統計数理(2006)
第 54 巻 第 2 号 375–403

特集「予測と発見」
[総合報告]

予測と発見：ゲノムデータ解析のための統計的方法を目指して

江口 真透†

(受付 2006 年 2 月 1 日；改訂 2006 年 4 月 10 日)

要　　旨

最尤法の欠点を補うために最小ダイバージェンス法のクラスが提示されているが，その統計的性能について総合報告する．ゲノムデータ解析のための統計的方法の基本課題から SNP，マイクロアレイ，プロテオームを含む幾つかのデータの統計解析の問題点が考察され，その一つのアプローチとして最小ダイバージェンス法の適用について紹介する．

キーワード：U-ダイバージェンス，U-モデル，U-ロス関数，遺伝子発現，ロバスト，情報幾何．

1. はじめに

データを取り巻く環境の急激な変化が指摘されて久しい．世の中の IT 化の普及と進展に伴っていわゆるネットワーク社会が形成され，それに伴いデータの質と量ともに大きな様変わりが遂げられた．不可能と思われていたような大規模のデータが生成され，それまでは想定されないような異なる質のデータの統合が行われている．従来は文章データ，音声データ，画像データ，映像データなどの解析のために，それぞれの文脈でそれぞれの分野で固有の方法論が開発される傾向にあった．それは，それぞれの専門化がそれぞれの展開に有用だったからだ．しかし現在ではデータの広範囲の共有化や階層化によって，より普遍的な方法論が要請されている．互いにネットワークで結ばれた異なるデータの集合は，それぞれの固有の方法論を展開するよりも一つの統一的な視点からそれぞれの方法論を統合する方がより有機的な進展が可能である．要するに，本質的には現実のデータは‘一つ’なので，従来の細分化された専門分野で棲み分けていた形態は役に立たなくなっているのである．

このような激しい変化を遂げつつある社会において，統計学も新しい局面を開くことが期待されている．ここに統計学の新しい使命がある．早急に新しい形式のデータにも統計的な考えを普遍化する必要がある．この過程の中で，近い将来，統計学，人工知能，機械学習，コンピュータ科学などの中で共有できる部分は一つの新しい分野として統合されるだろう．(統計的学習については Vapnik (1995), Hastie et al. (2001)を参照.)その中で統計学の特徴を明らかにし新分野への構築のためにも積極的に大きな役割を担って行かなければならない．2005 年度より発足した予測発見戦略研究センターの遺伝子多様性解析グループでは急速に進展を遂げている分野の一つであるゲノム科学からの多様なデータの学習と推論のための新しい方法論を築くことを目指している．このことが統計学の現代的な要請に答える一助になることを願ってい

† 統計数理研究所：〒106–8569 東京都港区南麻布 4–6–7

る．いろいろな事柄が未だ準備段階であり，近い将来像さえ描ききれていないが，幾つかの研究方向を調和させながら，統計学の進展に繋がる新分野の開拓を長期目標に，現状の研究の成果と今後の計画について論説したい．

本総合報告は次の構成に従う．2 節では，最尤法を含む推定法のクラスである最小ダイバージェンス法を導入しそのクラスの統計的な性質を Γ-ミニマックス性，ロバストネス，局所学習の統合の観点から考察する．3 節は，前節の一般的な考察から具体的に主成分分析と独立性分析への適用を論じ，4 節では，統計的パタン認識の適用を解説する．5 節において，それまでに紹介された統計方法のゲノムデータへの応用が，特に SNP データ，プロテオームデータ，DNA マイクロアレイデータに焦点を当て，考察された．最後に 6 節で，予測と発展の観点から今後の展開が展望された．

2. 最小ダイバージェンス法

統計学の基本ツールに最尤法とベイズ法がある．その提案と統計的な解明の基礎部分は，節 1 で触れられたデータの大きな変容より以前に成されている．従って，必ずしも最尤法とベイズ法では現在の大規模データの解析は想定されてないことから生じる問題が現れてきている．この節では，最尤法の優れた点と欠点について復習する．次に最尤法を含む推定法の広いクラスを紹介して，その統計的な性質を総合報告する．そのクラスの中で推定関数の不偏性，推定量の漸近効率，ロバストネス，局所学習の統合性について考察し，主成分分析，独立成分分析，判別分析への適用について再考察し，データの現代化に対応する方法について検討したい．

2.1 最尤推定の光と影

フィッシャーが提案した最尤法は統計学の中で発展し，20 世紀の輝かしい科学の遺産の一つとして数え挙げられるだろう．これ程までに，ありとあらゆる形式のデータに膨大な適用が繰り返されている科学的方法は他には例があるだろうか．一般に最尤推定量は，次の良い性質を持つ．

