統計数理(2010)
第 58 巻 第 2 号 167–183

特集「統計的機械学習」
[研究詳解]

指数型分布族の空間におけるデータ解析法について

赤穂 昭太郎[1] ・渡辺 一帆[2] ・岡田 真人[3]

(受付 2010 年 1 月 4 日；改訂 4 月 12 日；採択 4 月 16 日)

要　　旨

次元圧縮のために広く用いられている主成分分析は正規分布を仮定したもとで最適なものであるが，非正規的なデータに対しては必ずしも適切な低次元構造を抽出できない．本稿では，データが指数型分布族から生成されたもの，あるいは指数型分布族のパラメータとして与えられている場合の次元圧縮法の枠組みについて，著者らが行ってきた情報幾何的アプローチを中心に解説する．また，指数型分布族に属さない混合分布についても，指数型分布族の空間に埋め込む手法について述べる．さらに，確率モデルを導入し，ベイズ推定を行う拡張法や，それに基づいた次元圧縮とクラスタリングの同時最適化を行う手法についても述べる．

キーワード： 主成分分析，情報幾何，双対性，次元圧縮，クラスタリング，ベイズ推定．

1. はじめに

高次元のデータ集合が与えられたとき，そのままの形で回帰やクラスタリングといったデータ解析を行っても，思うような結果が得られないことがある． そのような場合でも，低次元空間に射影すれば結果が改善することがある．それは，高次元のデータ集合に潜在的に低次元の構造が隠れていて，ノイズなどの無駄な情報が加わったものとして高次元のデータが観測されていると考えられるからである．

そのような次元圧縮を行う手法のうち最も代表的なものとして主成分分析がある．主成分分析は，もとの高次元空間がユークリッド空間であり，データが多変量正規分布に従う場合，情報量を最大限保つアファイン部分空間を求めることができる手法である．また，計算アルゴリズムの観点からも，データ行列の特異値分解(または分散共分散行列の固有値問題)に帰着できるため解を求めるのが容易であるという利点もある．

ところが，テキストマイニングなどのデータマイニング，遺伝子情報処理など昨今重要性を増している応用課題においては，離散データなど必ずしもユークリッド空間上の点ではないデータを取り扱うことが多い上，たとえ連続値であっても分布の正規性が保証されないことも多い．そのような場合でも，形式的に主成分分析を適用することは可能だが，仮定が成り立っ

[1] 産業技術総合研究所 ヒューマンライフテクノロジー研究部門：〒305-8568 茨城県つくば市梅園 1–1–1 中央第 2
[2] 奈良先端科学技術大学院大学 情報科学研究科：〒630-0192 奈良県生駒市高山町 8916–5
[3] 東京大学大学院 新領域創成科学研究科：〒277-8561 千葉県柏市柏の葉 5–1–5

ていない以上，そうして得られた低次元構造は必ずしも適切なものになっているとは言えない．

そこで，いろいろなデータ形式や分布について個別に適切な低次元抽出を行う試みが提案されてきたが，近年，情報幾何的な観点から，データが確率分布のなす空間上の点とみなせる場合には，統一的な取り扱いが可能であることがわかった．

本稿ではまず，確率分布の中でも特に指数型分布族の空間でのデータ解析法について情報幾何に基づいた枠組みについて紹介し(2 節)，その中でも特に基本となる次元圧縮法について説明する(3，4 節)．次に，指数型分布族に属さない混合分布などについても，指数型分布族の空間に埋め込むというアプローチについて述べる(5 節)．続いて，次元圧縮に確率モデルを導入し，ベイズ推論などを可能にする枠組みへと拡張する(6 節)．そこでは，次元圧縮だけでなく，離散点への情報圧縮法であるクラスタリングについても考察し，低次元空間への射影とその空間でのクラスタリングを同時最適化する枠組みについて述べる．最後に，関連研究や今後の課題などについてまとめる(7 節)．

2. 指数型分布族の空間上のデータ解析

集合 $\mathcal{X}$ 上に値を取る確率変数 $x \in \mathcal{X}$ の確率密度(もしくは確率関数)が，パラメータ $\theta = (\theta_1,\ldots,\theta_M)^T \in \mathcal{S} \subset \mathbb{R}^M$ を用いて

$$p(x;\theta) = \exp(\theta^T F(x) + F_0(x) - \psi(\theta)) \tag{2.1}$$

の形に書けるとき指数型分布族であるという．ここで，$F(x) = (F_1(x),\ldots,F_M(x))^T$ であり，θ^T は θ の転置を表す．

パラメータ θ を座標とみなすことによって指数型分布族に属する分布の空間を考えることができる．これが情報幾何(Amari, 1985; Amari and Nagaoka, 2000)の出発点である．本稿では，観測されるデータがパラメータ θ のなす空間上の n 個の点 $\theta^{(1)},\theta^{(2)},\ldots,\theta^{(n)}$ として与えられる場合に，それらを次元圧縮したりクラスタリングしたりする手法について考える．

指数型分布族の確率変数の実現値 x ではなく，パラメータ θ が与えられるという設定には補足説明が必要であろう．まず文字通りパラメータの組が与えられるというのはどのような状況が考えられるかについて説明する．

例えば大量の顧客データなどを集めてデータ解析を行う場合に，それぞれの支店で集めたデータを中央のセンターに送るという状況を考える．このとき，すべてのデータをそのままセンターに送るのではなく，支店でデータの分布に関する統計的推測を行って，分布のパラメータだけをセンターに送り，センターでは各支店から送られたパラメータの集合に対してデータ解析を行うとする．これが，最初に考えた「パラメータがデータとして与えられる」という状況の一例である．このような処理スキームの利点は大まかに二つに分けられる．

まず第一に，データの量があまりにも多い場合，すべてのデータをセンターで処理するには負荷が大き過ぎるため，支店で事前に処理してから送ることにより処理の負荷分散が可能となる．また，支店からセンターに送る通信量も軽減できる．これは最近注目されている分散データマイニング(Agrawal and Srikant, 2000; Kumar et al., 2006)で想定されている問題設定であり，さらに支店をよりミクロなセンサーに置き換えて考えると，大量のセンサを用いてデータ処理を行うセンサーネット(Chong and Kumar, 2003)にも関連している．

第二の利点は，各支店で統計情報に集約することにより，データのプライバシーを保つことができるという点である．すべてのデータがセンターに集まっていると情報が漏洩した場合のリスクが大きいが，統計情報になっていればそのリスクは軽減する．特に顧客データや医療データなどプライバシー保護が必要なデータを集める際には有効な手法である．近年「プライバシー保護データマイニング」としていろいろな分野が関連して研究が進められているが

(Aggarwal and Yu, 2008)，上記の処理手法はその一つの形態とみなすことができる．

このように，サンプル点が確率分布のパラメータとして与えられるという問題設定は昨今注目されている新しいデータ解析手法と深く関係している．一方，得られるデータが指数型分布族の確率分布に従う確率変数の実現値として観測された場合も同じ枠組みで取り扱うことができることを説明しよう．

指数型分布族の空間は θ を座標系としてとることもできるが，

$$\eta_j = \mathrm{E}_{p(x;\theta)}[F_j(x)] = \int F_j(x)p(x;\theta)dx \tag{2.2}$$

によって定まるパラメータ $\eta = (\eta_1, \ldots, \eta_M)$ を座標系に取ることもできる．θ を自然パラメータ，η を期待値パラメータと呼び，これらはそれぞれ 1 対 1 写像によって座標変換される． 以下，その座標変換を $\theta(\eta), \eta(\theta)$ のように表す．

ここで，確率変数の実現値 x が観測されたとき，$r_j = F_j(x)$ を η_j 座標とみなすことにより，x を分布の空間上の点とみなすことができる．これは単一の観測値を十分統計量とみなした場合の最尤推定量になっている．これによって，実現値もパラメータに変換することによりパラメータ空間でのデータ解析法に帰着できるのである．

