Dimensionalty Reduction and Cluster Analysis

Applied to Spotify Data

Julian Albert, Fabio Fehr, James Nevin

University of Cape Town
Department of Statistics

October 10, 2019

Contents

Introduction

Exploratory Data Analysis

Correlation

Correlations are an important assumption for reduction techniques.

	Acoustic	Danceability	Energy	Instrumental	Loudness	Speechiness	Valence
Acoustic	1.00	-0.46	-0.54	0.05	-0.64	0.32	-0.38
Danceability	-0.46	1.00	0.40	-0.15	0.48	0.11	0.58
Energy	-0.54	0.40	1.00	-0.23	0.75	0.34	0.37
Instrumental	0.05	-0.15	-0.23	1.00	-0.21	-0.25	-0.11
Loudness	-0.64	0.48	0.75	-0.21	1.00	-0.07	0.41
Speechiness	0.32	0.11	0.34	-0.25	-0.07	1.00	-0.01
Valence	-0.38	0.58	0.37	-0.11	0.41	-0.01	1.00

Table: Correlation Matrix of Retained Variable

Boxplot

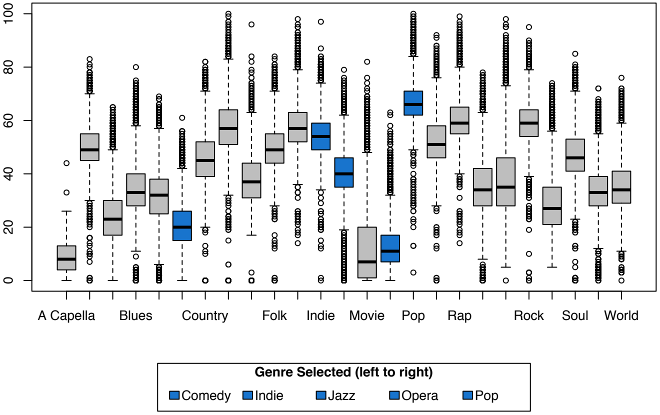


Figure: Boxplot of Popularity by Genre.

Sample Set

From the sampled set we also transform our response variable popularity into an ordinal grouping which categorises the popularity into "Popular", "Mediocre" and "Unpopular".

	Comedy	Indie	Jazz	Opera	Pop
Full	9681	9543	9441	8280	9386
Sampled	216	211	207	173	193

Table: Frequency Table of Retained Genres.

Dimension Reduction

The Methodology

- When dealing with high-dimensional data in a multivariate setting we can compress the data to a low-dimensional subspace that captures most of the variability.
- This is done to improve computational efficiency, inference and aid in visualisation.

PCA vs t-SNE

Method

PCA is a technique for deriving a reduced set of orthogonal linear projections of a single collection of correlated variables $\{X_1, X_2, \ldots, X_p\}$ where the projections are ordered by decreasing variances.

The method considers linear transformations $\mathbf{Z_j}$ for $j = 1, 2, \ldots, t$ representing the first t principal components of the form

$$Z_j = \vec{\phi}'\mathbf{X} = \phi_{j1}X_1 + \phi_{j2}X_2 + \cdots + \phi_{jp}X_p$$

and tries to minimise the “information" loss due to the transformation given by

$$\sum_{j=1}^{p} \text{var}\left(X_j\right) = \text{tr}\left(\mathbf{\Sigma}_{XX}\right) \tag{1}$$

Choosing No. of Components

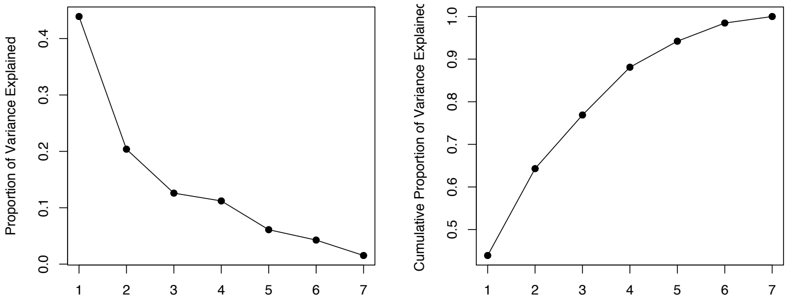
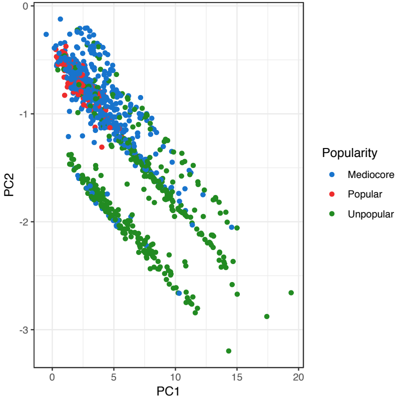
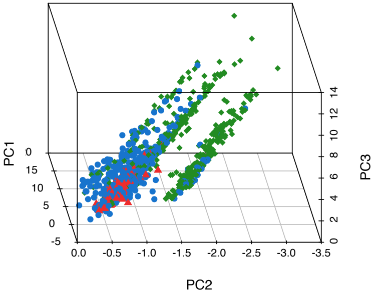


Figure: Variation Explained by Principal Components.

