# 多次元データの散布図表示のための次元選択手法の応用例

中林明日香† 伊藤 貴之†

† お茶の水女子大学大学院人間文化創成科学研究科 〒 112–8610 東京都文京区大塚 2–1–1
E-mail: †{g1520533,itot}@is.ocha.ac.jp

**あらまし** 有名な多次元データの可視化手法に散布図行列や平行座標法などがあるが，これらの手法では膨大な次元数を有するデータにおいて非常に大きな画面空間を必要とする問題点がある．この問題を解決するための一手段として我々は，多次元データ中の任意の 2 変数を 2 軸とする散布図の中から重要なものを単純かつ対話的なスライダー操作によって選出し，さらにその散布図を「例外点群」および「例外でない点群の包括領域」の 2 種類に分けて描画する手法を提案している．この可視化手法は，例外点をデータから削除するか否かの判断，例外でない点群のモデル化手法の検討などに有用であると考えられる．本手法を適用した新しい事例として本報告では，航空機設計の最適化過程のデータを題材にした実行例を示す．

**キーワード** 多次元データ，可視化手法，散布図

## 1 はじめに

多次元データは我々の日常生活や専門業務に幅広く存在している．このような多次元データから重要な知識を得るための方法論として，データの特徴や規則性を観察することがあげられる．その観察手段として多次元データの可視化が有効であると考えられる．

多次元データの可視化の代表的な手法に，散布図行列や平行座標法 (Parallel Coordinate Plots; 以下 PCP) があげられる．多次元データを構成する次元数を $n$ としたときに，散布図行列は全ての 2 次元ペアの散布図を作成し $n \times n$ の格子状に並べることで表現し，平行座標法は $n$ 本の平行な座標軸に変数の値をプロットしそれを折れ線で結ぶことで表現する．これらの手法は多次元データを構成する全ての次元を可視化するものであるが，膨大な次元数を有するデータにおいては非常に大きな画面空間を必要とする問題点がある．一方で，多次元データの全ての次元に興味深い特徴や規則性が見られるとは限らない．言い換えれば，多次元データの中から興味深い特徴や規則性を有する次元だけを事前選択したほうが，効率のよい可視化を実現できる場合もある．そこで近年では，多次元データから可視化する意義の高い次元だけを選択して表示する手法が多く提案されている．

我々は次元選択手法を搭載した多次元データ可視化手法として，低次元 PCP の集合で多次元データを可視化する Hidden [1] を提案した．さらに我々は，低次元 PCP の代わりに選択的な散布図集合による多次元データの可視化手法 [2] を提案してきた．この手法は具体的には以下の 2 つの処理工程から構成されるものである．

- 多次元データ中の任意の 2 変数を 2 軸とする散布図の中から重要ないくつかを，単純かつ対話的なスライダー操作によって選出する．
- 散布図に表示される点群を「例外点群」および「例外でない点群の包括領域」の 2 種類であるとして描画する．

本手法は複数の説明変数と目的関数を持ち，カテゴリ変数によってラベル付けされているようなデータに対しての適用が有効である．このような多次元データ可視化手法の有効活用が期待される場面の例として，多次元データのモデル化があげられる．多次元データを構成する数値群の中にどのようなノイズや例外値が含まれているかを理解し，適切なスクリーニング処理によってこれらを除去したのちに，どのようなモデルを適用できるかを検討する処理が必要となる場面が多い．例えば機械学習の訓練データに多次元データを利用する際に，このような工程が重要な意味を持つことが多い．このような工程にも多次元データの可視化手法が貢献できる可能性が期待される．

本報告ではこの研究の続報として，新たなデータでの応用例を紹介する．本報告の構成は以下の通りである．2 章では関連研究について，3 章では提案手法について述べる．そして 4 章で本手法の実行結果と考察について，5 章で本報告のまとめと今後の課題について述べる．

## 2 関連研究

### 2.1 次元選択を用いた多次元データの可視化

多次元データの中から重要な部分だけを可視化するためのアプローチとして，可視化する意義の高い低次元部分空間を事前に抽出する手法は従来から数多く提案されている．例として，多次元データから所定の基準を満たす複数の 2 次元ペアの散布図を生成し，各散布図間の類似度距離に基づいて配置する手法 [3] や，所定の基準を満たす次元間の低次元 PCP を生成し，各 PCP 間の次元の共有率から算出される類似度距離に基づいて配置する手法 [4] などがある．しかしこれらの手法による可視化結果は固定的なものであり，PCP や散布図の表示数を対話的に調節することができなかった．

