応用統計学　Vol. 33, No. 3 (2004), 225–241

研究論文

多次元空間への相対射影追跡法について

北海道大学大学院工学研究科　弘　新太郎
北海道大学情報基盤センター　小宮　由里子
北海道大学情報基盤センター　南　弘征
北海道大学情報基盤センター　水田　正弘

要　旨　近年，ゲノムデータやPOSデータのような変量の多いデータが増加し，そのような高次元データに対する解析手法の必要性が増している．一般に，データ解析において，データが高次元になるほど，有益な情報を抽出することは困難になる．そこで，多変量データ解析では解釈が容易な低次元空間にデータを次元縮小し，有益な情報を引き出す手法が数多く研究されている．なかでも，射影追跡法(Friedman and Tukey, 1974)は，興味深い構造が現れる低次元空間を探索する有効な次元縮小法である．従来の射影追跡法では興味深さを数値化する射影指標がいくつか提案されているが，その提案のすべてにおいて，興味深い構造を正規分布から最も離れている分布と定義しているため，正規分布を基準としないような興味深い構造の探索は難しい．

これに対して，解析者が参照とする標本を定義して，その標本の分布から最も離れている分布が現れる射影方向を探索する相対射影追跡法がMizuta(2002)によって提案されており，用いられる射影指標として，Area射影指標(弘・小宮・南・水田，2003)が既に提案されている．しかし，この指標は2次元以上の空間へ射影した場合に興味深さを測ることができない．そこで本論文では，2次元以上の空間へ射影する場合に対応したArea相対射影指標を作成する．また，従来の射影追跡法で使用されるHallの射影指標を相対射影追跡法を行うための射影指標に拡張し，新たなHall Type相対射影指標を作成する．この2つの相対射影指標を用いて高次元データを2次元空間へ次元縮小し，興味深い射影方向空間が得られるかを比較検討する．

1. はじめに

射影追跡法(Friedman and Tukey, 1974)は，高次元空間上の標本を低次元空間へ射影し，その射影が最も興味深い分布となる空間を探索する手法である．線形射影を用いた次元縮小法は，基本的な多変量解析手法であり，その例としてデータの分散が大きい低次元空間を探索する主成分分析等が挙げられる．

従来の射影追跡法では，射影の分布が正規分布から最も離れている構造を興味深いと仮定して探索している．しかし，興味深い構造とは個々のデータや解析目的によって異なるため，常に正規分布と比較して特徴的な構造を探索するのが適切と言えない．そこで，興味のない構造を一意

的に正規分布とは仮定しない相対射影追跡法(Mizuta, 2002)が提案された．この手法では興味のない構造を持っていると考える標本を事前に設定し，そこから最も離れた構造を有する射影方向を探索する．事前に設定された参照とする標本と，解析対象とする標本の分布間の離れ具合を示す指標は，従来の射影追跡法で用いられる射影指標と区別して相対射影指標と呼ばれ，その指標として Area 相対射影指標(弘・小宮・南・水田，2003)が提案されている．しかし，これは 1 次元空間への射影に関する相対射影指標のみが扱われており，2 次元以上の空間へ次元縮小する指標は作成されていない．実際に多変量データを解析する際には，1 次元空間への射影のみでデータの構造を捉えることは困難であり，2 次元以上の空間へ射影することで，より興味深い構造を捉えることが期待される．

本論文では，2 次元以上の空間における興味深い構造を測ることができるように，Area 相対射影指標を拡張する．また，Area 相対射影指標の他に，従来の射影追跡法で用いられている Hall の射影指標を参考にして，標本の密度関数間の距離によって興味深さを測る Hall Type 相対射影指標を提案する．この指標についても 2 次元以上，つまり，k 次元へ次元縮小する場合に使用可能な指標を作成する．さらに，これらの指標を用いた相対射影追跡法を人工データ及び実データに適用し，特徴的な構造を検出できるかを評価することにより，その有効性について考察する．

2. 従来の射影指標

興味深い構造を探索するためには興味深さを定式化する必要があり，定式化されたものを射影指標と呼ぶ．従来の射影指標は，興味のない構造を正規分布と定めて，そこからの離れ具合を定式化したものである．従来の代表的な指標として Friedman の射影指標(Friedman, 1987)，Hall の射影指標(Hall, 1989)が提案されている．本論文では Hall の射影指標の考え方に基づいて新たな相対射影指標を作成するので，これについて簡単に説明する．

Hall の射影指標

Hall の射影指標(Hall, 1989)は，解析の対象である標本の分布と正規分布の密度関数の差の 2 乗を測っている．標本を球化し，標本の持つ確率変数を p 次元ベクトル $\boldsymbol{Z}$，および p 次元射影方向ベクトルを $\boldsymbol{\alpha}$ とすれば，ベクトル $\boldsymbol{\alpha}$ 方向への $\boldsymbol{Z}$ の射影を表す確率変数 X は

$$X = \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Z}$$

と表すことができる．ただし，$\boldsymbol{\alpha}$ には $\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha} = 1$ という条件を課す．この X の確率密度関数を f_{α} とすると，1 次元の Hall の射影指標は

$$J \equiv \int_{-\infty}^{\infty} \{f_{\boldsymbol{\alpha}}(u) - \phi(u)\}^2 du$$

と定義される．ここで，$\phi(\cdot)$ は標準正規分布の密度関数である．この定義に従って，密度関数を Hermite 関数による直交関数展開で近似して射影指標を求めると，1 次元空間に射影する場合の Hall の射影指標は

$$I(\boldsymbol{\alpha}) = [\theta_0(\boldsymbol{\alpha}) - 2^{-1/2}\pi^{-1/4}]^2 + \sum_{j=1}^{J} {\theta_j}^2(\boldsymbol{\alpha})$$

となる．ここで，$\boldsymbol{Z}$ の実現値を $\boldsymbol{z}_i$ $(i = 1, \ldots, n)$ とすれば，

$$\theta_j(\boldsymbol{\alpha}) \;=\; n^{-1} \sum_{i=1}^{n} P_j(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{z}_i)\phi(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{z}_i), \qquad P_j(z) \;=\; \left(\frac{2}{j!}\right)^{1/2} \pi^{1/4} H_j(2^{1/2}z)$$

である．$H_j(\cdot)$ は j 次の Hermite 多項式で，

$$H_j(x) \;=\; (-1)^j\{\phi^2(x)\}^{-1}\frac{d^j}{dx^j}\phi^2(x)$$

と表される．

また，2 次元空間へ射影する場合の射影指標は

$$I(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) \;=\; \sum_{j=0}^{q}\sum_{k=0}^{q-j}\left\{\frac{1}{n}\sum_{i=1}^{n} h_j(\boldsymbol{\alpha}_1{}^{\mathrm{T}}\boldsymbol{z}_i)h_k(\boldsymbol{\alpha}_2{}^{\mathrm{T}}\boldsymbol{z}_i)\right\}^2 - \pi^{-1/2}\left\{\frac{1}{n}\sum_{i=1}^{n} h_0(\boldsymbol{\alpha}_1{}^{\mathrm{T}}\boldsymbol{z}_i)h_0(\boldsymbol{\alpha}_2{}^{\mathrm{T}}\boldsymbol{z}_i)\right\} + (4\pi)^{-1}$$

となる．ここで，

$$h_j(u) \;=\; (j!)^{-1/2}\pi^{1/4}2^{-(j-1)/2}H_j(u)\phi(u) \quad (-\infty < u < \infty)$$

である．

3. 相対射影指標

相対射影追跡法において従来の射影追跡法と異なる点は使用する射影指標のみである．従って，本論文では相対射影追跡法で用いられる相対射影指標について議論する．特に本章では，既に提案されている Area 相対射影指標と新たに作成した Hall Type 相対射影指標について説明する．また，この両指標について，2 次元以上の多次元空間へ射影を行った場合に，興味深さを測ることができるように拡張し，k 次元空間への相対射影指標を提案する．

