extend this work significantly in three ways. Firstly, we make use of the recently-available Million Song Dataset (MSD,[1]) to gather a large number of audio and lyrical features, verifying our previous work on a larget dataset. Secondly, we incorporate a third feature space based on social tags from Last.fm[1].

* This work was partially supported by the EPSRC grant number EP/E501214/1

[1] www.last.fm

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

On these three datasets we are able to conduct pairwise 2-dimensional CCA on the largest public dataset of this type currently available. Lastly, we demonstrate how 3-dimensional CCA can be used to investigate these data simultaneously, leading to a multi-modal analysis of three aspects of music. Whilst it was intuitive to us in our previous work that lyrics and audio would have mood in common, it is less clear to us what commonalities are shared between the other pairs of datasets. We therefore take a more serendipitous approach in this study, aiming to discover which features are most strongly related.

The rest of this paper is arranged as follows. In the remainder of this Section we discuss the relevant literature and background to our work. We detail our data collection methods, feature extraction, and framework in Section 2. Section 3 deals with the theory of CCA in 2 and 3 dimensions. In Section 4 we present our findings, which are discussed and concluded in Section 5.

1.1 The Core Affect Space

Although it may be the case that our CCA analysis leads to components other than emotion, we suspect that many will relate to the mood of the piece. We therefore review the analysis of mood in this Subsection.

Russell [17] proposed a method for placing emotions onto a two-dimensional *valence-arousal* space, known in psychology as the *core affect space* [18]. The valence of a word describes its attractiveness/aversiveness, whilst the arousal relates to the strength, energy or activation. An example of a high valence, high arousal word is ecstatic, whilst depressed would score low on both valence and arousal. A third dimension describing the dominance of an emotion has also been suggested [6], but rarely used by researchers. A more detailed visualisation of the valence/arousal space with example words is shown in Figure 1.

1.2 Relevant Works

The valence/arousal space has been used extensively by researchers in the field of automatic mood detection from audio. Harmonic and spectral features were used by [8], whilst in [5] they utilised low-level features such as the spectral centroid, rolloff, flux, slope, skewness and kurtosis. Time-varying features in the audio domain were employed by various authors [15, 20], which included MFCCs and short time Fourier transforms. For classification, many authors have utilised SVMs, which have been shown to successfully discriminate between features [9].

In the lyrical domain, [7] used bag-of-words (BoW) models as well as n-grams and term frequency-inverse document frequency (TFIDF) to classify mood based on lyrics, whilst [10] made use of the experimentally deduced affective norms of english words (ANEW) to assign valence and arousal scores to individual words in lyrics. Both of these studies were conducted on sets of 500-2,000 songs.

The first evidence of combining text and audio in mood classification can be seen in [21]. They employed BoW text features and psychological features for classification and demonstrated a correlation between the verbal emotion features and the emotions experienced by the listeners on a small set of 145

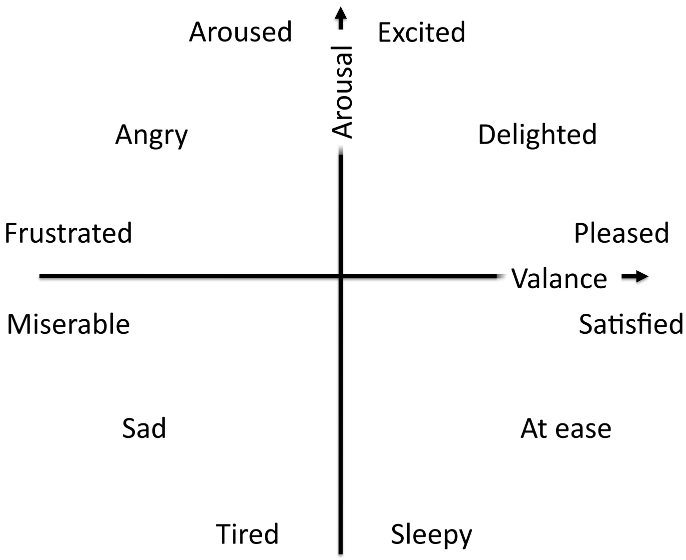

Fig. 1: The 2-dimensional valence/arousal space as proposed by Russell [17]. Words with high valence are more positive, whilst low valence words are pessimistic. High/low arousal words are energetic/restful respectively.

songs. A larger study was conducted in [13] where they classified 1,000 songs into 4 mood categories and found that by combining audio and lyrical features an increase in recognition accuracy was observed.

In the tag domain, [14] used the social website Last.fm to create a semantic mood space using latent semantic analysis. Via the use of a self-organising map, they reduce this high-dimensional space to a 2-D representation and compared this to Russell's valence/arousal space, with encouraging results.

