2000 年情報論的学習理論ワークショップ
2000 Workshop on Information-Based Induction Sciences (IBIS2000)
Izu, Japan, July 17-18, 2000.

カーネル正準相関分析
A kernel method for canonical correlation analysis

赤穂 昭太郎*
Shotaro Akaho

Abstract: Canonical correlation analysis is a technique to extract common features from a pair of multivariate data. In complex situations, however, it does not extract useful features because of its linearity. On the other hand, kernel method used in support vector machine is an efficient approach to improve such a linear method. In this paper, we investigate the effectiveness of applying kernel method to canonical correlation analysis.
Keyword: multivariate analysis, multimodal data, kernel method, regularization

1 まえがき

複数の多変量情報源からデータが得られるとき，それらに共通に含まれる特徴量を抽出するという問題を扱う．例えば，画像を用いて何か物を見せ，音声によってその名前を教えるというパターン認識の学習課題があったとしよう．新たに提示された画像データに対して，過去に覚えた訓練パターンから音声を想起して再生することにする． あるいは音声によって示された画像を想起する．すると，これは画像から音声への (あるいは音声から画像への) 回帰の問題である． しかしながら，画像や音声は一般に非常に次元が大きいので，通常の回帰分析などの手法が有効に働かない．そこで，一旦低次元の特徴空間に移してから，回帰の問題を解くということが考えられる．

従来，そのような目的のために正準相関分析と呼ばれる多変量解析手法が用いられてきた． これは二つの多変量の線形変換によって，相関係数が最大となる特徴量を求めるというものである． 情報論的には二つの多変量の同時正規性を仮定したときに相互情報量を最大にする特徴抽出法になっている． しかしながら，多変量間に非線形性の強い関係が存在する場合には必ずしも有効に働くとは限らないという問題点をもつ．

一方，数年前に Vapnik らによって提案されたサポートベクターマシンはパターン認識の問題に優れた能力を発揮し注目されるようになった [8]．そこで用いられている手法の中でカーネル法 (カーネルトリック) は，識別だけでなく他の線形手法にも適用可能な汎用手法であり，線形回帰モデルを拡張したカーネル回帰，主成分分析を拡張したカーネル PCA などが提案されている [6]．

そこで本稿では，正準相関分析にカーネル法を適用し，多変量のペアに対しても適用可能な解析手法を導く． また，カーネル法は高次元化に基づく手法であるゆえ，過学習を起こしやすい．そこで，サポートベクターマシンを始めとする種々の方法ではマージン最大化などの正則化法 (あるいは Bayes 法) を用いて complexity の調節を行っている．本稿でも，有効で計算が容易な正則化法について検討を行う．

2 正準相関分析

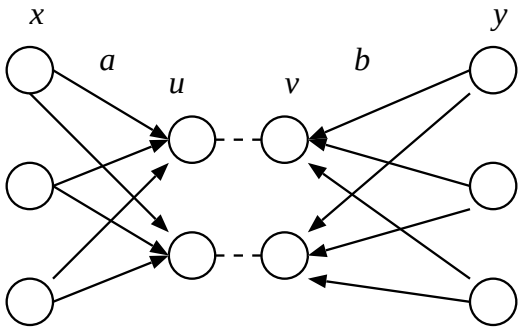

図 1: 正準相関分析

正準相関分析は，Hotelling によって 1935 年に提案された多変量解析手法である [3]．2 つの多変量 $\boldsymbol{x} \in \mathcal{R}^{n_x}$ と $\boldsymbol{y} \in \mathcal{R}^{n_y}$ をそれぞれ線形変換して，その間の相関係数ができるだけ大きくなるような空間を求める手法である (図 1)．ここでは簡単のため $\boldsymbol{x}$ と $\boldsymbol{y}$ の平均は $\mathbf{0}$ であ

*電子技術総合研究所, 〒 305-8568 茨城県つくば市梅園 1-1-4 tel. 0298-61-5549, e-mail akaho@etl.go.jp,
Electrotechnical Laboratory, 1-1-4 Umezono, Tsukuba, Ibaraki 3058568, Japan