この問題点を解決する多次元データ可視化手法として Itoh らは Hidden [1] を発表した．Hidden [1] は画面右部の次元散布

図上を対話的に操作することによって選択される低次元部分空間群を，画面左部で複数の PCP によって表示する．図 1 に Hidden による可視化の例を示す．また Hidden を拡張し PCP と散布図を併用して可視化した手法 [5] も発表されている．この手法では原則として PCP で多次元データを可視化しつつ，PCP では視認しにくい数値分布を有する 2 軸のみに対して散布図を適用して表示する．

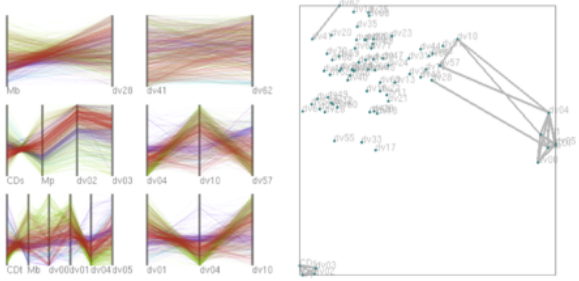

図 1　Hidden による可視化画面 [1]

### 2.2　散布図集合による多次元データの可視化

散布図を用いて多次元データを可視化する手法として，Wilkinson らの手法 [6] や Dang らの手法 [7] が挙げられる．Wilkinson らの手法では，多次元データの各 2 次元ペアから生成される散布図の形状などから Scagnostics と呼ばれる 9 種類の特徴を定量評価し，特定の傾向を持った散布図を生成する 2 変数を推薦する．Dang らの手法では，散布図行列から Scagnostics の基準により各散布図を特徴づけてクラスタリングし，リーダーとなる散布図を選出して類似するものを近くに配置する．これらの手法では，膨大な次元数の多次元データでも非常に大きい画面空間を使うことなく可視化することができる．我々の手法では 3 章にて後述する基準で散布図を選出しているが，Scagnostics を適用することも可能である．

### 2.3　密度を考慮した散布図による可視化

散布図を塗りつぶすように表現する手法の例として，Continuous Scatterplots [8] は，$n$ 次元の入力データセット上で定義された任意の密度を考慮して，この密度値を次元散布図にマッピングする手法である．この手法は散布図のような統計的可視化と流体の可視化のような科学的可視化を組み合わせたものである．我々の手法も散布図を連続的に描画するものであるが，密度を直接考慮して散布図を描画するわけではない．

### 2.4　点群を囲むように描画する可視化

Schreck らは 2 次元空間上の点群を囲むように描画する手法を提案している [9] [10]．これらの手法では，多次元データを主成分分析などの手法により 2 次元空間で表現し，クラスごとにデータ点を凸包やよりコンパクトな形状で囲むように描画することによって各クラスの分布を視認しやすくしている．我々の手法では 3 章にて後述するアルゴリズムで点群を囲む描画を実現しているが，Schreck らの手法の囲み方で描画することもできる．

## 3　選択的な散布図集合による多次元データの可視化

本章では我々が提案した多次元データ可視化手法 [2] について紹介する．1 章でも述べた通り，この手法では，多次元データ中の任意の 2 変数を 2 軸とする散布図の中から重要なものを選出し，さらにその散布図を構成する点群を「例外点群」および「例外でない点群の包括領域」の 2 種類であるとして描画する．

現時点での我々の実装では，散布図の選出基準には Itoh らの手法 [1] と同じく相関係数またはエントロピーによる基準を採用している．相関係数による基準を適用した際には，各散布図を生成する 2 次元間の相関係数を計算し，その絶対値の大きい散布図を優先的に表示する．任意の 2 次元 $d_1$, $d_2$ を与えられたとき，間の相関係数の絶対値 $d_{d_1,d_2}$ はスピアマンの順位相関係数 $f_c(d_1, d_2)$ を用いて以下の式のように表せる．

$$d_{d_1,d_2} = |1.0 - f_c(d_1, d_2)| \tag{1}$$

エントロピーによる基準を適用した際には，多次元データ中のカテゴリ型変数が各個体のラベルに相当するとみなして，点群がラベルごとによく分離されている散布図を優先的に表示する．具体的には以下の式より各次元ペアのエントロピー $H$ を算出する．

$$H(d_1, d_2) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} p(y_i = c|x_i^{d_1,d_2}) \log p(y_i = c|x_i^{d_1,d_2}) \tag{2}$$

ここで入力データセット $Ds$ を $Ds = \{x_1, \ldots, x_N\}$ としたとき，$x_i$ は $i$ 番目のプロット，$y_i$ は $i$ 番目のプロットに割り当てられたラベルを表し，$N$ と $C$ はそれぞれプロットの数とラベルの数，$p(y_i = c|x_i^{d_1,d_2})$ は $c$ 番目のクラス $Y_c$ が $i$ 番目のプロット $x_i$ に割り当てられている確率を表す．また $x_i^{d_1,d_2}$ はプロット $x_i$ について次元 $d_1, d_2$ を取り出してできる 2 次元ベクトルを表す．この式によって算出されるエントロピー $H$ は，次元 $d_1$ と $d_2$ によって生成される散布図において，クラスラベルが全体的にどの程度分離されているかを示すものである．

