統計数理(2007)
第 55 巻 第 1 号 69–83

特集「統計データの可視化」
[原著論文]

3次元平行座標プロット

本多 啓介[1] ・中野 純司[1,2]

(受付 2006 年 8 月 21 日；改訂 2007 年 3 月 19 日)

要　　旨

平行座標プロットは多変量データを視覚化するのに有用な統計グラフである．平行座標プロットで変数間の関係を調べるためには，brushed highlighting と呼ばれる対話型操作を用いることが一般的であるが，これだけでは変数間の関係を同時に把握することは難しい．本論文では，平行座標プロットを 3 次元に拡張し，変数の関係を全体的に表示することを目的とした手法を提案する．これは，平行座標プロットで，観測値を表す折れ線を直交した座標に配置することにより実現する．この視覚化手法で，条件変数の値でデータを分割することにより，非線形関係にある変数に対して，部分的な線形関係を視覚化することが可能である．例題を用いた解析で，この手法が探索的なデータ解析に有用であることも示す．

キーワード： 3 次元データ表示，条件付きグラフ，対話型操作，統計グラフ．

1. はじめに

平行座標プロット(parallel coordinate plot, PCP)は多変量データを視覚化するのに有用な統計グラフの一つである(Inselberg, 1985; Wegman, 1990)．平行座標プロットでは，変数の座標軸を平行に配置し，すべての変数で最小値が下端に，最大値が上端になるように観測値をプロットする．そして隣接する座標軸上の観測値をそれぞれ線分で結ぶ．平行座標軸の間隔を調整することにより，すべての変数の座標軸を一画面に表示できることから，同時にすべてのデータを表示することが可能である．観測値を表す折れ線が一般的なデータの表(行に観測値，列に変数をとる)の一行に対応している点で，基本的なグラフと言える．

平行座標プロットは，隣接する 2 変数間の特性を直接的に示すことが可能であるが，2 軸以上離れた変数間の関係は間接的に示すことしかできない．また，データ数が増大すると観測値を表す折れ線が表示領域全体を覆い，各観測値を区別することができないという問題もよく知られている．これらの問題を解決するため，平行座標プロットを用いたデータ解析では対話型操作を行う．例えば，グラフ上の選択した領域に含まれる観測値を強調して表示したり，平行座標の座標軸の順序の交換を行うことなどである．

このような統計グラフを用いたデータの視覚化は非常に有用であることが知られている(Symanzik, 2004)．さらに，平行座標プロットと散布図や散布図行列などの他の統計グラフとの間で，強調表示などを連動させる linked views と呼ばれる機能もデータの特徴を捉える上で重要な機能であり，多くの視覚化システムで実装されている．これらの機能を有した代表的なシステムとして，Mondorian (Theus, 2002)や GGobi (Swayne et al., 2004)がある．Mondrian は主

[1] 総合研究大学院大学 複合科学研究科統計科学専攻：〒106–8569 東京都港区南麻布 4–6–7
[2] 統計数理研究所：〒106–8569 東京都港区南麻布 4–6–7

にカテゴリカルデータ，地理データを扱うことを目的に開発されたシステムである．Mondrian は，α-Chanel と呼ばれる，透明度を用いた重ね合わせ処理を備え，大規模データの視覚化に優れている．GGobi は，grand tour と呼ばれるデータ表示手法を中心としたソフトウェアであり，フリーの統計解析システムである R 上でも動作する．

平行座標プロットを 3 次元空間に拡張する研究はいくつか行われている．Wegenkittl et al. (1997) は Extruded parallel coordinate と呼ばれる手法を提案した．これは特に多変量時系列データの視覚化を目的としたものである．平行座標の軸と直交した観測時刻を表す座標軸を追加することで，各時刻を表す観測値の折れ線が等間隔に 3 次元空間に配置される．観測時刻に沿って折れ線を並べることにより，データ全体の時間ごとの変化を視覚化できる．Barlow and Stuart (2004) は，観測値を表す線分を 3 次元空間で平行移動することにより平行座標プロットを表示している．Johansson et al. (2005) は clustered multi-relational parallel coordinates technique と呼ばれる手法を提案した．この手法は 3 次元空間で円の中心に，注目した変数の座標軸を置き，その円周上に他の変数の座標軸を配置する．そして円の中心と円周上の各変数の軸との間で観測値を線分で結ぶ．その際，クラスタリングの結果を反映させた色で観測値の線分を塗り分ける．また，各変数の軸を配置する弧の長さは中心の変数との相関係数の大きさに比例して決められる．この結果注目した変数と他の変数との関係を同時に表示することができる．また，Fanea et al. (2005) は Parallel Glyphs と呼ばれる，平行座標の各軸のかわりに star glyphs を用いた手法を提案した．star glyph は星型の図形で，一つの glyph は変数，もしくは観測値を表す．Parallel Glyphs の手法は変数ごとの star glyph を作成し，それらを 3 次元空間に平行に並べる．観測値の値を表す頂点は線分で結ばれる．また，各 glyph を容易に識別するため色で区別して表示する．Parallel Glyphs は平行座標プロットのデータ全体を視覚化する利点と star glyph の各変数の特徴を視覚化する利点の両方を持つ．

平行座標プロットを用いて変数間の関係を明らかにするためには，選択領域を一つの変数の平行座標の軸上で直線的に連続して移動させ，その移動の過程で，強調表示される折れ線の動きを調べることが多い．この操作は brushed highlighting と呼ばれるものである (Symanzik, 2004)．平行座標の軸上で選択領域を移動させるということは，注目した変数の大きさの順に観測値を並べ替えて表示することに相当する．これは画面上で選択領域を移動させる一種のアニメーション処理であるがゆえに，その操作の過程を同時に見ることはできない．データの構造を調べるためには，その構造を一つのグラフとして示すことが望ましい．さらに，データの局所的な構造を調べるためには，表示領域の拡大，視点の回転といった操作が可能な 3 次元空間によるグラフ表示が有効であろう．