るとし，移す空間の次元を 1 として定式化すると，

$$u = \langle \boldsymbol{a}, \boldsymbol{x} \rangle, \tag{1}$$

$$v = \langle \boldsymbol{b}, \boldsymbol{y} \rangle, \tag{2}$$

という変換により，相関係数

$$\rho = \frac{\mathrm{E}[uv]}{\sqrt{\mathrm{Var}[u]\mathrm{Var}[v]}}, \tag{3}$$

ができるだけ大きくなるような変換 $\boldsymbol{a}$, $\boldsymbol{b}$ を求めたい ($\langle \boldsymbol{a}, \boldsymbol{x} \rangle$ は内積を表す)．ただし，スケーリングの自由度を消すために

$$\mathrm{Var}[u] = \mathrm{Var}[v] = 1, \tag{4}$$

という制約を置く． $\boldsymbol{a}$ と $\boldsymbol{b}$ は一般化固有値問題の最大固有値に対する固有ベクトルとして得られる．2 次元以上の空間に移す場合には 2 番目以降の (非 0 の) 固有値に対する固有ベクトルを取ればよい．

正準相関分析が情報論的に重要なのは，$\boldsymbol{x}$ と $\boldsymbol{y}$ が同時正規分布に従う時に，それらの間の相互情報量を最大にするような変換を求めているという性質である．そのような仮定が満たされていない場合でも，正準相関分析は有効な解析手法として用いられるが，回帰が目的の場合は相関係数が小さいと使い物にならない．相関係数が小さいとき，具体的には次のようないくつかの場合が考えられる．

1. 正規性の仮定は満たされているが，本質的に $\boldsymbol{x}$ と $\boldsymbol{y}$ の間には関連性が薄い．
2. 正規性の仮定が満たされていないが，$\boldsymbol{x}$ と $\boldsymbol{y}$ には (非線形な) 強い関連性がある．
3. 正規性の仮定も満たされておらず，$\boldsymbol{x}$ と $\boldsymbol{y}$ の間にも関連性がない．

このうち，1 番目と 3 番目の点に関してはこれ以上改善できる余地は少ない．しかしながら，2 番目のように，正準相関分析では現われない何らかの関連性があると考えられる場合には何らかの手段によって関連性を顕在化できると考えられる．一つの立場は，変換に非線形性を許すというもので，Asoh らが提案したニューラルネットワークを用いた非線形正準相関分析がある [4]． しかし，この学習モデルは学習に要する時間がかかるという問題点があった．本稿では，カーネル法を適用することによって，非線形に拡張しながらも，線形手法を適用可能にし，学習の計算量を低減化することを目指す．

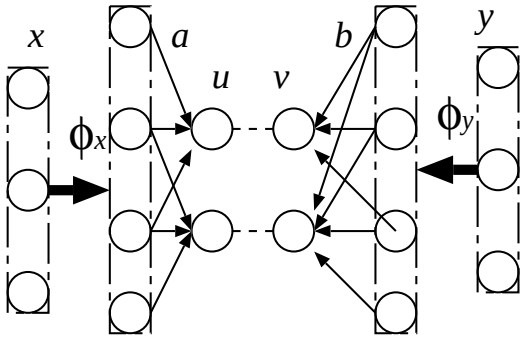

図 2: カーネル正準相関分析

3 カーネル正準相関分析

まず，$\boldsymbol{x}$ と $\boldsymbol{y}$ をそれぞれ Hilbert 空間の元 $\phi_x(\boldsymbol{x}) \in H_x$, $\phi_y(\boldsymbol{y}) \in H_y$ に移す． それらを Hilbert 空間の元 $a \in H_x, b \in H_y$ によって線形変換し，

$$u = \langle a, \phi_x(\boldsymbol{x}) \rangle, \tag{5}$$

$$v = \langle b, \phi_y(\boldsymbol{y}) \rangle, \tag{6}$$

を相関係数が最大になるようにする．

今 $\boldsymbol{x}$ と $\boldsymbol{y}$ の学習サンプル $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N$ が与えられているとしよう．u と v との相関係数を最大とする a と b は Lagrangean

$$\begin{aligned} \mathcal{L}_0 \quad = \quad & \mathrm{E}[(u - \mathrm{E}[u])(v - \mathrm{E}[v])] \\ & -\frac{\lambda_1}{2}\mathrm{E}[(u - \mathrm{E}[u])^2] \\ & -\frac{\lambda_2}{2}\mathrm{E}[(v - \mathrm{E}[v])^2], \end{aligned} \tag{7}$$