Area 相対射影指標

Area 相対射影指標は解析の対象とする標本の分布と，参照とする標本の分布の経験分布関数の差の面積を指標として計算する．既に提案されている 1 次元空間へ次元縮小する場合の Area 相対射影指標を数式で表すと以下のようになる．

$$I_A(\boldsymbol{\alpha}) \;=\; \int |F_n(x) - G_m(x)|\; dx\,.$$

ここで，$F_n(x)$ は解析対象とする標本データ $\boldsymbol{z}$（p 次元 × n 個）を p 次元射影方向ベクトル $\boldsymbol{\alpha}$ で射影したデータ $(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{z})$ に対する経験分布関数，$G_m(x)$ は参照とする標本データ $\boldsymbol{w}$（p 次元 × m 個）を同じ射影方向ベクトル $\boldsymbol{\alpha}$ で射影したデータ $(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{w})$ に対する経験分布関数とする．n 個のデータに対する経験分布関数は次式で定義される．

$$\hat{F}_n(x) \;=\; \frac{1}{n}\sum_{i=1}^{n} H(x - x_i)\,.$$

ここで，

$$H(x) = \begin{cases} 1 & (x \geq 0) \\ 0 & (x < 0) \end{cases},$$

とする．

次に 2 次元へ次元縮小する場合の Area 相対射影指標を考える．2 次元の Area 相対射影指標は，各標本を射影方向ベクトル $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2$ で射影したデータから 2 次元経験分布関数をそれぞれ作成し，その関数の差の体積を測ればよい．

データ数が n の 2 次元経験分布関数は

$$\hat{F}_n(x_1, x_2) = \frac{1}{n} \sum_{i=1}^{n} H(x_1 - x_{1i})\, H(x_2 - x_{2i})$$

で表される．これを用いると，2 次元の Area 相対射影指標は

$$I_A(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) = \int\int \left| F_n^{(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)}(x_1, x_2) - G_m^{(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)}(x_1, x_2) \right| dx_1 dx_2$$

となる．関数 $F_n^{(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)}(x_1, x_2)$, $G_m^{(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)}(x_1, x_2)$ は 1 次元と同様に，解析対象とするデータ $\boldsymbol{z}$，及び参照とするデータ $\boldsymbol{w}$ をそれぞれ射影方向 $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2$ で射影した時のデータを 2 次元の経験分布関数にしたものである．

k 次元への Area 相対射影指標の拡張は容易であり，n 個のデータに対する k 次元経験分布関数を

$$\hat{F}_n(x_1, x_2, \cdots, x_k) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \prod_{d=1}^{k} H(x_d - x_{di}) \right\}$$

と定義すれば，k 次元の Area 相対射影指標は

$$I_A(\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_k) = \int \ldots \int \left| F_n^{(\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_k)}(x_1, \cdots, x_k) - G_m^{(\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_k)}(x_1, \cdots, x_k) \right| dx_1 \cdots dx_k$$

となる．

Hall Type 相対射影指標

新たな相対射影指標として Hall Type 相対射影指標を提案する．従来の Hall の指標は密度関数間の距離の差の 2 乗を測っている．つまり，解析対象とする標本 $\boldsymbol{z}_i, i = 1, \ldots, n$ を射影ベクトル $\boldsymbol{\alpha}$ で 1 次元空間に射影したときの密度関数を $f_{\boldsymbol{\alpha}}(x)$ とすると，2 節で説明したように，1 次元の Hall の指標は

$$J \equiv \int_{-\infty}^{\infty} \{f_{\boldsymbol{\alpha}}(x) - \phi(x)\}^2 dx$$

と定義される．参照とする標本を $\boldsymbol{w}_j$, $j = 1, \ldots, m$ とし，この定義を相対射影指標へと拡張する．参照とする標本を射影ベクトル $\boldsymbol{\alpha}$ で 1 次元空間に射影したときの密度関数を $g_{\boldsymbol{\alpha}}(x)$ とすれば，1 次元の Hall Type 相対射影指標は

$$\begin{aligned} I(\boldsymbol{\alpha}) &= \int_{-\infty}^{\infty} \{f_{\boldsymbol{\alpha}}(x) - g_{\boldsymbol{\alpha}}(x)\}^2 dx \\ &= \int_{-\infty}^{\infty} f_{\boldsymbol{\alpha}}^2(x) dx + \int_{-\infty}^{\infty} g_{\boldsymbol{\alpha}}^2(x) dx - 2 \int_{-\infty}^{\infty} f_{\boldsymbol{\alpha}}(x) g_{\boldsymbol{\alpha}}(x) dx \end{aligned}$$

と書くことができる．ここで，解析対象とする標本と参照とする標本をそれぞれ射影ベクトル $\boldsymbol{\alpha}$ で射影したときの密度関数を，バンド幅 h_f, h_g，カーネル関数を正規分布の密度関数としたカーネル密度推定を用いて計算すると，

$$\hat{f}_{\boldsymbol{\alpha}}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sqrt{2\pi}h_f}\exp\left\{-\frac{(x-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{z}_i)^2}{2{h_f}^2}\right\},\ \hat{g}_{\boldsymbol{\alpha}}(x) = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{\sqrt{2\pi}h_g}\exp\left\{-\frac{(x-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{w}_j)^2}{2{h_g}^2}\right\}$$

となる．ここで，バンド幅 h_f, h_g は Scott(1992)により求められた最適なバンド幅

$$h_f = \left(\frac{4}{3}\right)^{1/5}\sigma_f n^{-1/5}, \quad h_g = \left(\frac{4}{3}\right)^{1/5}\sigma_g m^{-1/5}$$

を使用する．σ_f, σ_g はそれぞれ解析対象とする標本，および参照とする標本を射影したときの標準偏差を表す．これより，

$$\begin{aligned}
\int_{-\infty}^{\infty}\hat{f}_{\boldsymbol{\alpha}}^2(x)dx &= \frac{1}{2\sqrt{\pi}n^2h_f}\sum_{i=1}^{n}\sum_{j=1}^{n}\exp\left\{-\frac{(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{z}_i-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{z}_j)^2}{4{h_f}^2}\right\},\\
\int_{-\infty}^{\infty}\hat{g}_{\boldsymbol{\alpha}}^2(x)dx &= \frac{1}{2\sqrt{\pi}m^2h_g}\sum_{i=1}^{m}\sum_{j=1}^{m}\exp\left\{-\frac{(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{w}_i-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{w}_j)^2}{4{h_g}^2}\right\},\\
\int_{-\infty}^{\infty}\hat{f}_{\boldsymbol{\alpha}}(x)\hat{g}_{\boldsymbol{\alpha}}(x)dx &= \frac{1}{\sqrt{2\pi}nm\sqrt{{h_f}^2+{h_g}^2}}\sum_{i=1}^{n}\sum_{j=1}^{m}\exp\left\{-\frac{(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{z}_i-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{w}_j)^2}{2({h_f}^2+{h_g}^2)}\right\}.
\end{aligned}$$

従って，1 次元の Hall Type 相対射影指標は

$$\begin{aligned}
I(\boldsymbol{\alpha}) &= \int_{-\infty}^{\infty}\hat{f}_{\boldsymbol{\alpha}}^2(x)dx + \int_{-\infty}^{\infty}\hat{g}_{\boldsymbol{\alpha}}^2(x)dx - 2\int_{-\infty}^{\infty}\hat{f}_{\boldsymbol{\alpha}}(x)\hat{g}_{\boldsymbol{\alpha}}(x)dx\\
&= \frac{1}{2\sqrt{\pi}n^2h_f}\sum_{i=1}^{n}\sum_{j=1}^{n}\exp\left\{-\frac{(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{z}_i-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{z}_j)^2}{4{h_f}^2}\right\}\\
&\qquad + \frac{1}{2\sqrt{\pi}m^2h_g}\sum_{i=1}^{m}\sum_{j=1}^{m}\exp\left\{-\frac{(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{w}_i-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{w}_j)^2}{4{h_g}^2}\right\}\\
&\qquad - \frac{\sqrt{2}}{\sqrt{\pi}nm\sqrt{{h_f}^2+{h_g}^2}}\sum_{i=1}^{n}\sum_{j=1}^{m}\exp\left\{-\frac{(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{z}_i-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{w}_j)^2}{2({h_f}^2+{h_g}^2)}\right\}
\end{aligned}$$

