

点と折れ線による可視化を併用した 対話的な多次元データ可視化システム

太田 尚博^{†1,a)} 脇田 建^{†1,†2,b)}

概要：近年、情報技術の発達により、多数の属性を持つデータである多次元データを大量に収集することが可能となっている。これを可視化、分析するためにデータの各項目を折れ線形で表現する平行座標系 (Parallel Coordinate Plot, PCP) が考案されている。しかし、PCP は平面上で多数の属性の相関性を可視化することのできる利点があるがクラスターが散逸され可視化されてしまう問題や属性の順番の決定の問題があった。一方で多次元データを点の形で可視化する場合はクラスターを点の塊として表現できるものの、多次元データの数値を直接表現することはできない。このようにそれぞれの可視化のアプローチには長所、短所が存在する。

そこで本研究ではこの2つの可視化手法を並置連携させることとした。線による可視化としてPCPを拡張したものを使用し、点による可視化として利用するために射影による社会ネットワーク可視化手法 Active Graph Interface を応用した手法である多次元射影系を提案した。そして2つを併用することにより各項目間の類似性と各属性間との相関性の双方を連携した可視化画面で見ることが目指した。

この手法を可視化システム VisLAMP として実装し実際のデータに適用したところ両画面を連携させたことにより項目のクラスター構造と属性同士の関係性を同時に見ることができた。既存手法と比較した結果においても有用性を示すことができた。

キーワード：多次元データ可視化 並行座標系 対話的手法 クラスタ発見 階層型クラスタリング

NAOHIRO OHTA^{†1,a)} KEN WAKITA^{†1,†2,b)}

1. はじめに

近年、情報技術の発達により、大量のデータを計測し分析することが可能となっている。

これらの中でも1つの項目が多くの属性を持つものを多次元データと呼び、これらを可視化、分析するために並行座標系 (Parallel Coordinate Plot, PCP) をはじめとした様々な手法やそれを実装したシステムが存在する。

PCPの方法は属性軸を並列にならべそれらの軸上にある値を結ぶ折れ線としてそれぞれの項目を表すものである。このため、PCPは平面上の可視化でありながら高次元の属性間との関係性やデータの値をそのまま表すことができる。しかし、既存のPCPでは二つの問題が存在する。まず一

つ目は散逸の問題である。これは項目を線として描いているために散逸しやすくデータ内でのクラスター構造が見えにくいという問題である。またもう一つは軸を並べる順番の決定方法である。この順番により属性同士の関係性の見易さが大きく左右される。この順番は属性数の増大により膨大になるため最適な順番を求めるのは難しい。

また多次元データを表すには主成分分析を利用した散布図を用いるなど、各項目を点として表す方法もある。この方法では、クラスターを点の塊として表現することができるため前述の散逸による問題は起こらない。しかしこの場合は多数の次元からなるデータを人間が理解できる次元数まで縮約しなければならないため、値を直接表現することができない。

この二点を要約するとPCPは属性に関する可視化に長けており、点による可視化は項目間との関係性を可視化することに向いていると言える。よってこれらをユーザーのインタラクションに対して連携させることができれば必要に

^{†1} 現在、東京工業大学
Presently with Tokyo Institute of Technology

^{†2} 現在、JST/CREST

^{a)} ohta.n.aa@m.titech.ac.jp

^{b)} wakita@is.titech.ac.jp

応じて注目する可視化画面を切り替えることで互いの長所を両立させる可視化分析が行うことでこの双方の問題を解決することができるのではないかと考えられる。しかし既存の研究にこの課題に対処したものはほとんど存在しない。

そこで本研究ではこれら二つの可視化手法を組み合わせることで互いの短所を補い、長所を両立させた手法を提案することとした。このために本研究ではまず射影による社会ネットワーク可視化手法 AGI を多次元データに対応させるために応用する。AGI は高次元配置に配置したノードを低次元に射影し、さらにその射影されたノードをユーザーが移動操作を行うことで射影の更新ができ柔軟な可視化結果の更新が可能である。本研究ではこの性質が多次元データの分析に有用であると考え応用を行う。そしてこれを多次元射影系 (*Multi-dimensional Projection Plot*, MPP) として提案する。

次に PCP を拡張し軸間の距離を可変にする。そしてこの 2 つを併用することにより各項目間の類似性を MPP で可視化し、また各属性間の相関性を PCP で可視化する。さらに双方を連携させることによりインタラクティブ性を持たせた可視化手法を提案し、これを Visualization System Linked and Apposed MPP to PCP (VisLAMP) として実装することで実際に多次元データの可視化を行った。

本研究の貢献は主に三つある。一つ目は社会ネットワーク可視化手法である AGI を多次元データに応用した MPP を提案した点である。まず多次元データに導入するために高次元の配置計算を主成分分析 (PCA) による計算に切り替えることを行い、MPP 画面上の配置における属性の影響を可視化するために項目だけでなく属性軸を射影しこれを用いての射影の更新も行えるようにした。

次に二つ目はこの AGI と PCP を組み合わせた新たな手法を提案した点である。AGI で操作し高次元から低次元への射影を更新することにより、PCP の軸の並びと軸の順番を変化させる。クラスター発見のために複数の項目の同時選択やフィルタリングなどの機能を与え、階層型クラスタリングを行うためにデータの部分集合に対する提案手法の再適用を行うことができるようにした。そしてこれらを MPP と PCP 双方の可視化画面に反映させた。