の停留解として与えられる． ただし，このままでは ϕ_x 等の次元が高い場合に不良設定となるので，2 次の正則化項を導入し，

$$\mathcal{L} = \mathcal{L}_0 + \frac{\eta}{2}(\|a\|^2 + \|b\|^2), \tag{8}$$

を解く (η は正則化パラメータ)．ここで，u の期待値が，

$$\mathrm{E}[u] = \frac{1}{N}\sum_i \langle a, \phi_x(\boldsymbol{x}_i) \rangle, \tag{9}$$

また，uv の期待値が

$$\mathrm{E}[uv] = \frac{1}{N}\sum_{i,j} \langle a, \phi_x(\boldsymbol{x}_i) \rangle \langle b, \phi_y(\boldsymbol{y}_i) \rangle, \tag{10}$$

で与えられることなどに注意すると，$\mathcal{L}$ を a で微分した値が 0 であるという条件から，

$$a = \sum_i \alpha_i \phi_x(\boldsymbol{x}_i), \tag{11}$$

という形に書ける (α_i はスカラー)．すると，

$$u = \sum_i \alpha_i \langle \phi_x(\boldsymbol{x}_i), \phi_x(\boldsymbol{x}) \rangle, \tag{12}$$

が得られる．従って，u は H_x 上の内積だけで定義される値である．さて，カーネルトリックでは，ϕ_x を経由することなしに，$\phi_x(\boldsymbol{x}_1)$ と $\phi_x(\boldsymbol{x}_2)$ の間の内積を計算できるような関数 $k_x(\boldsymbol{x}_1, \boldsymbol{x}_2)$ が存在するような ϕ_x を採用する (ϕ_y についても同様である)．実際には，ϕ_x の陽な形は知る必要がないので，まず簡単に計算ができるような k_x を適当に選び，それが内積の形に分解できるかどうかを判定する．Mercer の定理によって，正定値であるような k_x を選べば内積の形に分解できることがわかっているので，そのようなカーネルを用いればよい．

カーネルを用いて $\mathcal{L}$ を書き直そう．まず，$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)^{\mathrm{T}}$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_N)^{\mathrm{T}}$ とし，カーネルを並べた行列を

$$(K_x)_{ij} = k_x(\boldsymbol{x}_i, \boldsymbol{x}_j), \tag{13}$$

$$(K_y)_{ij} = k_y(\boldsymbol{y}_i, \boldsymbol{y}_j), \tag{14}$$

と定義しておく．すると，$\mathcal{L}$ は，

$$\begin{aligned} \mathcal{L} &= \boldsymbol{\alpha}^{\mathrm{T}} M \boldsymbol{\beta} \\ &\quad - \frac{\lambda_1}{2} \boldsymbol{\alpha}^{\mathrm{T}} L \boldsymbol{\alpha} - \frac{\lambda_2}{2} \boldsymbol{\beta}^{\mathrm{T}} N \boldsymbol{\beta} \end{aligned} \tag{15}$$

という形に書ける．ここで，

$$M = \frac{1}{N} K_x^{\mathrm{T}} J K_y, \tag{16}$$

$$L = \frac{1}{N} K_x^{\mathrm{T}} J K_x + \eta_1 K_x, \tag{17}$$

$$N = \frac{1}{N} K_y^{\mathrm{T}} J K_y + \eta_2 K_y, \tag{18}$$

$$J = I - \frac{1}{N} \mathbf{1}\mathbf{1}^{\mathrm{T}}, \tag{19}$$

$$\mathbf{1} = (1, \ldots, 1)^{\mathrm{T}}, \tag{20}$$

また，$\eta_1 = \eta/\lambda_1$, $\eta_2 = \eta/\lambda_2$.