となる．

次に，2 次元の Hall Type 相対射影指標を考える．従来の 2 次元の Hall の射影指標は，解析対象とする標本 $\boldsymbol{z}_i, i = 1,\ldots,n$ を射影ベクトル $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2$ で 2 次元空間に射影したときの密度関数を $f_{\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2}(x_1, x_2)$ とし，正規分布の密度関数を $\phi(x_1, x_2)$ とすると，

$$J \equiv \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\{f_{\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2}(x_1,x_2)-\phi(x_1,x_2)\}^2dx_1dx_2$$

で表される．これより，参照とする標本を $\boldsymbol{w}_j$, $j = 1,\ldots,m$ とし，この標本を射影ベクトル $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2$ で 2 次元空間に射影したときの密度関数を $g_{\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2}(x_1, x_2)$ とすれば，2 次元の Hall Type 相対射影指標は

$$
\begin{aligned}
I(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \{f_{\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2}(x_1, x_2) - g_{\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2}(x_1, x_2)\}^2 dx_1 dx_2 \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f^2_{\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2}(x_1, x_2) dx_1 dx_2 + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g^2_{\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2}(x_1, x_2) dx_1 dx_2 \\
&\quad -2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2}(x_1, x_2) g_{\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2}(x_1, x_2) dx_1 dx_2
\end{aligned}
$$

となる．ここで，射影ベクトル $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2$ で射影したときの密度関数 $f_{\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2}$ の推定に用いるバンド幅をそれぞれ h_1, h_2，密度関数 $g_{\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2}$ の推定に用いるバンド幅をそれぞれ b_1, b_2 とし，カーネル関数を正規分布の密度関数としてカーネル密度推定を行うと，

$$
\begin{aligned}
\hat{f}_{\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2}(x_1, x_2) &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2\pi h_1 h_2} \exp\left\{-\frac{(x_1 - \boldsymbol{\alpha}_1{}^{\mathrm{T}} \boldsymbol{z}_i)^2}{2{h_1}^2}\right\} \exp\left\{-\frac{(x_2 - \boldsymbol{\alpha}_2{}^{\mathrm{T}} \boldsymbol{z}_i)^2}{2{h_2}^2}\right\}, \\
\hat{g}_{\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2}(x_1, x_2) &= \frac{1}{m}\sum_{j=1}^{m} \frac{1}{2\pi b_1 b_2} \exp\left\{-\frac{(x_1 - \boldsymbol{\alpha}_1{}^{\mathrm{T}} \boldsymbol{w}_j)^2}{2{b_1}^2}\right\} \exp\left\{-\frac{(x_2 - \boldsymbol{\alpha}_2{}^{\mathrm{T}} \boldsymbol{w}_j)^2}{2{b_2}^2}\right\}
\end{aligned}
$$

と表すことができる．ここで，使用した最適なバンド幅は

$$
h_i = \sigma_{fi} n^{-1/6} \quad (i = 1, 2), \qquad b_i = \sigma_{gi} m^{-1/6} \quad (i = 1, 2)
$$

で計算される．σ_{fi}, σ_{gi} は解析対象とする標本，参照とする標本それぞれを各射影方向ベクトルで射影したデータの標準偏差である．この密度推定を用いると，2 次元 Hall Type 相対射影指標は以下の式で示される．

$$
\begin{aligned}
I(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) &= \frac{1}{4\pi n^2 h_1 h_2} \sum_{i=1}^{n}\sum_{k=1}^{n} \exp\left(-\frac{(\boldsymbol{\alpha}_1{}^{\mathrm{T}} \boldsymbol{z}_i - \boldsymbol{\alpha}_1{}^{\mathrm{T}} \boldsymbol{z}_k)^2}{4{h_1}^2} - \frac{(\boldsymbol{\alpha}_2{}^{\mathrm{T}} \boldsymbol{z}_i - \boldsymbol{\alpha}_2{}^{\mathrm{T}} \boldsymbol{z}_k)^2}{4{h_2}^2}\right) \\
&+ \frac{1}{4\pi m^2 b_1 b_2} \sum_{j=1}^{m}\sum_{k=1}^{m} \exp\left(-\frac{(\boldsymbol{\alpha}_1{}^{\mathrm{T}} \boldsymbol{w}_j - \boldsymbol{\alpha}_1{}^{\mathrm{T}} \boldsymbol{w}_k)^2}{4{b_1}^2} - \frac{(\boldsymbol{\alpha}_2{}^{\mathrm{T}} \boldsymbol{w}_j - \boldsymbol{\alpha}_2{}^{\mathrm{T}} \boldsymbol{w}_k)^2}{4{b_2}^2}\right) \\
&- \frac{1}{\pi n m \sqrt{({h_1}^2 + {b_1}^2)({h_2}^2 + {b_2}^2)}} \sum_{i=1}^{n}\sum_{j=1}^{m} \exp\left(-\frac{(\boldsymbol{\alpha}_1{}^{\mathrm{T}} \boldsymbol{z}_i - \boldsymbol{\alpha}_1{}^{\mathrm{T}} \boldsymbol{w}_j)^2}{2({h_1}^2 + {b_1}^2)} - \frac{(\boldsymbol{\alpha}_2{}^{\mathrm{T}} \boldsymbol{z}_i - \boldsymbol{\alpha}_2{}^{\mathrm{T}} \boldsymbol{w}_j)^2}{2({h_2}^2 + {b_2}^2)}\right).
\end{aligned}
$$

k 次元へ射影する場合の Hall Type 相対射影指標も同様に求めることができる．射影ベクトル $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_k$ で射影したときの密度関数 $f_{\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\cdots,\boldsymbol{\alpha}_k}$ の推定に用いるバンド幅をそれぞれ $h_1, h_2, \cdots, h_k$，密度関数 $g_{\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\cdots,\boldsymbol{\alpha}_k}$ の推定に用いるバンド幅をそれぞれ $b_1, b_2, \cdots, b_k$ とし，カーネル関数を正規分布の密度関数としてカーネル密度推定を行うと，

$$
\begin{aligned}
\hat{f}_{\boldsymbol{\alpha}_1,\cdots,\boldsymbol{\alpha}_k}(x_1, \cdots, x_k) &= \frac{1}{(2\pi)^{k/2} n h_1 \cdots h_k} \sum_{i=1}^{n} \left\{\prod_{d=1}^{k} \exp\left(-\frac{(x_d - \boldsymbol{\alpha}_d{}^{\mathrm{T}} \boldsymbol{z}_i)^2}{2{h_d}^2}\right)\right\}, \\
\hat{g}_{\boldsymbol{\alpha}_1,\cdots,\boldsymbol{\alpha}_k}(x_1, \cdots, x_k) &= \frac{1}{(2\pi)^{k/2} m b_1 \cdots b_k} \sum_{j=1}^{m} \left\{\prod_{d=1}^{k} \exp\left(-\frac{(x_d - \boldsymbol{\alpha}_d{}^{\mathrm{T}} \boldsymbol{w}_j)^2}{2{b_d}^2}\right)\right\}
\end{aligned}
$$

となり，このときのバンド幅は

$$
h_i = \left(\frac{4}{k+2}\right)^{\frac{1}{k+4}} \sigma_{fi} n^{-\frac{1}{k+4}} \qquad (i = 1, 2, \ldots, k)
$$

$$b_i = \left(\frac{4}{k+2}\right)^{\frac{1}{k+4}} \sigma_{gi} m^{-\frac{1}{k+4}} \qquad (i = 1, 2, \ldots, k)$$