ここで，指数型分布族に対するデータ解析の適用範囲について述べておこう．指数型分布族は多項分布，Poisson 分布，正規分布など多くの基本的な分布を含む広いクラスの分布族である．従って，二値データや計数データなどさまざまな形のデータを扱うことができるようになる．また，近年盛んに研究が進んでいるカーネル法(赤穂, 2008)で用いられるグラム行列は，直観的には類似度行列を表しており，遺伝子情報処理やネットワーク解析などの応用分野において重要な役割を果たす．多くの類似度行列が与えられたときに，それらに対してデータ解析を行うということもマルチカーネルなどと呼ばれ注目されている．グラム行列は多変量正規分布の分散共分散行列とみなすことにより，指数型分布族のデータ解析の枠組みで扱うことができる(Ohara, 1999)．

3. e-PCA と m-PCA

n 個のパラメータの値が得られたとき，それを次元圧縮するために，θ (あるいは η)をユークリッド空間の座標系とみなして通常の主成分分析を行うということが考えられるが，それには二つの問題がある．第一の問題点は，ユークリッド空間での射影を取った点が必ずしも θ の定義域 $\mathcal{S}$ に入っているとは限らないという問題である(例えば多項分布で確率値が $[0,1]$ に入っていなかったり，正規分布で分散が負になるなど)．特に，部分空間に射影した点が分布としてどのような分布になっているかを知りたい場合には，定義域に入っていなければ分布として定義されないということになってしまう．第二の問題点は，ユークリッド空間での射影が必ずしも確率分布の空間上で適切なものであるとは限らないという点である． ユークリッド空間上の射影は，点から部分空間への二乗距離を最小にするものであるが，パラメータ間の二乗距離は必ずしも分布間の距離を表す尺度としては適切なものではない．

これらの問題点は情報幾何的に空間の構造を考えることにより解決できる．ここで情報幾何について本稿で必要となる基礎概念についてまとめておく．情報幾何では統計的な不変性などに基づいて微分幾何的に空間の構造を導入する．局所的には各点の近傍は線形空間(接空間)とみなし，その構造を Riemann 計量によって定める． 統計的な不変性から Riemann 計量は Fisher 情報行列

$$G_{ij}(\theta) = \mathrm{E}_\theta \left[\frac{\partial}{\partial \theta_i} \log p(x;\theta) \frac{\partial}{\partial \theta_j} \log p(x;\theta) \right] \tag{3.1}$$

によって定めればよいことが導かれる．従って θ の近傍におけるベクトル $d\theta$ の長さは $\sum_{i,j} G_{ij}(\theta) d\theta_i d\theta_j$ で測ることになる．直観的には，Cramer-Rao の不等式により，Fisher 情報行列はパラメータの推定における精度の度合いを表していると考えられるから，Fisher 情報行列の(固有)値の大きい θ では $d\theta$ が小さくても統計的には大きな距離となる．

一方，大域的な構造は，接空間と接空間の間で接ベクトルを平行移動させる際のずれの度合いを表現するための(アファイン)接続(係数)によって定まるが，情報幾何の特徴的な点は，実数 α をパラメータとする接続の族(α-接続)があることである．通常の Riemann 幾何では計量的な接続，つまり平行移動によって長さを保つような Levi-Civita 接続のみを考えるが，それは $\alpha=0$ の場合に相当する．しかしながら，情報幾何で重要なのはむしろ $\alpha=\pm 1$ の場合で非計量的である．

接続係数はテンソル的な変換をしないので，座標系をうまく取ると 0 にできる場合がある．α-接続に対するそのような座標系を α-アファイン座標系と呼び，その座標系について α-平坦であるという．α-アファイン座標系では，接ベクトルをその向きに平行移動してできる測地線が，単なる直線となり，感覚的にはユークリッド空間に近い単純な構造を持つことがわかる．

指数型分布族では，$\alpha=1$ に対しては θ 座標が 1-アファイン座標系であり，$\alpha=-1$ に対しては η 座標が -1-アファイン座標系となっている．このように，$\alpha=\pm 1$ には特別な性質があるので，$\alpha=1$ に対応する 1-接続や 1-アファイン座標系を e-接続とか e-(アファイン)座標系と呼び，$\alpha=-1$ に対応する -1-接続や -1-アファイン座標系を m-接続とか m-(アファイン)座標系と呼ぶ(e，m は exponential，mixture の頭文字で，mixture については混合分布族の自然パラメータが -1-平坦であることに由来する)．θ と η は Legendre 変換で変換される互いに双対な座標系である．また上にも述べたように，$\alpha\neq 0$ の接続は非計量的であるが，α と $-\alpha$ は互いに双対的な関係にあり，双対的な内積である $\sum_{j=1}^{M} d\theta_j d\eta_j$ は平行移動に関して不変な量になっている．

さて，次元圧縮を考える際には，部分空間を考え，点から部分空間への射影を取るという操作が必要である．e-座標系 θ (または m-座標系 η)では，測地線が直線で表されるため，アファイン部分空間はその上の 2 点間の測地線を自分自身の中に含む．そのような部分空間のことを e-自己平行な(または m-自己平行な)部分空間と呼ぶ．

部分空間の外の点から部分空間に引いた e-測地線(または m-測地線)で，部分空間の交点において(Riemann 計量の意味で)直交するとき，その点を e-射影(または m-射影)と呼ぶ．情報幾何における射影定理は一見非線形で複雑そうに見える射影をすっきりと見通しのよいものにする．

射影定理．点 $\theta\in\mathcal{S}$ から e-自己平行な部分空間 $\mathcal{T}$ への m-射影 $\hat{\theta}$ は一意的に存在する．それは $\mathcal{T}$ 上の点のうち m-ダイバージェンス

$$K_m(\theta;\hat{\theta}) = (\theta-\hat{\theta})^T\eta(\theta) - \psi(\theta) + \psi(\hat{\theta}) \tag{3.2}$$

の最小とする点として与えられる．この命題は e と m を入れ替えてもそのまま成立し，その場合 e-ダイバージェンスは $K_e(\theta;\hat{\theta}) = K_m(\hat{\theta};\theta)$ で定義する．

射影定理から，指数型分布族の空間 $\mathcal{S}$ における次元圧縮法として，双対的な二つのものが考えられる(赤穂, 2003; Akaho, 2004)．一つは $\mathcal{S}$ の e-自己平行な部分空間で，各点からの m-ダイバージェンスの総和が最小になるものを見つけるというもので，これを e-PCA と呼ぶ．もう一方は e と m を入れ替えて e-ダイバージェンスの総和を最小にする m-自己平行な部分空間を見つけるもので，m-PCA と呼ぶ．

e-PCA および m-PCA では最初に挙げた二つの問題点が解決されていることに注意する．す

なわち，射影定理により，各点から部分空間への射影は(定義域の外にはみ出したりせず)一意的に必ず存在することが保証される．また，e-(m-)ダイバージェンスは Kullback-Leibler ダイバージェンスに一致し，統計的にも分布間の隔たりを表すのに自然な量である．ちなみに e-(m-)ダイバージェンスは $\phi(\theta)$ をポテンシャル関数とする Bregman ダイバージェンスでもある．

e-PCA や m-PCA の特別な場合として重要なのが 0 次元空間への射影，すなわち，データを 1 点に集約する場合である．0 次元の e-PCA で求める点を e-中心，m-PCA の場合を m-中心とすると，それらは閉じた形で表現でき，

$$\theta_c^e = \theta\left(\frac{1}{n}\sum_{i=1}^{n}\eta(\theta^{(i)})\right), \quad \eta_c^m = \eta\left(\frac{1}{n}\sum_{i=1}^{n}\theta(\eta^{(i)})\right). \tag{3.3}$$