そこで本論文では，2 次元の平行座標プロットを 3 次元に拡張することで，変数間の関係を同時に明らかにすることを目的とした視覚化手法を提案する．この視覚化手法を 3 次元平行座標プロットと呼び，特に部分的な線形関係を明らかにすることを目的とする．

次節では本論文で提案する視覚化手法の基本的な考え方について述べる．3 節で変数間の非線形関係を見つけるための条件変数の考え方とその視覚化の方法について人工データを用いて説明する．4 節では解析事例で本手法の有用性を示す．5 節は簡単なまとめである．

2. 2 次元平行座標プロットの 3 次元への拡張

平行座標プロットでは 2 変数間の線形関係は隣接する平行座標軸間の観測値を表す線分のパターンとして現れる (Inselberg, 1985)．2 変数間の正の相関が強くなるとそれらは平行に近くなり，逆に負の相関が強くなると平行座標軸間の中心で交差する様な形が現れる．例えば，平行座標プロットの例である図 1 を見ると，右端の変数 medv とその隣にある変数 lstat ではそ

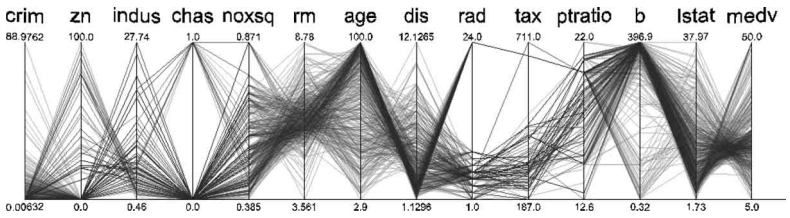

図 1. ボストンの住宅事情データを表示した平行座標プロット.

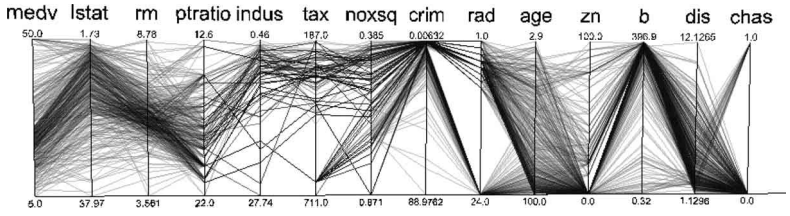

図 2. 目的変数の相関係数で座標軸を並べ替えた平行座標プロット.

のパターンが見られる．これは 2 変数間に負の相関関係があることを示している．

ところで，変数間の線形関係を見る場合，まず，相関係数を調べることが一般的である．そこで，変数 medv に対して，それ以外の変数との相関係数を計算する．そして，相関係数の絶対値が大きい順に他の変数を並べ替える．この時，相関係数の値が負である変数に対して，平行座標の軸の大小の方向を逆転させると，データ全体の線形関係が見やすくなる場合がある．すなわち，大小を逆転させた変数では，平行座標の下部が最大値になり，上部が最小値になる．図 2 はこの操作の結果を示したものである．この図を見ると，medv に対して相関係数の絶対値が比較的大きかった変数は lstat，rm，ptratio であった（相関係数の値はそれぞれ $-0.738, 0.695, -0.51$ であった）．左端の medv から順に ptratio まで，隣接する線分の多くが平行に近い．

このように 2 次元の平行座標プロットにおいては 2 変数間の線形関係が，平行座標軸間の線分のパターンとして直接現れる．従って，目的変数と説明変数との関係を見るためには，隣接する変数を交換する操作を繰り返すことが必要になる．ただし，brushed highlighting を目的変数に対して行うことで，隣接する変数間だけでなく，データ全体の特徴も間接的に観察することはできる．

しかし，brushed highlighting によっても，目的変数と説明変数の関係を同時に見ることは難しい．それは，brushed highlighting による動的な処理は選択した領域をスライドさせることでデータの変動を見ていくが，変数の値の小さい場所から大きな場所までの変数の動き全体を同時に表示することはできないからである．そこで，これを 3 次元空間を利用し表示することを考える．すなわち平行座標プロットにおいて，注目した変数を基準とし，その変数の観測値を，新たに追加した平行座標に対して直交な座標に配置し，その結果を線分で結び，3 次元空間に表示することを考える．3 次元空間で X 軸，Y 軸，Z 軸を図 3 のようにとる直交座標系を考える．この Z 軸に対し，基準とした目的変数の観測値を大きさの順に配置する．この状態で観測値を表す折れ線を表示すると，基準変数の大きさで並べ替えられた平行座標プロットになる．このように 3 次元空間に表示された統計グラフを 3 次元平行座標プロット（3 Dimensional Parallel

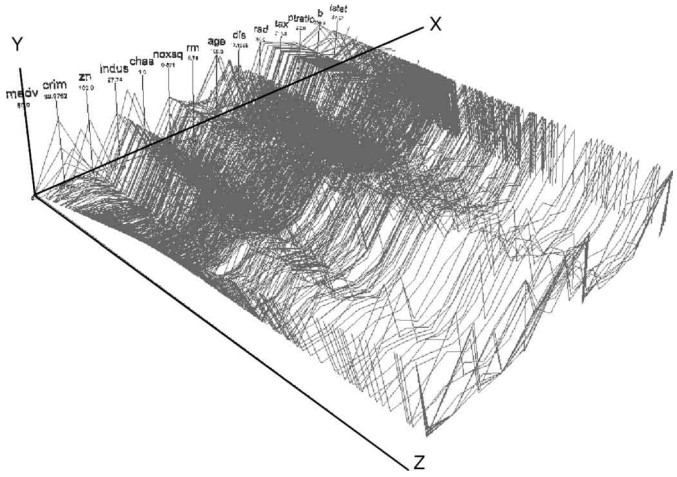

図 3. 折れ線を観測値ごとに結んだ状態の 3D PCP.