である．σ_{fi}, σ_{gi} は解析対象とする標本，参照とする標本それぞれを各射影方向ベクトルで射影したデータの標準偏差である．この推定された密度関数を用いると，k 次元の Hall Type 相対射影指標は

$$\begin{aligned} I(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_k) = & \frac{1}{2^k \pi^{k/2} n^2 h_1 h_2 \cdots h_k} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\{ \prod_{d=1}^{k} \exp\left(-\frac{(\boldsymbol{\alpha}_d{}^{\mathrm{T}} \boldsymbol{z}_i - \boldsymbol{\alpha}_d{}^{\mathrm{T}} \boldsymbol{z}_j)^2}{4 {h_d}^2} \right) \right\} \\ & + \frac{1}{2^k \pi^{k/2} m^2 b_1 b_2 \cdots b_k} \sum_{i=1}^{m} \sum_{j=1}^{m} \left\{ \prod_{d=1}^{k} \exp\left(-\frac{(\boldsymbol{\alpha}_d{}^{\mathrm{T}} \boldsymbol{w}_i - \boldsymbol{\alpha}_d{}^{\mathrm{T}} \boldsymbol{w}_j)^2}{4 {b_d}^2} \right) \right\} \\ & - \frac{1}{2^{k/2-1} \pi^{k/2} nm \prod_{d=1}^{k} ({h_d}^2 + {b_d}^2)^{1/2}} \sum_{i=1}^{n} \sum_{j=1}^{m} \left\{ \prod_{d=1}^{k} \exp\left(-\frac{(\boldsymbol{\alpha}_d{}^{\mathrm{T}} \boldsymbol{z}_i - \boldsymbol{\alpha}_d{}^{\mathrm{T}} \boldsymbol{w}_j)^2}{2({h_d}^2 + {b_d}^2)} \right) \right\} \end{aligned}$$

となる．

3.1. 数値実験

本節では人工データを用意し，多次元へ拡張した Area 相対射影指標および新たに提案した Hall Type 相対射影指標を用いた相対射影追跡法によって，参照とする標本と比べて興味深い構造を検出できるかを調べ，両指標を比較する．本実験では 10 次元データを 2 次元空間へ次元縮小する．

有効性の評価

得られた射影方向ベクトルがどの程度，真の射影方向に近いかを測る評価式として，ここでは重相関係数の 2 乗を用いる．この評価式は Li(1991) で使用されており，値が 1 に近ければ近いほど良い結果が得られたといえる．重相関係数の 2 乗を

$$R^2(\hat{\boldsymbol{\alpha}}_i) = \max_{\boldsymbol{\alpha} \in A} \frac{(\hat{\boldsymbol{\alpha}}_i{}^{\mathrm{T}} \Sigma_{xx} \boldsymbol{\alpha})^2}{\hat{\boldsymbol{\alpha}}_i{}^{\mathrm{T}} \Sigma_{xx} \hat{\boldsymbol{\alpha}}_i \cdot \boldsymbol{\alpha}^{\mathrm{T}} \Sigma_{xx} \boldsymbol{\alpha}}$$

と表す．ここで，A は真の射影方向空間を示し，$\boldsymbol{\alpha}$ は真の射影方向空間上のベクトルであり，$\hat{\boldsymbol{\alpha}}_i$ は得られた射影方向ベクトルである．また，Σ_{xx} は標本の分散共分散行列である．

実験方法

以下の方法で 2 次元射影方向ベクトルを求め，そのベクトルの重相関係数の 2 乗を評価する．

1. 用意した人工データを平均 0，分散 1，共分散 0 に球化する．
2. 射影方向ベクトル $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2$ の初期値を一様乱数を用いて設定する．
3. 射影指標が最大となる $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2$ を求める．ここでは，非線形最適化の手法である準ニュートン法のアルゴリズムを用いて求める．準ニュートン法においてヘッセ行列の更新式は Davidon-Fletcher-Powell 公式を使用する．
4. 求めた射影方向ベクトル $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2$ の重相関係数の 2 乗を計算する．
5. 2 の初期値を 100 回変化させ，3〜4 を繰り返し，射影指標の値が最も大きいベクトル $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2$ を求める．

人工データ　各変数が混合比率 1 : 5 の混合正規分布 $N(-1.8, 0.4^2)$, $N(1.8, 0.4^2)$ に従う 10 変数 $(x_1, \cdots, x_{10})$ のデータを 1000 個用意する．このデータに対して $\sin(x_1)+\cos(x_2)+\varepsilon$; $\varepsilon \sim N(0, 0.2^2)$ の値を計算し，$-\frac{2}{3} < \sin(x_1)+\cos(x_2)+\varepsilon < \frac{2}{3}$ を満たす標本のみを取り出す．取り出された標本数は 430 である．この取り出した標本を解析対象とし，参照とする標本を 1000 個全てのデータと設定して，相対射影追跡法を適用する．解析対象の標本の分布は，参照とする標本の分布と，変数 x_1, x_2 で張る空間において異なる特徴を持つ．この x_1, x_2 で張る空間を検出できるかを評価し，有効性を示すことが本実験の目的である．以下に用意した人工データの対散布図を表示する．

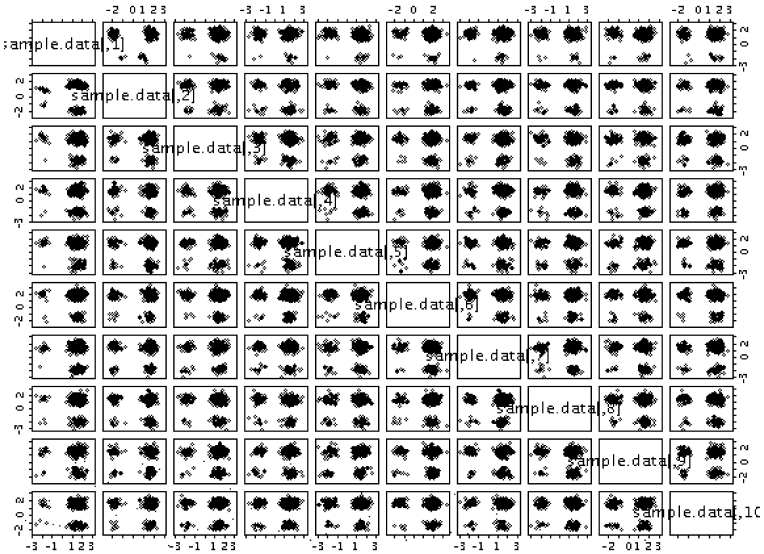


図 **1.**　解析対象とする人工データの対散布図

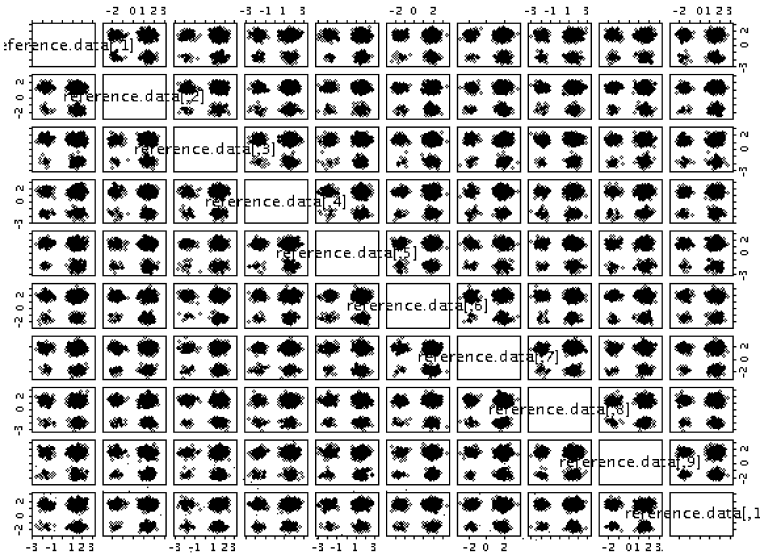


図 **2.**　参照とする人工データの対散布図

実験結果　従来の Friedman の射影指標と Hall の射影指標，及び先に述べた Area 相対射影指標と Hall Type 相対射影指標それぞれの指標を用いた射影追跡法を適用した．結果として，求めた射影方向ベクトルの重相関係数の 2 乗と計算時間（1 初期点が局所解に収束するまでの時間）を示す．この実験における真の射影方向空間は 2 つのベクトル $\boldsymbol{\beta}_1 = (1,0,0,0,0,0,0,0,0,0)$，および $\boldsymbol{\beta}_2 = (0,1,0,0,0,0,0,0,0,0)$ で張られる空間であり，この空間との重相関係数の 2 乗 $R^2(\boldsymbol{\alpha}_1), R^2(\boldsymbol{\alpha}_2)$ を計算する．なお，この計算は Linux 搭載計算機（Intel Xeon 2.8 GHz）上で S-PLUS（一部は C 言語）を用いて行った．