と書ける． つまり，指数型分布族の空間上で距離(ダイバージェンス)や中心を求めることは比較的容易である． これを使えば，k-平均法などのクラスタリング手法も構築可能である．

ちなみに，通常の主成分分析は e-PCA や m-PCA の特殊な場合とみなすことができる．すなわち，指数型分布族の空間として，多変量正規分布で分散共分散行列を単位行列に固定し，平均ベクトルだけをパラメータとした空間を考える． このとき $\theta=\eta$ であり，e-ダイバージェンスと m-ダイバージェンスはいずれも $K_m(\theta;\hat{\theta}) = K_e(\theta;\hat{\theta}) = \|\theta-\hat{\theta}\|^2$ となる．そこで各データ点 x を η 座標に対応づけ e-PCA (この場合 m-PCA も同じ)を行うと，点から部分空間への二乗距離が最小となる部分空間を求めることになり，これは通常の主成分分析にほかならない．

通常の主成分分析では，最適な L 次元アファイン部分空間は，より高次元の最適な L' 次元アファイン部分空間に含まれる．しかしながら，e-PCA や m-PCA は一般の場合，そのような階層性がないことに注意する． 例えば e-中心や m-中心は，より高次元の e-PCA や m-PCA で得られるアファイン部分空間に一般には含まれない．これは，情報幾何で扱う空間が本質的には非線形であることに由来する．

4. e (m)-PCA のための最適化法

e-PCA と m-PCA は双対な関係にあるので, 本節では e-PCA についてのみ説明する． m-PCA については説明の e- と m-，θ と η を入れ替えるだけでよい．

4.1 交互最適化

まず e-PCA の定式化を行う．e-座標系 θ のアファイン部分空間 $\mathcal{T}$ の基底ベクトル $u_1,\ldots,u_L$ と $\mathcal{T}$ 上の一点 u_0 を取り，

$$\tilde{\theta}(w) = \sum_{j=1}^{L} w_j u_j + u_0 \tag{4.1}$$

によって $\mathcal{T}$ 上の点を表現する． e-PCA では n 個の点 $\theta^{(1)},\ldots,\theta^{(n)}$ が与えられたとき，各点から $\mathcal{T}$ への m-射影を

$$\arg\min_{w^{(i)}} K_m(\theta^{(i)};\tilde{\theta}(w^{(i)})) \tag{4.2}$$

によって求め, このダイバージェンスの総和ができるだけ小さくなるように, $U=(u_1,\ldots,u_L,u_0)$ を求める． つまり，$W=(w^{(1)},\ldots,w^{(n)})$ と置き，

$$l(U,W) = \sum_i K_m(\theta^{(i)};\tilde{\theta}(w^{(i)})) \tag{4.3}$$

を U,W について最小化することが e-PCA の目標である．

一般に解は閉じた形では求められないので，適当な初期値からはじめて最急勾配法や Newton 法などの繰り返し法によって解を求めることになる．射影定理より，自己平行な部分空間を固定すれば各点からの双対的な射影は一意的である，すなわち，U を固定して W を最適化する問題はただひとつの最適解を持つ．一方，W を固定して U を求める問題も $\theta^{(1)},\ldots,\theta^{(n)}$ のなす $\mathcal{S}^n$ という大きな空間での e-自己平行な部分空間への m-射影とみなせるため，やはりただひとつの最適解を持つ．このように，U と W を交互に最適化するというアルゴリズムが自然に考えられる．ここで，それぞれのステップでの最適化はただ一つの最適解をもつ問題であるが，アルゴリズムを反復して収束した解は大域的な最適解とは限らないことに注意する．

e-自己平行な部分空間への m-射影は，混合座標系(Amari, 2001)というものを考えると陽な表現をもつ．m-座標系や e-座標系は線形変換の自由度をもつので，e-座標をうまく取れば，$\theta=(\theta_I,\theta_{II})$ のように $M-L$ 個と L 個の成分に分け，部分空間は $\theta_I=\theta_I^0$ という制約式で表現できる(θ_I^0 は定数)．一方それに双対な m-座標を $\eta=(\eta_I,\eta_{II})$ とすると，$(\theta_I;\eta_{II})$ という両方の座標系が混在した新たな座標系をとることができ，これも $\mathcal{S}$ 上の点を一意に定める．この混合座標系を用いると，点 $(\theta_I;\eta_{II})$ から，部分空間 $\theta_I=\theta_I^0$ への m-射影は $(\theta_I^0;\eta_{II})$ として陽に表現できる．ただし，混合座標を m-座標だけの表現，e-座標だけの表現に変換するには一般に非線形な方程式を解く必要があることに注意する必要がある．Akaho (2004)では，混合座標系表現を用いて各ステップの最適化が 2 次収束する反復アルゴリズムを導出している．

さて，部分空間の基底 $u_1,u_2,\ldots,u_L$ の表現には任意性がある．部分空間を定めるには，$\mathbb{R}^M$ 中の L 個の互いに直交するベクトルを定めればよいが，一つの部分空間を定める基底の表現は一意的ではない．同じ部分空間を定める直交枠を同一視した L 次元直交枠全体のなす空間は Grassmann 多様体と呼ばれ，その上での最急勾配法は微分幾何の観点から研究されている(Edelman et al., 1998; Fiori, 2001)．

L 本の正規直交ベクトルからなる行列 $\tilde{U}=(u_1,u_2,\ldots,u_L)$ を Grassmann 多様体上の点の表現とみなすと，目的関数 $l(\tilde{U})$ に対する最急勾配は

$$\mathrm{grad}l(\tilde{U})=(I-\tilde{U}\tilde{U}^T)\nabla_{\tilde{U}}l(\tilde{U}) \tag{4.4}$$

で与えられる．ただし，この最急勾配の方向に $\tilde{U}$ を変化させても，正規直交基底であるという制約条件から少しずつずれていくので，何ステップおきかに Schmidt の直交化法などを適用するか，Grassmann 多様体上の測地線に沿って修正を行う必要がある．

4.2 初期解

繰り返し法では必ずしも大域的最適解に収束するとは限らないため，よい初期解を選ぶことが重要な問題となる．最も単純な方法は通常の主成分分析の解を用いて部分空間の基底の初期化を行うことだが，空間に関する計量の情報などを用いればもっとよい初期解が得られる可能性がある．

Fujiki and Akaho (2007), 藤木・赤穂(2009)は，空間の各点で異なるリーマン計量をもつ場合に，一般に $a^Tg(\theta)=b$ の形の超曲面(もとの空間より 1 次元低い曲面)をあてはめるためのユークリッド化と名付けた近似解法を提案している．具体的には，点 θ から低次元空間 $a^Tg(\theta)=b$ へのユークリッド距離の二乗は近似的に

$$d(\theta,a)=\frac{(a^Tg(\theta)-b)^2}{a^T\nabla_\theta g(\theta)G^{-1}(\theta)\nabla_\theta g(\theta)^Ta} \tag{4.5}$$

で表される．

本稿では自己平行な部分空間あてはめを考えるので $g(\theta)=\theta$ だから，$\nabla_\theta g(\theta)=I$ (単位行列)となる．ただし，サンプルに対する総和 $l=\sum_{i=1}^n d(\theta^{(i)},a)$ は a に関して複雑な形をしている

ため，これでも簡単には最小化できない．

そこで，さらに近似を入れて，分子がすべての i について共通となるようにすれば，l は Rayleigh 商の形となり，(一般化)固有値問題として解くことができる．