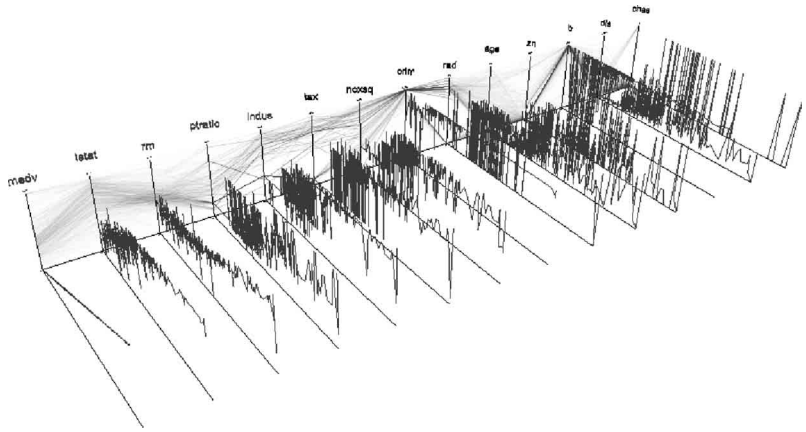

図 4. 目的変数の相関係数で座標軸を並べ替えた状態の 3D PCP.

Coordinate Plot, 3D PCP) と呼ぶ．3D PCP において観測値を表す点を結ぶ折れ線の表示には 2 種類の方法が考えられる．観測値ごとに結ぶ方法と変数ごとに結ぶ方法である．図 3 は折れ線を観測値ごとに結び表示した状態であり，図 4 は折れ線を変数ごとに結び表示した状態である．観測値ごとに結んだ折れ線は 2 次元の平行座標プロットと同じ折れ線を基準変数を用いて Z 軸上に配置した状態である．3D PCP では観測値を表す折れ線と変数を表す折れ線を同時に 3 次元空間に表示できる．

ここでは，変数ごとに結んだ折れ線が基準とした変数との関係を表すことに注目する．Y 軸と Z 軸で構成される平面内で各観測値を変数ごとに線分で結ぶと，基準とした変数の場合，その線分は観測値を最小値から最大値までを結ぶ直線になる．そして，他の変数に対しても基準とした変数の観測値の順に，観測値を線分で結ぶと，基準とした変数との関係を表す折れ線となる．図 4 では左端の medv を基準とした．このとき，各変数の折れ線は，基準とした変数と

強い線形関係にある場合，直線に近いものとなる．

図 4 では medv との関係を見るため，2 次元の平行座標プロットの場合の図 2 と同様に 3D PCP 上で平行座標の変数の順序を相関係数の大きさによって並べ替えて表示している．また相関係数の値が負である変数の平行座標の大小の方向を逆転させている．この操作により，相関係数の正負に関わらず，基準とした変数の直線と近い形状の折れ線をもつ順に平行座標が並ぶことになる．これにより，基準とした変数との関係を表す折れ線を比較しやすくなる．

図 4 を見ると，左端に medv の直線があり，2 番目以降順に，lstat，rm，ptratio といった順で変数を表す折れ線が並ぶ．左端の medv から順に ptratio まで，各折れ線が medv の直線と近い形で表されている．図 2 でも示したように medv に対して相関係数の絶対値が最も大きい lstat の折れ線を見てみると，lstat は medv の観測値が小さいときに，大きくなり，medv の観測値が大きくなるにつれ，徐々に大きくなる．lstat は座標軸の逆転を行ったので，これは medv との負の相関関係を示している．

次に，2 番目に相関係数の絶対値が大きかった rm の折れ線を見てみる．rm は画面中央にある．rm は medv の値が小さいときに，小さい観測値が並び，観測値が大きくなるにつれ，rm も徐々に大きくなる．これは medv との正の相関関係を示している．

さらに右端にある変数 crim の折れ線を見ると，medv の値が小さい部分に集中して値が大きなものが現れている．2 次元の平行座標プロットでは，主に隣接する 2 変数間の関係を見ることが出来たが，3D PCP は，すべての変数を Y-Z 平面に折れ線として配置することで，brushed highlighting を行わずにすべての変数の関係を同時に示すことが出来る．

2 次元の統計グラフではデータ全体を視覚化するものとして散布図行列があり，対話型操作を用いて 3D PCP で行ったような解析を行うことができる．しかし，変数の数が増大すると，2 次元平面で描画する散布図行列では観測値を同時に表示することは困難になる．とくに現在の計算機の表示環境では 20 変数以上になると，散布図行列では有効なデータ解析を行うことは難しいと思われる．変数間に非線形関係がある場合，散布図行列では見えない関係が 3D PCP では特徴的なパターンとして現れる場合がある．これについて人工データを用いて次節で説明する．

3. データの分割による条件付きグラフ

3D PCP によってデータを視覚化し，折れ線の比較をしながらデータの関係を観察すると，線形関係だけでなく，非線形関係を見つけることができる場合がある．例えば，ある値を境に変数の増加傾向が明らかに変わる場合や，増減が逆転する場合である．このような非線形関係をもつ変数の場合，部分的な線形関係として近似できる場合がある．いくつかの変数で構成される条件のもとで観測値を分割することで，簡単な線形関係として説明できるグループに分割することができる．また，多変量データで因果関係がある場合，共変量に基づいてデータを分割した上で，解析を行うことが一般的である．3D PCP ではデータの部分的な線形関係の探索や層別化のために，ある変数の値によるデータの分割（条件付け）を行った上で視覚化する．

統計グラフを作成する際の条件付けの有用性は Cleveland（1993）によって示された．このような彼の提案は，S の Trellis Graphics や R の coplot コマンド（Murrell, 2005）として実装されている．

本節では人工データを用いて，3D PCP における条件付けの有用性について示す．2 つの確率変数 X_1，X_2 は区間 $[-1,1]$ 上の独立な一様分布に従うとする．$\boldsymbol{X}=[X_1,X_2]'$ の実現値を $\boldsymbol{x}=[x_1,x_2]'$ とするとき，関数

$$f_1(\boldsymbol{x}|\theta)=x_1\sin\theta-x_2\cos\theta$$
$$f_2(\boldsymbol{x}|\theta)=x_1\cos\theta+x_2\sin\theta$$

