DEIM2022 F31-4

多次元データ可視化のための次元削減結果の散布図選択

岡田　佳也†　　伊藤　貴之†

† お茶の水女子大学 人間文化創成科学研究科理学専攻　〒 112–8610 東京都文京区大塚 2 丁目 1-1
E-mail: †{okada.kaya,itot}@is.ocha.ac.jp

あらまし　機械学習をはじめとした多次元・多変数のデータ活用においては，次元数の大きさがもたらす「次元の呪い」という現象を防ぐなどの目的から，事前に次元削減を適用した上でデータの特徴を抽出することが多い．さらに，次元削減は多次元・多変数データを散布図として可視化する目的にも有効である．次元削減手法については既に数々の手法が確立されており，データの種類によって適用すべき次元削減手法や削減後の次元数は全く異なる．一方で，各々のタスクにとって最適な次元削減手法を選択する工程は，依然としてデータ分析専門家の経験に頼ることが多く，ビジネス等の目的で機械学習に取り組む初心者にとっては非常に困難である．そこで我々は，多次元データに対しさまざまな次元削減手法を適用した結果を 2 次元散布図上での点群の分布に沿って数値評価し，評価の高い散布図を形成する最適な次元削減手法および次元数を提示する手法を提案する．本報告では 1,315 曲の J-POP ヒット曲の音響データから音楽特徴量を抽出した結果に対する適用事例を示す．

キーワード　散布図，次元削減，多次元データ，可視化手法，音楽特徴量

1 はじめに

データ蓄積システムの大規模化や機械学習の発展といった技術の発展により，多次元データの記録，活用が進んでいる．トランザクションや人間行動などの多様で膨大な多次元データが記録されるようになった上に，音声，画像，動画などのマルチメディアデータからも多次元データの形式で特徴量を抽出する技術が発達した．一方で記録されたままの状態の多次元データでは人間が閲覧し有用な知見を得ることは難しい．加えて多次元データには「次元の呪い」と呼ばれる問題があり，多次元のまま扱うことが望ましいとは限らない．そのため多次元データの処理において，次元削減を適用した上で特徴を抽出することで，データに潜む知見を発見しやすくなる．

近年では既にさまざまな次元削減手法が普及している．主成分分析 (PCA: Principal Component Analysis) や多次元尺度構成法 (MDS: Multidimensional Scaling) のような線形分析，t-SNE(t-distributed Stochastic Neighbor Embedding) や UMAP(Uniform Manifold Approximation and Projection) のような非線形分析が主要な手法として知られている．適用すべきデータの種類や傾向や特徴は，Nanga ら [1] が整理している通り手法によってさまざまである。このような手法を適切に使い分け，データごとに最適な次元削減手法を選択することは，データ分析の専門家であれば可能であるが，非専門家にとっては依然として難しい．今日では専門家以外のユーザをターゲットにした機械学習のツールが多々提供されており，ビジネス等の業務目的で利用する非専門家が増えてきている．手軽にデータ分析に着手したい非専門家にとって，ツール利用前のデータの適切な加工や理解が障壁となることが多い．次元削減手法の選択もその障壁となりえることが考えられる．

本報告では，任意の多次元データに対する適切な次元削減手法を半自動的に取捨選択できることを目的とした手法を提案する．具体的には，多次元データに対し各次元削減手法を適用したものから 2 軸を選択し，2 次元散布図を生成する．生成された各散布図について，点群の分布や位置関係によりスコアを算出する．これにより，人間が閲覧した際に何らかの知見が得られると考えられる，意味づけが可能な散布図のスコアが高くなり，有用な散布図の選択が可能になる．

本報告では，楽曲分類を目的とした機械学習手法が算出した音響特徴量 [15] を適用する．この適用例では，楽曲発表年などのメタ情報を付与された 1,315 曲の J-POP ヒット曲を対象とし，その音響データから抽出した 1000 次元以上の音響特徴量に次元削減手法を適用している．