例えばすべての i について $G(\theta^{(i)})$ を単位行列に置き換えれば，主成分分析そのものに帰着される．また，$G(\theta^{(i)}) \simeq \lambda_i G^*$ のように，i 毎に異なるスカラー λ_i と i に依存しない行列 G^* の積で近似すれば，より近似の度合いを上げることができる．この λ_i と G^* の取り方にはいろいろな可能性が考えられるが，例えば $\lambda_i = |\det G(\theta^{(i)})|^{1/M}$，$G^* = \sum_{i=1}^n G(\theta^{(i)})$ とする手法などが提案されている(Akaho, 2004; 藤木・赤穂, 2009)．

また，本稿では自己平行な部分空間あてはめのみを扱ったが，$a^T g(\theta) = b$ の形をしたより一般の曲面のあてはめについても同じように議論できる．

5. 指数型分布族以外への拡張

指数型分布族は正規分布など多くの基本的な分布を含んでいるが，実際の応用でよく用いられる混合分布や隠れマルコフモデルなどは指数型分布族に含まれていないため，ここではそれらの分布に対する一つのアプローチを紹介する(Akaho, 2008)．

5.1 指数型分布族への埋込

以下では指数型分布族の混合分布

$$p(x) = \sum_{k=0}^{K} \pi_k f_k(x;\xi_k), \quad f_k(x;\xi_k) = \exp(\xi_k^T F_k(x) - \psi_k(\xi_k)), \quad k = 0,\dots,K \tag{5.1}$$

を考える．ここで，$\pi_k \geq 0$，$\sum_{k=0}^K \pi_k = 1$ で，$\{\pi_k\}_{k=0}^K$ の自由度は K であるから，$\pi_1,\dots,\pi_K$ をパラメータとし，π_0 はそれ以外の π_k から決まる関数 $\pi_0 = 1 - \sum_{k=1}^K \pi_k$ とする．このモデルを生成モデルとして解釈すると，まず，π_k に比例する確率で要素分布 f_k をランダムに選び，それから $f_k(x;\xi_k)$ に従って x を生成する．混合指数型分布族自体は指数型分布族ではないが，どの要素分布が選ばれるかを潜在確率変数 $z \in \{0,1,2,\dots,K\}$ として導入すると，(x,z) の同時分布は指数型分布族になる(Wang et al., 2003; Amari, 1995)．具体的には，

$$\begin{aligned} p(x,z) &= \pi_z f_z(x;\xi_z) \\ &= \exp\left[\sum_{k=1}^K \xi_k^T F_k(x)\delta_k(z) + \xi_0^T F_0(x)\left(1 - \sum_{k=1}^K \delta_k(z)\right) + \sum_{k=1}^K \nu_k \delta_k(z) - \psi_*\right] \end{aligned} \tag{5.2}$$

という形の指数型分布族になる． ただし，$\delta_k(z) = 1$ は $z = k$ のとき 1，それ以外で 0 を取る関数で，

$$\nu_k = \log\pi_k - \psi_k(\xi_k) - (\log\pi_0 - \psi_0(\xi_0)), \quad \psi_* = -\log\pi_0 + \psi_0(\xi_0). \tag{5.3}$$

である．

これによって指数型分布族の混合分布を別の指数型分布族の空間に埋め込むことが可能となるが，二つの問題点がある．一つ目は，潜在変数 z の入れ方には置換の自由度があるという問題で，例えば $ap(x) + bq(x)$ という分布は $\pi_0 = a, \pi_1 = b, f_1(x) = p(x), f_2(x) = q(x)$ と埋め込むこともできるが $\pi_0 = b, \pi_1 = a, f_1(x) = q(x), f_2(x) = p(x)$ と埋め込むこともできる． 一つの分布を埋め込むだけならそのうちの任意の埋め込みを選べばよいが，複数の分布を埋め込む場合には互いの位置関係が変化してしまうので適切に選ぶ必要がある．二つ目は，実際の応用場面を考えると混合分布の要素数の異なる混合分布を扱いたい場合もあるが，パラメータ数が変化して

しまうためそのままでは要素数の異なる分布を一つの空間の中では扱うことができないという問題である.

5.2 埋込みアルゴリズム

まず,要素分布の数が等しい場合に適切な埋め込みを行う手法を考える.以下に述べるように,二つの分布であればそれらがダイバージェンスの意味で最も近く配置される埋め込み法を求めることが可能なので,一般の n 個の場合には greedy に分布を追加していくことによって埋め込みを行うという方法が考えられる.

潜在変数を導入した二つの混合分布

$$p_1(x,z)=\alpha_z f_z(x;\xi_z),\qquad p_2(x,z)=\beta_z f_z(x;\zeta_z) \tag{5.4}$$

が与えられたとき,p_1 と p_2 の m-ダイバージェンスは

$$K_m(p_1;p_2)=\sum_{k=0}^{K}\alpha_k\left[K_m(f_k(x;\xi_k);f_k(x;\zeta_k))+\log\frac{\alpha_k}{\beta_k}\right] \tag{5.5}$$

で与えられる.これは対応する成分毎の関数の和に分離した形をしている.従って,グラフマッチングによって最適な対応関係を見つければ最もダイバージェンスの小さい埋め込みを見つけることができる.つまり,p_1,p_2 それぞれの分布の各成分をノードとし,分布間を結んだ二部グラフを考え,p_1 の k 番目と p_2 の k' 番目のノードを結ぶエッジの重みは(5.5)式の

$$w_{kk'}=\alpha_k\left[K_m(f_k(x;\xi_k);f_{k'}(x;\zeta_{k'}))+\log\frac{\alpha_k}{\beta_{k'}}\right] \tag{5.6}$$

に取る.これによって定義された二部グラフの最小重み最大マッチングを求めれば,2 つの分布の m-ダイバージェンスの意味での最適な埋め込み法が得られる.すなわち,

$$\min\sum_{k=0}^{K}\sum_{k'=0}^{K}w_{kk'}x_{kk'},\quad \sum_{k=0}^{K}x_{kk'}=\sum_{k'=0}^{K}x_{kk'}=1,\quad x_{kk'}\geq 0 \tag{5.7}$$

を満たす $x_{kk'}\in\{0,1\}$ を求め,$x_{kk'}=1$ となる p_1 の k 番目と p_2 の k' 番目を対応づければよい(実行可能解は区間 $[0,1]$ にあるが,最適解では $\{0,1\}$ の 2 値になることがわかっている).これは最小費用流問題を解くアルゴリズムによって効率よく解くことができる.また,m-ダイバージェンスと e-ダイバージェンスの定義から,p_1 と p_2 を入れ替えれば e-ダイバージェンスの意味での最適な埋め込み法も得られる.

n 個の分布を埋め込む場合はこのいずれかを greedy に繰り返す方法が提案されているが,e と m のどちらがよいかという議論を含め,より適切な埋め込み法を見つけることも今後の検討課題である.

5.3 コンポーネント数が異なる場合

次に,要素分布の数が異なる場合の埋め込み手法について考える.基本的なアイディアは,少ない要素数でも要素分布 $p(x)$ を $ap(x)+(1-a)p(x)$ のように分割すれば要素数を見かけ上増やすことができるというものである.これは,混合分布のもつ特異性(Fukumizu et al., 2003; Watanabe and Watanabe, 2007)を逆に利用したものである.

しかしながら,この場合,「どの要素分布を分割するか」と「分割したとしてその分割の割合をどれだけにするか」という二つの問題を解く必要がある.本来この二つは相互に絡み合っているが,近似的にこれらの問題を二段階に分割して解く.