In combining tag and audio data, [3] demonstrated that tag features were more informative than audio, whilst the combination was more informative still. This was conducted on a set of 1,612 songs and up to 5 mood or theme categories. Finally, a recent study considered regression of musical mood in continuous dimensional space using combinations of audio, lyrics and tags on a set of 2,648 UK pop songs [19].

Whilst insightful in terms of features and classification techniques, all of the studies previously mentioned were conducted on small datasets by todays standards (all significantly less than 10,000 songs). In this paper we address this issue in a truly large-scale, multi-modal analysis. We discuss our feature extraction and framework for our analysis in the following Section.

2 Data Collection & Framework

This section details our data collection methods and the motivation for our approach. We found the overlap of the Million Song, MusiXmatch and Last.fm datasets to be 223,815 songs in total, which was comprised of 197,436 training songs and 26,379 test songs. After removing songs which contained empty features, no lyrics or no tags, as well as those not in English, we were left with 101,235 (88%) training songs and 13,502 test songs (12%).

2.1 The Million Song Dataset

Devised as a way for researchers to conduct work on musical data without the need to purchase a large number of audio files, the Million Song Dataset was released on Feb 8^{th}, 2011. We downloaded this dataset in its entirety and extracted from it features relating to the audio qualities of the music. The features we specifically computed are shown in Table 1. We also give our interpretation of the features extracted, although there are some (e.g. danceability) where we are unsure of the feature extraction process.

Table 1: List of audio features extracted from the million song dataset, with interpretations.

Feature	Interpretation
Mean Bar Confidence	Average bar stability
Std Bar Confidence	Variation in bar stability
Mean Beat Confidence	Average beat stability
Std Beat Confidence	Variation in beat stability
Danceability	Danceability of track
Duration	Total track time in seconds
Key	Track harmonic centre (major keys only)
Key Confidence	Confidence in Key
Loudness	Loudness of track
Mode	Modality (major or minor) of track
Mode Confidence	Confidence in Mode
Mean Sections Confidence	Average confidence in section boundaries
Std Sections Confidence	Variation in section boundary confidences
Mean Seg. Conf.	Average confidence in segment boundaries
Mean Timbres 1-12	12 features relating to average sound quality
Std Timbres 1-12	12 features related to variation in sound quality
Tempo	Speed in Beats Per Minute
Loudness Max	Total maximum of track loudness
Loudness Start	Local max of loudness at the start of the track
Tatums Confidence	Confidence in tatum prediction
Time Signature	Predicted number of beats in a bar
Time Signature Confidence	Confidence in time signature

2.2 MusiXmatch

An addition to the MSD, the MusiXmatch dataset contains lyrical information for a subset of the million songs. The features are stored in bag-of-words format (for copyright reasons), and are stemmed versions of the top 5,000 words in the database. In order to ensure we had meaningful words, we restricted ourselves to the words which were part of the ANEW dataset [4], which reduced our dataset to 603 words. We converted the BoW data to a term frequency-inverse document frequency (TFIDF) score [11] via the following transformation.

Let the term frequency of the i^{th} feature from the j^{th} song be simply the BoW feature normalised by the count of this lyric's most frequent word:

$$TF_{i,j} = \frac{|\text{word } i \text{ appears in lyric } j|}{\text{maximum word count of lyric } j}$$

where $|\cdot|$ denotes 'number of'. The inverse document frequency measures the importance of a word in the database as a whole and is calculated as:

$$IDF_i = \log \frac{\text{total number of lyrics}}{|\text{lyrics containing word } i| + 1}$$

(we include the +1 term to avoid potentially dividing by 0). The TFIDF score is then the product of these two values:

$$TFIDF_{i,j} = TF_{i,j} \times IDFi$$

The TFIDF score gives an indication of the importance of a word within a particular song and the entire database. Note that we used the ANEW database simply to construct a dictionary of words which contain some emotive content - no experimental valence/arousal or mood scores were incorporated into our feature matrix.

2.3 Last.fm Data

The Last.fm data contains information on user-generated tags and artist similarities, although we neglect the latter for the purpose of this study. The dataset contains information on 943,347 tracks matched to the MSD and tag counts for each song. We discovered 522,366 unique tags although only considered tags which appeared in at least 1,000 songs, which resulted in 829 features. The top tags from the dataset were *Rock, Pop, Alternative, Indie* and *Electronic*. We constructed a TF-IDF score for each tag in each song analogously to the previous section. Although it would have been possible to filter the tags according to the ANEW database as per the lyrics, we know that tags contain information other than mood, such as genre data. We are optimistic that our algorithm may pick up such information, and so did not filter the Last.fm tags.

2.4 Framework

In our previous work [16] we introduced an exploratory framework for the use of CCA in correlating audio and lyrical features. We briefly recap this framework for 2-way CCA before extending it to use in 3 datasets.