2 関連研究

2.1 多次元データ可視化のための次元削減手法

多次元データの次元削減は，データの要約や理解のため既に多くの手法が確立されている．次元削減手法をレビューした代表的な論文 [1–3] では，主要な方法として現在用いられている PCA や LDA(Linear Discriminant Analysis) などのさまざまな次元削減手法について，特徴，パラメータ，適用に向いているデータ種類などを整理して紹介している．

また，次元削減手法の応用事例として，Lee ら [4] は MDS に基づき，点群の距離関係がより正確にレイアウトへ表現される手法を提案している．Liu ら [5] は，高次元データの次元削減後の 2 次元散布図に対し，探索のための可視化インタラクションを提示している．

2.2 散布図の数値評価

2 次元散布図の評価方法についても，いくつかの指標が考案されている．Wilkinson ら [6] が提唱した Scagnostics では，散

布図を構成する点群の分布にもとづいて、形状や異常値の判別、相関といった 9 種類の包括的な指標を示している．この手法にもとづき，Nakabayashi ら [7] はいくつかの指標を算出し，小売店の販売情報と気象情報に関する散布図群への適用結果を示している．

Wilkinson らが示した各指標についての改良算出手法の研究もいくつか報告されている．散布図における点群のクラスの分離性については Aupetit ら [8] および Sedlmair ら [9] が，点群の相関については Harrison ら [10] が，高次元データの低次元散布図へのマッピング時のクラスの整合性や一貫性については Sips ら [11] がそれぞれ述べている．特に Sedlmair ら [9] は散布図におけるクラス分離性のデータや分布の特性を定義し，次元削減後のデータに適用している．

Dang ら [12] および Matute ら [13] は多数の散布図群の整理，要約，探索のための手法を提案している．さらに Wang ら [14] は人間の視覚認知にもとづいた，散布図に見られる異常値やクラスタの判断に関する Scagnostics の改良手法を示す．

それに対して本報告では，上記の評価基準のうち相関，クラスタの分離性，クラスの分離性，クラスの連続性の各指標について実装し，重み付けした各指標のスコアによって，多数の散布図群を評価し選択する手法を提案する．

3 提 案 手 法

提案手法では多次元データに対して，3.1 章で紹介する各次元削減手法を適用し 2 次元まで削減したのち，3.3 章以降で示す散布図のスコア算出手法によって次元削減結果を評価する．

3.1 次元削減手法

本報告で適用する次元削減手法を列挙する．いずれも一般的な手法であるため，詳細については本報告では割愛する．

（1） PCA (Principal Component Analysis)
（2） t-SNE (T-distributed Stochastic Neighbor Embedding)
（3） LDA (Linear Discriminant Analysis)
（4） MDS (Multi Dimensional Scaling)
（5） UMAP (Uniform Manifold Approximation and Projection)
（6） SVD (Singular Value Decomposition)
（7） NMF (Non-negative Matrix Factorization)
（8） ICA (Independent Component Analysis)
（9） KernelPCA (Kernel Principle Component Analysis)

3.2 データ構造

各手法による次元削減後のデータ構造を以下の通り定式化する．削減後の m 次元データ A は n 個 (本報告では 1,315 曲) の標本 $A = \{a_1, a_2, \cdots, a_n\}$ を持っており，i 番目の標本 a_i は m 次元ベクトル $a_i = (a_{i1}, a_{i2}, \cdots, a_{im})$ を有するとする．また各標本には 1 個以上のクラスが割り当てられている場合もあるものとする．そして，このデータから N 個の散布図 $S = \{s_1, s_2, \cdots, s_N\}$ が生成されるとする．また，各散布図について M 種類の指標に沿ってスコアを算出し，そのスコアを M 次元ベクトル $s_i = (s_{i1}, s_{i2}, \cdots, s_{iM})$ として格納する．