まず,「どの要素分布を分割するか」という問題については,前節で述べた要素分布が等しい場合のマッチングを取るところで,1 対 1 のマッチングを 1 対多の対応を見つける問題に拡

張することによって近似的に解くことができる．つまり，(5.7)式を

$$\min \sum_{k=0}^{K}\sum_{k'=0}^{K} w_{kk'}x_{kk'}, \quad \sum_{k=0}^{K} x_{kk'} \geq 1, \quad \sum_{k'=0}^{K} x_{kk'} = 1, \quad x_{kk'} \geq 0 \tag{5.8}$$

と拡張し，(ただし一般性を失うことなく $k \geq k'$ と仮定) $x_{kk'} = 1$ を満たす k が複数ある場合には，p_2 の k' 番目の要素を分割して対応づけることにする．ただし，ここで得られる対応付けは分割の割合を考慮していないので厳密には最適なものではないことに注意する．

上記のアルゴリズムによって，どの要素分布を分割するかが決まってしまえば，最適な分割の重みは m-ダイバージェンス(または e-ダイバージェンス)を最小にするという意味で解析的に求められる(詳細な式の形は Akaho, 2008 を参照)．

さて，以上のアルゴリズムによって指数型分布族の混合分布を指数型分布族の空間に埋め込むことができたので，後は e-PCA または m-PCA を適用することによって次元圧縮を行うことが可能となる．

ここでは混合分布についてのみ紹介したが，隠れマルコフモデルなど離散的な潜在変数をもつ確率モデルなら上記に示したアルゴリズムや近似法はほぼそのまま適用可能である．しかしながら，隠れマルコフモデルでは潜在変数についての可能な組み合わせが想定する系列の長さに関して指数的に増大するため動的計画法を利用するなど計算上の工夫が必要であると考えられる．また，近年ノンパラメトリックベイズなどの枠組みで注目されているモデルでは潜在変数が無限個あるため，そのままの形では適用できず，今後の課題として残されている．

6. 指数分布族上の確率モデル

主成分分析は明示的な生成モデルを持たず，統計的な推測としては解釈できない．そこで，生成モデルに基づいて次元圧縮する手法として因子分析や確率的主成分分析というものが提案されている．今まで述べた e-PCA や m-PCA も主成分分析と同じくそのままでは明示的な生成モデルはもたない．単なる次元圧縮の手段としては問題がないが，ベイズ的な推論を行ったり，確率モデルとしての解釈を行うためには生成モデルを考えたモデル化が有効である．

ここでは Watanabe et al. (2008, 2009) によって導入された e-PCA の確率モデルと，それを発展させた次元圧縮とクラスタリングの同時最適化について紹介する．

6.1 e-PCA の生成モデル

しばらくの間，部分空間を規定するパラメータ U は固定して考えることにする．部分空間上の局所座標 w に対応する $\mathcal{S}$ における e-座標を

$$\tilde{\theta}(w) = \sum_{j=1}^{L} w_j u_j + u_0 \tag{6.1}$$

とおくと，$\mathcal{S}$ 上のデータ点 $\theta^{(i)}$ の m-座標 $\eta(\theta^{(i)})$ は十分統計量とみなせるので，その尤度は

$$p(\eta(\theta^{(i)}) \,|\, w^{(i)}) = \exp(\eta(\theta^{(i)})^T \tilde{\theta}(w^{(i)}) + F_0(\eta(\theta^{(i)})) - \psi(\tilde{\theta}(w^{(i)}))) \tag{6.2}$$

と書ける． ここで，$\theta^{(i)}$ に対応する w の値を $w^{(i)}$ とおいた．上の式は $\theta^{(i)}$ がこの分布に従うただ一つの観測値から生成されたパラメータである場合の式である(2 節の最後の部分を参照)．ちなみに $\theta^{(i)}$ が一般の N 個のデータから得られた十分統計量から推定されたパラメータの場合は exp の中身が N 倍されるのみで本質的には同様に議論できる． ここでは簡単のため $N = 1$ の場合のみ説明する．

w を確率変数とみなし，事前分布 $p(w)$ を定めると，n 個の観測値 $\Theta^n=\{\theta^{(i)}\}$ に対する潜在変数 $W^n=\{w^{(i)}\}_{i=1}^n$ の事後分布は簡単な式変形により

$$p(W^n\,|\,\Theta^n)\propto\exp\left(-\sum_{i=1}^n K_m(\theta^{(i)};\tilde{\theta}(w^{(i)}))\right)\prod_{i=1}^n p(w^{(i)}) \tag{6.3}$$

となることを示すことができる．この式は e-PCA の損失関数である m-ダイバージェンスにマイナスをつけて指数の肩にのせ，W^n についての事前分布をかけた形をしており，e-PCA の確率モデルとみなせる．

ここで，低次元空間でのクラスタリングをモデル化するための事前分布として

$$p(w\,|\,a,V)=\sum_{k=1}^K a_k\delta(w-v_k),\quad \sum_{k=1}^K a_k=1,\quad a_k\geq 0 \tag{6.4}$$

という形の分布を仮定する． ただし，$a=(a_1,\ldots,a_K)^T, V=\{v_k\}_{k=1}^K$ は確率変数で，それぞれ共役事前分布である Dirichlet 分布と指数型分布族

$$p(a;\phi)\propto\prod_k a_k^{\phi-1},\quad p(v_k;\xi,\alpha)=\exp\Big(\xi\{\alpha^T v_k-\psi(\tilde{\theta}(v_k))\}-\Phi_k(\alpha,\xi)\Big) \tag{6.5}$$

を仮定する． ただし，ϕ,ξ,α はハイパーパラメータであり，

$$\Phi_k(\alpha,\xi)=\log\int\exp\Big(\xi\{\alpha^T v_k-\psi(\tilde{\theta}(v_k))\}\Big)\,dv_k \tag{6.6}$$

とおいた．

6.2 クラスタリングアルゴリズム

上で定義したモデルにおいて v_k をクラスタ中心とみなし，v_k に関するベイズ推論を行うことにより低次元空間でのクラスタリングを行う．その計算アルゴリズムの詳細は式が複雑になるのでここでは導出過程の概要を説明する． 詳細については Watanabe et al. (2009) を参照されたい．

(6.4)式は，一種の混合分布の形をしており，そのままでは扱いが困難である．そこで，各サンプルデータ $\theta^{(i)}$ がどのクラスタに属するかを新たに潜在変数として導入する．すなわち，$\theta^{(i)}$ に対応する低次元空間上の局所座標 $w^{(i)}$ はクラスタ中心 $v_1,\ldots,v_K$ のどれかの値を取るが，その添え字を $z^{(i)}$ と置く． すると $w^{(i)}$ について周辺化した $\eta(\theta^{(i)}), z^{(i)}$ の同時分布は

$$p(\eta(\theta^{(i)}),z^{(i)}\,|\,a,V)=a_{z^{(i)}}p(\eta(\theta^{(i)})\,|\,v_{z^{(i)}})=\prod_{k=1}^K a_k p(\eta(\theta^{(i)})\,|\,v_k)^{\delta_k(z^{(i)})} \tag{6.7}$$

という形になる． ここで，$p(\eta(\theta^{(i)})\,|\,v_k)$ は(6.2)式で述べた v_k を定めたときの $\eta(\theta^{(i)})$ の尤度である．

しかしながら，これでも事後分布は計算が困難なため，変分ベイズ法(またはナイーブ平均場近似(Attias, 1999)による事後分布の近似を行う．潜在変数のうち W^n は含まない形になったので，それ以外の潜在変数を a,V と $Z^n=\{z^{(i)}\}$ の二つの組に分け，全体の事後分布をそれぞれの組の分布の積の形

$$p(a,V,Z^n\,|\,\Theta^n)\simeq q(a,V)q(Z^n) \tag{6.8}$$

で近似する． このとき，近似の評価規準として

$$K_m(q(a,V)q(Z^n);p(a,V,Z^n\,|\,\Theta^n)) \tag{6.9}$$

を取り，これが最小となるように $q(a,V)$ と $q(Z^n)$ についての変分を取ると

$$q(a,V) \propto p(a,V) \exp\left(\sum_{i=1}^{n} \mathrm{E}_{q(Z^n)} \left[\log p(\eta(\theta^{(i)}), z^{(i)} \mid a,V) \right] \right) \tag{6.10}$$

$$q(Z^n) \propto \exp\left(\sum_{i=1}^{n} \mathrm{E}_{q(a,V)} \left[\log p(\eta(\theta^{(i)}), z^{(i)} \mid a,V) \right] \right) \tag{6.11}$$