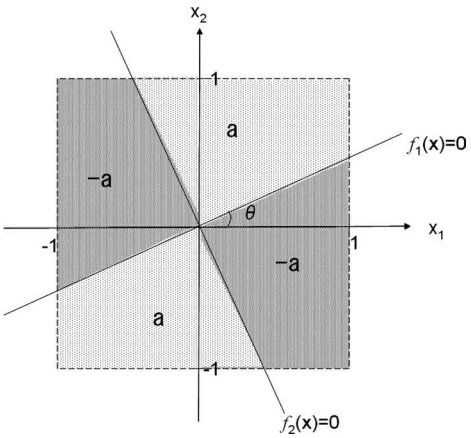

図 5. $y = g(\boldsymbol{x}|\theta)$.

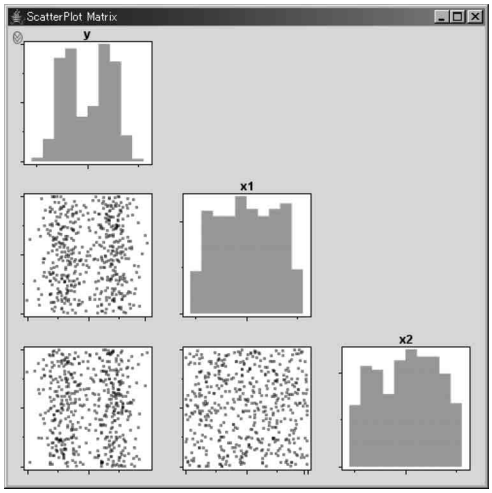

図 6. 散布図行列による人工データ ($a = 2$, $\theta = 0$) の表示.

を定義する．さらに，$a > 0$ として関数

$$g(\boldsymbol{x}|\theta) = \begin{cases} a & f_1(\boldsymbol{x}|\theta) f_2(\boldsymbol{x}|\theta) > 0 \text{ のとき} \\ -a & f_1(\boldsymbol{x}|\theta) f_2(\boldsymbol{x}|\theta) \leq 0 \text{ のとき} \end{cases} \tag{3.1}$$

を定義して，変数 Y の実現値を $y = g(\boldsymbol{x}|\theta) + \epsilon$ とする．ここで ϵ は平均 0，分散 1 の正規雑音とした．$y = g(\boldsymbol{x}|\theta)$ を図示すると図 5 になる．

まず，a=2，$\theta = 0$ としてデータを $[x_1, x_2, y]'$ を 500 組生成した．生成されたデータを散布図行列で表示すると図 6 のようになる．図 6 を見る限り，X_1 と X_2 がどのような影響を Y に与えているかについてはわからない．また，X_1 と X_2 に交互作用があるかについても図 6 からは読み取れない．次に，このデータを 3D PCP で表示する．ここでは，X_2 に注目し，X_2 を基

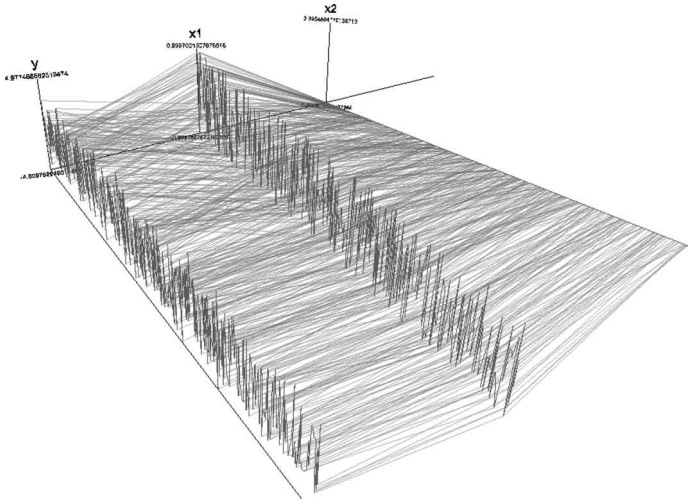

図 7. 3D PCP による変数 X_2 を基準とした人工データ ($a=2$, $\theta=0$) の表示.

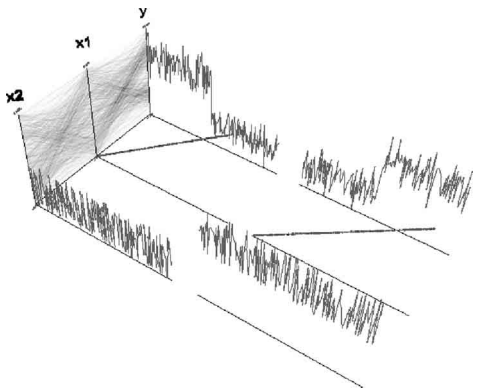

図 8. 変数 X_2 の条件によって分割された人工データ ($a=2$, $\theta=0$) の表示.

準として並べ替えて表示すると図 7 のようになる．図 7 において，Y と X_1 の間の実現値ごとに結んだ折れ線に注目すると，平行座標軸の間で交差した折れ線が多く現れる部分と平行に近い折れ線が多く現れる部分があることがわかる．

そこで X_2 の値の大きさでデータを分割して表示する．この場合は X_2 の値の中央付近, $x_2=0$ で分割する．次に，分割したグループごとに，変数の折れ線を配置する．ただし，このとき，複数のグループを Z 座標方向に配置する順序として，条件付けを行った際の変数の値の大小関係で順序付ける．すなわち，小さい値のグループは原点に近くなり，大きい値は原点から離れた位置に配置する．2 変数以上の変数で条件付けを行った場合は，条件変数の順序に合わせ，辞書式にグループを配置する．このような方法を用いて，X_1 を基準とする変数として表示すると図 8 のようになる．

この図から，X_2 の値の大きさによってデータが 2 つのグループに分割されているのがわか

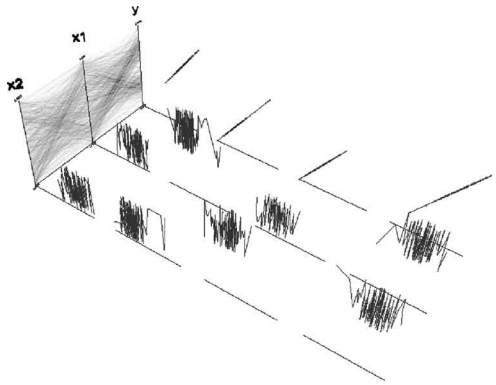

図 9. 変数 X_1 と X_2 の条件によって分割された人工データ ($a=2$, $\theta=0$) の表示.