1. データの 1 対 1 変換に不変である．
2. パラメータ変換に正確に共変的である．
3. 漸近一致推定量である．
4. 漸近有効推定量である．

さらにその極めて優れた点を示す数理が，指数型分布族のモデルにおいて発揮される．指数モデルの諸性質については Barndorff-Nielsen (1978) に詳細な考察がある．最尤推定量は

5. 最小十分統計量である．
6. 期待値パラメータに対して最小分散不偏推定量である．

これは，ガウスの最小 2 乗法を拡張するものである． 最小 2 乗法は正規分布の平均ベクトルに線形モデルを仮定するとき，ユークリッド幾何の線形射影で特徴付けられる．一方で最尤法は指数型分布族のモデルの上では共役凸幾何の m-射影によって特徴付けられる (Amari, 1985 及び Amari and Nagaoka, 2000 を参照)．

この最尤推定の最適性はモデルの正確さを反映する点では唯一，優れた方法であろう．しかし，この最適性は実に壊れやすいものであることが指摘され，改良が議論され統計学はさらに進展した． 最尤推定の欠点は主に次の点に挙げられる．

A. データ分布のモデルからの微小な乖離に対してロバストでない．

いとも容易く，モデルの近傍で 3 の漸近一致性が不連続的に崩れる．最尤法に代わってロバストな M-推定量が Huber (1981)によって提案されている． これは最尤推定そのものを改良するのではなくて仮定されたモデルを変更するアプローチといえる． たとえば正規分布の平均の推定において，正規分布ではなく，両側指数分布を考えると，最尤推定量は標本平均ではなく標本中央値となる．このように最尤推定であることは固定して，その代わりにモデルを変える発想である．しかし平均パラメータ(位置パラメータ)を超えて，一般に仮定されたモデルを変更することは困難である．ここに M-推定の限界がある．2 番目の欠点は

B. データに過剰適合する．

よく知られているように最尤法はデータを過剰に学習する傾向がある．例えば，正規混合モデルにおいて，分散が異なる場合では尤度は無限大に発散する．要するに，最尤法とは，モデルを正確に映す鏡だと思えば，良い性質 1−6 と悪い性質 A と B は表裏一体の表現であって，矛盾するものでは全くない． このように，最尤推定の評価を巡って生じた光と影の再確認の下で，以下の考察では推定法の広いクラスを導入して，その統計的な性質について紹介する．

2.2 ダイバージェンスのクラス

最尤法の原理は非常に簡明な関数の持つ特別な性質の上に構築されている．一つは指数関数で，もう一方では対数関数である．指数関数は指数モデルを定義し，対数関数は対数尤度を定義する．言うまでもなく同時に次のように共役な凸関係で結ばれている．

$$\log(s) = \operatorname*{argmin}_{t\in\mathbf{R}}\{\exp(t) - ts\} \tag{2.1}$$

この共役な凸関係は指数モデルと対数尤度の背後にあるクルバック・ライブラーダイバージェンスを考察するとより深い理解が得られることを紹介する．この考察を通して $\exp(t)$ と逆関数 $\log(t)$ の対から一般の単調増加関数 $u(t)$ と，その逆関数 $\xi(u)$ の対へ変更することによって指数モデルと対数尤度は U-モデルと U-ロスに自然に変更できることを主張する．

データ空間 $\mathcal{Z}$ 上の考えうる全ての関数の全体を $\mathcal{F}$ とおく．$\mathcal{Z}$ 上のある σ-有限な測度 Λ に関して有限なマスを持つ非負関数を，$\mathcal{M} = \{\mu : \langle\mu\rangle < \infty\}$ と定める．ここで $\langle\mu\rangle$ は Λ に関する積分 $\int_{\mathcal{Z}} \mu(z)\Lambda(dz)$ を表す．更に部分集合 $\mathcal{P} = \{p \in \mathcal{M} : \langle p\rangle = 1\}$ は確率密度関数の全体となっている．一般に $\mathcal{M}\times\mathcal{M}$ 上で定義された関数 D が距離の第一公理を満たすとき D を $\mathcal{M}$ 上のダイバージェンス関数，あるいはコントラスト関数と呼ぶ．すなわち，$D(\mu,\nu) \geq 0$ でかつ $D(\mu,\nu) = 0 \Longleftrightarrow \mu = \nu$ (a.e.Λ) を要請する．