現時点での我々の実装では，「例外点群の抽出」および「例外でない点群の包括領域の生成」に Delaunay 三角分割法を利用している．Delaunay 三角分割法は与えられた点群を連結して三角メッシュを生成する手法であり，三角メッシュを構成する三角形の最小角度が最大になるように三角メッシュを生成するものである．本手法では，各散布図に対して，散布図中の全ての点群を包括する大きな四角形を作成し三角形に分割し，散布図中の点群を 1 つずつ追加して頂点として連結していくことで三角メッシュを逐次的に更新し，全ての点群を追加したら最初に作成した大きな四角形とその頂点に連結される辺を削除する，というインクリメンタルなアルゴリズムを採用している．

このような処理によって生成された三角メッシュから，ユーザ指定の閾値を超える長さの辺を削除することで，図 2 のようにどの点群とも連結されていない点を例外点として抽出する．ユーザによる対話操作で閾値を調節することで，例外点と判定された点の数を調節できる．そして，例外点以外の点で構成される三角形群の領域境界を構成する辺のみを濃い色で描画し，

三角メッシュを薄い色で塗りつぶすことによって，点群の包括領域を表示する．

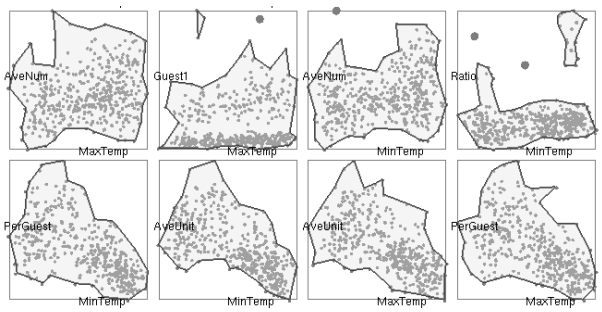


図 2　散布図を「例外点群」と「例外でない点群の包括領域」の 2 種類として描画した例

また，図 3 のようにユーザが入力データセットから任意の点を削除し散布図を再描画することもできる．図の左画像の右上の例外を削除すると，右画像のように縦軸の縮尺が更新され，例外でない点群の領域も再描画される．このように離れた位置にある点を削除することで例外でない点群の領域を拡大表示できる．これにより，それまでは発見しにくかった新しい特徴や傾向，あるいは新しく例外となり得る点を発見できることが期待される．

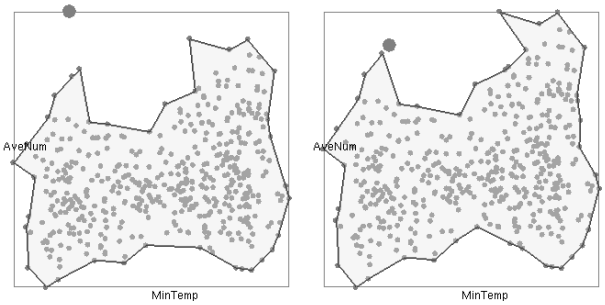


図 3　例外削除の例 (左は例外削除前，右は例外削除後)

## 4　実　行　例

本報告では小売店の気象と売上の関係のデータ，および新しい適用事例として航空機設計の最適化過程のデータを本手法に適用した可視化の実行例を示す．

### 4.1　小売店の気象と売上の関係のデータの実行結果

本節では，2016 年 5 月 1 日から 2017 年 7 月 31 日までの 457 日間のアパレルの小売店における各日の来客数や売上と，その各日の気象値との関係のデータを題材にして，本手法を用いた可視化の実行例を示す．データ中の説明変数 (気象の数値・横軸) と目的関数 (売上の数値・縦軸) の対応表を表 1 に示す．なお本章で用いるデータは現実のデータに乱数を加算したものであり，現実の数値をそのまま可視化したわけではない点に注意されたい．

図 4 では点群の一属性 (平日) が他方の属性 (休日) を内包するような形状になっている数値分布の例である．この結果から，客単価や平均買上商品単価は休日よりも平日の方がばらついていることがわかる．

図 5 は 2 月と 7 月と 11 月の買上率の数値分布を示している．2 月上旬と 7 月下旬のみ突出して買上率が高い期間があり，これは売り尽くしセールなどの特殊なイベントのために，ウィンドウショッピングとして来店した人よりも，最初から商品を購入するつもりで来店する人が多かった可能性が考えられる．また 11 月中に 1 日だけ特に買上率の高い日があることが読み取れる．