We are interested in what is consistent between the audio, lyrics and tags of a song. In previous work, researchers have searched for a function f which maps audio to mood [$f(audio) = mood$], else from lyrics or tags [$g(lyrics) = mood$, $h(tags) = mood$]. In our 2-way CCA we seek functions which satisfy one of:

$$\begin{aligned} f(audio) &\approx g(lyrics) \\ f(audio) &\approx h(tags) \\ g(lyrics) &\approx h(tags) \end{aligned}$$

to a good approximation and for a large number of songs. Previously, we assumed that the first relationship in the above equations captured some aspect of mood, knowing of no other commonalities between the audio and lyrics of a song. This was verified by using 2-way CCA to find such functions f and g. In this study, we take a more serendipitous approach. We will use 2-way CCA on each pair of datasets and see which kinds of commonalities are found. Perhaps they will relate to mood, but we hope to discover other relationships and correlations within the data. The extension of this work to 3 dimensions follows a similar framework. We now seek functions f, g and h such that:

$$f(audio) \approx g(lyrics) \approx h(tags) \tag{1}$$

simultaneously. Again, these functions will not hold true for every song, but we hope they are approximately true for a large number of songs. The next Section deals with the theory of canonical correlation analysis.

3 Canonical Correlation Analysis and a 3-Way Extension

3.1 2-Way CCA

Given two datasets $X \in \mathbb{R}^{n \times d_x}$ and $Y \in \mathbb{R}^{n \times d_y}$, canonical correlation analysis finds what is consistent between them. This is realised by finding projections of X and Y through the dataset which maximise their correlation. In other words, we seek weight vectors $w_x \in \mathbb{R}^{d_x}$, $w_y \in \mathbb{R}^{d_y}$ such that the angle θ between Xw_x and Yw_y is minimised:

$$\{w_x^*, w_y^*\} = \underset{w_x, w_y}{\operatorname{argmin}} \ \theta(Xw_x, Yw_y)$$

Conveniently, this can be realised as a generalised eigenvector problem (a full derivation can be found in, for example, [2]):

$$\begin{pmatrix} 0 & X^TY \\ Y^TX & 0 \end{pmatrix} \begin{pmatrix} w_x \\ w_y \end{pmatrix} = \lambda \begin{pmatrix} X^TX & 0 \\ 0 & Y^TY \end{pmatrix} \begin{pmatrix} w_x \\ w_y \end{pmatrix} \tag{2}$$

In our experiments, X and Y will represent data matrices formed from the MSD, MusiXmatch or Last.fm datasets. The eigenvalue λ is the achieved correlation between the two datasets and (w_x, w_y) are the importance of each vector in the corresponding data space. The eigenvectors corresponding to λ can be sorted by magnitude to give a rank of feature importance in each of the data spaces.

3.2 3-Way CCA

Whilst it will be insightful to see the pairwise 2-way correlations between the three datasets, it would be more satisfying to investigate what is consistent between all 3 simultaneously. Various ways of exploring this have been explored in [12] - a natural extension in our setting can be motived as follows. Consider three datasets $X \in \mathbb{R}^{n \times d_x}$, $Y \in \mathbb{R}^{m \times d_y}$, $X \in \mathbb{R}^{p \times d_z}$. We motivate the correlation of these three variables graphically. Consider 3 datasets and (for ease of plotting) 3 songs within this set. A potential set of projections Xw_X, Yw_Y, and Zw_Z is shown in Figure 2.

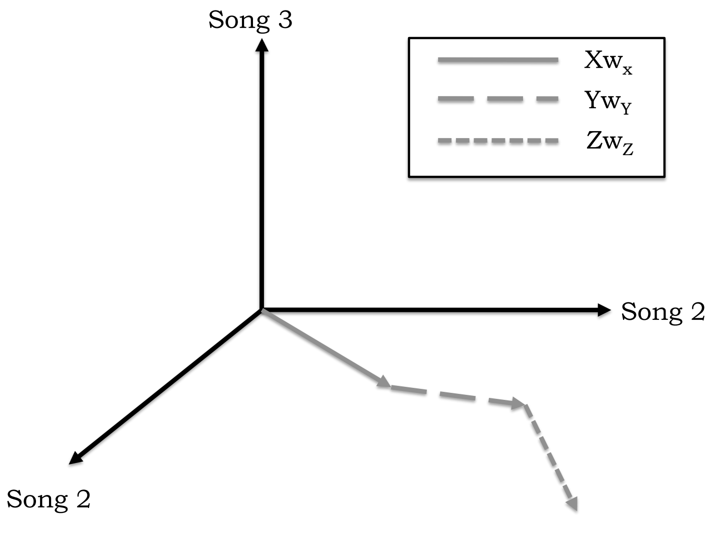

Fig. 2: Motivation for 3-way CCA on 3 example songs, showing the projections Xw_X, Yw_Y, Zw_Z.