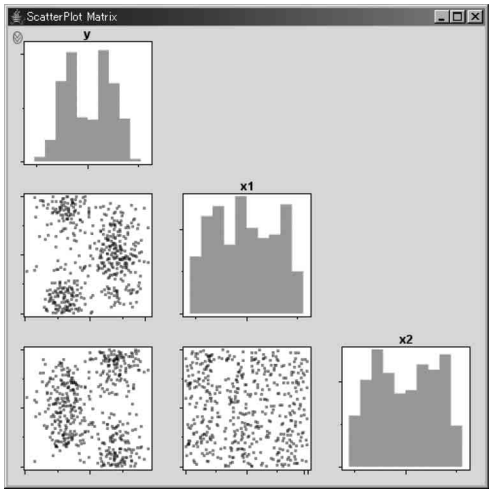

図 10. 散布図行列による人工データ ($a=2$, $\theta=\pi/4$) の表示.

る．ここで Y を表す折れ線に注目すると 2 つのグループとも X_1 の中央付近で大きい値と小さい値に分かれている．このことから X_1 と X_2 に Y に対する交互作用があると考え，X_2 の値で条件をつけた上で更に，X_1 の値の中央付近，$x_1=0$ という値でデータを分割する．2 変数で条件をつけることによりデータは 4 つのグループに分割される．それぞれ X_2 と X_1 が，小と小，小と大，大と小，大と大というグループである．この状態で Y を基準とする変数として表示すると図 9 のようになる．

この図から，X_1 と X_2 の値によって与えられた条件によって，内側に配置されている Y の値が大きい 2 グループと，外側に配置された Y の値が小さい 2 グループに分けられたことがわかる．このように，3D PCP では各グループがどのように分割されたかを折れ線の位置の高さとして視覚的に表現している．この特徴により各グループがどのような条件に従うか直感的に確認しながら，変数間の関係を同時に見ることができる．

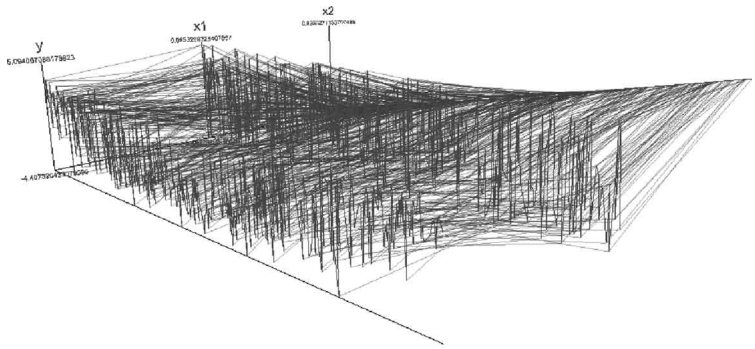

図 11. 3D PCP による変数 X_2 を基準とした人工データ ($a=2$, $\theta=\pi/4$) の表示.

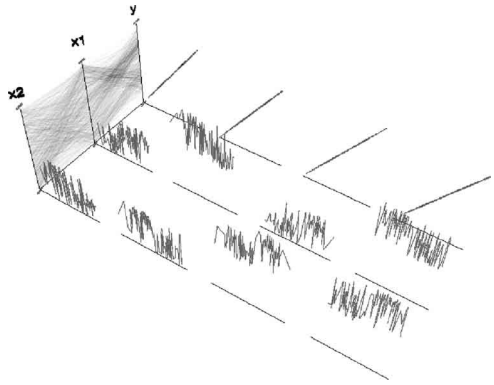

図 12. 変数 X_1 と X_2 の条件によって分割された人工データ ($a=2$, $\theta=\pi/4$) の表示.

ここで用いたデータでは，比較的簡単な条件付けでデータを分割することにより Y と X_1, X_2 との非線形関係を説明できた．次に，条件付けが容易でない例として $\theta=\pi/4$ の場合を考える．θ 以外はさきほどと同じ設定でデータを生成し，どのような視覚的パターンが現れるかについて見てみる．生成されたデータを散布図行列で表示すると図 10 のようになる．次に 3D PCP による解析を行ってみる．X_2 に注目し，X_2 を基準として並べ替えて表示する (図 11)．この図を見ると Y の変数ごとに結んだ折れ線が X_2 の値の中央付近で低く，下に凸のパターンを示していることがわかる．

ここでさきほどのデータと同じ条件，X_2=0，X_1=0 を与えてデータを分割し，Y を基準として表示してみる (図 12)．この図において変数 Y を表す各グループの線分に注目すると，Y の値の範囲に大きな違いはなく，与えた条件がうまくデータを分割していないことがわかる．

3D PCP ではこのように空間上に表示された観測値の折れ線のパターンにより，データ間の関係を視覚的に見つけることができる場合がある．その際，観測値を表す折れ線と変数を表す折れ線の両方をすべての変数で同時に表示できる点で散布図行列を用いた解析よりも多くの情報を得ることができる．これらの情報を用いて，簡単な条件の組み合わせでデータ分割を行うとデータ全体の関係がより明らかになる場合がある．基準変数としてどの変数を選ぶかは解析