表 1　データの説明変数と目的関数の対応表

| 説明変数 (気象数値) | | 目的関数 (売上数値) | |
|---|---|---|---|
| MinTemp | 最低気温 | Revenue | 売上 |
| MaxTemp | 最大気温 | Guest1 | 購入人数 |
| SumRain | 降水量 | Guest2 | 来客人数 |
| SumSnow | 降雪量 | Ratio | 買上率 |
| SumSnowC | 積雪量 | PerGuest | 客単価 |
| SumSunTime | 日照時間 | AveUnit | 平均買上商品単価 |
| MaxWind | 最大風速 | AveNum | 平均買上点数 |

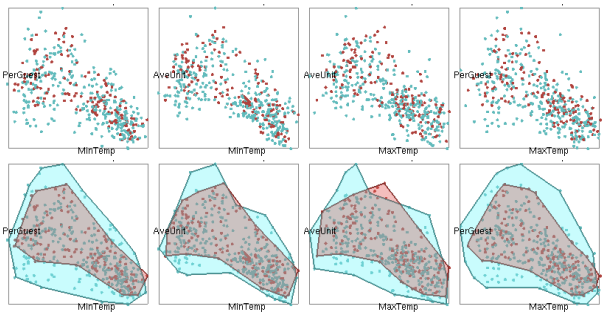


図 4　ある属性が別の属性を内包するような数値分布になっている例 (水色は平日，赤色は休日を示しており，上は包括領域を囲む前，下は囲んだものである)

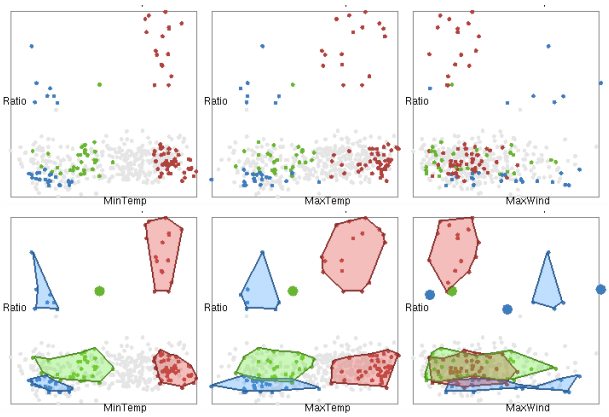


図 5　2 月と 7 月と 11 月の買上率の数値分布 (青色は 2 月，赤色は 7 月，緑色は 11 月，灰色はその他の月を示しており，上は包括領域を囲む前，下は囲んだもの)

### 4.2 航空機設計の最適化過程データの実行結果

我々は航空機の翼形状設計の最適化過程を可視化した．この事例では 72 個の説明変数により翼形状を設計し，流体力学シミュレーションにより 4 個の目的関数を算出した．この処理を多目的遺伝的アルゴリズムによって反復することで 776 個のパレート解を得た [11]．これを 76 次元ベクトルを構成する 776 個の点群として可視化した．

本章では説明変数を $dv00 \sim dv71$ と記述する．この中でも以下の 6 種類の説明変数は最適解の発見に特に重要な説明変数であることが知られている．

- $dv00$, $dv01$: 内翼および外翼のスパン長
- $dv02$, $dv03$: 後進角
- $dv04$, $dv05$: 翼根の翼弦長

他の説明変数には以下が含まれる．

- $dv06 \sim dv25$: 翼の反りに関する変数
- $dv26 \sim dv32$: 翼の捻りに関する変数
- $dv33 \sim dv71$: 翼の厚さに関する変数

4 個の目的関数は以下のとおりである．

- $CD_t$: 遷音速巡航の抵抗係数
- $CD_s$: 超音速巡航の抵抗係数
- $M_b$: 超音速巡航時の翼根にかかる曲げモーメント
- $M_p$: 翼先端部にかかる捻りモーメント

図 6 は，航空機設計の説明変数と目的関数との組み合わせの中で，特に相関の高い散布図を構成する次元から生成される 24 組の散布図を表示した例である．この中でも特に，説明変数 $dv02$,$dv03$ と目的関数 $Mp$,$CDs$ との間には非常に高い相関がみられる．また，説明変数 $dv00$,$dv01$,$dv04$,$dv05$ と目的関数 $Mb$,$CDt$ との間にも高い相関がみられる．これらの相関は航空機設計の分野では既に有名な知見であり，また我々による先行研究 [1] でもこれらの相関を可視化している．一方で，我々の先行研究が採用した PCP では，どのように両変数が相関しているか，外れ値はどのように分布しているか，といった点を視認するのは容易ではない．本手法を用いて散布図の集合で表示することで，これらの非常に高い相関を有する変数間は線形に近い (ただし厳密に線形ではない) 相関を有し，非常に少ない外れ値を有することを容易に観察できる．同時に，説明変数 $dv02$,$dv03$ と目的関数 $Mb$,$CDt$，あるいは説明変数 $dv00$,$dv01$,$dv04$,$dv05$ と目的関数 $Mp$,$CDs$ との間には相関がみられないことも同時に観察できる．