$\eta > 0$ ならば，L および N はほぼ確率 1 で正定値対称行列となり，また，制約条件を考慮すると $\lambda_1 = \lambda_2 = \lambda$ となるので，最終的に

$$M\boldsymbol{\beta} = \lambda L \boldsymbol{\alpha}, \tag{21}$$

$$M^{\mathrm{T}} \boldsymbol{\alpha} = \lambda N \boldsymbol{\beta}, \tag{22}$$

という一般化固有値問題を解くことにより，$\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ が求まる．具体的には一般化固有値問題のパッケージを用いるか，L および N の Cholesky 分解などを用いて解くことができる．

4 計算機実験

4.1 実験 1

学習データとテストサンプルを独立かつランダムに以下のように作成した．まず，$[-\pi, \pi]$ 上の一様分布から θ を生成し，2 次元の $\boldsymbol{x}$ と $\boldsymbol{y}$ を以下のように作成した．

$$\boldsymbol{x} = \begin{pmatrix} \theta \\ \sin 3\theta \end{pmatrix} + \boldsymbol{\epsilon}_1, \tag{23}$$

$$\boldsymbol{y} = \mathrm{e}^{\theta/4} \begin{pmatrix} \cos 2\theta \\ \sin 2\theta \end{pmatrix} + \boldsymbol{\epsilon}_2. \tag{24}$$

ただし，$\boldsymbol{\epsilon}_1$, $\boldsymbol{\epsilon}_2$ は標準偏差 0.05 の独立な正規ノイズを二つ並べたベクトルである．

学習サンプル数 40, テストサンプル数 100 で学習を行った結果を示す．まず，正準相関分析を行った場合の x–y 散布図を図 3 に示す．特徴量の間の相関係数は次のようになっている (かっこ内はテストサンプルに対する値)．

	v_1		v_2	
u_1	0.71	(0.40)	0.00	(0.09)
u_2	0.00	(0.00)	0.27	(0.19)

続いて，同じデータセットに対してカーネル正準相関分析を適用した結果の x–y 散布図を図 4 に示す．ここではカーネルとして，$\boldsymbol{x}$ についても $\boldsymbol{y}$ についてもガウスカーネル

$$k(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp(-\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2}{2\sigma^2}), \tag{25}$$

を用いた．パラメータは，$\eta = 1.0$, $\sigma = 1.0$ に取った．特徴量の間の相関係数は次のようになっている (かっこ内はテストサンプルに対する値)．

	v_1		v_2	
u_1	0.98	(0.95)	0.00	(0.02)
u_2	0.00	(0.02)	0.97	(0.93)

カーネル正準相関分析の場合は第 3 成分以降の相関も存在するが，ここでは第 2 成分まで示した．

4.2 実験 2

次に，はじめに述べたマルチモーダルのパターン認識の学習を想定し，以下のような学習セットを作成した．

まず，学習サンプルは，$[0, 1]^2$ 上の一様乱数で $\boldsymbol{x}$ と $\boldsymbol{y}$ を独立に生成し対応づけた．各学習サンプルがクラス中心に相当する．テストサンプルは等確率で選んだ学