の目的によって 2 通りの場合が考えられる．ひとつは目的変数と説明変数の関係を見るために基準変数として目的変数を選ぶ場合であり，もうひとつはデータ分割を行う際の条件を見つけるために解析者が注目した変数を選ぶ場合である．

4. 解析事例

本節では 3D PCP の手法を用いた解析事例について示す．これまでの図や例題でも用いてきたが，解析するデータとして Harrison and Rubinfeld (1978) のボストン住宅事情データを利用する．このデータはボストンの住宅価格 (人口調査地区単位の住宅価格の中央値，medv) と大気汚染の影響 (大気中の窒素酸化物濃度，noxsq) を調査する目的で採取され，これまでも多くの解析がなされている (Härdle et al., 2000)．

まず，目的変数 medv と説明変数間の関係を見るため，medv を基準とする変数とし，medv に対する相関係数で平行座標の順序を並べ替える．図 4 がその状態を表示したものである．medv が左端にあり，medv との相関係数が大きな順に変数が並べ替えられている．この状態で視点移動を行いながら，全体像を俯瞰してみる．2 次元の平行座標プロットで左から変数間がより直線に近い形の順に並んだように，3 次元の場合も左から順に medv とより線形性が強い変数が並ぶことになる．図 2 の操作の結果から低所得者の割合 (lstat) と 1 軒あたりの部屋数 (rm) がそれぞれ，-0.738 と 0.695 と相関が強いことがわかっているが，図 4 を見ると確かに lstat と rm が medv と線形的な動きをしていることが確認できる．

次に，このデータの解析目的である，住宅価格と大気汚染の関係を見るため，medv と noxsq について見てみる (図 13)．この図を見ると，medv が低いときに noxsq は特に大きく，medv が上がるにつれ徐々に noxsq は小さくなるが，全体としてはっきりと線形関係があるようには見えない．特に，medv の最小値から中間付近までは負の相関が見られ，中間付近から medv の最大値までは線形関係が見られないということがわかる．すなわち，大気汚染が住宅価格に影響を与える度合いが住宅価格の値の範囲によって異なるということが言える．

このような非線形関係に対し medv，noxsq 以外の変数が関係があるかどうか，あるならばどの変数か，について調べたい．そこで，次に medv と相関の強かった lstat と rm を詳細に観察してみる．再度，図 4 で lstat と rm を見てみると，どちらも部分的な線形関係があることがわかる．lstat は medv の値が小さい部分からから中央までと中央から値の大きい部分で増加傾向が変化していることがわかる．すなわち，前半部分と後半部分を比較すると，前半部分のほうが増加する傾向が強い．また rm は中央を除いた前半部分と後半部分で増加する傾向が弱いことがわかる．

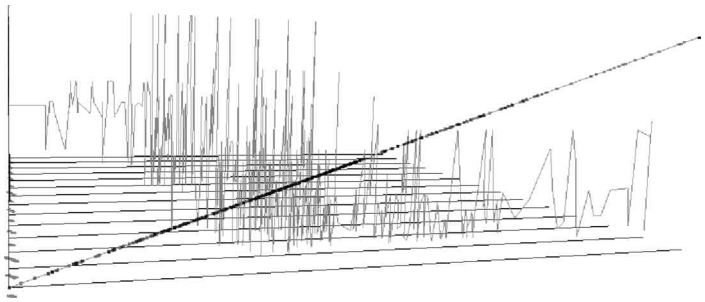

図 13. 変数 medv を基準とした場合の変数 noxsq.

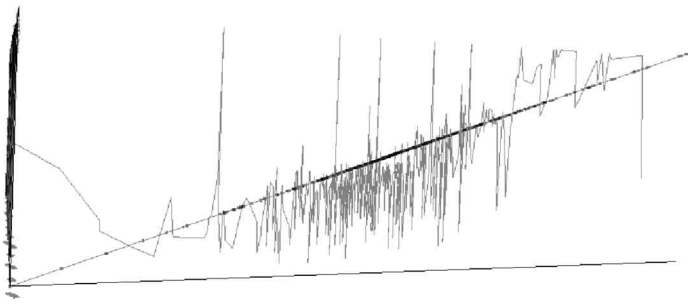

図 14. 変数 rm を基準とした場合の変数 medv.

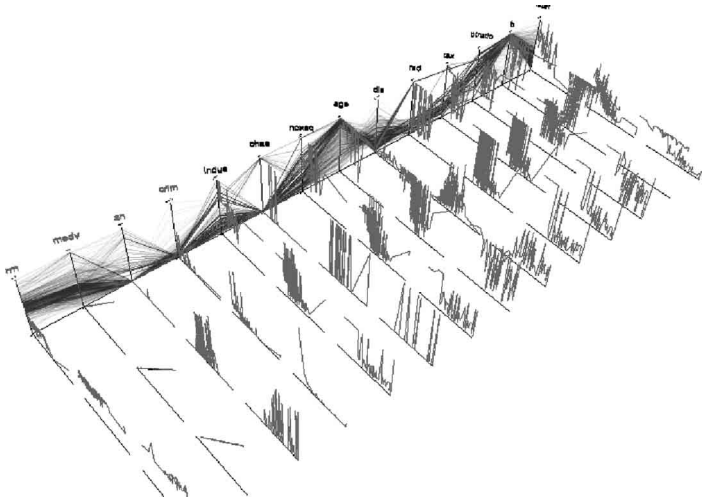

図 15. 変数 rm で条件をつけた状態の 3D PCP.