3.3 散布図選択手法

N 個の散布図に対し，複数の指標に従いスコアを算出する．提案手法における指標を本章で示す．

3.3.1 相　　関

散布図の相関が大きい場合，閲覧するに値する散布図であると評価される場合が多い．k 番目の散布図の 1 番目のスコアを以下の式で算出する．

$$s_{k1} = |Spear(i,j)^2|$$

ここで $Spear(i,j)$ は散布図の 2 軸となる i 番目と j 番目の次元間の Speaman 順位相関係数であり，スコアが大きい方が正または負の相関が高くなる．

3.3.2 クラスタの分離性

次元削減後，散布図を構成する点群が有限個のクラスタに明確に分離する場合，閲覧によって知見が得られることがある．例えば LDA のような，クラス間の分離性を最大化するような手法の適用結果に合致する．まず点群に対し階層型クラスタリングを適用し，3〜10 個程度のクラスタを形成する．散布図 s_i のクラスタリング結果を c_i とする．各クラスタ数 u の場合に対し，Calinski-Harabasz 基準値 [16] によるクラスタリング評価を適用し，$u = 3, 4, \cdots, 10$ の中で算出したスコアの最大値を散布図 s_i のスコアとする．クラスタ内の凝集性，複数のクラスタ間の離散性がそれぞれ高いものが高スコアとなる．

$$s_{k2} = max(CalinskiHarabasz(c_k)_u)$$

3.3.3 クラスの分離性

特定のクラスを有する点群が，ほかの点群から明確に分離している散布図も，閲覧することで知見を得られる場合が多い．提案手法では情報エントロピーにもとづいてクラスの分離性を数値化している．具体的には，i 番目と j 番目の次元から構成される散布図について，以下の値を算出する．

$$H(i,j) = -\frac{1}{n}\sum_{k=1}^{n}\sum_{c=1}^{C} p(y_k = c|(a_{ki}, a_{kj}))\log p(y_k = c|(a_{ki}, a_{kj}))$$

ここで y_k は k 番目の標本のクラス，(a_{ki}, a_{kj}) は k 番目の標本の座標値，C はクラスの数である．筆者らの実装では散布図を L 個の部分領域に分割し，l 番目の部分領域におけるエントロピー $H(i,j)_l$ を上述の式により算出する．そして最終的に，以下の式により k 番目の散布図のスコアを算出する．

$$s_{k3} = (H_{max} - \sum H(i,j)_l)/H_{max}$$

ここで H_{max} は $H(i,j)_l$ の最大値である．

3.3.4 クラスの連続性

点群が有するクラスに順序がある場合，散布図上で各クラスに所属する点群が順序通りに並んでいるもの (順序を考慮した

	1986年~1991年		2002年~2007年
	1992年~1996年		2008年~2012年
	1997年~2001年		2013年~2018年

図 1　年代ごとの色分類.

色付けを行なった際，グラデーションを示すようなもの) を判別できれば，散布図の相関をクラスで説明することができる．i 番目の標本 a_i に割り当てられているクラス番号を c_i とする．クラス番号を有する点群を接続する Delaunay 三角メッシュを生成し，各辺の両端点 (辺で接続された頂点を順番に並べた時，v 番目と $v+1$ 番目の点) のクラス番号の差分の絶対値を合計したものをスコアとする．

$$s_{k4} = \sum_{v=1}^{m} |(c_{v+1} - c_v)|$$

3.3.5　各評価値の正規化と重み付け

各散布図に対し M 種類の評価指標でスコアを算出した M 次元ベクトル $s_i = (s_{i1}, s_{i2}, \cdots, s_{iM})$ に対し，MinMax 法により正規化する．さらに，各指標ごとに重み $w = \{w_1, w_2, \cdots, w_M\}$(ただし $w_1 + w_2 + \cdots + w_M = 1.0$) を設定し，各 k 番目の散布図の 1 番目のスコアは下記の通り更新される．

$$s'_{k1} = w_1(s_{k1} - min(s_{i1}))/(max(s_{i1}) - min(s_{i1}))$$

4　適 用 事 例

本報告では 1986 年から 2018 年までの日本のヒット曲 1,315 曲からなる楽曲群を使用した．RP_extract [17] で抽出した各曲の特徴量 (rp1440 次元，rh60 次元，ssd168 次元) において，全特徴量に対し次元削減手法を適用した．