最後に三つ目はこの手法を VisLAMP として実装し可視化分析を行った点である。このシステムを用いて Cars データセットを分析した結果、MPP 画面と PCP 画面の 2 つを見比べることでクラスター構造の把握とそのクラスターの特徴の理解を同時に行うことを可能にした。この結果上記の課題を解決することができた。

2. 関連研究

多次元データの可視化の既存手法として有名なものとして並行座標系 (*Parallel Coordinate Plot*, PCP)[6][7] がある。これは各属性を軸として並列にならべ、それらの

値を通る折れ線で一つの項目を表現している。すなわち、属性 $X_1, X_2, \dots, X_m$ を持つ n 項目のデータを $m \times n$ 行列 $D = \{d_{ij}\}$ を用いて表したとき、平面上において属性 X_i の軸を $x = (i-1)/(m-1)$ の直線とし、 $(d_{1j}, d_{2j}, \dots, d_{mj})$ と表現できる j 番目の項目は $(d_{1j}, 0), (d_{2j}, 1/m-1), \dots, (d_{mj}, 1)$ を結ぶ折れ線として表す。

しかしながら PCP には幾つかの問題が存在する。まず一つ目としては散逸問題がある。これは PCP では項目を折れ線として表示するために可視化結果での配置が散逸してしまいやすいという点である。この問題により、PCP 上では属性間の関係を見ることには向いているものの、データ間の類似性やクラスタ構造を見ることには向いていない。また軸の順番をいかに決定するかといった問題が存在する。PCP では軸が並列に並んでおり、属性間の関係性を可視化可能なものは隣接したものの同士のみである以上、どの軸とどの軸を隣接させるかは大きな問題である。

そこで、PCP を応用し、これらの課題に対処した研究が多く存在する。一例としてエッジの束化といった複数の項目を 1 本に束ねてしまう方法がある。例えば Palmas らは軸間ごとにクラスタリングを行ったのちそのクラスタの中心に辺を集め Bezier 曲線を用いて一本の帯状の図形にしている [11]。

一方で項目ではなく軸を束ねる可視化も存在する。渡辺らは属性軸に対し Bicustering を PCP に適用している [14]。軸のクラスタリングとしては k-means を元にしたアルゴリズムを使い円周上で類似した属性ベクトルを一本のベクトルにまとめている。このようにして類似した軸を隣接させ合成することで散逸の問題に対処している。

軸の順番に対してはより適切な順番を決定する手法も存在する [1][3]。ただし、PCP において関係性を可視化可能なのは軸が隣接した属性同士のみである。この点に関しては複数の PCP を平面上で結合した可視化を行っている Classen らの研究がある [2]。だがこれは可視化結果の散逸性には対処していない。

また多次元データの各項目を点として可視化した研究も存在する。Cheng らは GBC[10] と呼ばれる可視化手法とヒートマップと PCP を組み合わせた可視化手法を提案している [12]。GBC は円周上に属性を配置し、円内に項目を点として配置する。彼らはこれを改良し、実際のデータと可視化結果上での配置の差を減らす研究を行った。またこの手法と PCP やヒートマップを併用した可視化システムを実装した。

これらを組み合わせた研究も存在する。Yuan らは SPCC[15] という PCP に散布図を埋め込むような可視化を行っている。この研究では、MDS を基にした手法で配置を計算したのち散布図を作成し、これを PCP に埋め込んでいる。PCP 上で各項目を表す折れ線がこの散布図上では曲線となり、対応する点を通っている。

また Giménez らは階層型の可視化を行い PCP と star coordinates の可視化を組み合わせている [4]. この可視化手法では左側の PCP と右側の star coordinates が連動している. そして双方の対応する領域には同じ色が付けられている.

AGI[5] は多次元データではなく, グラフの中でも特に人間同士のつながりを表している社会ネットワークのグラフを可視化する手法であり, データ間に存在する距離をもとに多次元尺度法 (*Multi-Dimensional Scaling, MDS*) と呼ばれる手法を用いて高次元での配置を決め, それを 2 次元空間上に射影することで可視化を行うものである. これは Kruskal らの手法 [8] をさらに発展させたものである. 具体的にはグラフのノードの移動や回転などを射影のしかたに対応させることによって, 元の高次元配置を破壊的に変更することなく可視化結果を変えることができる. のちに高見らによって AGI は三次元可視化へ発展され, 社会ネットワークの可視化への有効性が確認されている. [13]

しかしこれらの研究は先に挙げた可視化結果の散逸性への対処と多次元データの属性の表示を両立させてはいない. これに加えてより多次元データの分析を行いやすくする為にはクラスタの中の構造や項目などを可視化できるようにし, また必要に応じてユーザがインタラクションを行うことによって可視化結果を変化させ, 多次元データに対して注目している部分を変更できるようにするべきであると考えられる.

3. 提案手法概要

前述したように多次元データの可視化手法には PCP のような折れ線による可視化と点による可視化の 2 つのアプローチが存在する. これらにはそれぞれ利点, 欠点が存在する. これをまとめると 表 3 として表すことができる.