関数空間 $\mathcal{M}$ の上にダイバージェンスを実数値関数 $U(t)$ から生成しよう．$\mathbf{R}$ 上の凸関数 $U(t)$ を採ってくれば，空間 $\mathcal{F}$ の任意の f, g に対して

$$d_U(f,g) = \big\langle U(f) - U(g) - U'(g)(f-g)\big\rangle$$

は U が凸関数であるという仮定から d_U は空間 $\mathcal{F}$ 上で距離の第一公理を満たす．これより，写像 $\varphi : \mathcal{M} \to \mathcal{F}$ を採ってくると，$d_U(\varphi(\mu),\varphi(\nu))$ と定義すれば，$\mathcal{M}$ 上のダイバージェンスが構成できる．これをブレグマン・ダイバージェンスと呼ぶことがある．凸関数 $U(t)$ が非負の導関数 $u(t) = U'(t)$ を持つと仮定する．このとき $u(t)$ の逆関数 $\xi(u)$ は，自然に $\mathcal{F}$ から $\mathcal{M}$ への写像と見られる．従って $\varphi = \xi$ と選んだとき

$$\begin{aligned} D_U(\mu,\nu) &= d_U(\xi(\mu),\xi(\nu)) \\ &= \big\langle U(\xi(\nu)) - U(\xi(\mu))\big\rangle - \big\langle \mu, \xi(\nu) - \xi(\mu)\big\rangle \end{aligned} \tag{2.2}$$

を U-ダイバージェンスと呼ぶ．ここで $\langle A, B\rangle = \langle AB\rangle$ を表す．このとき，定義から $u(\varphi(\mu)) = \mu$ となることが U-ダイバージェンスの豊かな応用を生むことの本質的な理由となる．次に U-ク

ロスエントロピー

$$L_U(\mu,\nu)=\big\langle U(\xi(\nu))\big\rangle-\big\langle\mu,\xi(\nu)\big\rangle \tag{2.3}$$

と U**-エントロピー**

$$H_U(\mu)=L_U(\mu,\mu)=\big\langle U(\xi(\mu))\big\rangle-\big\langle\mu,\xi(\mu)\big\rangle \tag{2.4}$$

を定めると $D_U(\mu,\nu)=L_U(\mu,\nu)-H_U(\mu)$ **となることから**

$$L_U(\mu,\nu)\geq H_U(\mu) \tag{2.5}$$

が成立する．

生成関数 $U(t)=\exp(t)$ **を考えると** $u(t)=\exp(t)$**，** $\xi(u)=\log(u)$ **から対応する** U**-ダイバージェンスは**

$$D_{\mathrm{KL}}(\mu,\nu)=\big\langle\nu-\mu\big\rangle-\big\langle\mu,\log(\nu)-\log(\mu)\big\rangle \tag{2.6}$$

となり，クルバック・ライブラーのダイバージェンスと呼ばれる．通常は D_{KL} **は，確率密度関数全体** $\mathcal{P}$ **上に制限して利用されるので(**2.6**)の右辺の初めの項は打ち消しあうことに注意する．次の典型例は，生成関数が**

$$U_\beta(t)=\frac{1}{\beta+1}(1+\beta t)^{\frac{\beta+1}{\beta}}$$

で与えられる場合である．これより

$$u_\beta(t)=(1+\beta t)^{\frac{1}{\beta}},\quad \xi_\beta(u)=\frac{u^\beta-1}{\beta},\quad U_\beta(\xi_\beta(u))=\frac{u^{\beta+1}}{\beta+1}$$

に注意すると U**-ダイバージェンスは**

$$D_\beta(\mu,\nu)=\frac{\big\langle\nu^{\beta+1}-\mu^{\beta+1}\big\rangle}{\beta+1}-\frac{\big\langle\mu,\nu^\beta-\mu^\beta\big\rangle}{\beta}$$

となり，ベータダイバージェンスと呼ぶ．特に $\lim_{\beta\downarrow0}U_\beta=\exp$ **より，** $\lim_{\beta\downarrow0}D_\beta=\mathrm{KL}$ **となるので** $\beta=0$ **のときベータダイバージェンスは** KL **ダイバージェンスに帰着されることが云える．また，** L_2 **ノルムを** $\|\ \|$ **と書くとき** $\beta=1$ **のときは** $D_\beta(\mu,\nu)=\frac{1}{2}\|\mu-\nu\|^2$ **となる．これに基づくロバスト推定については** Scott (2001) **が参照される．データ解析のデータの前処理として，各々の成分に対して対数変換を含む** Cox-Box **のベキ変換が行われるときがある．このベキ変換は，ベータダイバージェンスを定義するときの** ξ_β **に一致する．この文脈で云えば，クルバック・ライブラー ダイバージェンスは頻度を対数変換して得られるもの，ベータダイバージェンスは頻度をベキ変換して得られるものである．後のディスカッションにおいて一般の統計モデルの推測から，主成分分析，独立成分分析までロバストネスの観点から研究が紹介されるだろう．**