It is clear that the three projections are strongly correlated if the norm of their sum is large. However, this is easy to obtain if each of the projections is arbitrarily large. We therefore enforce the constraint that the individual lengths are bounded, and solve the following optimization problem:

$$\max_{w_x,w_y,w_z} \|Xw_x + Yw_y + Zw_z + 1\|^2$$
$$s.t. \quad \|Xw_X\|^2 + \|Yw_Y\|^2 + \|Zw_Z\|^2 = 1$$

Solving the above via the method of Lagrange multipliers, we obtain

$$\frac{1}{2}\frac{\partial}{\partial w_*}\Big[\|Xw_X + Yw_Y + Zw_Z\|^2 - \lambda\Big(\|Xw_X\|^2 + \|Yw_Y\|^2 + \|Zw_Z\|^2\Big)\Big] = 0$$

where the asterix $*$ represents partial differentiation with respect to the appropriate variable. This leads to the simultaneous equations

$$X^TXw_X + X^TYw_Y + X^TZw_Z - \lambda X^TXw_X = 0$$
$$Y^TXw_X + Y^TYw_Y + Y^TZw_Z - \lambda Y^TXw_Y = 0$$
$$Z^TXw_X + Z^TYw_Y + Z^TZw_Z - \lambda Z^TZw_Z = 0$$

which, in matrix form, is

$$\begin{pmatrix} 0 & X^TY & X^TZ \\ Y^TX & 0 & Y^TZ \\ Z^TX & Z^TY & 0 \end{pmatrix}\begin{pmatrix} w_X \\ w_Y \\ w_Z \end{pmatrix} = (\lambda - 1)\begin{pmatrix} X^TX & 0 & 0 \\ 0 & Y^TY & 0 \\ 0 & 0 & Z^TZ \end{pmatrix}\begin{pmatrix} w_X \\ w_Y \\ w_Z \end{pmatrix} \tag{3}$$

Substituting $\lambda \to \lambda - 1$, we see that 3-dimensional CCA is an obvious extension of the 2-dimensional set-up seen in Equation 2. Note however that the λ is now a generalisation of the notion of correlation, and is not necessarily bounded in absolute value by 1. In our setting, the datasets X, Y and Z represent the MSD, MusiXmatch and Last.fm datasets and our aim will be to maximise the correlation between them. Our experimental results using pairwise CCA and 3-way CCA are presented in the next Section.

4 Experiments

4.1 Audio - Lyrical CCA

We begin with a reproduction of our previous work [16] which uses CCA on audio and lyrical datasets. This will serve to verify our method scales to datasets of realistic sizes. The projections of the Audio and Lyrical datasets, ranked by test correlation magnitude, are shown in Table 2. In each pairwise CCA experiments we found the significance of the correlations under a χ^2 distribution to be numerically 0, owing to the extremely large data sizes. It is therefore more important to look at the magnitude of the correlations rather than their significance in the following experiments.

These projections agree with our previous finding that mood is one of the common components between audio and lyrics. In the first component, words

Table 2: Features with largest weights using Audio and Lyrical features in 2-way CCA, first 3 CCA components. Training correlations on the first three components were 0.5032, 0.4484 and 0.2409 whilst the corresponding test correlations were 0.5034, 0.4286 and 0.2875.

CCA	Lowest		Highest	
Comp.	**Lyrical Feature**	**Lyrical Weight**	**Lyrical Paper**	**Lyrical Weight**
	Death	-0.0358	Love	0.0573
	Dead	-0.0274	Baby	0.0394
1	Burn	-0.0239	Blue	0.0197
	Hate	-0.0219	Girl	0.0190
	Pain	-0.0204	Man	0.0170
	Audio Feature	**Audio Weight**	**Audio Feature**	**Audio Weight**
	Loudness Max	-0.6824	Mean Timbre 1	0.6559
	Loudness	-0.0711	Mean Seg. Conf.	0.1638
1	Duration	-0.0413	Loudness Start	0.1539
	Mean Timbre 10	-0.0311	Mean Timbre 5	0.0698
	Std Timbre 6	-0.0222	Mean Timbre 6	0.0649
	Lowest		**Highest**	
	Lyrical Feature	**Lyrical Weight**	**Lyrical Feature**	**Lyrical Weight**
	Dream	-0.0182	Man	0.0354
	Love	-0.0177	Hit	0.0325
2	Heart	-0.0142	Girl	0.0303
	Fall	-0.0117	Rock	0.0291
	Lonely	-0.0113	Baby	0.0268
	Audio Feature	**Audio Weight**	**Audio Feature**	**Audio Weight**
	Loudness Max	-0.5568	Mean Timbre 1	0.7141
	Loudness Start	-0.2846	Loudness	0.1424
2	Std Seg. Conf.	-0.0855	Std Timbre 8	0.1233
	Std Timbre 4	-0.0525	Mean Seg. Conf.	0.1227
	Std Timbre 1	-0.0402	Mean Timbre 8	0.0446
	Lowest		**Highest**	
	Lyrical Feature	**Lyrical Weight**	**Lyrical Feature**	**Lyrical Weight**
	Baby	-0.0304	Man	0.0572
	Fight	-0.0281	Love	0.0409
3	Hate	-0.0223	Dream	0.0341
	Girl	-0.0223	Child	0.0301
	Scream	-0.0199	Dark	0.0295
	Audio Feature	**Audio Weight**	**Audio Feature**	**Audio Weight**
	Mean Timbre 1	-0.6501	Loudness Max	0.5613
	Loudness Start	-0.2281	Duration	0.1874
3	Std Timbre 6	-0.1507	Loudness	0.1377
	Std Seg. Conf.	-0.0898	Std Timbre 8	0.1050
	Tatums Conf.	-0.0850	Std Timbre 10	0.0891