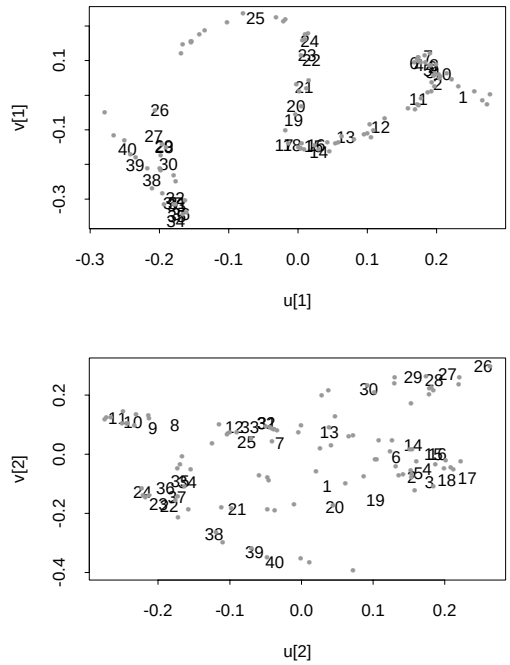

図 3: 実験 1，正準相関分析の x–y 散布図． 番号は θ を昇順に並べた学習サンプルの通し番号

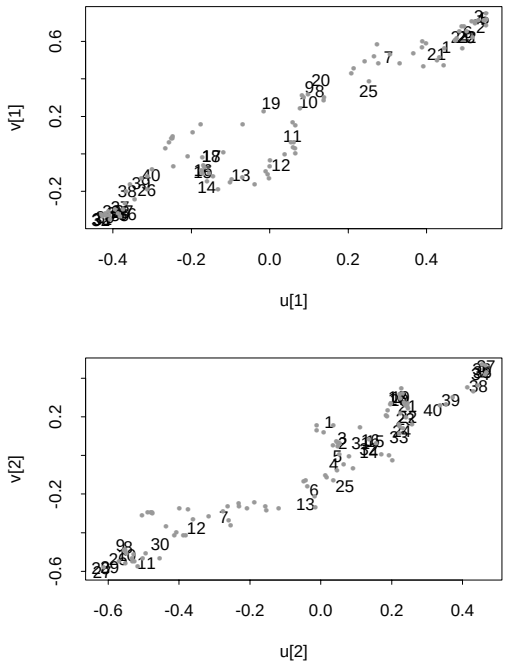

図 4: 実験 1，カーネル正準相関分析の x–y 散布図． 番号は θ を昇順に並べた学習サンプルの通し番号

習サンプルに，標準偏差 0.05 の独立な正規ノイズを加えて作成した．

学習サンプル数 (クラス数) 10, テストサンプル数 100 で学習を行った結果を示す．

まず，正準相関分析を行った x–y 散布図を図 5 に示す．特徴量の間の相関係数は次のようになっている (かっこ内はテストサンプルに対する値)．

	v_1	v_2
u_1	0.40 (0.44)	0.00 (−0.10)
u_2	0.00 (−0.05)	0.13 (0.19)

続いて，同じデータセットに対してカーネル正準相関分析を適用した結果の x–y 散布図を図 6 に示す．この場合もやはりガウスカーネルを用い，パラメータは，$\eta = 0.1$, $\sigma = 0.1$ に取った．特徴量の間の相関係数は次のようになっている (かっこ内はテストサンプルに対する値)．

	v_1	v_2
u_1	0.97 (0.90)	0.00 (0.04)
u_2	0.00 (0.01)	0.95 (0.88)

5 まとめと課題

5.1 カーネル法と正則化

正準相関分析にカーネル法の考え方を導入したカーネル正準相関分析を提案した． サポートベクターマシンと同様に，その本質はカーネル法による線形モデルによる非線形性の実現と，正則化項の導入によるオーバーフィットの防止にある．

一般に正則化では，正則化パラメータの設定が重要になる．さらに，どのようなカーネルを用いるか (本稿の実験ではガウスカーネルの分散の設定) も結果に影響を与える．本稿で行った計算機実験では，これらのパラメータを実験者がが適当に決めたが，よりシステマティックな方法として，クロスバリデーションのようなリサンプリング法や Mackay による Laplace 近似による経験 Bayes 的なアプローチが考えられる [7]．これらの計算には一般に反復的なアルゴリズムが伴い，学習に時間がかかることと，ローカルミニマムなどの問題が生じる．これらは今後の課題である．