一方で α**-ダイバージェンス(**cf. Amari, 1985**)は** $\alpha=\pm1$ **以外ではこのクラスに入らない．** U**-ダイバージェンスの一番の特徴は，それを構成する** U**-クロスエントロピー** $L_U(\mu,\nu)$ **が** μ **に関して線形であることである．このことが，後で議論されるように** μ **をデータの従う分布に選んだときデータに関して自然な量が導かれることを可能にする．一般には** α**-ダイバージェンスはこの性質を享受していない点に注意する．**

2.3 U-ダイバージェンスの情報幾何

関数空間 $\mathcal{M}$ **の中に有限個のパラメータで指定されるモデル**

$$N=\{\mu(\boldsymbol{z},\boldsymbol{\theta}):\boldsymbol{\theta}\in\Theta\}$$

を考えよう．ここでモデル N を座標 $\theta=(\theta^1,\dots,\theta^d)$，座標空間 Θ を持つ d 次元微分可能多様体と見られる場合を考えよう．統計モデルのリーマン空間として考察は Rao (1945)から始まった．そのための滑らかさの仮定，例えばモデル関数 $\mu(z,\theta)$ に対して，積分記号 $\int_{\mathcal{Z}}\cdot\, d\Lambda(z)$ の下での微分可能性を仮定する．一般にダイバージェンス D に対して定義域を $N\times N$ へ制限することによって D は多様体 N 上にリーマン計量 $g^{(D)}$ 及び 2 つの線型接続 $\nabla^{(D)}$ と ${}^*\nabla^{(D)}$ を次のように連想する(Eguchi, 1983; Eguchi, 1992 を参照)．多様体 N のベクトル場 X,Y に対して，

$$g^{(D)}(X,Y)(\mu)=-D(X|Y)(\mu)\quad(\forall\mu\in N)$$

と定める．ここで，記号 $D(X|Y)$ は，一般に，ベクトル場 $X,Y,\dots,Z,W,\dots$ に対して

$$D(X,Y,\dots|Z,W,\dots)(\mu)=X_\mu Y_\mu\cdots Z_\nu W_\nu\cdots D(\mu,\nu)|_{\nu=\mu}$$

を表す．ベクトル場 $\nabla^{(D)}_X Y$ と ${}^*\nabla^{(D)}_X Y$ を任意のベクトル場 Z に対して

$$g^{(D)}(\nabla^{(D)}_X Y,Z)=-D(XY|Z),\quad g^{(D)}({}^*\nabla^{(D)}_X Y,Z)=-D(Z|XY)$$

を満たすと定める．リーマン計量 $g^{(D)}$ の非退化性から 2 つの線型接続 $\nabla^{(D)}$ と ${}^*\nabla^{(D)}$ が一意に定義できることに注意する．リーマン計量 $g^{(D)}$ との関係として，$\bar\nabla^{(D)}=\frac{1}{2}(\nabla^{(D)}+{}^*\nabla^{(D)})$ は $g^{(D)}$ に関するリーマン接続であることが示される．この意味で $\nabla^{(D)}$ と ${}^*\nabla^{(D)}$ は共役であるという．

この公式を使って(2.2)で定義した U-ダイバージェンス D_U が導く 3 つの幾何量 $g^{(U)},\nabla^{(U)},{}^*\nabla^{(U)}$ は

$$g^{(U)}(X,Y)(\mu)=\big\langle X\mu,Y\xi(\mu)\big\rangle,$$
$$g^{(U)}(\nabla^{(U)}_X Y,Z)(\mu)=\big\langle XY\mu,Z\xi(\mu)\big\rangle,\quad g^{(U)}({}^*\nabla^{(U)}_X Y,Z)(\mu)=\big\langle Z\mu,XY\xi(\mu)\big\rangle.$$

で与えられる．ここで $\mu\in N$．このように U-ダイバージェンスの導く 3 つの幾何量はモデル N と ξ にのみに依存する積分表現で与えられる．ここで ξ は U の導関数の逆関数であった．KL ダイバージェンスは $U=\exp$ で，$\xi=\log$ なので，情報計量 g，m-接続 ∇，e-接続 ∇^* に対して

$$(g^{(\exp)},\nabla^{(\exp)},{}^*\nabla^{(\exp)})=(g,\nabla,\nabla^*)$$

と還元される．

一般の U-ダイバージェンスでは，モデル N が親空間 $\mathcal{M}$ で平坦であれば $\nabla^{(U)}$-平坦である．空間 $\xi(\mathcal{M})$ の中で $\{\xi(\mu):\mu\in N\}$ が平坦であるとき，モデル N は ${}^*\nabla^{(U)}$-平坦である．このことを使って次の小節で U-モデルを導入しよう．ここで最も注目すべきことは任意の U に対して $\nabla^{(U)}$ は恒等的に ∇ に等しいことである．一方で ${}^*\nabla^{(U)}$ は $U=\exp$ のときに限り e-接続 ∇^* に等しくなる．