with low weights appear more aggressive, whilst more optimistic words have the highest weights. This suggests that this CCA component has captured the notion of valence. Audio features in this domain show that high valence songs are loud, whilst low valence words have important timbre features.

The second CCA component appears to have identified relaxed lyrics at one extreme and more active words at the other. We consider this to be a realisation of the arousal dimension. In the audio domain, loudness and timbre again seems to play an important role. It is more difficult to interpret the third CCA component, although the sharp decay of test correlation values show that the first two CCA components dominate the analysis.

4.2 Audio - Tag CCA

We now investigate 2-way CCA on audio/tag data, using Last.fm tags in place of the lyrical data from Subsection 4.1. Components 1-3 are shown in Figure 3.

The first component of this CCA analysis seems to have found that the maximal correlation can be obtained by having tags associated with metal tags at one extreme and more serene tags at the other. The audio features in this CCA component seems to be well described by the later timbre features.

In the second component, we also see an obvious trend, with modern urban genre tags receiving high weights and more traditional music at the other. In the audio space, these genres seem to be associated with timbre and audio features.

The correlations between these two sets is so strong that we can even interpret the third CCA component, which has identified modern electronic music and acoustic blues/country as strongly opposing tags in this dimension. Interestingly, components 2 and 3 appear to have identified two distinct types of 'oldies' music (folk/blues respectively). In the audio domain these are accompanied by structural stability (segment/tatum confidence) features.

4.3 Lyrical - Tag CCA

The first three CCA components of this experiment are shown in Figure 4.

In the first component it seems we are distinguishing heavy metal genres from less aggressive styles. In the lyrical domain we see that the words with low weights hold strongly negative valence; those with high weights are more optimistic. The authors find the notion of Melodic Black Metal somewhat oxymoronic.

The second component also has a clear trend - extremes in this dimension appear to be hip-hop/rap vs. worship music. We postulate that this represents the dominance dimension mentioned in the Introduction, with the lyrical weights corroborating this. In the third component we see no particular trend, which is supported by the low correlation of 0.1807. Comparison with the training correlation of 0.4826 suggests that this component is suffering from overfitting.

4.4 3-way Experiment

We display our results from 3-way CCA in Table 5.

Table 3: Features with largest weights using Audio and Tag Features in 2-way CCA, first 3 CCA components. Training correlations on the on these components were 0.7361, 0.6432 and 0.5725 whilst the corresponding test correlations were 0.5685, 0.5237 and 0.3428 respectively.

CCA	Lowest		Highest	
comp.	**Tag Feature**	**Tag Weight**	**Tag Feature**	**Tag Weight**
1	Female Vocalists	-0.0352	Metal	0.0672
	Acoustic	-0.0304	Death Metal	0.0542
	Singer-Songwriter	-0.0289	Brutal Death Metal	0.0425
	Classic country	-0.0271	Punk rock	0.0378
	Folk	-0.0265	Metalcore	0.0371
	Audio Feature	**Audio Weight**	**Audio Feature**	**Audio Weight**
1	Mean Timbre 1	-0.5314	Loudness Max	0.7460
	Loudness Start	-0.1700	Std Timbre 6	0.0988
	Mean Timbre 6	-0.1558	Mean Timbre 2	0.0500
	Mean Seg. Conf.	-0.1469	Mean Timbre 3	0.0491
	Mean Timbre 5	-0.1021	Std bar Conf.	0.0267
	Lowest		**Highest**	
	Tag Feature	**Tag Weight**	**Tag Feature**	**Tag Weight**
2	Oldies	-0.0153	Hip-Hop	0.0418
	Beautiful	-0.0132	Dance	0.0355
	60s	-0.0126	Hip hop	0.0353
	Singer-Songwriter	-0.0116	Rap	0.0351
	Folk	-0.0110	Rnb	0.0231
	Audio Feature	**Audio Weight**	**Audio Feature**	**Audio Weight**
2	Loudness Start	-0.5069	Mean Timbre 1	0.7522
	Loudness Max	-0.3506	Loudness	0.1248
	Mean Timbre 6	-0.0631	Std Timbre 8	0.0578
	Std Timbre 1	-0.0374	Mean Timbre 4	0.0497
	Std Seg. Conf.	-0.0360	Mean Timbre 10	0.0415
	Lowest		**Highest**	
	Tag Feature	**Tag Weight**	**Tag Feature**	**Tag Weight**
3	Electronic	-0.0284	Oldies	0.0335
	Dance	-0.0220	Classic Blues	0.0325
	Vocal Trance	-0.0198	Classic country	0.0290
	Epic	-0.0186	50s	0.0279
	Pop	-0.0181	Delta blues	0.0250
	Audio Feature	**Audio Weight**	**Audio Feature**	**Audio Weight**
3	Mean Timbre 1	-0.6988	Loudness Max	0.6416
	Mean Timbre 4	-0.1141	Mean Timbre 3	0.1404
	Tatums Conf.	-0.0649	Mean Seg. Conf.	0.0757
	Duration	-0.0589	Mean Timbre 6	0.0732
	Std Segs Conf.	-0.0556	Loudness Start	0.0507