図 7 は，航空機設計の説明変数と目的関数との組み合わせの中でも弱い相関をもつ 14 組を散布図として表示した例である．この図では説明変数 $dv16$, $dv17$, $dv21$, $dv33$, $dv60$ が登場するが，航空機設計の研究者いわく，説明変数 $dv06$ 以降と目的関数の関係はシミュレーション実行当時はまだあまり研究されていなかったとのことである．このような可視化により，まだあまり研究されていない説明変数のシミュレーションおよび最適化計算への寄与について解明できる可能性があると考えられる．

例えば説明変数 $dv16$,$dv17$ に着目すると，小さめの値 (散布図の左側) に点群が集中していること，また大きめの値 (散布図の右側) では目的関数の値が一定の範囲に集中していることがわかる．このことから，$dv16$ および $dv17$ は値を小さくしたほうがパレート解を導きやすいか，あるいは最適化計算の過程で小さめの値ばかりが探索されてきたかのいずれかが示唆される．前者が該当する場合には，$dv16$ および $dv17$ の値の範囲を最初から小さめに設定したほうが効率よく多数のパレート解を導出できる可能性が高い．後者が該当する場合には，$dv16$ および $dv17$ の大きめの値も探索するように最適化計算の過程を調整することが望ましいかもしれない．説明変数 $dv21$, $dv33$, $dv60$ についても同様に，やや小さめの値 (散布図の中央よりやや左) に点群が集中していることから，$dv16$ および $dv17$ と同様な議論が可能である．