表 1 から従来の Friedman の射影指標及び Hall の射影指標を用いた場合，得られた両方のベクトルの重相関係数の 2 乗の値が 1 に近い値ではなく，真の射影方向は検出できなかったことがわかる．

表 1.　人工データに対する射影追跡法と相対射影追跡法の結果

射影指標	$R^2(\boldsymbol{\alpha}_1)$	$R^2(\boldsymbol{\alpha}_2)$	計算時間（秒）
Friedman の射影指標	0.635	0.362	133.0
Hall の射影指標	0.286	0.710	91.8
Area 相対射影指標	0.176	0.979	1199.9
Hall Type 相対射影指標	0.998	0.996	1722.1

これに対して，Hall Type 相対射影指標を用いた場合は，重相関係数の 2 乗の値が両方のベクトルともに 1 に近い値を示しており，真の射影方向空間をよく検出していることがわかる．Area 相対射影指標を用いた場合では，得られた 2 次元ベクトルの片方のベクトルのみ真の射影方向空間に含まれるが，もう片方のベクトルでは真の射影方向空間と異なる方向が探索された．この結果は，2 次元空間のうち 1 方向は真の射影方向が探索されているという意味で従来の 2 つの射影指標と比べると優れていると言えるが，Hall Type 相対射影指標と比べると劣っている．

1 次元空間への射影を考えたとき，非特徴的な構造が正規分布とならない場合において，Area 相対射影指標は，従来の Friedman の射影指標と比べて有効であることが示されている（弘・小宮・南・水田，2003）．しかし，本実験により，2 次元に拡張した Area 相対射影指標では，Hall Type 相対射影指標で求めることのできる真の射影方向空間を必ずしも探索できないことが示された．この理由は，両指標の解析対象とする標本と，参照とする標本の分布の差の測り方にある．Hall Type 相対射影指標は標本の射影の密度関数をカーネル密度推定を用いて推定し，密度関数の差の 2 乗を計算している．これに対して，Area 相対射影指標は標本の射影の経験分布関数を作成して，分布関数の差の体積を計算する．密度関数の推定という観点で考えれば，Hall Type 相対射影指標は 1 個の標本の周辺に正規分布を仮定してその重ね合わせによって推定しているが，Area 相対射影指標は 1 個の標本点にデルタ関数を仮定して重ね合わせているため，推定の精度があまり良くない．さらに，Area 相対射影指標は経験分布関数の差の体積を測っているので，各分布の裾における差が指標の値に大きく影響する．このため，次元数の増加に伴い，解析対象とする標本と，参照とする標本の本質的な分布の差を検出できない場合があると考察される．

一様乱数により発生させた 1 つの初期点の最適化が終わるまでの計算時間に関しては，従来の Friedman の射影指標と Hall の射影指標では 1 分半から 2 分程度であるのに対して，Area 相対射影指標は 20 分，Hall Type 相対射影指標は 30 分程度と時間がかかる．これは，相対射影指標の

場合，参照とする標本の分だけ，扱うデータ数が従来の射影指標よりも多くなることや非特徴的な構造を正規分布と定めずに標本から推定していることが原因であると考える．また，これらのプログラムでは，従来の射影指標の計算よりもループが多くなるが，S 言語でループを不用意に利用すると計算時間が増大することが知られている．本実験では，できる限りループ文は C 言語を用いているが S 言語で書かれている部分もあるため，その部分で計算時間が多めにかかっていると考察される．従って，各相対射影指標の本質的な計算時間は表 1 の結果より多少短くなると考えられる．また現在の計算機の計算速度を考えればこの計算時間は問題となる遅さではない．

3.2. 実データへの適用例

実データへの適用例として AAUP Faculty Salary Data（1994）を利用する．これは，American Association of University Professors（AAUP）に所属するアメリカの大学ごとの教員の年俸調査のデータであり，16 変数 1161 個からなる．16 変数の中で解析に用いた 10 変数を表 2 に示す．

表 2. AAUP Faculty Salary Data における解析に用いた変数

変数名	意味	
x_1	教授の年間平均基本給	（Average salary of full professors）
x_2	助教授の年間平均基本給	（Average salary of associate professors）
x_3	助手の年間平均基本給	（Average salary of assistant professors）
x_4	教授の年間平均年俸	（Average compensation of full professors）
x_5	助教授の年間平均年俸	（Average compensation of associate professors)
x_6	助手の年間平均年俸	（Average compensation of assistant professors）
x_7	教授の人数	（Number of full professors）
x_8	助教授の人数	（Number of associate professors）
x_9	助手の人数	（Number of assistant professors）
x_{10}	講師の人数	（Number of instructors）

このデータは 1161 の大学がそれぞれ Type I, Type IIA, Type IIB の 3 Type に分けられており，Type I は Doctoral-Level Institutions，Type IIA は Comprehensive Institutions，Type IIB は General Baccalaureate Institutions である．

実験方法

この AAUP Faculty Salary Data から，Type I である大学を抽出し，表 2 の変数 $x_1, \cdots, x_{10}$ を持った部分集合に対して，Friedman の射影指標と Hall の射影指標を用いた従来の 2 次元射影追跡法と，Area 相対射影指標と Hall Type 相対射影指標を用いた 2 次元相対射影追跡法を適用する．つまり，Type I の大学の興味深い次元縮小空間を探索する．相対射影追跡法で定める参照とする標本には，全大学のデータを設定する．ただし，1161 個のデータのうち欠損値のない 1074 個を用いる．

実験結果

Friedman の射影指標を適用した場合，Hall の射影指標を適用した場合，Area 相対射影指標を適用した場合，Hall Type 相対射影指標を適用した場合の順で結果を示し，それぞれの指標で興味

深い構造が検出されているかを比較する．

各指標を適用した場合について，得られた射影方向ベクトルを表示し，その射影方向空間へ射影したときの分布を散布図によって表示する．この射影方向ベクトルと散布図を用いて「興味深い」構造を捉えられているかを考察する．また，以後表示する散布図の座標は $(x, y) = (\boldsymbol{\alpha}_1{}^{\mathrm{T}}\boldsymbol{z}, \boldsymbol{\alpha}_2{}^{\mathrm{T}}\boldsymbol{z})$ である．ここで，$\boldsymbol{z}$ は標本とする．