Table 4: Features with largest weights using Lyrical and Tag Features in 2-way CCA, first three CCA components. Training correlations on these components were 0.5828, 0.4990 and 0.4826 whilst test correlations were 0.3984, 0.3713 and 0.1807 respectively.

CCA	**Lowest**		**Highest**	
comp.	**Lyrical Feature**	**Lyrical Weight**	**Lyrical Feature**	**Lyrical Weight**
1	Death	-0.1851	Love	0.2330
	Dead	-0.1201	Baby	0.1807
	Human	-0.1049	Girl	0.0792
	God	-0.0993	Christmas	0.0726
	Pain	-0.0925	Blue	0.0679
	Tag Feature	**Tag Weight**	**Tag Feature**	**Tag Weight**
1	Brutal Death Metal	-0.3029	Xmas	0.0785
	Death Metal	-0.2470	Female Vocalists	0.0718
	Metal	-0.2449	Oldies	0.0688
	Melodic black metal	-0.2029	Pop	0.0680
	Black metal	-0.1338	Rnb	0.0652
	Lowest		**Highest**	
	Lyrical Feature	**Lyrical Weight**	**Lyrical Feature**	**Lyrical Weight**
2	Hit	-0.1448	Christmas	0.4082
	Man	-0.1267	Snow	0.0907
	Rock	-0.1180	Glory	0.0607
	Money	-0.1073	Joy	0.0549
	Brother	-0.0999	Angel	0.0530
	Tag Feature	**Tag Weight**	**Tag Feature**	**Tag Weight**
2	Hip hop	-0.2312	Xmas	0.4111
	Rap	-0.2014	Christmas	0.1679
	Hip-Hop	-0.1927	Christian	0.0662
	Gangsta Rap	-0.1460	Female Vocalists	0.0501
	Underground hip hop	-0.1143	Worship	0.0480
	Lowest		**Highest**	
	Lyrical Feature	**Lyrical Weight**	**Lyrical Feature**	**Lyrical Weight**
3	Love	-0.0273	Christmas	0.6031
	Heart	-0.0262	Snow	0.0992
	Rain	-0.0247	Man	0.0800
	Alone	-0.0229	Rock	0.0716
	Dream	-0.0224	Hit	0.0702
	Tag Feature	**Tag Weight**	**Tag Feature**	**Tag Weight**
3	Love	-0.0399	Xmas	0.5932
	Female vocalists	-0.0257	Christmas	0.2381
	Alternative rock	-0.0252	Hip hop	0.1265
	Rain	-0.0237	Rap	0.0975
	Oldies	-0.0227	Hip-Hop	0.0906

Table 5: Summary of 3-way CCA analysis. CCA components are shown in rows, with the highest and lowest-weighted features of each data space (audio, lyrical, tag) occupying the columns. The generalised training correlations on the first three components were found to be 2.1749, 2.0005, and 1.76559 whilst the generalised test correlations were found to be 2.1809,2.0036 and 1.7595 (recall that these generalised correlations are not necessarily bounded in absolute value by 1). Abbreviations: DM = Death Metal, BM = Black Metal, SS = Singer-Songwriter, FV = Female Vocalists, AR = Alternative Rock, UHH = Underground hip hop.