説明変数 $dv00 \sim dv05$ は各目的関数と非常に高い相関が見

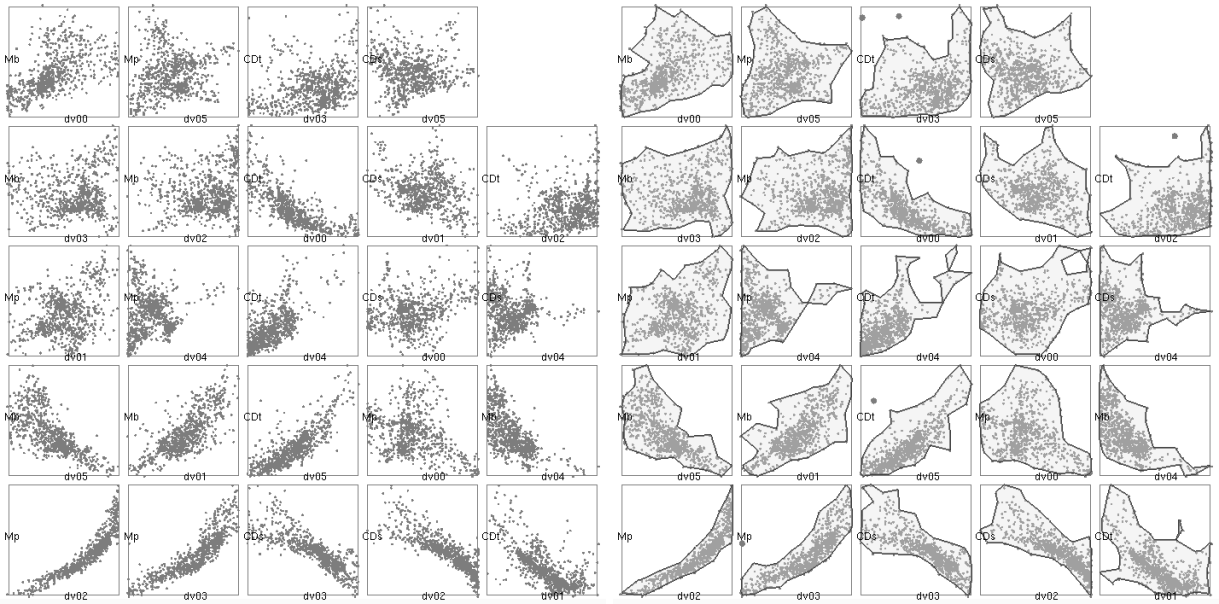

図 6　航空機設計の説明変数と目的関数の間で最も相関の高い組み合わせを集めた散布図群

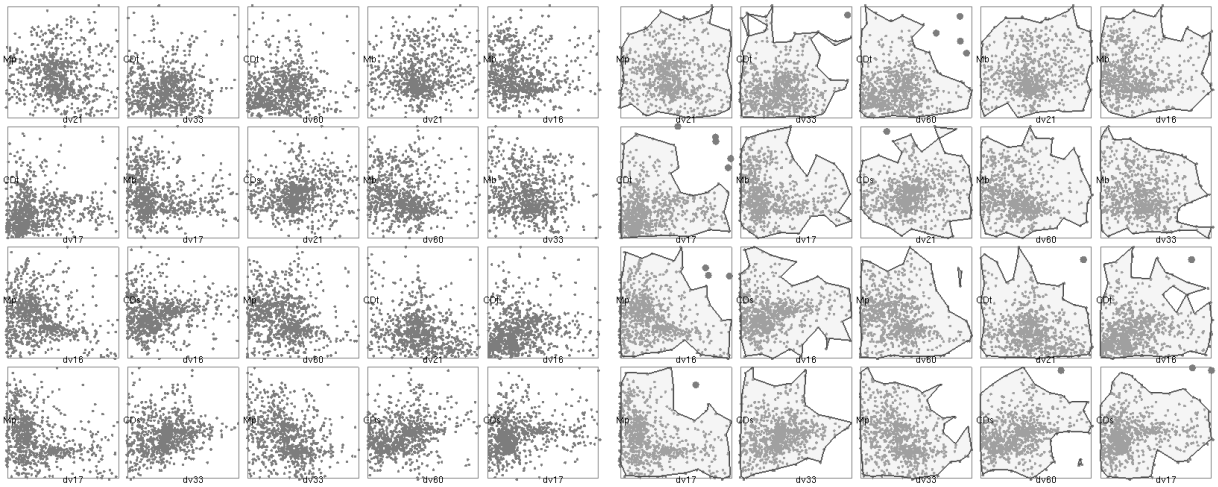

図 7　航空機設計の説明変数と目的関数の間で弱い相関をもつ組み合わせを集めた散布図群

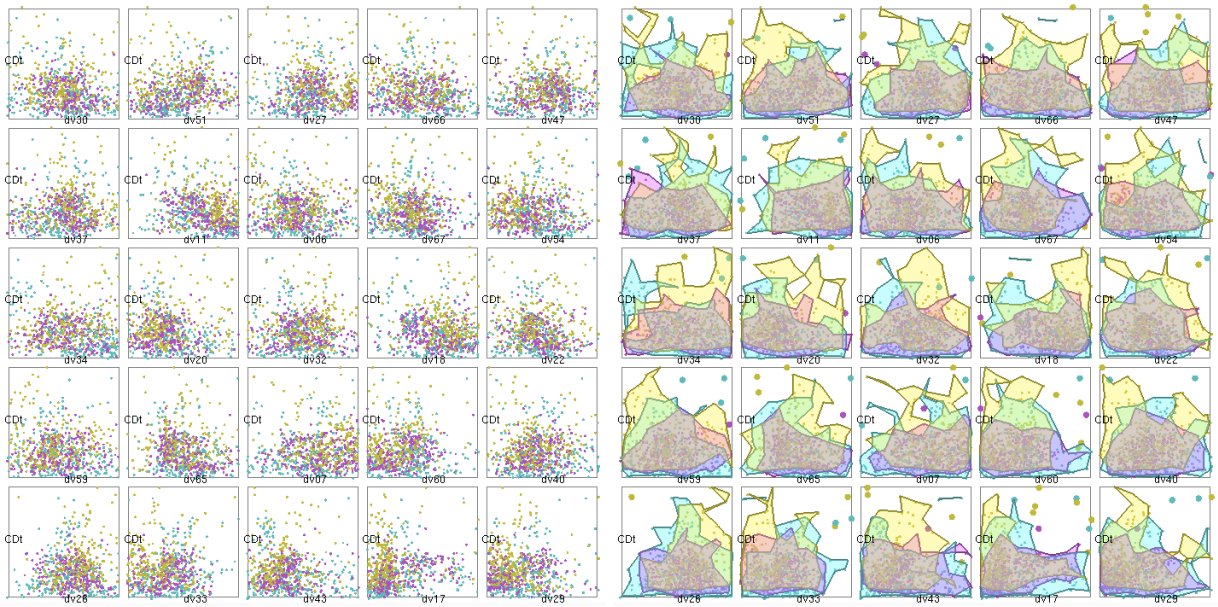

図 8　$dv00 \sim dv05$ を除く説明変数群と目的関数 $CDt$ の関係を示した散布図群

られ，かつこれは既知であることから，これらを散布図の横軸または縦軸に割り当てて可視化する必要はさほどないと判断される状況も考えられる．そこで，これらの 6 つの説明変数をもとにカテゴリ変数を作成し，そのいずれかで散布図に色をつけて可視化した．図 8 は $dv00 \sim dv05$ を除く説明変数群と目的関数 $CDt$ の関係を示したものであり，$dv03$ の値が小さい方から 259 個のパレート解の点を水色，大きい方から 259 個のパレート解の点を黄色，その他のパレート解の点を紫色で描いている．$dv33$, $dv43$, $dv18$, $dv51$, $dv66$ を横軸にした散布図において，上部にごく薄い水色の三角形がある．そして，この三角形を構成する 3 点は同じ点であった．このことから $dv03$ が小さめの値においては $CDt$ の値が高くなることは稀であるが，類似する解が 3 つあることから最適化計算の過程で $CDt$ が小さくなるような値ばかりが探索されてきた可能性が考えられる．この場合に，$dv03$ が小さめの値で $CDt$ の値が高くなるような解を探索するように最適化計算の過程を調整することが望ましいかもしれない．