という方程式を得る．これを解くために通常は $q(a,V)$ と $q(Z^n)$ の一方を固定し，上記の方程式に代入して他方の最適解を得るという交互最適化によって局所最適解を求める．このとき，$q(a,V)$ は

$$q(a,V) = q(a) \prod_{k=1}^{K} q(v_k) \tag{6.12}$$

の形に自然に分離し，(6.5)式と同じ形の分布，つまり $q(a)$ は Dirichlet 分布，$q(v_k)$ は 指数型分布族になり，$q(Z^n)$ を固定して $q(a)$，$q(v_k)$ を求める変分ベイズ法の手続きも書き下すことができる(式の詳細は Watanabe et al., 2009, を参照)．

一方，$q(Z^n)$ も $\prod_{i=1}^{n} q(z^{(i)})$ の形に分離することがわかるが，変分ベイズ法の手続きが書き下せるのは，U が空間 $\mathcal{S}$ の全体を張る場合，すなわち全く次元圧縮を行わない場合のみである．我々の興味の対象である次元圧縮を行う場合には一般に困難な多次元数値積分，具体的には(6.6)式の α,ξ に関する微分の計算が必要となってしまう．

そこで Watanabe et al. (2009)では，さらに(6.6)式の exp の中身を v_k について 2 次までのテーラー展開で近似するという Laplace 近似を適用する．これによって低次元空間を固定したもとでの潜在変数の事後分布を(近似的に)計算するアルゴリズムが完成する．

さて，上記のアルゴリズムはクラスタに関する情報，つまりクラスタ中心 V と各データがどのクラスタから生成されたかの $z^{(i)}$ に関する分布を求めたが，各データを低次元化する際には潜在変数 $w^{(i)}$ に関する分布を求めることが必要である．このためには逆に，潜在変数全体の事後分布を a,V について積分消去して $w^{(i)}$ だけの分布を求めることができ，さらに上記の Laplace 近似によって求めた関数を用いて事後平均も計算できることが示されている．

6.3 次元圧縮とクラスタリングの同時最適化

ここまでは部分空間を固定して考えてきたが，ここでは部分空間を最適化することを考える．部分空間の基底に対して完全にベイズ的な扱いをすることは計算量などの点から困難がある．

Watanabe et al. (2009) では以下のような手法を提案している．変分ベイズ法において，各サンプルがどのクラスタに属するかの事後確率 $q(z^{(i)})$ が求められているので，それに関するデータの m-座標系での重み付き平均

$$\nu_k = \frac{1}{n_k} \sum_{i=1}^{n} q(z^{(i)} = k) \eta(\theta^{(i)}) \tag{6.13}$$

とおく．ただし $n_k = \sum_{i=1}^{n} q(z^{(i)} = k)$ である．これが部分空間内でのクラスタ中心の事後平均

$$\bar{v}_k = \int v_k q(v_k) dv_k \tag{6.14}$$

に近づくように部分空間を学習する．近さの規準としては

$$\sum_{k=1}^{K} n_k K_m(\theta(\nu_k); \tilde{\theta}(\bar{v}_k)) \tag{6.15}$$

を取り，最急降下法により U を更新していく．変分ベイズ法とこの手続きを更新することにより次元圧縮とクラスタリングの両方を最適化できる．

なお，モデル選択としてより自然なのは，ベイズ自由エネルギーを最小にするような部分空間を選ぶという方法であろう．変分ベイズ法では，ベイズ自由エネルギー $\mathcal{F}(\Theta^n)$ そのものは計算が困難であるが，これに対する一つの上界である変分自由エネルギー

$$\hat{\mathcal{F}}(q) = \sum_{Z^n} \int q(Z^n) q(a,V) \log \frac{q(Z^n)q(a,V)}{p(a,V)\prod_{i=1}^n p(\eta(\theta^{(i)}), z^{(i)} \mid a, V)} \tag{6.16}$$

は計算可能なのでこれを部分空間を選ぶモデル選択に用いるという手法が考えられる（なお，$\hat{\mathcal{F}}(q)$ と真の自由エネルギーとの差が変分ベイズ法で最小化した m-ダイバージェンスの値に一致する）．

すると，変分自由エネルギーを部分空間の基底について最小化することになるが，最急降下法で導かれた更新式には計算が困難な部分が残ってしまう．(6.15)式で提案した規準はこの変分自由エネルギーの規準の近似とみなすことができ，得られる更新式の形も類似している．

クラスタリングにおいては，適切なクラスタ数 K を決めるのも重要なモデル選択の問題であるが，これに対してはクラスタ数を変えて変分自由エネルギーを計算し，それを最小にするクラスタ数を選ぶことによってモデル選択するという手法が考えられる．

6.4 m-PCA の確率モデル

e-PCA の確率モデルはもとの指数型分布族の尤度から m-ダイバージェンスが出てくることで容易に導かれたが，m-PCA については異なる確率モデルを導入する必要がある．まず，(6.2)式の尤度と類似の以下の関数を考える

$$p_m(\theta^{(i)} \mid w) = \exp(\theta^{(i)T}\tilde{\eta}(w) + \tilde{F}_0(\theta^{(i)}) - \varphi(\tilde{\eta}(w))) \tag{6.17}$$

ここで，

$$\varphi(\eta) = \theta(\eta)^T \eta - \psi(\theta(\eta)) \tag{6.18}$$

であり，$\tilde{\eta}(w)$ は m-座標系の L 次元アファイン部分空間上の点の助変数表現

$$\tilde{\eta}(w) = \sum_{j=1}^{L} w_j u_j + u_0 \tag{6.19}$$

を表し，$\tilde{F}_0$ は正規化条件

$$\int p_m(\theta \mid w) d\theta = 1 \tag{6.20}$$

を満たすように決められる関数とする．

このような $\tilde{F}_0(\theta)$ が存在すれば(6.17)式は $\theta^{(i)}$ の確率密度とみなせる．φ の定義式を代入すると，

$$p(\theta^{(i)} \mid w) = \exp\{-K_e(\eta(\theta^{(i)}); \tilde{\eta}(w)) + \psi(\theta^{(i)}) + \tilde{F}_0(\theta^{(i)})\} \tag{6.21}$$

となり，これを $\theta^{(i)}$ の尤度として w を求める最尤法が e-ダイバージェンスを最小にする w を求める m-PCA と等価になる．w に事前分布を考えれば，e-PCA の場合と同じようにベイズ推論を行うことが可能となる．

上記の解釈が成立するためには $\tilde{F}_0$ が存在する必要があるが，存在条件など理論的な特徴付けは今後の課題として残されている．

ここでは例として，1 次元指数分布

$$p(x \mid \lambda) = \exp(-\lambda x + \log \lambda) \tag{6.22}$$

を考える $(x \geq 0, \lambda > 0)$．この分布において，$\theta = -\lambda$，$\psi(\theta) = -\log(-\theta)$，$\eta = -1/\theta$ ゆえ，$\varphi(\eta) = -\log \eta - 1$ であり，(6.17)式を計算すると $\tilde{F}_0(\theta^{(i)}) = -1$ とすればよいことがわかり，

$$p_m(\theta^{(i)} \mid w) = \exp(\theta^{(i)} \tilde{\eta}(w) + \log \tilde{\eta}(w)) \tag{6.23}$$