この楽曲群の音響特徴量に対し，第 3 章に示す全ての手法で 2 次元まで削減し，2 次元散布図として可視化した結果が図 2 である．なお適用例では，年代と音響特徴量の相関を観察するため，図 1 に示す基準で各楽曲の発表年を 6 つに分類し，それぞれの分類に固有の色を割り当てた．

図 2 からわかるように，LDA を適用した場合に顕著な結果が現れた．右上方に，年代ごとのクラスタがいくつか現れていることが分かる．また t-SNE，UMAP では，全体に点群が広がっているが，比較的色相がグラデーションになっており，年代ごとの傾向が示される次元が抽出されている．さらに SVD では，細くグラデーションのある，相関関係を示す形状となっている．次節では，これらの可視化結果の評価スコアの算出結果を示す．

4.1　クラスの分離性とクラスタの分離性に関する評価結果

クラスごとのクラスタが明確に分離されている散布図は，分類に適したデータの特徴量をよく捉えていると考えられる．そこで，評価指標のうちクラスの分離性とクラスタの分離性にそれぞれ 0.5 ずつの重み付けを行いスコアを算出した．結果は表

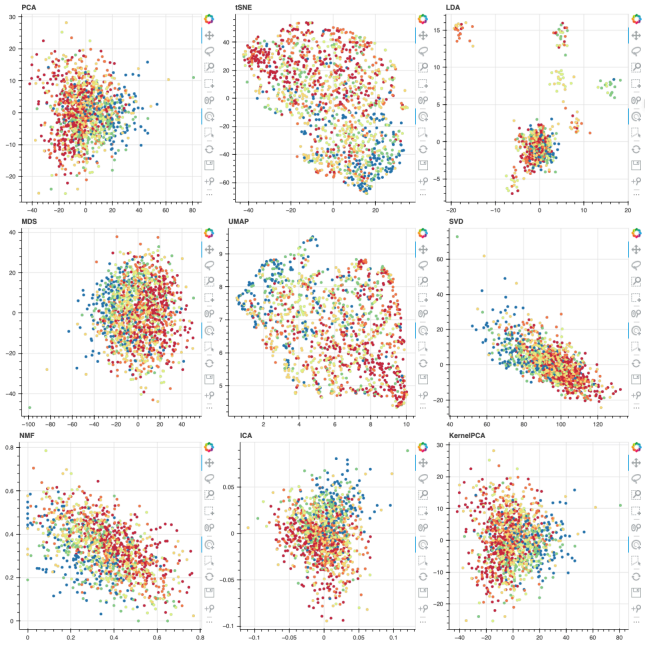

図 2　散布図生成結果．上段左から，PCA，t-SNE，LDA，中段左から MDS，UMAP，SVD，下段左から NMF，ICA，KernelPCA を適用し 2 次元まで削減した結果.

1 の通りとなった．

点数が高い手法は上から UMAP，t-SNE，LDA となった．t-SNE と UMAP は年代 (クラス) ごとのグラデーションになっているため，クラスの分離性のスコアが高く，LDA はクラスタが明確に分類されていたことから高得点となった．ただし LDA においては，下部にクラスが混在したクラスタが 1 つ存在したことから，クラスの分離性のスコアは低くなってしまっている．

4.2　相関とクラスの分離性，連続性に関する評価結果

相関関係を示す散布図も，人間が閲覧した際に知見を得られる 2 つの特徴量を選択できている可能性が高い．ただしクラスと無関係な相関である可能性も高いため，ここではクラスの分離性とクラスの連続性にも重みを設定することで，色によって得られる知見を数値に反映した．重みはそれぞれ，相関に 0.5，クラスに分離性に 0.25，クラスの連続性に 0.25 とした．