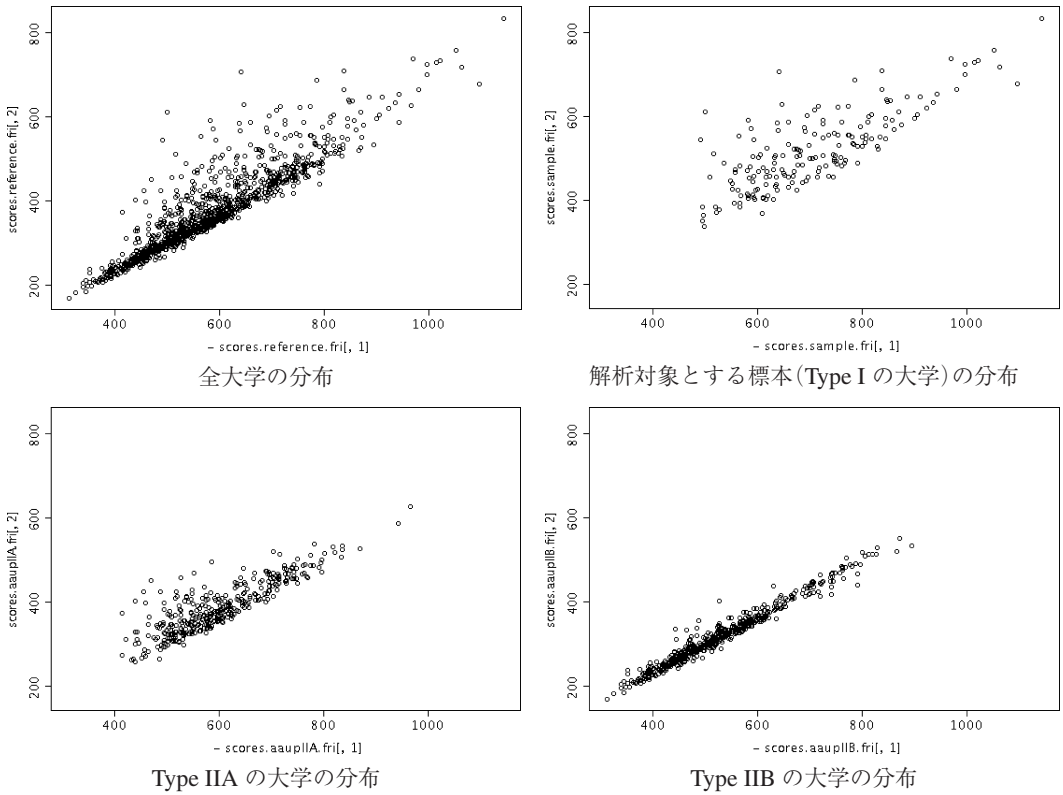
Friedman の射影指標を適用した場合

従来の Friedman の射影指標を用いた射影追跡法を行い，そのとき得られた射影方向ベクトルを表 3 に示す．

表 3.　Friedman の射影指標を用いた射影追跡法の結果

射影方向ベクトル	$(a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8, a_9, a_{10})$
$\boldsymbol{\alpha}_1$	$(0.291, 0.114, 0.106, 0.580, -0.173, 0.165, 0.131, -0.324, 9.86 \times 10^{-6}, -0.613)$
$\boldsymbol{\alpha}_2$	$(0.296, 0.039, 0.073, 0.439, -0.269, 0.013, 0.138, -0.136, 0.172, 0.757)$

この実験結果は，従来の射影追跡法を適用したので，正規分布と Type I の大学のデータの 2 次元射影とを比較して，正規分布から最も離れた構造を検出した空間を得たものとなる．得られた射影方向空間へ射影したときの散布図を図 3 に示す．



全大学の分布　　解析対象とする標本（Type I の大学）の分布

Type IIA の大学の分布　　Type IIB の大学の分布

図 3.　Friedman の射影指標で得られた射影方向で射影したときの散布図

この散布図から，解析対象とする標本の分布は確かに正規分布とは異なっており，正規性から離れているという意味で興味深い構造である．しかし，全大学の分布と比較して Type I の大学特有の興味深い構造は捉えられていない．なぜなら，全大学の分布が正規分布に従うのであれば，ここで得られた射影方向空間は Type I の大学特有の興味深い構造を捉えていることになるが，この射影方向空間上の全大学の分布は正規分布に従っていないからである．

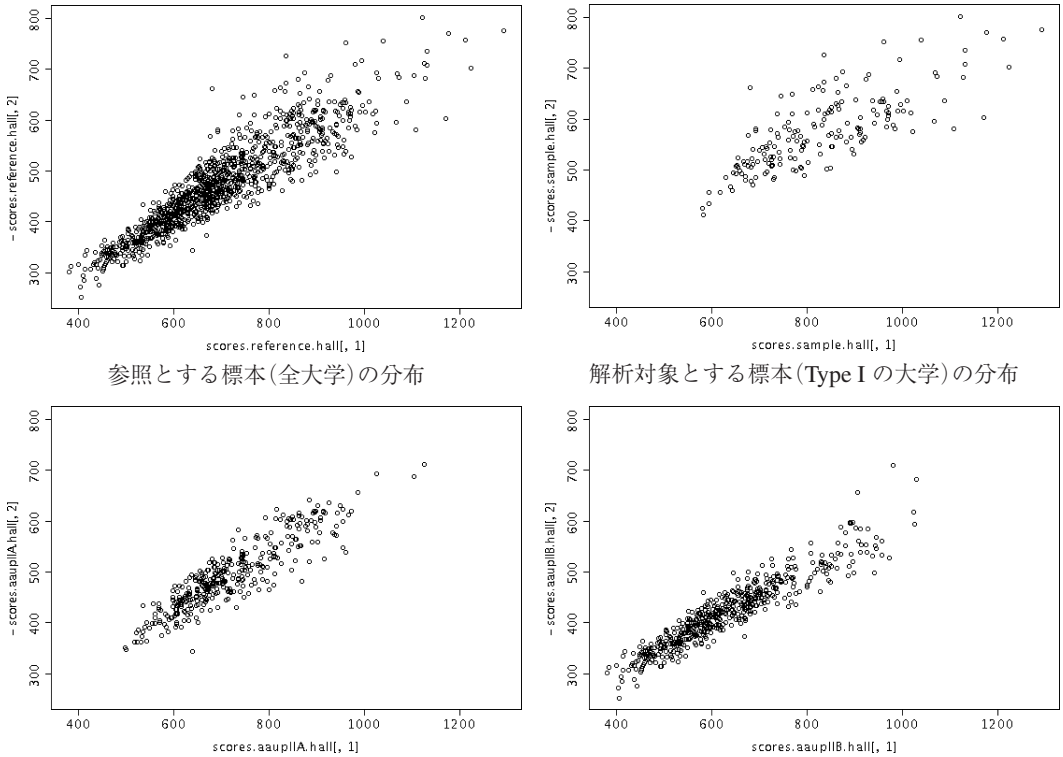
Hall の射影指標を適用した場合

従来の Hall の射影指標を用いた射影追跡法によって得られた射影方向ベクトルを，表 4 に示す．

表 4.　Hall の射影指標を用いた射影追跡法の結果

射影方向ベクトル	$(a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8, a_9, a_{10})$
$\boldsymbol{\alpha}_1$	(−0.007, 0.182, 0.221, 0.739, 0.405, −0.315, 0.129, −0.282, −0.097, −0.056)
$\boldsymbol{\alpha}_2$	(0.235, −0.175, 0.103, −0.314, 0.820, 0.310, 0.153, 0.025, −0.0021, 0.109)

この実験結果も，Friedman の指標の場合と同様，正規分布から最も離れた構造を検出した 2 次元射影空間となっている．得られた射影方向空間へ射影したときの散布図を図 4 に示す．



参照とする標本（全大学）の分布　　解析対象とする標本（Type I の大学）の分布

Type IIA の大学の分布　　Type IIB の大学の分布

図 4.　Hall の射影指標で得られた射影方向で射影したときの散布図

図 4 から，Friedman の射影指標の場合と同様の理由で，全大学の分布が正規分布に従っていないために，得られた射影方向空間では，全大学の分布と比較して Type I の大学特有の興味深い構造が捉えられているとは言えない．