	折れ線による可視化	点による可視化
利点	属性の傾向を表すのに適している	クラスタ構造を表しやすい
欠点	散逸の問題がありクラスタ構造を見ることができない	その項目やクラスタの値の属性を見ることができない

表 1 点と折れ線による多次元データ可視化の利点および欠点

そこで本研究ではこの二つのアプローチの利点を両立させ, 欠点を補うために PCP と点による可視化を組み合わせることを提案する. 双方の可視化結果を必要に応じて連携させることで必要に応じて着目する可視化結果を柔軟に切り替えることができるようになる.

PCP と組み合わせる可視化手法としては, 本研究では前述した AGI に注目した. AGI は高次元空間に配置されたグラフを射影することにより平面上で可視化を行っている. これは高次元空間上であればグラフの複雑な構造や各

頂点間の距離をより反映した配置が行いやすいからである. これを必要に応じてユーザがインタラクションを行い射影を変えることによりグラフの特徴を人間が直接見える形で表現していた.

本研究で扱うものは多次元データであり, グラフではない. しかしながらこのようなアプローチは有効でないかといえる. 多次元データの各項目の値を元に高次元空間上の配置を行うことで, データ全体の特徴や個々のデータ間の類似性を表現することが可能であると考えられる. これを同じように平面上に点として射影することができれば高次元上での情報を失う量が最低限で済む. ここから, AGI を応用して項目を点の形で可視化することにより可視化結果の散逸性を軽減できると考える. さらにユーザがインタラクションを行い可視化結果を変えることにより注目する属性や項目を変更することができる.

そこで本研究ではまず AGI を多次元データに応用した手法として多次元射影系 (*Multi-dimensional Projection Plot, MPP*) を提案する. 次に PCP に対し拡張を行い, 軸の順番だけでなくそれぞれの軸間の距離を可変にしたものを幅可変 PCP と呼ぶこととする. 幅を可変とすることで, PCP 上で軸間の関係性がより強調して可視化されるものと思われる. 本研究ではこの 2 手法を併用することにより多次元データの可視化を行うことで前述した欠点を互いの長所で補うことができると考える.

加えて両方の可視化結果に何らかの対応関係が必要である. そのため, AGI を応用した MPP がユーザのインタラクションによる可視化結果の更新ができることを利用し, これを幅可変 PCP に反映する方法を提案する. 大まかな方法としては MPP 画面上に射影されている属性軸のなす角度を元に幅可変 PCP の軸の並びおよび軸間の距離を決定するものである. MPP 画面を元に決定することで, PCP の軸の順番の決定問題に対処できる.

またクラスタ発見を行うため, フィルタリングや項目の複数選択も実装し, これを互いの可視化結果に反映させるようにする. また発見されたクラスタに対して提案手法を再度適用することで階層型クラスタリングを行えるようにする.

前述の MPP と 幅可変 PCP にこの対応関係を加え統合した手法が本研究における提案手法である. この提案手法の詳細については次節以降で述べる. そしてこれを Visualization System Linked and Apposed MPP to PCP (VisLAMP) として実装した. このシステムによる可視化結果については 6 節で述べる.

4. AGI を応用した多次元データ可視化手法 MPP

前節で述べたように MPP は AGI と同様にデータ項目を高次元空間上の頂点として扱い, それを二次元平面に射影

する可視化手法である。さらに、射影された点のドラッグ操作にともなう、射影に高次元回転を施すことで、可視化された頂点群全体を操作できる。一方、MPP は多次元データを頂点として可視化する点において、グラフの頂点を可視化する AGI とは異なる。MPP は、項目を頂点として表すとともに、属性軸も併せて表示し、操作の対象とする。次節で述べるように MPP は PCP の軸配置を変更するための操作系としても機能する。

4.1 多次元データの多次元配置の計算

ここで、 n 項目、 m 属性の多次元データの、各項目を縦ベクトルとして扱い、 $D = \{d_{ij}\} : m \times n$ で表す。多次元データはすべて数値データであり、 $[0, 1]$ 区間で正規化されているものと仮定する。

このような多次元データ D は多次元空間上のベクトルと見做すことができる。これを可視化するために、MPP はまず多次元データに主成分分析 (Principal Component Analysis, PCA) を施し、そこで得られた主成分が張る空間に沿って多次元データを配置する。すなわち、主成分分析から得られる基底 W を用いて、 D を同じく m 次元空間に WD によって写すことができる。

MPP における射影を定める基底は、前述の PCA で求めた基底である分散、共分散行列の固有ベクトルに対応する固有値を用いる。これを λ_i とする。

高次元配置を PCA で求めた後は射影行列 P を高次元配置を表す行列にかけることによって可視空間への射影 $P = (e_1, e_2)$ が得られる。

$$e_i = f_i / \|f_i\| \quad (1)$$

$$f_1 = (\sqrt{\lambda_1}, 0, \sqrt{\lambda_3}, 0, \dots) \quad (2)$$

$$f_2 = (0, \sqrt{\lambda_2}, 0, \sqrt{\lambda_4}, \dots) \quad (3)$$

この射影行列 P を乗ずることによって、多次元データの二次元配置が得られる。つまり、多次元データ中の第 j 項目 $\hat{p} = (d_{1j}, d_{2j}, \dots)^T$ に対応する、二次元平面上の位置は $p = PW\hat{p}$ で与えられる。さらに、多次元データの全項目の射影も PWD によって計算できる。