Results



(a) 2 PCs



(b) 3 PCs

Figure: First 2 and 3 Principal Component Reduction.

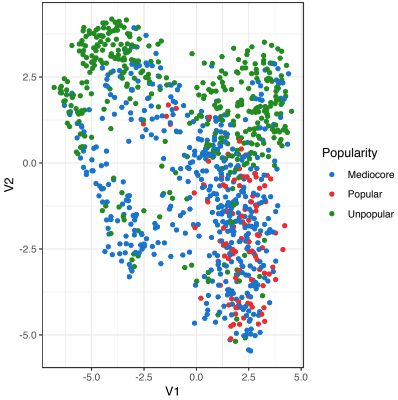
Method

t-distributed Stochastic Neighbour Embedding (t-SNE) is a non-linear dimension reduction technique. [MH08] explain the details of Stochastic Neighbour Embedding (SNE) and the extension introduced by t-distributed SNE (t-SNE). The cost function to be minimised is given by

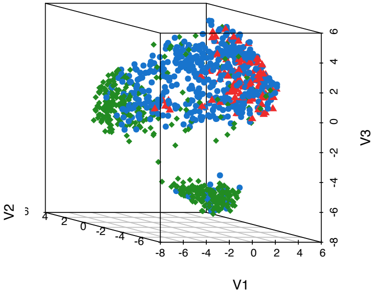
$$C = KL(P\|Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}, \tag{2}$$

where p_{ij} and q_{ij} are probabilities. These are the probabilities of datapoint i picking datapoint j as its neighbour, with full and reduced dimensions respectively. The p_{ij} use normal distributions while the q_{ij} use t-distributions with 1 degree of freedom.

Results



(a) 2D



(b) 3D

Figure: 2 and 3-Dimensional t-SNE Reduction.

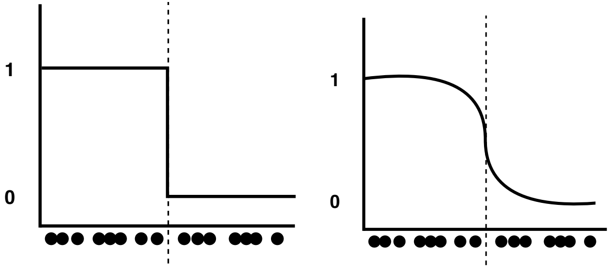
Cluster Analysis

The Methodology

- A cluster is generally thought of as a group of items (objects, points) in which each item is “close" (in some appropriate sense) to a central item of a cluster and that members of different clusters are “far away" from each other.
- unsupervised method for organizing data into homogeneous subgroups for descriptive analytics.

Hard vs Soft

We explore and contrast hard (represented by k-means) and soft (represented by fuzzy) clustering techniques.



(a) Hard Clustering

(b) Soft Clustering

Figure: Hard and Soft Clustering Assignment.

How many clusters?

A key problem here is choosing the optimal number of clusters *apriori*. To do this we can use methods such as minimising the total within-clusters sum of squares objective given by Equation 3

$$\min \left(\sum_{k=1}^{K} \mathrm{ESS}_K \right) \tag{3}$$

We can calculate the measure for different numbers of clusters to generate Figure 6 and choose a cluster configuration that corresponds to an “elbow" in the plot — in this case 3 clusters.

Scree Plot

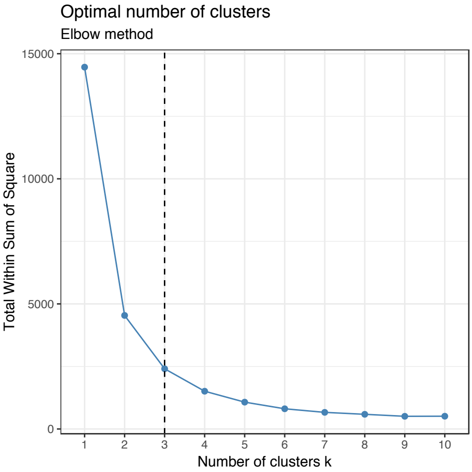


Figure: WSS Method for Determining Optimal Number of Cluster.