正則化項としては，本稿で導入した 2 次の正則化項を入れるほかに，$\|\boldsymbol{\alpha}\|^2$ や $\|\boldsymbol{\beta}\|^2$ などを導入することが考えられ，以下で述べるカーネル判別分析においてもこのような正則化を行っている．アルゴリズムとしては，N や L といった行列に単位行列の定数倍が足し合わされた形となり，計算量が増えることはない．しかしながら，

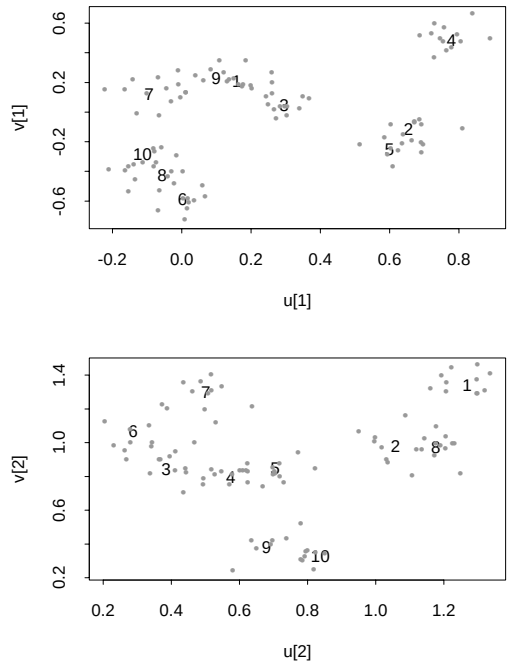

図 5: 実験 2，正準相関分析の x–y 散布図． 番号は学習サンプルの通し番号

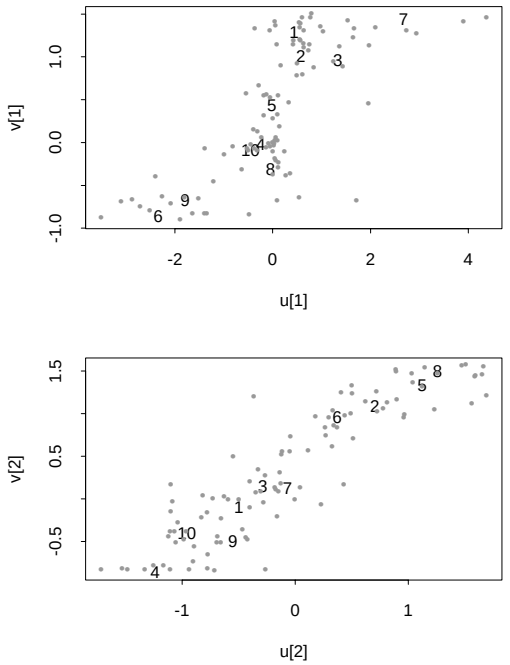

図 6: 実験 2，カーネル正準相関分析の x–y 散布図． 番号は学習サンプルの通し番号

本稿の計算機実験のデータセットでも試みたが，$\|a\|^2$ の正則化だけを行った場合と本質的に違いが見られなかった． これについては，より多くの実験例を必要とするであろう．

5.2 カーネル判別分析との関係

次に，正準相関分析と関連が深い判別分析について述べておく．判別分析はパターン認識のための多変量手法で，クラス内分散を最小にし，クラス間分散を最大にするような空間への写像を与える． これは，正準相関分析の特殊な場合とみなすことができる． 判別分析に対してカーネル法を適用したものとして Mika らの研究 [5] があるが，カーネル法を適用すると，判別分析と正準相関分析の間の包含関係は成り立たなくなる． すなわち，カーネル正準相関分析では，$\boldsymbol{y}$ をカーネルで非線形変換するが，カーネル判別分析では $\boldsymbol{y}$ についてはそのままにしておく点で異なっている．ただし，定性的な性質として，どちらの方法でも通常の正則化の入れ方ではスパースな表現が得られにくいという特徴があり，これは計算量的には不利な場合があり得る． 従って，スパース性を評価基準として採用することも考えられる．