結果として，年代が連続する t-SNE と UMAP に加え，細く相関の形状を示す SVD も高スコアとなった（表 2)．クラスの分離性とクラスタの分離性でスコアの高かった LDA は飛び抜けて低いスコアが算出された．前節と同様，散布図から視認できる数値分布と評価結果が概ね一致する結果になったと考えられる．

4.3　次元削減手法を組み合わせた場合の可視化結果とスコア

前節までに，単一の次元削減手法で一気に 2 次元まで次元削減した結果を散布図として提示した．続いて本節では，2 種類の次元削減手法を組み合わせて適用した結果を散布図とスコアで紹介する．ここではまず，1 度目の次元削減として PCA を適用し，累積寄与率 80%となる次元数（本報告で適用したデータでは 16 次元）まで削減した．この結果に対して，前節まで

指標	PCA	t-SNE	LDA	MDS	UMAP	SVD	NMF	ICA	KPCA
クラスの分離性	0.2953	0.4499	0.0000	0.2204	0.5000	0.1090	0.3285	0.2585	0.2953
クラスタの分離性	0.1054	0.3230	0.5000	0.0474	0.3028	0.1777	0.0986	0.0000	0.1054
合計点	0.4006	0.7730	0.5000	0.2677	0.8028	0.2867	0.4272	0.2585	0.4006

表 1　クラスの分離性とクラスタの分離性のスコアに 0.5 ずつの重みを設定し算出したスコア.

指標	PCA	t-SNE	LDA	MDS	UMAP	SVD	NMF	ICA	KPCA
相関	0.0000	0.1548	0.3217	0.0121	0.3444	0.5000	0.2773	0.0000	0.0000
クラスの分離性	0.1377	0.2500	0.0000	0.1060	0.2112	0.0508	0.1533	0.1199	0.1377
クラスの連続性	0.1268	0.2500	0.0000	0.1477	0.2496	0.1379	0.1326	0.1238	0.1268
合計点	0.2645	0.6548	0.0322	0.2658	0.8052	0.6888	0.5632	0.2437	0.2645

表 2　相関のスコアに 0.5，クラスに分離性に 0.25，クラスの連続性に 0.25 ずつの重みをそれぞれ設定し算出したスコア.

指標	PCA	t-SNE	LDA	MDS	UMAP	SVD	NMF	ICA	KPCA
相関	0.0001	0.0439	0.0000	0.0090	0.0000	0.2419	0.2500	0.0001	0.0001
クラスの分離性	0.2264	0.2030	0.0601	0.2350	0.2174	0.0000	0.0090	0.2500	0.2264
クラスタの分離性	0.0867	0.0945	0.0000	0.0688	0.1155	0.2500	0.2185	0.0303	0.0867
クラスの連続性	0.1815	0.2500	0.0000	0.1960	0.2184	0.1852	0.2104	0.1876	0.1815
合計点	0.4948	0.5914	0.0601	0.5087	0.5513	0.6771	0.6879	0.4680	0.4948

表 3　UMAP で 16 次元まで削減し，その後各手法で 2 次元に削減した結果のスコア.

と同様に各手法を適用し，2 次元まで削減した．さらに，1 度目の次元削減として PCA 以外の手法も適用してみた．

以下，特に結果が顕著だった 2 パターンを紹介する．まず UMAP で 16 次元まで削減し，その後に各手法で 2 次元に削減した (図 3)．図 2 と比較し，いずれの手法を組み合わせた場合においても年代ごとのまとまりが明確になっている．この時のスコア (表 3) は，総合して SVD と NMF が高くなった．

次に LDA では，累積寄与率が 70%となる 8 次元まで削減後，各手法を適用して 2 次元に削減した (図 4)．その結果，LDA の分離性を最大化するという特性が出てしまい，いずれも似た可視化結果となった．ただ NMF においては，LDA のみで次元削減した場合には分離されなかった青色 (1986 年〜1991 年) のクラスタが分離されている．