CCA	**Lowest**		**Highest**		**Lowest**		**Highest**		**Lowest**		**Highest**	
Comp.	Audio Feature	Weight	Audio Feature	Weight	Word	Weight	Word	Weight	Tag Feature	Weight	Tag Feature	Weight
1	Loudness Max	-0.7008	Mean Timbre 1	0.6064	Death	-0.0346	Love	0.0572	Metal	-0.0581	FVs	0.0242
	Loudness	-0.0442	Loudness Start	0.1906	Dead	-0.0272	Baby	0.0382	DM	-0.0517	Pop	0.0189
	Std Timbre 6	-0.0426	Mean Segs Conf.	0.1553	Hate	-0.0220	Blue	0.0179	Brutal DM	-0.0510	Classic Country	0.0183
	Duration	-0.0305	Mean Timbre 6	0.0925	Burn	-0.0216	Girl	0.0171	Melodic BM	-0.0348	Oldies	0.0176
	Mean Timbre 10	-0.0255	Mean Timbre 5	0.0766	Pain	-0.0191	People	0.0147	Metalcore	-0.0267	Soul	0.0167
2	Loudness Max	-0.4676	Mean Timbre 1	0.6797	Dream	-0.0161	Hit	0.0372	Beautiful	-0.0183	Hip Hop	0.0630
	Loudness Start	-0.4107	Loudness	0.2063	Love	-0.0131	Man	0.0343	FVs	-0.0130	Hip-Hop	0.0625
	Std Segs Conf.	-0.0899	Std Timbre 8	0.1238	Heart	-0.0123	Rock	0.0320	Ambient	-0.0113	Rap	0.0596
	Std Timbre 1	-0.0568	Mean Segs Conf.	0.1031	Home	-0.0114	Girl	0.0291	Christian	-0.0112	Gangsta Rap	0.0333
	Std Timbre 4	-0.0469	Mean Timbre 10	0.0480	Sad	-0.0111	Baby	0.0283	Mellow	-0.0106	UHH	0.0260
3	Mean Timbre 1	-0.7004	Loudness Max	0.6402	Girl	-0.0124	Man	0.0268	Dance	-0.0266	Folk	0.0239
	Loudness Start	-0.1666	Loudness	0.0756	Fight	-0.0124	Blue	0.0177	Rock	-0.0207	Brutal DM	0.0208
	Mean Timbre 4	-0.0797	Mean Timbre 6	0.0672	Crash	-0.0107	Death	0.0176	Pop	-0.0183	Melodic BM	0.0195
	Std Timbre 6	-0.0683	Duration	0.0558	Alive	-0.0104	Christmas	0.0156	AR	-0.0174	Acoustic	0.0176
	Std Segs Conf.	-0.0547	Mean Timbre 3	0.0472	Baby	-0.0100	Dark	0.0136	Electronic	-0.0162	Classic Country	0.0171

In this incorporated experiment, the most prevalent dimension appears to relate to arousal - highly weighted tags and features are gentle in nature, with aggressive tags, lyrics and audio features. The second component seems to represent arousal. We struggle to find an explanation for the third component.

5 Discussion & Conclusions

In this Section, we discuss some of the findings from the previous Section, summarise the conclusions of our study and suggest areas for future work.

5.1 Discussion

It is clear there are similar components in this study across different experiments. For instance, the first component of the audio/lyrical 2-way CCA experiment in the lyrical domain (first few rows of Table 2) were very similar to the first component in the lyrical domain in the 3-way experiment (first rows of Table 5, second cell). It appears that both of these discovered dimensions are capturing the valence of the lyrics. To verify that these projections were indeed similar, we computed the correlation between them (ie Yw_Y from Table 2 with Yw_Y from Table 5), and found it to be 0.9979. The conclusion to be drawn is that the valence of lyrics is very easily captured, by comparing with audio and/or tag information.

We now turn our attention to the second CCA component. Interested in what 3-Way CCA analysis offered over pairwise CCA experiments, we investigated the correlations between each pair of lyrical and tag projections from all three experimental set-ups (2 pairwise and 3-Way). These are shown in Table 6.

Table 6: Comparison of Lyrical and Tag projections in pairwise and 3-way experiments.

(a) Lyrical Projections

CCA comp. 2	$\mathbf{YW_Y}$ Lyrics/Tags	3-Way CCA
Lyrics/Audio	0.8679	0.9899
Lyrics/Tags	-	0.8886

(b) Tag Projections

CCA comp. 2	$\mathbf{ZW_Z}$ Tags/Lyrics	3-Way CCA
Tags/Audio	0.7534	0.8853
Tags/Lyrics	-	0.9434

The first of these tables can be interpreted as follows. Recall that in the lyrics-audio CCA experiment we found the second component to describe the arousal of the lyrics. In the lyrics-tag space we found the second lyrical component related to the dominance of the lyrics. Recall that the correlations are equivalent to the angles between the projected datasets. Table 6(a) therefore shows that the cosines of the angles between these vectors and the third CCA component are 0.9899 and 0.8886 respectively, but that the cosine of the angle between themselves is 0.8679. This shows that the 3-Way CCA component sits somewhere between arousal and dominance, which can be verified by looking at the top and bottom-ranked words in Tables 2, 3 and 5.