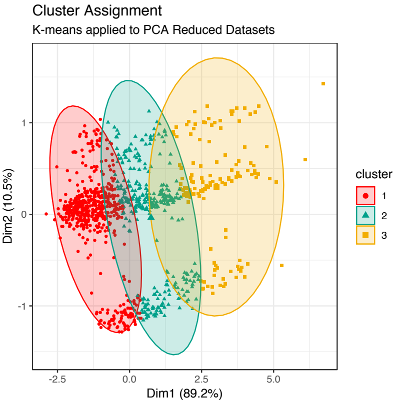
Method

K-Means Algorithm

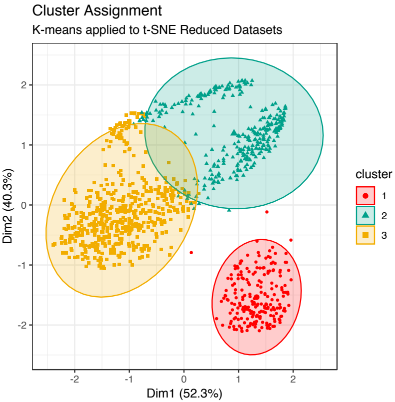
1. The algorithm starts by generating random centroid(s) and assigns each data point to the closest centroid.
2. Each collection of points assigned to the same centroid is now a cluster.
3. Update the centroid of each cluster based on the points assigned to the cluster.
4. Repeat the assignment and update steps until no point changes clusters, or equivalently, until the centroids remain the same (subject to some threshold δ)

No unique solution... Local minima

Results



(a) PCA



(b) t-SNE

Figure: K-Means Clustering Assignment for Alternate Dimensionality Reduction Techniques.

Method

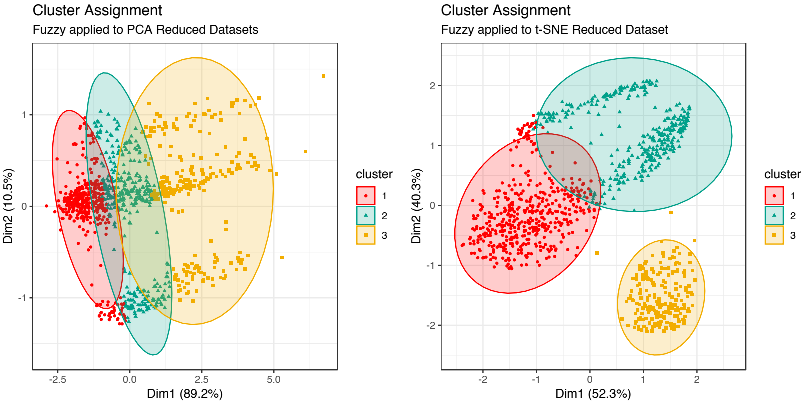
For each object i and each cluster v there will be a *membership* u_{iv} which indicates the strength of membership. Memberships are subject to the following conditions:

$$\begin{aligned} & u_{iv} \geq 0 \; \forall \; i = 1, \ldots, n \; \forall \; v = 1, \ldots, k \\ & \textstyle\sum_{v=1}^{k} u_{iv} = 1 \; \forall \; i = 1, \ldots, n \end{aligned} \tag{4}$$

The fuzzy clustering algorithm returns a membership score of each observation which takes on a value between zero and one. The memberships u_{iv} are defined through the minimisation of the objective function in Equation 5.

$$\sum_{v=1}^{k} \frac{\sum_{i,j=1}^{n} u_{iv}^2 u_{jv}^2 d(i,j)}{2 \sum_{j=1}^{n} u_{jv}^2} \tag{5}$$

Results



(a) PCA (b) t-SNE

Figure: Fuzzy Clustering Assignment for Alternate Dimensionality Reduction Techniques.

Cluster Validation

What is it?

The procedure of evaluating the goodness of clustering algorithm results.

Internal Validity	External Validity
• Utilises internal information of the clustering process to evaluate the goodness of a clustering structure without reference to external information. • Silhouette width, Dunn Idx and Sum of Squares	• Consists of comparing the results of a cluster analysis to an externally known result, such as externally provided class labels (our popularity scores). • Missclass Table

Table: Internal and External Validity Comparisons

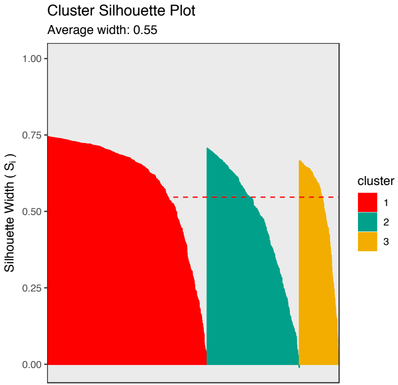
Silhouette Width

In silhouette width validation we compute $s_i\left(C_k\right)$ for cluster K given by Equation 6