となる． これは $\theta^{(i)} < 0$ で定義される指数分布であり，$\eta(\theta^{(i)}) = x^{(i)}$ の対応によって $x^{(i)}$ の分布に変数変換すると，

$$p(x^{(i)} \mid w) = \exp\left(-\frac{\tilde{\eta}(w)}{x^{(i)}} - 2\log x^{(i)} + \log \tilde{\eta}(w)\right) \tag{6.24}$$

という分布モデルとなり，これはもとの指数分布とは異なる分布からの生成モデルを仮定していることを意味している．

7. 関連研究とまとめ

次元圧縮やクラスタリングはデータ解析の基本的手法であり，非常に広範にわたるためそのすべてにわたるレビューをするのは困難である．ここでは著者らの知見に基づいてその主なもの，特に本研究と関連の深いものについて述べる．

離散値に関する主成分分析の拡張は，数量化 III 類（林, 1993）や Correspondence analysis (Benzecri, 1992)などに始まる研究があるが，これらは本質的には潜在的に正規性を仮定したものになっている．

離散確率変数の特別な場合にではあるが，確率分布の幾何的な構造に着目して，ダイバージェンスを距離においた次元圧縮法として提案されたのが pLSA (probabilistic Latent Semantic Analysis)であり(Hofmann, 1999)，テキストマイニングなどで広く用いられている．このほか一般化線形モデルに基づいた二値データの PCA の提案(Schein et al., 2003)や多項分布(Buntine, 2002)の場合の研究がある．このほか NMF (Nonnegative Matrix Factorization)なども確率モデルに基づいた次元圧縮法の一種と考えられる．

これらを包含する形で指数型分布族の空間での情報幾何に基づく PCA に一般化した最初の研究は Collins et al. (2002) である．Collins らの研究は，本稿で説明した双対的な次元縮小法のうち e-PCA のみで，かつ，観測値が確率変数の実現値の場合のみである．

PCA に確率モデルを導入するという話はもともと因子分析として広く研究されてきた．確率的 PCA として知られる最近の研究(Tipping and Bishop, 1999; Bishop, 1999a, 1999b; Oba et al., 2003a)も枠組みとしては因子分析と共通する部分が多く，最近の進展は主に最適化アルゴリズムの高速化にある．Collins らの提案した情報幾何的な枠組みに確率モデルを導入した研究に Sajama and Orlitsky (2004)がある．

一方，クラスタリングについては，離散データのクラスタリング(Dhillon et al., 2003)をはじめさまざまな研究があり，Collins らの枠組みからクラスタリング法を研究したものとして Banerjee et al. (2005)がある．本稿で取り上げたクラスタリングと次元圧縮の同時最適化は比較的新しい話題である．Ding and Li (2007)は伝統的な多変量解析法である判別分析と k-means 法とを同時最適化する手法を提案した．松本 et al. (2008a, 2008b)は正規性の仮定のもとで次元圧縮とクラスタリングを同時最適化するアルゴリズムを変分ベイズ法を用いて構成した．本稿で解説した指数分布族の空間での同時最適化アルゴリズム(Watanabe et al., 2008, 2009)はその延長線上に位置づけられる．

次元圧縮とクラスタリングの同時最適化と関連の深いモデルに混合因子分析モデルがある (Ghahramani and Beal, 2000; Oba et al., 2003b). これは, クラスタ毎にそれぞれ別の低次元構造を見つけるというモデルであり, 局所的に低次元構造が異なる場合にはより柔軟なモデルと言える. しかしながら, 高次元でデータ数が比較的少ない場合には全体を同じ低次元空間に射影する本稿の手法の方が次元の呪いを受けにくいと考えられ, 場合に応じて適切な方を選ぶ必要があるだろう. なお, 学習アルゴリズムの導出のしやすさという意味では混合因子分析モデルの方が単純であり, 次元圧縮とクラスタリングの同時最適化はまだ未知の問題を含むなど多くの研究の余地が残されている手法である.

本稿以外の次元圧縮法について重要と思われるものに二つ言及しておく. 一つは外れ値が含まれるようなデータについても頑健性を持たせたロバスト PCA (Higuchi and Eguchi, 1998) である. もう一つはカーネル法の一種であり, 非線形の低次元構造を抽出するカーネル PCA (Schöelkopf et al., 1998) である. 多次元尺度構成法などとも関係し, さまざまな関連研究がある (赤穂, 2008; 藤木・赤穂, 2009; Tenenbaum et al., 2000; Fletcher et al., 2004). ロバスト性やカーネル法による非線形化と, 本稿で述べた指数分布族 PCA やそのベイズ的な拡張との融合も興味深い研究対象である.

さて近年情報幾何は, 様々な機械学習の手法を幾何学的に見通しよく解釈することによって, 理論解析やアルゴリズムの一般化に有効に働いてきた (例えば Tanaka, 2001; Ikeda et al., 2002; Murata et al., 2004 など). 本稿で解説したのもそのような事例の一つと考えられ, 主成分分析を指数型分布族の空間に対して自然に一般化することができた. 現在のところ用いているのは情報幾何の基本的な性質のみなので, 今後はより進んだ理論を用いて, 本稿で述べた未解決の問題が解決されることが期待される.

謝 辞

本稿で紹介した研究のうち, 著者らの共同研究者である大町真一郎氏, 藤木淳氏, 松本有央氏, 福水健次氏, 菅生康子氏に感謝する. 本研究は文部科学省科学研究費補助金 特定領域研究「統計力学の深化と発展」, 基盤研究 (C) 19500136, 若手研究 (スタートアップ) 20800012 の補助を受けて行われた.

参 考 文 献

Aggarwal, C. and Yu, P. S. (eds.) (2008). *Privacy-preserving Data Mining*: *Models and Algorithms*, Springer, New York.

Agrawal, R. and Srikant, R. (2000). Privacy-preserving data mining, *Proceedings of the ACM, Special Interest Group on Management of Data* (*SIGMOD*), 439–450.

赤穂昭太郎 (2003). 指数分布族の空間における平坦な部分空間あてはめ, 情報論的学習理論ワークショップ (*IBIS2003*), 131–136.

Akaho, S. (2004). The e-PCA and m-PCA: Dimension reduction by information geometry, *Proceedings of International Joint Conference on Neural Networks* (*IJCNN*), 129–134.

Akaho, S. (2008). Dimension reduction for mixtures of exponential families, *International Conference on Artificial Neural Networks* (*ICANN*), 1–10.

赤穂昭太郎 (2008). 『カーネル多変量解析—非線形データ解析の新しい展開』, 確率の情報と科学, 岩波書店, 東京.

Amari, S. (1985). *Differential Geometrical Methods in Statistics*, Springer-Verlag, Berlin.

Amari, S. (1995). Information geometry of the EM and em algorithms for neural networks, *Neural*

Networks, **8**(9), 1379–1408.

Amari, S. (2001). Information geometry on hierarchy of probability distributions, *IEEE Transactions on Information Theory*, **47**(5), 1701–1711.

Amari, S. and Nagaoka, H. (2000). *Methods of Information Geometry*, AMS and Oxford University Press, New York.

Attias, H. (1999). Inferring parameters and structure of latent variable models by variational Bayes, *Proceedings of Uncertainty in Artificial Intelligence* (*UAI*), 21–30.

Banerjee, A., Merugu, S., Dhillon, I. and Ghosh, J. (2005). Clustering with Bregman divergences, *Journal of Machine Learning Research*, **6**, 1705–1749.

Benzecri, J. (1992). *Correspondence Analysis Handbook*, Marcel Dekker, New York.

Bishop, C. M. (1999a). Bayesian PCA, *Advances in Neural Information Processing Systems* (*NIPS*) *11*, 382–388.

Bishop, C. M. (1999b). Variational PCA, *Proceedings of International Conference on Artificial Neural Networks* (*ICANN*), 509–514.