5.3 情報論的観点からの課題

著者らは正準相関分析に情報論的な考え方を取り入れ拡張したマルチモーダル独立成分分析を提案した [2]．その変換は線形に限られているため，本質的に非線形な変換で得られた多変量同士からは適切な特徴抽出ができないことがあった．しかしながら，相互情報量を用いた評価基準はより一般的であり，これをカーネル正準相関分析に取り入れられないだろうか．

これに対する答えは与えられたデータの性質によって有効な場合とそうでない場合があるように思える．まず，本稿で行った計算機実験のように，比較的ノイズが少ない場合には正則化パラメータを小さな値にして，相関係数がほぼ 1 に近くなるのが望ましい状態である． この状態ではマルチモーダル独立成分分析による改善はあまり期待できない． なぜなら，相関係数が 1 に近いというのは高い相互情報量である十分条件になっているからである．一方で，ノイズが大きくて正則化パラメータを大きな値に設定した方が望ましい場合には，相関係数は低くなるので，そのままマルチモーダル独立成分分析が有効に働く可能性がある． しかしながらこの場合は通常の正準相関分析で十分であることも多いので，有効性には疑問が残る． ただし，$\boldsymbol{x}$ から $\boldsymbol{y}$ への写像が多価関数になっているような問題 (例えば複数属性概念の学習 [1]) では，ノイズは小さいが，相関係数が低くなる場合と考えられるので，検討の余地はある．

そこで，ノイズが小さい場合について更に考えてみよう． 計算機実験の結果を見ると，相関係数は 1 に近くても，サンプルの分布が数ヶ所に固まってしまっていて，

$\boldsymbol{u}$ から $\boldsymbol{v}$ への回帰には成功しても，$\boldsymbol{x}$ から $\boldsymbol{y}$ への想起は必ずしもうまくいくとは限らない． そこで，できるだけ $\boldsymbol{u}$ や $\boldsymbol{v}$ の空間では分布がばらけていた方が望ましい． より情報論的な見方をすれば最もエントロピーが高くなるような空間に移す方がよいのではないかという仮説が立つ． 平均と分散を固定した場合に最大エントロピーを実現するのは正規分布の場合だから，正規分布に近い場合に小さな値を取るような損失関数を設定し，それを小さくするように学習を行えばよい．例えば 3 次や 4 次のキュムラントをできるだけ 0 に近くすればよい．これは射影追跡や独立成分分析とはむしろ逆の立場であるが，それは可視化を主な目的としたこれらの手法と，回帰を目的とした正準相関分析の差であると考えられる． また，ノイズが小さいという仮定も異なっている．これについては，前節でのスパース性の基準とも絡み，今後の検討課題である．

参考文献

[1] 赤穂, 速水, 長谷川, 吉村, 麻生: ＥＭ法を用いた複数情報源からの概念獲得, 電子情報通信学会論文誌, Vol. J80-A, No. 9, pp. 1546-1553, 1997.

[2] 赤穂，梅山: マルチモーダル独立成分分析, 電子情報通信学会論文誌, A, 2000, in press.

[3] T. W. Anderson: An Introduction to Multivariate Statistical Analysis — Second edition, John Wiley & Sons, 1984.

[4] H. Asoh, O. Takechi: An approximation of Nonlinear Canonical Correlation Analysis by Multilayer Perceptrons, *Proc. of Int. Conf. Artificial Neural Networks*, pp. 713–716, 1994.

[5] S. Mika, G. Rätsch, J. Weston, B. Schölkpf, K-R. Müller: Fisher discriminant analysis with kernels, In *Y.-H. Hu, et al.(eds.): Neural Networks for Signal Processing IX*, pp. 41-48, IEEE, 1999.

[6] B. Schölkopf, A. Smola and K. Müller: Kernel principal component analysis, In B. Schölkopf et al. (eds), *Advances in Kernel Methods, Support Vector Learning*, MIT Press, 1998.

[7] M. E. Tipping: The relevant vector machine, to appear in *Advances in Neural Information Processing Systems (NIPS) 12*, 2000.

[8] V.N. Vapnik : *Statistical Learning Theory*, John Wiley & Sons, 1998.