5　まとめ・今後の課題

本報告では，多次元データに対し各種の次元削減手法を適用し，その散布図の形状や傾向を数値評価することで，適切な次元削減手法を選択する手法を提案した．適用事例では，さまざまな年代の J-POP から抽出した音響特徴量について，適用した結果とスコアを示した．

今後の課題として，3 点を検討中である．

まず 1 点目は散布図の色付けである．適用例では楽曲群の年代に沿って散布図を 6 色で描画した．ここで同じ楽曲群のデータを適用するにしても，年代以外のクラス，例えば歌手や作曲者に沿って色を割り当てる場合には，色数が大きく異なる場合もあるし，また年代と違って順列をもたない場合もある．よって，異なるクラスを適用した際にはクラスタやグラデーションの視認性が異なってくる可能性がある．このような事例について考察を深めたい．

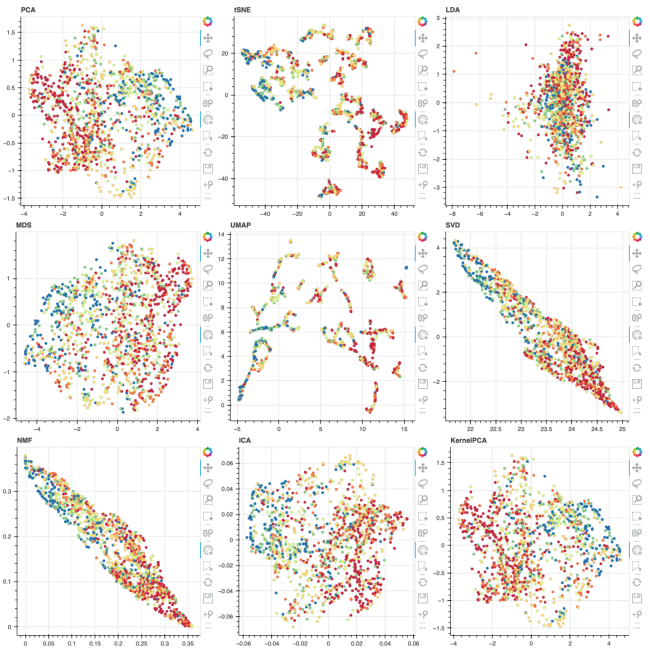

図 3　UMAP で 16 次元まで削減後，各手法で 2 次元に削減した結果.

2 点目は異なる形式のデータへの適用である．本報告では J-POP の音響特徴量を対象としたが，異なる形式のデータとして例えば，テキストデータを word2vec などのツールによって多次元ベクトル化して適用した場合，今回とは異なる次元削減手法のほうがより良い結果が得られると考えられる．そこで，いくつかの異なる形式のデータで，本手法の有効度を確認したい．

3 点目は各指標の重み付けに対する評価である．適用例では著者自身の仮説と主観にもとづいて，各指標の重みを比較的均等に設定している．より信頼性の高い散布図選択を実現する最