図 9 は $dv00 \sim dv05$ を除く説明変数群と目的関数 $Mp$,$CDs$ の関係を示したものであり，$dv00$ の値が小さい方から 259 個のパレート解の点を水色，大きい方から 259 個のパレート解の点を黄色，その他のパレート解の点を灰色で描いている．いくつかの散布図において，水色の点群が右下がりかつ黄色の点群が右上がり (または水色の点群が右上がりかつ黄色の点群が右下がり) になっている傾向が見られる．この傾向に沿うように最適化計算の過程を調整することで，パレート解を導きやすくなる可能性があると考えられる．

## 5　まとめと今後の課題

本報告では，散布図の選択と描画に関する我々の提案手法を紹介し，その新しい適用事例として航空機設計の最適化過程のデータを可視化した事例を示した．我々の可視化手法では，多次元データ中の任意の 2 変数を 2 軸とする散布図の中から重要と思われる散布図を選出し，さらにその散布図を構成する点群

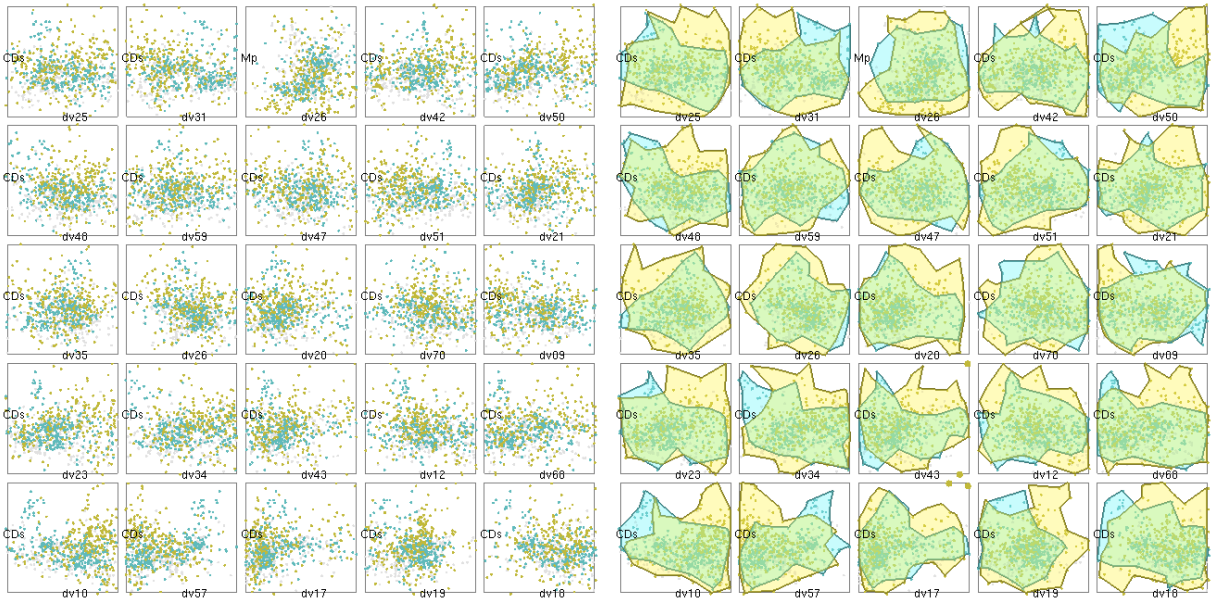


図 9　$dv00 \sim dv05$ を除く説明変数群と目的関数 $Mp$,$CDs$ の関係を示した散布図群

を「例外点群」および「例外でない点群の包括領域」の 2 種類であるとして描画する手法を提案している．我々の可視化手法は，例外点をデータから削除するか否かの判断や，例外でない点群のモデル化手法の検討などに有用であると考えられる．