ここでまず，部屋数が小さいところと大きいところで住宅価格が異なるという考えにもとづいて，データ分割を行う際の条件を見つけるため，rm を基準とする変数として，medv との関係を見てみる(図 14)．この図を見ると，rm に全体的に強い増加傾向があることがわかるが，rm の値が小さいところと大きいところでははっきりと線形関係があるようには見えない．そこで，rm の値で条件付けを行いグループに分割してみる．このとき，rm の値を小，中，大となるよう 3 分割するが，medv との線形関係が明らかに見えるグループ(中)と，そうでないグループ(小，大)となるように分割する．rm の値 5.71 と 6.77 付近で分割し，medv を基準とした状態が図 15 である．画面の奥から値が小さいグループ，中央のグループ，大きいグループと 3 分割されている．このときのそれぞれのグループに含まれる観測値の個数は順に 79，328，98 である．この図で再度 medv と noxsq を観察してみる．rm の値が小さいグループと中央のグループでは medv に対して noxsq は減少傾向の線形関係が見られるが，大きいグループに関してはこれらと比較して明らかな線形関係は見られない．これは部屋数が多い住宅については少なくとも，大気汚染が住宅価格に与える影響は少ない傾向にあると言える．

次に違う観点から条件変数を選択する．図 4 から犯罪発生率(crim)の値の大きい場所が medv

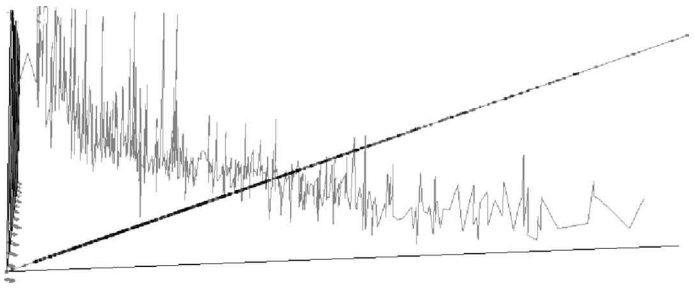

図 16. 変数 lstat を基準とした場合の変数 medv.

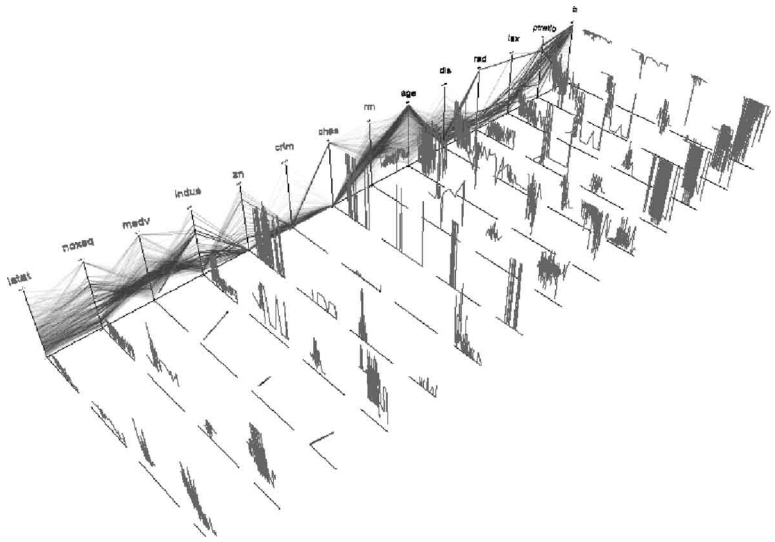

図 17. 2 変数 (lstat, noxsq) で条件をつけた状態の 3D PCP.

の値が小さいところで集中して現れることがわかっており，どのような変数が影響をしているかについて調べる．条件変数として lstat と noxsq を選ぶ．lstat は 2 節で観察したように，medv に対して，負の相関がある変数であり，ある点を境に減少傾向に変化が見られた．この境界の値を図 16 から lstat の値を 10.7 付近とし，この値で分割する．次に，medv と noxsq を強調表示した図 13 に注目し，noxsq の値が 0.52 で分割する．

この 2 変数でそれぞれ 2 つに分割することで 4 つのグループに分割される．4 つのグループはそれぞれ lstat と noxsq が，小と小，小と大，大と小，大と大というグループである．それぞれ，順に第 1 グループから第 4 グループとする．このような lstat と noxsq の条件のもとで再度，medv を基準として表示する (図 17)．この図を見ると，画面の奥に近いほうから 4 つのグループに分割されているのがわかる．そしてこの状態で，medv と各変数の関係をそれぞれのグループごとのグラフに注目しながら観察してみる．

crim の値の大きな地域はほとんど第 4 グループに集中している．このことから lstat と noxsq の条件で犯罪発生率の高い地域と低い地域にデータを分割したということがわかる．また，第

4 グループは medv との負の相関(第 1 グループから順に，0.15，0，0，−4.8)が見られる．つぎに職業紹介センターまでの距離(dis)を見てみる．4 つのグループのなかで第 4 グループのみ正の相関(第 1 グループから順に，$-0.23, -0.52, 0.61, 0.32$)が見られる．これは，lstat と noxsq が高い地域では職業紹介センターまでの距離という条件が住宅価格に比較的高い影響を与えているということがわかる．これは第 4 グループを説明する上で妥当な情報であると言える．次に非小売業の割合(indus)を見てみる．4 つのグループとも medv に対して明らかな線形関係は見られないが，第 1 グループに値の小さなものが集まり，第 4 グループに値の大きなものが集まっている．次に 25,000 平方フィート以上の宅地の保有率(zn)を見てみる．4 つのグループとも medv に対して明らかな線形関係を見ることはできないが，zn の値が大きな観測値が第 1 グループに集まっている．特に第 2 グループと，第 4 グループにはまったく見られない．これは，lstat と noxsq が低い地域が宅地面積が他と比べて大きいということで，これは第 1 グループを説明する上で妥当な情報ということがわかる．次に環状高速道路への利便性(rad)を見てみる．4 つのグループとも medv に対して線形関係は見られないが，第 1 グループと第 3 グループには値の大きなものは見られない．

このようなことから，第 1 グループは，非小売業の割合が小さく，住宅面積が大きく，高速道路から離れた地域を表している．すなわち lstat，noxsq とも小さいグループというのは高級住宅が多い地域で，特に空気のきれいな郊外と言える．また，第 4 グループは，非小売業の割合が大きく，住宅面積が小さく，高速道路に近い地域を表している．すなわち lstat，noxsq とも大きいグループというのは所得の少ない住民が多く，産業が盛んな都市の中心部と言える．そして都市の中心部では犯罪発生率が高く，その地域では犯罪発生率が住宅価格に与える影響が大きいということが言える．