4.2 対話機能：射影の更新

MPP は同様に対話的な可視化技術である。ユーザは MPP が可視化したデータ点をドラッグすることで、わずかな操作によって多次元データのレイアウトを大規模に変更することができる特徴は AGI のものを受け継いでいる。MPP はさらに多次元データの属性を属性軸として可視化し、そのドラッグ操作をも許している。これら二種の操作体系はともに多次元データの射影 (P) を更新する。このためデータ点のレイアウトの変更は属性軸のレイアウトの変更と自然に連携し、一貫性も維持される。

まず、データ点とそのドラッグに伴う射影行列の更新については細部が提案した AGI の仕組み [5] をそのまま利用しているため、ここでの説明は概略に留める。画面上に映し出されたデータ点 p は高次元データ $\hat{p}$ が射影行列 P によって二次元平面上に射影されているものである。ユーザが、このデータ点を別の場所 p' にドラッグ操作で移動したとき、AGI は高次元空間内で射影行列に高次元回転を施し、 $p' = P'\hat{p}$ となるようなものに変換することを試みる。ここで得た新しい射影行列をすべてのデータ項目に適用することによって、表示されているすべてのデータ点の位置を更新することができる。

MPP では、AGI によるデータ点の移動にともなう射影の更新とともに、多次元データの座標軸についても同様の対話性を与える。主成分分析が与える行列 C の (j, i) 成分は、第 i 主成分が第 j 属性に与える影響と見做すことができる。このため、 C 自体をもって、各属性の高次元属性と扱えば、データ点の高次元表現と一貫性をとることができる。さらに、属性ベクトルの二次元可視平面への射影には単に射影行列 P と乗算すればよい。つまり、 PC の各列成分が属性軸の可視化ということになる。

MPP では、以上のように与えられた属性軸を表すベクトルを原点から伸びる線分として描画し、その端点のドラッグを許している。属性軸をドラッグ操作に伴う射影の更新の扱いは AGI と同様である。

5. PCP の拡張と MPP との連携

前節では AGI を多次元データに応用し、グラフ描画にはなかった軸の扱いも含めた多次元データの対話的可視化手法について説明した。この節では、PCP の軸の描画順序などの操作系として MPP を活用する手法について述べる。MPP を利用することで PCP の軸の配置問題に対して柔軟な操作体系を提供できるだけでなく、PCP が苦手としてきたクラスタ分析の機能を MPP に与えることで両者を相補的に連携できることについて述べる。

5.1 幅可変 PCP

2 節に述べたように、PCP において軸の表示順は重要な意義がある。相関の高い軸が隣接する場合は、相関係数の符号に応じて軸間を結ぶ線分群は特徴的な模様を織り成すからだ。このため、過去の研究においては PCP の軸の順序をユーザが手動で、半自動的なサポートを用いて、あるいは他の可視化手法と組み合わせて間接的に制御する方法が提案されてきた。しかしながら、軸の順序の並びの組み合わせは膨大であるため、従来の手法は軸数の増加にともない爆発的な困難をともなう。

軸の順序だけでなく軸間の距離もあわせて調整するべき能力である。隣接した軸間に相関が高い場合、それらの情報はかなり重複していると見做せる。そのような軸の組に

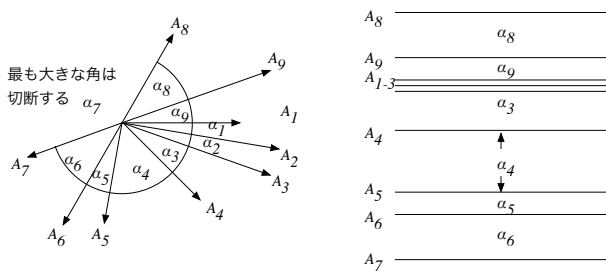
大きな空間を割り当てることは、空間の情報密度の面から望ましくない。そのため、軸間の距離も必要に応じて調整する機能は有用と考えられる。しかし、軸間の距離を調整する機能を与える研究はほとんどない。

本研究では、個々の軸の位置ではなく、軸の並び全体をMPPを通して包括的に操作対象とすることで、これら二つの問題をともに解決することを目指す。

このための準備として、通常は固定間隔で描画されるPCPに軸間距離を可変とする拡張を提案する。まず、通常のPCPを用いて n 項目、 m 属性の多次元データ行列 $D = \{d_{ij}\} : m \times n$ を縦に等間隔で並んだ属性軸を用いて描画するものとする。このとき、等間隔に並んだ属性軸の Y 座標は $Y = (i-1)/(m-1)$, $1 \leq i \leq m$ と表すことができる。

一方、我々が提案する軸間距離可変なPCPにおいては、属性軸は等間隔ではない。 i 番目と $(i+1)$ 番目の軸の間の距離を $0 \leq r_i \leq 1$ とすると、 $\sum_{i=1}^{m-1} r_i = 1$ 。そして j 番目のデータ項目 $(d_{1j}, d_{2j}, \dots, d_{mj})$ に対応する折れ線は、 $((d_{1j}, 0), (d_{2j}, r_1), (d_{3j}, r_1 + r_2), (d_{4j}, r_1 + r_2 + r_3), \dots, (d_{mj}, 1))$ となる。