今後の課題として，散布図の選出基準について引き続き検討したい．現時点で実装している相関係数やエントロピーにもとづいた散布図選出手法では，我々が重要であると主観的に判断しているような散布図が選出されないことがある．そこで新しい散布図選出基準を実装することで，このような散布図が選出されるようにしたい．

また，例外の検出方法についても引き続き検討したい．現在の実装では，Delaunay 三角メッシュを利用して，個々の散布図において他の点群から距離がある点を例外とみなしている．多次元データの例外の抽出方法に関しては多数の研究 [12] [13] [14] が発表されている．これらを利用した例外の抽出についても再考したい．

そしてこれらの機能を実装した後に，より多様なデータセットを本手法に適用し，さらに汎用性に富んだ実装になるように開発を進めたい．

## 謝　　辞

小売店の気象と売上に関するデータセットを提供して頂いた株式会社 ABEJA 様に感謝いたします．航空機設計の最適化過程のデータセットを提供して頂いた東北大学流体科学研究所大林茂教授に感謝いたします．

## 文　　献

[1] Takayuki Itoh, Ashnil Kumar, Karsten Klein, and Jinman Kim, “ High-Dimensional Data Visualization by Interactive Construction of Low-Dimensional Parallel Coordinate Plots,” Journal of Visual Languages and Computing, Vol. 43, pp. 1–13, 2017.

[2] Asuka Nakabayashi, Takayuki Itoh, “ A Technique for Selection and Drawing of Scatterplots for Multi-Dimensional Data Visualization,” Proceedings of 23rd International Conference on Information Visualisation (IV2019), pp. 62–67, 2019.

[3] Yunzhu Zheng, Haruka Suematsu, Takayuki Itoh, Ryohei Fujimaki, Satoshi Morinaga, and Yoshinobu Kawahara, “ Scatterplot layout for high-dimensional data visualization,” Journal of Visualization, Vol. 18, No. 1, pp. 111–119, 2015.

[4] Haruka Suematsu, Yunzhu Zheng, Takayuki Itoh, Ryohei Fujimaki, Satoshi Morinaga, and Yoshinobu Kawahara, “ Arrangement of Low-Dimensional Parallel Coordinate Plots for High-Dimensional Data Visualization,” Proceedings of 17th International Conference on Information Visualisation (IV2013), pp. 59–65, 2013.

[5] Ayaka Watanabe, Takayuki Itoh, Masahiro Kanazaki, and Kazuhisa Chiba, “ A Scatterplots Selection Technique for Multi-Dimensional Data Visualization Combining with Parallel Coordinate Plots,” Proceedings of 21st International Conference on Information Visualisation (IV2017), pp. 78–83, 2017.

[6] Leland Wilkinson, Anushka Anand, and Robert Grossman, “ Graph-Theoretic Scagnostics,” Proceedings of IEEE Symposium on Information Visualization, pp. 157–164, 2005.

[7] Dang Tuan Nhon and Leland Wilkinson, “ ScagExplorer: Exploring Scatterplots by Their Scagnostics,” Proceedings of IEEE Pacific Visualization Symposium (PacificVis 2014), pp. 73–80, 2014.

[8] Sven Bachthaler and Daniel Weiskopf, “ Continuous Scatterplots,” IEEE Transactions on Visualization and Computer Graphics, Vol. 14, No. 6, pp. 1428–1435, 2008.

[9] Tobias Schreck and C. Panse, “ A new metaphor for projection-based visual analysis and data exploration,” Proceedings of IS&T/SPIE Conference on Visualization and Data Analysis, 2007.

[10] Tobias Schreck, Michael Schuessler, Katja Worm, and Frank Zeilfelder, “ Butterfly plots for visual analysis of large point cloud data,” Proceedings of the 16th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG’08), pp. 33–40, 2008.

[11] Daisuke Sasaki, Shigeru Obayashi, and Kazuhiro Nakahashi, “ Navier-Stokes Optimization of Supersonic Wings with Four Objectives Using Evolutionary Algorithm,” Jour-

nal of Aircraft, Vol. 39, No. 4, pp. 621–629, 2002.

[12] Denis Cousineau and Sylvain Chartier, " Outliers detection and treatment: a review," International Journal of Psychological Research, Vol. 3, No. 1, pp. 58–67, 2010.

[13] Chang-Tien Lu, Dechang Chen, and Yufeng Kou, " Algorithms for Spatial Outlier Detection," Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03), 2003.

[14] Kay I. Penny and Ian T. Jolliffe, " A comparison of multivariate outlier detection methods for clinical laboratory safety data," Journal of the Royal Statistical Society: Series D (The Statistician), Vol. 50, No. 3, pp. 295–308, 2001.