このように 3D PCP を用いて，変数の区間による条件付けを行いつつ，基準変数と各変数との線形関係を見ていくことで，新しい情報を見つけることができる．

5. おわりに

多変量データ解析の初期段階では変数間の関係を視覚化することが重要である．3D PCP は，変数すべてを同時に視覚化できるので，そのような目的に適した統計グラフといえる．さらに，変数の値によってデータを分割した上で視覚化することにより，変数間の非線形関係を空間上に現れるパターンによって見つけることを可能にしている．このような変数間の関係は散布図行列では見つけられないことがある．

なお，本論文で提案した 3D PCP は筆者らが開発中の Java による統計グラフライブラリ Jasplot (http://jasp.ism.ac.jp/jasplot) の一部として実装した．Jasplot は複数の統計グラフを連携して動作させることが可能である．

今回，3 次元グラフィックを扱うため Java3D ライブラリ(Chen and Wegman, 2006)を採用した．Java3D は他の 3 次元描画ライブラリより比較的学習しやすい上，計算機能力に最適化された実行ファイルを出力することが特徴として知られている (Selman, 2002)．これまで，パーソナルコンピュータによる統計データの 3 次元処理を用いた視覚化は計算機の描画処理能力が不足し，実用的でないことがあった．しかし，このようなライブラリを用いることで，近年の計算機による 3 次元描画処理性能の大幅な向上により，3 次元視覚化ソフトウェアがパーソナルコンピュータ上で十分現実的な速度で動作することを示した．

謝　辞

本研究は部分的に情報・システム研究機構，新領域融合研究センター，機能と帰納プロジェ

クトの研究費補助を受けた．本論文の審査と編集を担当いただいた先生方には，丁寧な査読のうえに，不備な点のご指摘と有益なご意見を頂戴しました．ここに記して謝意を表します．

参 考 文 献

Barlow, N. and Stuart, L. J. (2004). Animator: A tool for the animation of coordinates, *Proceedings of the Eighth International Conference on Information Visualisation*, 725–730, IEEE Computer Society, Washington, DC.

Chen, J. X. and Wegman, E. J. (2006). *Foundations of 3d Graphics Programming*: *Using Jogl and Java3d*, Springer-Verlag, London.

Cleveland, W. S. (1993). *Visualizing Data*, Hobart Press, Summit, New Jersey.

Fanea, E., Carpendale, S. and Isenberg, T. (2005). An interactive 3D integration of parallel coordinates and star glyphs, *Proceedings of the 2005 IEEE Symposium on Information Visualization*, 149–156, IEEE Computer Society, Washington, DC.

Härdle, W., Klinke, S. and Müller, M. (2000). *Xplore-Learning Guide*, Springer-Verlag, Berlin, Heidelberg, NewYork.

Harrison, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air, *Journal of Environmental Economics and Management*, **5**, 81–102.

Inselberg, A. (1985). The plane with parallel coordinates, *The Visual Computer*, **1**, 69–91.

Johansson, J., Cooper, M. and Jern, M. (2005). 3-dimensional display for clustered multi-relational parallel coordinates, *Proceedings of the Ninth IEEE International Conference on Information Visualisation*, 188–193, IEEE Computer Society, Washington, DC.

Murrell, P. (2005). *R Graphics*, Chapman & Hall/CRC, Boca Raton, Florida.

Selman, D. (2002). *Java 3d Programming*: *A Guide to Key Concepts and Techniques*, Manning Publications, Greenwich, Connecticut.

Swayne, D. F., Buja, A. and Lang, D. T. (2004). Exploratory Visual Analysis of Graphs in GGobi, *COMPSTAT2004 Proceedings in Computational Statistics*, 477–488, Physica-Verlag, Heidelberg.

Symanzik, J. (2004). Interactive and dynamic graphics, *Handbook of Computational Statistics Concepts and Methods* (eds. J. Gentle, W. Härdle and Y. Mori), 294–336, Springer-Verlag, Berlin, Heidelberg, New York.

Theus, M. (2002). Interactive data visualization using Mondrian, *Journal of Statistical Software*, **7**, 1–9.

Wegenkittel, R., Löffelmann, H. and Gröller, E. (1997). Visualizing the behavior of higher dimensional dynamical systems, *Proceedings of the 8th Conference on Visualization '97. IEEE Visualization '97*, 119–125, IEEE Computer Society, Washington, DC.

Wegman, E. J. (1990). Hyperdimensional data analysis using parallel coordinates, *Journal of the American Statistical Association*, **85**, 664–675.

Proceedings of the Institute of Statistical Mathematics Vol. 55, No. 1, 69–83 (2007) 83

3 Dimensional Parallel Coordinate Plot

Keisuke Honda[1] and Junji Nakano[1,2]

[1]The Graduate University for Advanced Studies
[2]The Institute of Statistical Mathematics

Parallel coordinate plot (PCP) is a standard 2-dimensional (2D) graphical tool for visualizing multivariate data at a glance. We often need to use interactive operations such as highlighting and brushed highlighting to identify each observation in 2D PCP. These operations, however, are not very useful for understanding interrelations among variables at the same time.

In this paper, we propose to extend 2D PCP to one in 3-dimensional (3D) space to show relationships among variables intuitively. Our basic idea is to use the third spatial orthogonal axis to express the results of brushed highlighting of 2D PCP. We locate line segments that represent observations with respect to values of a selected reference variable in 3D space. We illustrate that rearrangements of the order and directions of axes are useful to clearly see piece-wise linear relationships between the reference variable and other variables. We also propose to divide observations according to the values of selected variables for conditioning into several groups and draw 3D PCP separately. Such 3D PCP is useful to show non-linear interaction by the variables for conditioning. We also show the usefulness of this technique by applying it to artificial and real data.

Key words: 3 dimentional data visualization, conditional graphics, interactive operation, statistical graphics.