以下では、MPPにおける属性軸の配置によって、PCPの属性軸の順序と隣接属性軸間の距離がどのように与えられるかについて述べる。今、9つの属性を持つ多次元データを可視化しているものとし、以下の図の左右がそれぞれMPPとPCPによる可視化とする。MPPによる可視化において、属性軸が原点を中心に時計回りに $A_1, A_2, \dots$ の順序で配置されており、属性軸 A_i と A_{i+1} がなす角度が α_i であったとする。このとき、MPPの可視化において隣接属性軸がなす角度をもって、PCPの可視化の属性軸間の距離とする。



たとえば、左図のMPP表示において属性軸が時計回りに $A_8, A_9, A_1, A_2, \dots, A_7$ と配置されているが、右図のPCPにもそれとまったく同じ順序で属性軸を配置する。さらに、PCP表示の属性軸間の距離はMPP表示において隣接属性軸がなす角度で定める。たとえば、MPP表示において A_4 と A_5 がなす角 α_4 を、PCP表示の属性軸 A_4 と A_5 の間の距離として反映している。なお、現在の実装においてはMPP表示において最も大きな角度をなす属性対をPCP表示の両端に配置している。

すでに述べたように、MPPは頂点や軸のドラッグによつ

て、個々の頂点や軸を移動するのではなく、高次元空間において回転操作を行っている。このため、わずかなドラッグ操作の系列が、可視化結果の大きな変化をもたらす点が特長である。この性質によって、属性数が増加した場合も、比較的少ない対話操作によって効果的にクラスタリングを実施でき、さらにこの操作によって相関の強い属性軸が隣接して表示される傾向にある。

この性質はPCP表示にとって重要な意義がある。MPP表示の操作で相関の高い軸が隣接し、そのなす角度が小さくなった場合、それらの相関の高い軸はPCP表示においても隣接し、それらの間隔が狭く割り当てられる。すでに述べたように相関の高い軸間には情報の重複が多いため、これらの描画に狭い領域が割り当ててことは合理的である。

5.2 項目間類似性と次数中心性にもとづくフィルタ

MPPは高次元空間に配置された多次元データを射影したものを表示する。このため、MPP表示において接近した2頂点が必ずしも高次元空間において近接しているとは限らない。このような誤解は、AGIに由来する頂点や座標軸のドラッグ操作によって修正することができる。MPPやAGIは3次元可視化を行っているわけではないが、ドラッグ操作の様子が多くのユーザーに3次元的な幻想を与えるためである。

この機能に加えて、ここではより直截的に高次元空間上における近接する頂点間を辺で接続する可視化手法を提案する。具体的には、多次元データ $D = d_{ij}$ 上の項目 $\hat{p}_j = (d_{1j}, d_{2j}, \dots)$ について、ノルムが閾値未満の場合 $(\|\hat{p}_j - \hat{p}_k\| < \theta)$ に辺を与えることとする。

社会ネットワーク解析において中心性と呼ばれる概念は研究されている。多次元データの多くも社会ネットワークに由来するものであり、中心性の可視化は重要である。本システムでも社会ネットワークに関連づけられた多次元データの解析を助けるために、次数中心性に応じたフィルタを装備している。

ここで述べたMPP表示におけるフィルタの仕組みは、全頂点対が織り成す完全グラフ $G = (V = \{1, 2, \dots, m\}, E = V^2)$ に対するフィルタ機能を見做せる。すなわち、高次元空間における頂点間距離についての閾値 θ と次数中心性についての閾値 φ を変数として、 G へのフィルタ機能は以下で表される。

$$V' = \{v \in V \mid \deg_\theta(v) > \varphi\}$$

$$E' = \{(v_j, v_k) \in V^2 \mid \|\hat{p}_j - \hat{p}_k\| < \theta\}$$

$$G' = (V', E')$$

MPP上の操作によって多次元データの項目にクラスタ構造を発見できることがしばしばある。ここで見つかるクラスタに属するデータ項目に共通する性質をPCP表示で

見つけたい。あるいは、逆に PCP の特定の座標軸において、値の分布に明瞭な傾向が見られた場合、その軸の特定の範囲の値を持つデータ項目群の多次元空間における分散状況を MPP で観察したい場合もある。

このような要請に答えるために、MPP, PCP それぞれの表示画面においてデータ項目群を選択し、自動的に着色する機能を用意した。一方の表示画面において選択されたデータ項目群については、他方の画面において対応するデータ項目を同じ色で彩色する。

MPP 表示においてはデータ項目は点として表示され、データ群是一群の点である。本システムにおいては、ドラッグ操作で指定される矩形領域、あるいは多角形領域内の点群が選択される。これらの点群には、自動的に彩色がなされる。このため、複数の領域を順次選択した場合、それらの分布を色の違いとして認知できる。

PCP 表示においてはデータ項目は折れ線として表示され、データ群は折れ線の集合である。MPP 表示において指定されたデータ項目群に対応する折れ線群はそれぞれ、MPP 表示におけるのと同じ色で着色される。この連携機能により、MPP で発見されたクラスタに属するデータ項目群の属性ごとの傾向を観察することができる。

PCP 表示においても I ツール [7] と同様の方法でデータ項目群を選択することができる。PCP 表示において、ある属性軸に沿ってドラッグ操作を行うと、その属性に関して、ドラッグした範囲の値を持つデータ項目が選択され、MPP における選択機能と同様に彩色される。PCP で選択されたデータ項目群への彩色は MPP 画面にも反映される。