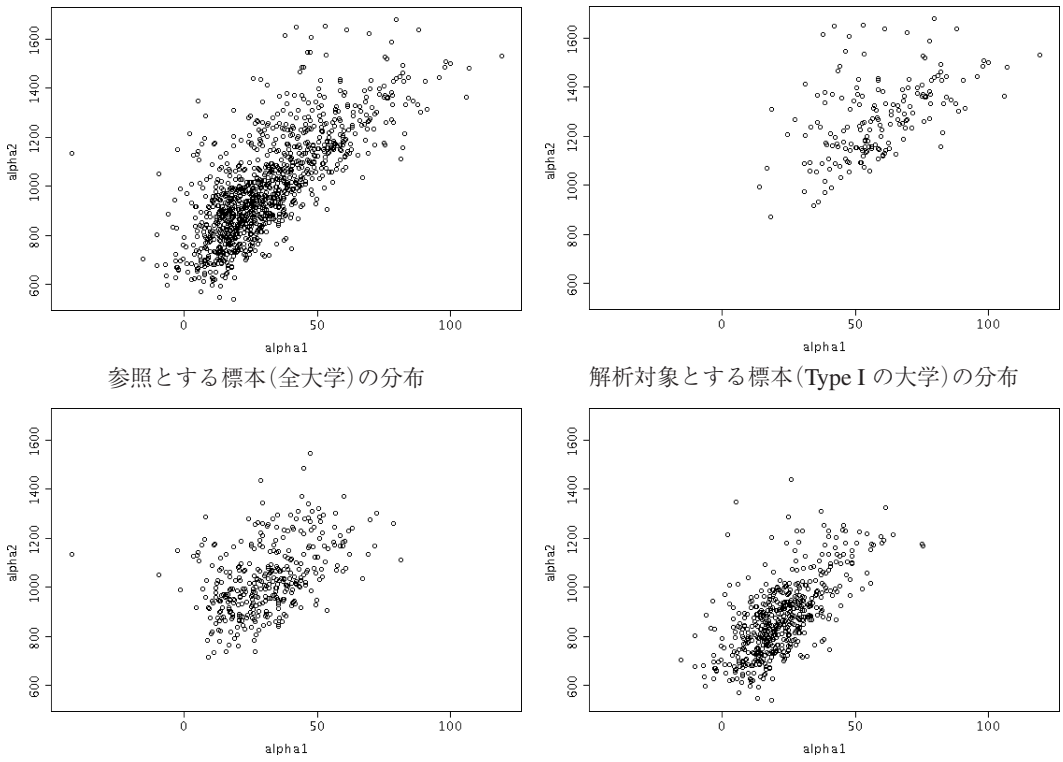
Area 相対射影指標を適用した場合

Area 相対射影指標を用いた相対射影追跡法によって得られた射影方向ベクトルを表 5 に示す．

表 5.　Area 相対射影指標を用いた相対射影追跡法の結果

射影方向ベクトル	$(a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8, a_9, a_{10})$
$\boldsymbol{\alpha}_1$	(−0.035, −0.233, 0.577, 0.166, 0.242, −0.686, −0.083, 0.109, 0.175, −0.071)
$\boldsymbol{\alpha}_2$	(0.154, 0.464, 0.351, 0.392, 0.161, 0.433, 0.017, 0.279, 0.439, 0.021)

この実験結果は，3 つの Type 全ての大学のデータを 2 次元空間へ射影したときのデータの分布と比較して，そこから最も離れた構造を検出した空間である．この射影方向空間へ射影したときの散布図を図 5 に示す．



参照とする標本（全大学）の分布　　解析対象とする標本（Type I の大学）の分布

Type IIA の大学の分布　　Type IIB の大学の分布

図 5.　Area 相対射影指標で得られた射影方向で射影したときの散布図

まず，表 5 に示した射影方向ベクトルからそれぞれの方向が何を意味しているかを考える．$\boldsymbol{\alpha}_1$

の要素の中で $a_1 = -0.035, a_4 = 0.166, a_2 = -0.233, a_5 = 0.242, a_3 = 0.577, a_6 = -0.686$ に注目する．a_1, a_2, a_3 はそれぞれ教授，助教授，助手への基本給にかかる係数であり，a_4, a_5, a_6 はそれぞれ教授，助教授，助手への最終年俸にかかる係数である．a_1 と a_4, a_2 と a_5, a_3 と a_6 の係数の絶対値がほぼ同じであると考えれば，$\boldsymbol{\alpha}_1$ は教授，助教授，助手それぞれの基本給と最終年俸との差を表している．最終年俸が基本給より少なくなることはないので，要素の正負を考慮すれば，$\boldsymbol{\alpha}_1{}^{\mathrm{T}}\boldsymbol{z}$ の値が正になれば，教授と助教授の最終年俸と基本給の差が大きいということになり，負になれば，助手の最終年俸と基本給の差が大きいということになる．$\boldsymbol{\alpha}_2$ で注目する要素は $a_2 = 0.464, a_3 = 0.351, a_4 = 0.392, a_6 = 0.433, a_9 = 0.439$ である．これらの係数は，他の係数と比べて値が大きく，同程度の値で，それぞれ助教授と助手の基本給，教授と助手の最終年俸，助手の人数にかかっている．ここから，$\boldsymbol{\alpha}_2$ は各大学の規模を表す射影方向になっていると推測される．

次に，これらの射影方向の意味を考慮して図 5 の散布図をみると，Type I の博士課程を持つ大学は他の Type の大学と比べ，最終年俸と基本給の差，つまり賞与に当たるものが多いことがわかる．

以上の考察から，Area 相対射影指標を用いた相対射影追跡法によって求められた射影方向空間は，全大学の標本の分布と比較して，解析対象である Type I の大学特有の興味深い構造を捉えていると言える．また，この低次元空間は正規分布との離れ具合で探索する従来の射影追跡法では得ることができない．

Hall Type 相対射影指標を適用した場合

Hall Type 相対射影指標を用いた相対射影追跡法を適用し，そのとき得られた射影方向ベクトルを表 6 に示す．

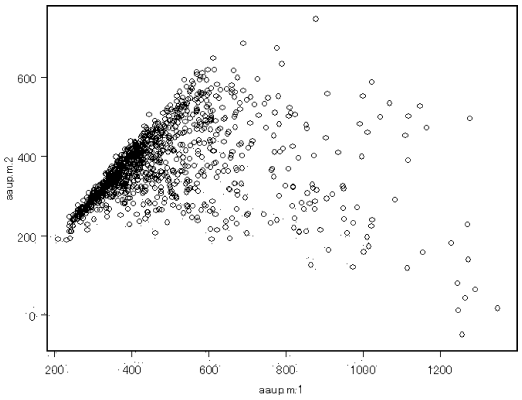
表 6.　Hall Type 相対射影指標を用いた相対射影追跡法の結果

射影方向ベクトル	$(a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8, a_9, a_{10})$
$\boldsymbol{\alpha}_1$	(0.216, −0.003, −0.027, 0.358, 0.135, −0.022, 0.883, −0.029, −0.098, −0.121)
$\boldsymbol{\alpha}_2$	(0.164, −0.207, −0.181, 0.642, 0.328, −0.197, −0.414, 0.049, −0.400, −0.072)

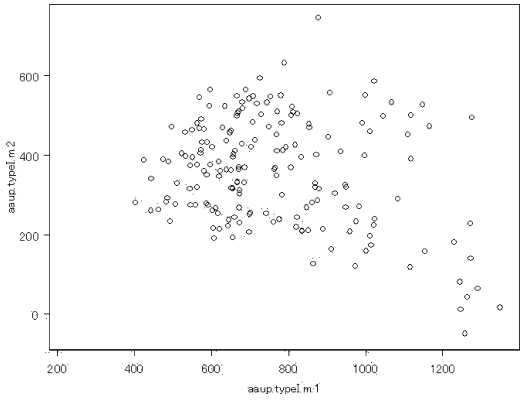
これまでと同様に，得られた射影方向空間に射影したときの散布図を図 6 に示す．

まず，表 6 の得られた射影方向ベクトルが何を意味しているかを考える．$\boldsymbol{\alpha}_1$ の各要素の値を見ると，$a_4 = 0.358$ と $a_8 = 0.833$ が他と比べて大きな値となっている．これらの要素はそれぞれ教授の給料と教授の人数に重み付けされており，ここから $\boldsymbol{\alpha}_1$ は各大学の教授に対する待遇を表す射影方向と考えることができる．$\boldsymbol{\alpha}_2$ の各要素の値では，$a_4 = 0.642, a_5 = 0.328, a_7 = -0.414, a_9 = -0.400$ が他と比べて大きな値になっており，これらの要素はそれぞれ教授の給料，助教授の給料，教授の人数，助手の人数に重み付けされる．これより，$\boldsymbol{\alpha}_2$ は各大学の規模を表す方向になっていると推測される．