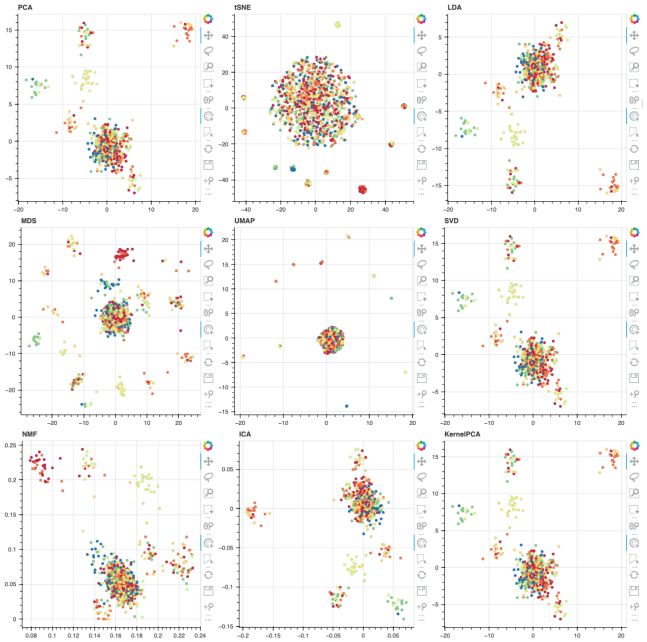

図 4　LDA で 8 次元まで削減後，各手法で 2 次元に削減した結果．

適な重みを同定するために，今後の課題として評価実験を実施したい．

文　　献

[1] S. Nanga, and others. "Review of Dimension Reduction Methods." Journal of Data Analysis and Information Processing 9.3, pp. 189-231, 2021.

[2] O. Saini, and S. Sharma. "A review on dimension reduction techniques in data mining." Computer engineering and intelligent systems 9, pp. 7-14, 2018.

[3] D. Engel, L. Huttenberger, and B. Hamann. "A survey of dimension reduction methods for high-dimensional data analysis and visualization." Visualization of Large and Unstructured Data Sets: Applications in Geospatial Planning, Modeling and Engineering-Proceedings of IRTG 1131 Workshop 2011. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2012.

[4] H. Lee, and others. "A structure-based distance metric for high-dimensional space exploration with multidimensional scaling." IEEE transactions on visualization and computer graphics, 20.3, pp. 351-364, 2013.

[5] J. H. Liu, and others. "Distortion- Guided Structure Driven Interactive Exploration of High- Dimensional Data." Computer Graphics Forum, Vol. 33, No. 3, 2014.

[6] L. Wilkinson, A. Anand, and R. Grossman. "Graph-theoretic scagnostics." Information Visualization, IEEE Symposium on. IEEE Computer Society, 2005.

[7] T. Itoh, A. Nakabayashi, and M. Hagita. "Scatterplot Selection Applying a Graph Coloring Algorithm." The 14th International Symposium on Visual Information Communication and Interaction, 2021.

[8] M. Aupetit, and M. Sedlmair. "Sepme: 2002 new visual separation measures." 2016 IEEE Pacific Visualization Symposium (PacificVis), IEEE, 2016.

[9] M. Sedlmair, and others. "A taxonomy of visual cluster separation factors." Computer Graphics Forum. Vol. 31. No. 3pt4. Oxford, UK: Blackwell Publishing Ltd, 2012.

[10] L. Harrison, and others. "Ranking visualizations of correlation using weber's law." IEEE transactions on visualization and computer graphics, 20.12, pp. 1943-1952, 2014.

[11] M. Sips, and others. "Selecting good views of high- dimensional data using class consistency." Computer Graphics Forum. Vol. 28. No. 3. Oxford, UK: Blackwell Publishing Ltd, 2009.

[12] T. N. Dang, and L. Wilkinson. "Scagexplorer: Exploring scatterplots by their scagnostics." 2014 IEEE Pacific visualization symposium. IEEE, 2014.

[13] J. Matute, A. C. Telea, and L. Linsen. "Skeleton-based scagnostics." IEEE transactions on visualization and computer graphics, 24.1, pp. 542-552, 2017.

[14] Y. Wang, and others. "Improving the robustness of scagnostics." IEEE transactions on visualization and computer graphics, 26.1, pp. 759-769, 2019.

[15] M. Watanabe, and others. "Visualization of song collections for understanding of the relationship between metadata and features." ITE Technical Report; ITE Tech. Rep. 45, 2021.

[16] T. Calinski, and J. Harabasz, "A Dendrite Method for Cluster Analysis." Communication in Statistics, 3, pp. 1-27, 1974.

[17] Vienna University of Technology, Music Information Retrieval, http://ifs.tuwien.ac.at/mir/musicbricks/#RPextract