Buntine, W. (2002). Variational extensions to EM and multinomial PCA, *Proceedings of European Conference on Machine Learning* (*ECML*), Lecture Notes in Artificial Intelligence, Vol. 2430, 23–34, Springer-Verlag, London.

林知己夫(1993).『数量化ー理論と方法』, 朝倉書店, 東京.

Chong, C.-Y. and Kumar, S. (2003). Sensor networks: Evolution, opportunities and challenges, *Proceedings of the IEEE*, **91**, 1247–1256.

Collins, M., Dasgupta, S. and Schapire, R. (2002). A generalization of principal component analysis to the exponential family, *Advances in Neural Information Processing Systems* (*NIPS*) *14*, 617–624.

Dhillon, I., Mallela, S. and Kumar, R. (2003). A divisive information-theoretic feature clustering algorithm for text classification, *Journal of Machine Learning Research*, **3**, 1265–1287.

Ding, C. and Li, T. (2007). Adaptive dimension reduction using discriminant analysis and k-means clustering, *Proceedings of International Conference on Machine Learning* (*ICML*), 521–528.

Edelman, A., Arias, T. and Smith, S. (1998). The geometry of algorithms with orthogonality constraints, *SIAM Journal on Matrix Analysis and Applications*, **20**(2), 303–353.

Fiori, S. (2001). A theory for learning by weight flow on Stiefel-Grassmann manifold, *Neural Computation*, **13**(7), 521–531.

Fletcher, P., Lu, C., Pizer, S. and Joshi, S. (2004). Principal geodesic analysis for the study of nonlinear statistics of shape, *IEEE Transactions on Medical Imaging*, **23**(8), 995–1005.

藤木　淳, 赤穂昭太郎(2009). 一次元正規分布のなす空間への曲線あてはめ, 情報論的学習理論ワークショップ(*IBIS2009*), 68–73.

Fujiki, J. and Akaho, S. (2007). Spherical PCA with Euclideanization, *Proceedings of Subspace 2007* (*Asian Conference on Computer Vision* (*ACCV*) 2007 *Workshop*), 61–68.

Fukumizu, K., Akaho, S. and Amari, S. (2003). Critical lines in symmetry of mixture models and its application to component splitting, *Advances in Neural Information Processing Systems* (*NIPS*) *15*, 856–872, MIT Press, Cambridge.

Ghahramani, Z. and Beal, M. J. (2000). Variational inference for Bayesian mixtures of factor analysers, *Advances in Neural Information Processing Systems* (*NIPS*) *12*, 449–455, MIT Press, Cambridge.

Higuchi, I. and Eguchi, S. (1998). The influence function of principal component analysis by self-organizing rule, *Neural Computation*, **10**(6), 1435–1444.

Hofmann, T. (1999). Probabilistic latent semantic analysis, *Proceedings of Uncertainty in Artificial Intelligence* (*UAI*), 289–296.

Ikeda, S., Tanaka, T. and Amari, S. (2002). Information geometrical framework for analyzing belief propagation decoder, *Advances in Neural Information Processing Systems* (*NIPS*) *14*, 407–414, MIT Press, Cambridge.

Kumar, A., Kantardzic, M. and Madden, S. (2006). Distributed data mining: Framework and implementations, *IEEE Internet Computing*, **10**, 15–17.

松本有央，赤穂昭太郎，福水健次，菅生康子，岡田真人(2008a). IT 野ニューロン集団の時間相関を取り入れたクラスタリング，信学技法 *NC*, **105**(658), 7–12.

松本有央，赤穂昭太郎，菅生康子，岡田真人(2008b). 側頭葉ニューロン集団活動のクラスタリングと次元圧縮の同時最適化，日本神経回路学会第 16 回全国大会公演論文集，18–19.

Murata, N., Takenouchi, T., Kanamori, T. and Eguchi, S. (2004). Information geometry of U-boost and Bregman divergence, *Neural Computation*, **16**(7), 1437–1481.

Oba, S., Sato, M. and Ishii, S. (2003a). Prior hyperparameters in Bayesian PCA, *Proceedings of International Conference on Artificial Neural Networks* (*ICANN*), 271–279.

Oba, S., Sato, M. and Ishii, S. (2003b). Variational Bayes method for mixture of principal component analyzers, *Systems and Computers in Japan*, **34**(11), 55–66.

Ohara, A. (1999). Information geometric analysis of an interior point method for semidefinite programming, *Geometry in Present Day Science* (eds. O. Barndorff-Nielsen and E. V. Jensen), 49–74, World Scientific, Singapore.

Sajama and Orlitsky, A. (2004). Semi-parametric exponential family PCA, *Advances in Neural Information Processing Systems* (*NIPS*) *16*, 1177–1184.

Schein, A., Saul, L. and Ungar, L. (2003). A generalized linear model for principal component analysis of binary data, *Proceedings of the Ninth International Workshop on AI & Statistics*, 14–21.

Schölkopf, B., Smola, A. and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem, *Neural Computation*, **10**, 1299–1319.

Tanaka, T. (2001). Information geometry of mean-field approximation, *Advanced Mean Field Methods— Theory and Practice* (eds. M. Opper and D. Saad), 259–273, MIT Press, Cambridge.

Tenenbaum, J., de Silva, V. and Langford, J. (2000). A global geometric framework for nonlinear dimensionality reduction, *Science*, **290**, 2319–2323.

Tipping, M. and Bishop, C. (1999). Probabilistic principal component analysis, *Journal of the Royal Statistical Society, Series B*, **61**(3), 611–622.

Wang, S., Schuurmans, D., Peng, F. and Zhao, Y. (2003). Learning mixture models with the latent maximum entropy principle, *Proceedings of the Twentieth International Conference on Machine Learning* (*ICML-2003*), 784–791.

Watanabe, K. and Watanabe, S. (2007). Stochastic complexities of general mixture models in variational Bayesian learning, *Neural Networks*, **20**(2), 210–219.

Watanabe, K., Akaho, S. and Okada, M. (2008). Clustering on a subspace of exponential family using variational Bayes method, *Proceedings of Worldcomp2008/Information Theory and Statistical Learning*, 10–16.

Watanabe, K., Akaho, S., Omachi, S. and Okada, M. (2009). Variational Bayesian mixture model on a subspace of exponential family distributions, *IEEE Transactions on Neural Networks*, **20**(11), 1783–1796.

Proceedings of the Institute of Statistical Mathematics Vol. 58, No. 2, 167–183 (2010) 183

Data Analysis Method on Space of Exponential Family Distributions

Shotaro Akaho[1], Kazuho Watanabe[2] and Masato Okada[3]

[1]The National Institute of Advanced Industrial Science and Technology
[2]Nara Institute of Science and Technology
[3]Graduate School of Frontier Sciences, The University of Tokyo

Principal component analysis (PCA) is widely used for dimension reduction, but it is only optimal for Gaussian distributed data and cannot extract a desired lower dimensional structure for non-Gaussian data. In this paper, we review research about dimension reduction for data generated from an exponential family or are given as parameters of an exponential family from an information geometrical point of view. As an extention of coventional PCA, we propose dually coupled methods for dimension reduction called e-PCA and m-PCA, in which the affine subspace of a dually coupled autoparallel coordinate system is extracted so as to minimize the sum of Kullback-Leibler divergence. We also consider the treatment for a mixture distribution that does not belong to an exponential family. The basic idea is to embed the mixture distribution into the exponential family. Further, we introduce a probabilistic model for the proposed framework and derive a clustering algorithm constrained on a lower dimensional subspace. The variational Bayes method and the Laplace approximation technique are applied in order to obtain a tractable computation time.

Key words: Principal component analysis, information geometry, dimension reduction, clustering, Bayesian estimation.