5.3 部分データ項目群の可視化

ここまで、多次元データに MPP と PCP を組合せた可視化を実施し、それらを連携させて複数のデータ項目群に分離できることを述べた。複雑な多次元データの場合、選択したデータ項目群のなかにさらに部分構造が見つかる場合がある。ここでは、このような多次元データに内在する階層的な構造を分析するための機能について述べる。

基本的には前述した方法で選択されたデータ項目群を新たな多次元データと見做し、これを表す行列 D' を構成し提案手法を再適用する。ただしこの部分集合において全項目での値が一定値である属性、つまり任意の j_1, j_2 に関して $d_{ij_1} = d_{ij_2}$ をみたす属性 X_i は MPP において PCA で高次元配置を求める際に無意味なので取り除く。このような属性を取り除いた多次元データ行列 D' を MPP, PCP 双方を用いて再び可視化する。

6. 評価, 考察

前節までで提案した手法を可視化システム VisLAMP として実装することで可視化を行う。これらの結果について評価および議論し、本手法の有用性を示す。

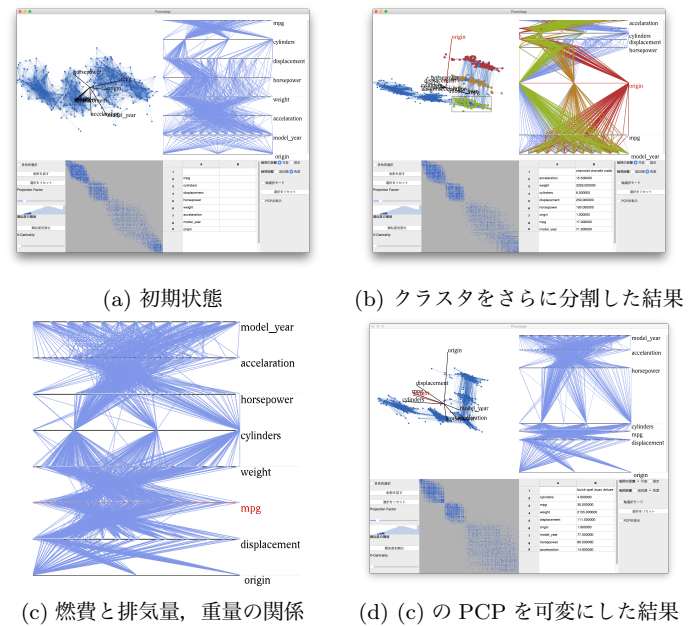


図 1 Cars データセットを可視化した結果

6.1 計算量に対する考察

まず可視化を行う前に本手法における計算量について考察する。最初に初期状態を生成するための計算量について考える。最初の高次元配置においては主成分分析で求めた際は固有値の計算量と等しくなるので $O(m^3)$ になる。項目同士の類似度の計算は m 次元の配置に対し n 個の項目間について計算するので $O(mn^2)$ となる。初期の射影を求める際は m 個の固有値を成分とした 2 本の射影ベクトルを求めるため $O(m)$ である。項目数を n としたとき通常 $n > m$ と考えてよいので、よって初期配置を求めるのに必要な計算量は $O(mn^2)$ となる。

次にユーザの操作により射影の更新が行われた場合の操作について考える。MPP における射影の更新については AGI と違い、項目だけでなく属性についても更新を行う必要がある。よって $O(m(n+m))$ である。次に幅可変 PCP の更新の際に計算量について考える。これはどちらも属性数 m のみが関係する。角度を使った計算方法の場合、ソートの計算量が更新の中において最も多くなるので $O(m \log m)$ であるとしてよい。よって更新にかかる計算量は $O(m(n+m))$ である。また提案手法全体の計算量は $O(mn^2)$ となる。

6.2 Cars データセットの可視化

本研究で提案した手法による可視化結果の評価検証にあたり、多次元データ可視化の研究においてよく用いられる Cars データセット [9] を扱う。このデータセットの項目は車種を表しており、その数は 398 である。また 8 の属性を持つ。まず本システムで可視化を行い、それを元に既存手

法と比較する。

本システムを用いて Cars データセットを可視化した初期状態が図 1(a) である。この状態において既に項目が 3 つのクラスタに分割されている。この三つのクラスタを左側から順にクラスタ 1, クラスタ 2, クラスタ 3 とする。

ここで属性軸に注目すると重量, 排気量, シリンダー数の軸が同方向を向いており, この正反対の方向に燃費が描かれており, これらの軸はクラスタ 1, 2, 3 と平行に描かれている。この特徴からクラスタ 1, 2, 3 はこの四属性で分割されており, 四属性に相関があることが考えられる。これらの属性は車の大きさや重量に依存すると考えられる。つまりクラスタ 1 が大型車であり, 2, 3 となるに従って大きさや馬力が小さくなっている。

さらに, クラスタ 3 内のノードをマウスで移動させながら射影行列を変化させることで可視化結果を詳しく観察すると, 他のクラスタに比べサイズが大きい上に項目の分布が広く, さらにサブクラスタに分割可能ではないかと興味を持った。