A similar, and in fact stronger pattern can be observed in tag space by investigating Table 6(b). Again, the 3-way CCA analysis seems to be an intermediate between the 'old vs new' dimension observed in the audio-tag space (Table 3, second component) and the dominance discovered in the lyrical-tag space (Table 4, second component).

5.2 Conclusions & Further Work

In this paper, we have conducted a large-scale study of the correlations between audio, lyrical and tag features based on the Million Song Dataset. By the use of pairwise 2-dimensional CCA we demonstrated that the optimal correlations between these datasets appear to have reconstructed the valence/arousal/dominance dimensions of the core affect space, even though this was in no way imposed by the algorithm. In some cases, we discovered components which appeared to capture some genre information, such as the third component of Table 3.

By using 3-dimensional CCA, we studied the 3 datasets simultaneously and discovered that valence and arousal were the most correlated features. The correlations beyond 2 or 3 components are difficult to interpret, which fits well studies which describe the core affect space as a 2 or 3 dimensional space.

In our future work we would like to investigate different multiway CCA extensions such as those seen in [12], perhaps on new datasets as they are released. We also would like to more thoroughly investigate regularization techniques to avoid overfitting.

References

1. Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.
2. T. De Bie, N. Cristianini, and R. Rosipal. Eigenproblems in pattern recognition. In E. Bayro-Corrochano, editor, *Handbook of Computational Geometry for Pattern Recognition, Computer Vision, Neurocomputing and Robotics*. Springer-Verlag, 2004.
3. K. Bischoff, C.S. Firan, R. Paiu, W. Nejdl, C. Laurier, and M. Sordo. Music mood and theme classification-a hybrid approach. In *Proc. of the Intl. Society for Music Information Retrieval Conf., Kobe, Japan*, 2009.
4. M.M. Bradley and P.J. Lang. Affective norms for english words (anew): Instruction manual and affective ratings. *University of Florida: The Center for Research in Psychophysiology*, 1999.
5. J.J. Burred, M. Ramona, F. Cornu, and G. Peeters. Mirex-2010 single-label and multi-label classification tasks: ircamclassification09 submission. *MIREX 2010*, 2010.

6. R. Cowie, E. Douglas-Cowie, B. Apolloni, J. Taylor, A. Romano, and W. Fellenz. What a neural net needs to know about emotion words. *Computational intelligence and applications*, pages 109–114, 1999.
7. H. He, J. Jin, Y. Xiong, B. Chen, W. Sun, and L. Zhao. Language feature mining for music emotion classification via supervised learning from lyrics. *Advances in Computation and Intelligence*, pages 426–435, 2008.
8. X. Hu and J.S. Downie. When lyrics outperform audio for music mood classification: a feature analysis. In *Proceedings of ISMIR*, pages 1–6, 2010.
9. X. Hu, J.S. Downie, C. Laurier, M. Bay, and A.F. Ehmann. The 2007 mirex audio mood classification task: Lessons learned. In *Proceedings of the 9th International Conference on Music Information Retrieval*, pages 462–467. Citeseer, 2008.
10. Y. Hu, X. Chen, and D. Yang. Lyric-based song emotion detection with affective lexicon and fuzzy clustering method. In *Proceedings of ISMIR*, 2009.
11. K.S. Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.
12. J.R. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 1971.
13. C. Laurier, J. Grivolla, and P. Herrera. Multimodal music mood classification using audio and lyrics. In *Machine Learning and Applications, 2008. ICMLA'08. Seventh International Conference on*, pages 688–693. IEEE, 2008.
14. C. Laurier, M. Sordo, J. Serra, and P. Herrera. Music mood representations from social tags. In *Proceedings of the 10th International Society for Music Information Conference, Kobe, Japan.* Citeseer, 2009.
15. M.I. Mandel. Svm-based audio classification, tagging, and similarity submissions. *online Proc. of the 7th Annual Music Information Retrieval Evaluation eX-change (MIREX-2010)*, 2010.
16. M. McVicar, T. Freeman, and T. De Bie. Mining the correlation between lyrical and audio features and the emergence of mood. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.
17. J.A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.
18. J.A. Russell. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145, 2003.
19. B. Schuller, F. Weninger, and J. Dorfner. Multi-modal non-prototypical music mood analysis in continous space: reliability and performances. In *Proc. of the 12th International Society for Music Information Retrieval (ISMIR) Conference*, pages 759–764, 2011.
20. K. Seyerlehner, M. Schedl, T. Pohle, and P. Knees. Using block-level features for genre classification, tag classification and music similarity estimation. *Submission to Audio Music Similarity and Retrieval Task of MIREX 2010*, 2010.
21. D. Yang and W.S. Lee. Disambiguating music emotion using software agents. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR04)*, pages 52–58, 2004.