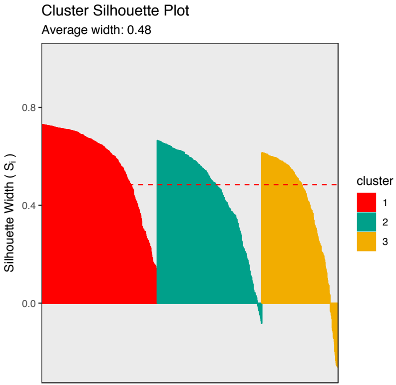
$$s_i\left(C_k\right) = s_{ik} = \frac{b_i - a_i}{\max\left\{a_i, b_i\right\}}, \tag{6}$$

- a_i is the average dissimilarity of the ith item to all other members of the same cluster $c(i)$.
- b_i can be thought of as the average distance between i and the observations in the "nearest neighboring cluster".

Silhouette Plots: PCA Reduced



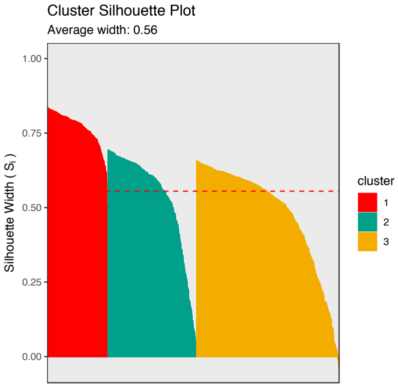
(a) K-means PCA



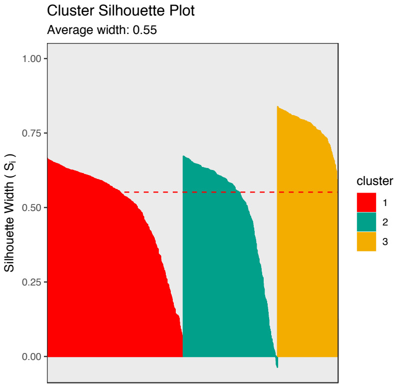
(b) Fuzzy PCA

Figure: Silhouette Plots for K-means and Fuzzy Clustering on PCA Reduced datasets.

Silhouette Plots: t-SNE Reduced



(a) K-Means t-SNE



(b) Fuzzy t-SNE

Figure: Silhouette Plots for K-means and Fuzzy Clustering on t-SNE Reduced datasets.

Sum of Squares: Counter-intuitive?

	PCA		t-SNE	
	K-Means	Fuzzy	K-Means	Fuzzy
Within-SS	2258	2269	6349	6350
Between-SS	12131	12564	20211	20925
Total-SS	14390	14833	26560	27275

Table: Sum of Squares (SS) for Comparing Cluster Performance

Missclassification Table

		K-Means			Fuzzy		
		1	2	3	1	2	3
PCA	Med	335	136	17	251	187	50
	Pop	99	5	0	85	18	1
	Unpop	114	175	119	54	148	206
t-SNE	Med	26	114	348	332	130	26
	Pop	0	5	99	99	5	0
	Unpop	182	185	41	37	189	182

Table: External Validity of Correct/Incorrect Cluster Assignment

Conclusions

Findings

- Often, the simple methods rely on *iid*. and correlation assumptions which tend to break down for real-world applications where the data is not fully specified for the particular task.
- t-SNE can capture possible non-linearities in the Spotify data making it a better candidate for the problem.
- Fuzzy clustering is necessary to understand the ambiguity that arises when using K-Means and clusters overlap.

Findings

- Often, the simple methods rely on *iid*. and correlation assumptions which tend to break down for real-world applications where the data is not fully specified for the particular task.
- t-SNE can capture possible non-linearities in the Spotify data making it a better candidate for the problem.
- Fuzzy clustering is necessary to understand the ambiguity that arises when using K-Means and clusters overlap.

Extensions and Caveats

- Predictions of the cluster variables (either genre or popularity) using the retained predictors in lower-dimensions.
- Although we observe three distinct groups, these could represent genre similarity rather than popularity as these two variables are by construct serially correlated.

Thank you! Questions?

References

- Alan Julian Izenman, **Modern multivariate statistical techniques: Regression, classification, and manifold learning**.
- Laurens van der Maaten and Geoffrey Hinton, **Visualizing data using t-sne**, Journal of machine learning research **9** (2008), no. Nov, 2579–2605.