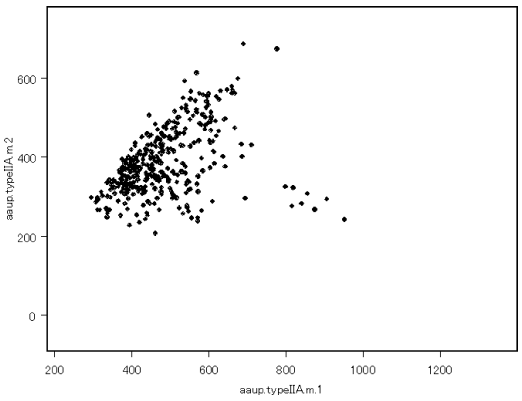
次に，これらの射影方向の意味を考慮して図 6 の散布図をみると，全大学の分布と比べて，Type I の博士課程を持つ大学は $\boldsymbol{\alpha}_1{}^{\mathrm{T}}\boldsymbol{z}$ と $\boldsymbol{\alpha}_2{}^{\mathrm{T}}\boldsymbol{z}$ の間に強い相関がないことが見てとれる．Type IIB の大学の分布を見ると $\boldsymbol{\alpha}_1{}^{\mathrm{T}}\boldsymbol{z}$ と $\boldsymbol{\alpha}_2{}^{\mathrm{T}}\boldsymbol{z}$ の間，つまり，教授に対する待遇と大学の規模の間には明らかな正の相関がある．これに対して Type I の大学では，教授へ支払う給料や教授を雇う人数は多くても助教授への給料などを含めた全体としての大学の規模は小さい所もある．実際，アメリカでは



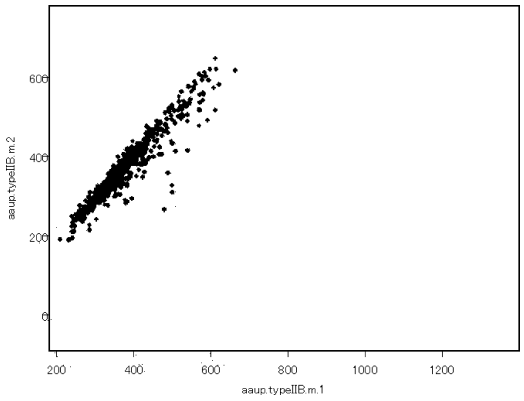
参照とする標本(全大学)の分布



解析対象とする標本(Type I の大学)の分布



Type IIA の大学の分布



Type IIB の大学の分布

図 6. Hall Type 相対射影指標で得られた射影方向で射影したときの散布図

博士課程を持つ大学機関はその規模に関わらず権威のある教授を高い給料で雇うことがあり，この散布図からそのことを読み取ることができる．

従って，Hall Type 相対射影指標を用いた相対射影追跡法によって求められた 2 次元空間は，他の Type の大学では相関が見られるが，解析対象である Type I の大学では強い相関が見られないという興味深い空間であると言える．散布図によって Area 相対射影指標を用いた場合と比較しても，全大学の分布とより離れた Type I の大学の構造を探索している．このことから，Hall Type 相対射影指標の方が興味深い構造を探索したと考えられる．また，Area 相対射影指標の場合と同様，従来の射影追跡法ではこのような低次元空間を探索することはできない．

4. おわりに

相対射影追跡法は予め参照とする標本を得ている場合に，興味深い低次元射影方向を探索する手法であり，本論文では k 次元へ射影する場合の Area 相対射影指標，Hall Type 相対射影指標を提案し，人工データと実データを用いた実験によりその有効性を示した．人工データによる数値実験では，射影する空間の次元が高くなった場合，Hall Type 相対射影指標の方が Area 相対射影

指標よりも興味深い構造を捉え得ることが示された．また，実データへ適用した結果，Hall Type 相対射影指標を用いた相対射影追跡法によって，2 次元空間上での興味深い構造が検出されることを確認した．

今後の課題としては，本論文で示した実データの解析以外に，正規分布から離れた構造ではなく参照とする標本の分布から離れた構造を捉えるという利点を生かした応用方法を考える必要がある．また，本論文で提案した相対射影指標は，参照とする標本が得られていると想定して作成しているが，参照として設定したい分布が予めわかっている場合にも相対射影追跡法は同様に定義できる．すなわち，標本からではなく，ある特定の分布からの離れ具合を測る相対射影指標を作成することも検討課題である．

参 考 文 献

AAUP faculty salary data (1994): (`http://lib.stat.cmu.edu/index.php`).

Friedman, J.H. (1987): Exploratory projection pursuit. *Journal of the American Statistical Association* **82**, 249–266.

Friedman, J.H. and Tukey, J.W. (1974): A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. on Computers* C**23**(9), 881–890.

Hall, P. (1989): On polynomial-based projection indices for exploratory projection pursuit. *The Annals of Statistics* **17**(2), 589–605.

Li, K.C. (1991): Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86**(414), 316–327.

Mizuta, M. (2002): Relative projection pursuit. *Data Analysis, Classification and Related Methods* (Edited by Andrzej Sokotowski and Krzysztof Jajuga), Cracow University of Economics, 131.

Scott, D.W. (1992): *Multivariate Density Estimation*. Wiley-InterScience.

弘 新太郎，小宮由里子，南 弘征，水田正弘 (2003): 経験分布関数を用いた新たな射影指標の提案．応用統計学 32(1), 17–28.

（2004 年 2 月 3 日受付　4 月 21 日採択）

著者連絡先：〒 060–0811 札幌市北区北 11 条西 5 丁目
北海道大学 情報基盤センター南館
E-mail: mizuta@cims.hokudai.ac.jp

Japanese J. Appl. Statist. 33 (3) (2004), 225–241

Multidimensional Relative Projection Pursuit

Shintaro Hiro[1], Yuriko Komiya[2], Hiroyuki Minami[2] and Masahiro Mizuta[2,*]

[1] Graduate School of Engineering, Hokkaido University
[2] Information Initiative Center, Hokkaido University

Abstract

We propose a new multidimensional projection index for relative projection pursuit (RPP; Mizuta, 2002). RPP is a dimension reduction method that is an extension of conventional projection pursuit (Friedman and Tukey, 1974). Conventional projection pursuit finds 'interesting' structures which differ from the normal distribution. RPP finds structures that differ from a reference data set predefined by the user as having 'uninteresting' structure. We have already proposed a one-dimensional projection index for RPP, the area index, which measures the difference between target data and reference data as a degree of 'interestingness'. However, it cannot be applied when a user wants to reduce high dimensional data into spaces of more than one dimension. Therefore, we extend the area index so that it can be applied even when the target data set is projected into multidimensional space. In addition, we develop a new index for RPP, which is based on the Hall index (Hall, 1989), called the Hall type relative projection index.

We demonstrate the effectiveness of multidimensional RPP using artificial and actual data. In the numerical example with artificial data, it is shown that with the Hall type relative projection index we can detect more 'interesting' multidimensional spaces than that with Area index. When we apply multidimensional RPP to actual data, we can obtain 'interesting' structures of high dimensional data that cannot be derived using conventional projection pursuit.

Key words: area index, Hall index, reduction of dimensionality

*Corresponding author
E-mail address: mizuta@cims.hokudai.ac.jp (Masahiro Mizuta)

Received February 3, 2004; Accepted April 21, 2004.