そこでこのクラスタ 3 に対し分割を試み, 得ることができた結果が図 1(b) である。この際 MPP 画面に描かれている製造地の軸を選択し, 先程の複数の属性と垂直に近い形で描かれるように移動操作を行った。それに伴い, 各項目の位置も更新されている。この結果, 先程のクラスタ 3 がさらに三分割された形で可視化されている。これらをクラスタ 3-1, クラスタ 3-2, クラスタ 3-3 として分析する。

ここで再び属性軸に注目すると MPP 画面では製造地の軸が非常に長く描かれている。この点から, このクラスタ 3-1, 3-2, 3-3 は製造地の値で分割されているのではないかと予想できる。ここで製造地の値は 1 が米国, 2 がヨーロッパ, 3 が日本を意味している。実際にクラスタの各要素について PCP 画面を併用しながら分析したところ製造地の値だけが異なり, それ以外の属性の値の傾向は似ていることが可視化されていた。またクラスタ 1, 2 に含まれる項目の製造地はほぼ全てが 1 で僅かに 2 があることもわかった。一方, PCP 画面の可視化結果としては製造地が非常に強調されている。これは他の属性との軸間の角度の差が大きいためである。そのため PCP 画面における隣接軸との距離も大きくなっている。

これらを総括すると Cars データセットはまず前述の 4 属性によって車のサイズでクラスタリングされており, このうち小型車はさらに製造地でのクラスタリングが可能であるといえる。逆にこのデータセット内の大型車, 中型車はアメリカ製が大多数である。

なお, PCP 画面においては燃費を除く三属性間が隣接しており, 軸間の幅が小さくなっている。ただ PCP を固定にし軸の位置を調整し, PCP 部分だけを抜粋した図 1(c) を見ると燃費と重量, 排気量には負の相関関係がある。しかし, これを前述の画面から読み取ることはできない。こ

れは軸の決定を MPP 画面上におけるベクトルの角度で計算しているため, ほぼ 180 度反対側にある図 1(b) では軸が離れてしまうためである。この状態で PCP の幅を変えると図 1(d) となり, MPP 側はさほど変化がないが PCP の軸の順番は大きく変化していることがわかる。このように MPP 画面に比べ PCP の画面の変化は大きくなりやすい。

6.3 Cars データセットにおける既存手法との比較

この結果を既存の可視化手法と比較する。まず項目についての可視化結果について考える。この点においては提案手法は強力なクラスタ発見能力を持っていると考えられる。例えば Cheng らの研究 [12] においても Cars データセットを扱っているがここで発見できているクラスタは本研究で最初に分割した三つのクラスタのみである。また Palmas らが同じ Cars データセットを可視化した結果 [11] と比較する。ここで彼らの可視化結果の製造地の軸に注目するとある一つのクラスタのみが日本やヨーロッパに存在している。この点に関しては本研究と同一であると言えるが彼らはあくまでも PCP における隣接した二軸のみを用いた局所的なクラスタリングを元にした結果このクラスタリングが行えているのに対し本研究はグローバルなクラスタとして発見することができている。この点から考えると本研究において MPP 画面で特定の属性軸を操作し, その影響を大きくする可視化を行うことがその属性に注目したクラスタリングに対応していることがわかる。

次に複数の手法を組み合わせたものとの比較に移る。SPPC[15] とは MDS による高次元からの射影による可視化と PCP を組み合わせている点は本研究と類似しているといえる。またこの研究においてもインタラクションの結果製造地を座標計算に使うことで 5 のクラスタに分けることに成功している。

しかし, 本研究において必要なインタラクションは MPP 上に表示してある属性軸を操作するのみである。また本研究において, 可視化結果を変更するためには射影行列の更新のみであることに対し, この研究では改良してあるとはいえず, MDS で高次元配置の計算をやり直さなければならない。本研究で射影の更新に必要な計算量は 6.1 節でみたように $O(m(m+n))$ であるのに対し, この研究では $O(n^3)$ の計算量が必要である。このインタラクションの簡便さと可視化結果の更新にかかる計算量は本研究の方が優れているといえる。

次に PCP に関する評価に移る。具体的には属性の特徴や相関性について適切に可視化されているかを評価する。この観点に関しては軸の順番が重要である。

まず本研究においては前節で述べた通り, MPP での属性軸の配置に強く依存する。図 1(b) では正の相関を示す重量, 排気量, シリンダー数がまとまっているものの, こ

れらと燃費との負の相関関係が可視化できていない。

しかし、例えば前述した Palmas らの研究 [11] ではこのような負の相関関係も判るように可視化されている。この点から考えると本研究における PCP の軸の順番の決定方法は類似した軸を束ねることができてはいるもののさらなる改良が必要であると考えられる。

とはいえ本研究には強力なインタラクション性があり、必要に応じてマウスを用いて属性軸を操作するのみで PCP の順番を変えることができる点、そして実際に可視化結果を変更することにより図 1(c) のようにこの負の相関関係を可視化することができる点を考えればこの事項は本研究の有用性を否定する程ではないといえるだろう。

7. おわりに

本研究では PCP をはじめとした項目を折れ線として表示可視化と散布図など項目を点として表示可視化についての長所と短所を分析した。そしてその結果からそれぞれの長所を両立し、短所を補うような多次元データ可視化手法に用いるために MPP を AGI を応用する形で提案した。

そしてこの MPP と軸間距離を可変にした PCP を並置し連携させることで実際に長所を両立させる手法を考案した。MPP で射影した軸の配置を元に PCP の軸の順番と軸間距離を決定した。クラスタ発見のための機能を双方の画面に実装し、そのクラスタ内の構造を詳細に分析するために再度提案手法を適用する方法を考案した。

そして提案手法を可視化システム VisLAMP として実装した。Cars データセットに対し、このシステムによる両画面を連携させた可視化分析によって車の大きさに関するクラスタリングだけでなく製造地によるクラスタ発見を行うことができたことから既存手法の結果と比較しても本研究の手法の有用性を示すことはできたといえる。

一方で今後の課題も幾つか挙げられる。まず本研究では MPP 側で射影を更新することから PCP の軸間距離を変更することを提案したが、この逆として PCP の軸の位置を変更することで MPP 上の属性軸を移動させ射影を更新するといった手法も考えられる。実装面の問題としては大規模なデータに対して計算時間の増加など性能面での問題が大きいことが挙げられる。また評価手法として既存手法との比較に加えユーザーテストを行うべきである。

謝辞 本研究の一部は科学技術振興機構戦略的創造研究推進事業 (JST, CREST) の支援を受けています。

参考文献

[1] Ankerst, M., Berchtold, S. and Keim, D.: Similarity clustering of dimensions for an enhanced visualization of multidimensional data, *Information Visualization, 1998. Proceedings. IEEE Symposium on*, pp. 52–60, 153 (online), DOI: 10.1109/INFVIS.1998.729559 (1998).
[2] Claessen, J. and van Wijk, J.: Flexible Linked

Axes for Multivariate Data Visualization, *Visualization and Computer Graphics, IEEE Transactions on*, Vol. 17, No. 12, pp. 2310–2316 (online), DOI: 10.1109/TVCG.2011.201 (2011).
[3] Ferdosi, B. J. and Roerdink, J. B.: Visualizing High-Dimensional Structures by Dimension Ordering and Filtering using Subspace Analysis, *Computer Graphics Forum*, Vol. 30, No. 3, pp. 1121–1130 (online), DOI: 10.1111/j.1467-8659.2011.01961.x (2011).
[4] Giménez, A., Rosenbaum, R., Hlawitschka, M. and Hamann, B.: Using R-Trees for Interactive Visualization of Large Multidimensional Datasets, *Proceedings of the 6th International Symposium on Visual Computing*, Springer, pp. pp 554–563 (2010).
[5] Hosobe, H.: A High-dimensional Approach to Interactive Graph Visualization, *Proceedings of ACM Symposium on Applied Computing (SAC 2004)*, New York, NY, USA, ACM, pp. 1253–1257 (online), DOI: 10.1145/967900.968155 (2004).
[6] Inselberg, A. and Dimsdale, B.: Parallel coordinates: a tool for visualizing multi-dimensional geometry, *Visualization, 1990. Visualization '90., Proceedings of the First IEEE Conference on*, pp. 361–378 (online), DOI: 10.1109/VISUAL.1990.146402 (1990).
[7] Inselberg, A. and Dimsdale, B.: Parallel coordinates, *Human-Machine Interactive Systems, Languages and Information Systems*, Springer US, pp. 199–233 (1991).
[8] Kruskal, J. B. and Seery, J. B.: Designing network diagrams, *In proceedings of the First General Conference on Social Graphics*, Leesburg, VA, USA, pp. 22–50 (1978).
[9] Lichman, M.: UCI Machine Learning Repository (2013).
[10] Meyer, M., Barr, A., Lee, H. and Desbrum, M.: Generalized Barycentric Coordinates on Irregular Polygons, *Graphics Tools*, Vol. 7, No. 1, pp. 13–22 (online), DOI: 10.1080/10867651.2002.10487551 (2002).
[11] Palmas, G., Bachynskyi, M., Oulasvirta, A., Seidel, H. and Weinkauff, T.: An Edge-Bundling Layout for Interactive Parallel Coordinates, *Pacific Visualization Symposium (PacificVis), 2014 IEEE*, pp. 57–64 (online), DOI: 10.1109/PacificVis.2014.40 (2014).
[12] S. Cheng, K. M.: Improving the Fidelity of Contextual Data Layouts Using a Generalized Barycentric Coordinates Framework, *Proceedings of IEEE Pacific Visualization Symposium 2015*, IEEE Pacific Visualization Symposium 2015 (2015).
[13] Wakita, K., Takami, M. and Hosobe, H.: Interactive high-dimensional visualization of social graphs, *Visualization Symposium (PacificVis), 2015 IEEE Pacific*, pp. 303–310 (online), DOI: 10.1109/PACIFICVIS.2015.7156391 (2015).
[14] Watanabe, K., Wu, H.-Y., Niibe, Y., Takahashi, S. and Fujishiro, I.: Biclustering Multivariate Data for Correlated Subspace Mining, *Proceedings of IEEE Pacific Visualization Symposium 2015* (2015).
[15] Yuan, X., Guo, P., Xiao, H., Zhou, H. and Qu, H.: Scattering Points in Parallel Coordinates, *Visualization and Computer Graphics, IEEE Transactions on*, Vol. 15, No. 6, pp. 1001–1008 (online), DOI: 10.1109/TVCG.2009.179 (2009).