小節 2.2 で定義された U-ダイバージェンス(2.2)の別ルートの導出は $\psi(f)=\langle U(f)\rangle$ と定義される U-ポテンシャル汎関数の凸解析から与えられる（Eguchi, 2005 を参照）．定義から $\psi(f)$ は凸汎関数で，その共役凸汎関数は

$$\psi^*(\mu)=\max_{f\in\mathcal{F}}\{\langle\mu,f\rangle-\psi(f)\}$$

が連想される．このとき，$f^*=\mathrm{argmax}_{f\in\mathcal{F}}\{\langle\mu,f\rangle-\psi(f)\}$ は $\mu=u(f^*)$ または $f^*=\xi(\mu)$ の関係で結ばれる．従って，この凸関係から $\mathcal{F}\times u(\mathcal{F})$ 上に自然なダイバージェンス

$$D(f,\mu)=\psi^*(\mu)-\langle\mu,f\rangle+\psi(f)$$

が導かれる．実際には，これは U-ダイバージェンスに他ならない．すなわち，$D_U(\mu,\nu) = D(\xi(\nu),\mu)$ が成立する．この双対構造が自然に上で考察された $\nabla^{(U)}$ と ${}^*\nabla^{(U)}$ の双対性を誘導している．

2.4 最小 U-ダイバージェンス推定量

確率密度関数の空間 $\mathcal{P}$ の有限次元モデル $M = \{p(\boldsymbol{z};\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ を考えよう．データの基礎分布 p は必ずしもモデル M の中にあると仮定しないが，何らかの意味で M によって近似できるとしよう．このとき U-ロス関数を

$$\ell_U(\boldsymbol{\theta}, p) = L_U(p, p(\cdot,\boldsymbol{\theta})) = b_U(\boldsymbol{\theta}) - E_p\{\xi(p(\boldsymbol{z};\boldsymbol{\theta}))\} \tag{2.7}$$

と定める．ここで L_U は(2.3)で定義された U-クロスエントロピー，E_p は密度関数 p に関する期待値，$b_U(\boldsymbol{\theta}) = \langle U(\xi(p(z;\boldsymbol{\theta})))\rangle$ とする．データ $z_1,\ldots,z_n$ が与えられたとき，経験 U-ロス関数は

$$\ell_U^{(\mathrm{emp})}(\boldsymbol{\theta}) = b_U(\boldsymbol{\theta}) - \frac{1}{n}\sum_{i=1}^{n} \xi(p(\boldsymbol{z}_i;\boldsymbol{\theta})) \tag{2.8}$$

となり，$\hat{\boldsymbol{\theta}}_U = \mathrm{argmin}_{\boldsymbol{\theta}\in\Theta} \ell_U^{(\mathrm{emp})}(\boldsymbol{\theta})$ を最小 U-ダイバージェンス推定量と呼ぶ．U-ダイバージェンス(2.2)の基本不等式 $L_U(p,q) \geq H_U(p)$ で，等号は $q = p(\Lambda\text{-a.e.})$ に限ることから $\theta = \mathrm{argmin}_{\boldsymbol{\theta}'\in\Theta} L(p(;\boldsymbol{\theta}), p(;\boldsymbol{\theta}'))$ となる．従って $\hat{\theta}_U$ は θ の漸近一致推定量になることが分かる．推定関数は

$$s_U(\boldsymbol{z};\boldsymbol{\theta}) = \frac{\partial}{\partial\boldsymbol{\theta}}\xi(p(\boldsymbol{z};\boldsymbol{\theta})) - \frac{\partial}{\partial\boldsymbol{\theta}} b_U(\boldsymbol{\theta}) \tag{2.9}$$

で与えられる．これより，

$$E_p\{s_U(\boldsymbol{z};\boldsymbol{\theta})\} = \left\langle p - p(\cdot;\boldsymbol{\theta}), \frac{\partial}{\partial\boldsymbol{\theta}}\xi(p(\cdot;\boldsymbol{\theta}))\right\rangle$$

と表されることに注意すると，もし真の分布がモデルに含まれるとき，すなわち $p = p(\cdot;\boldsymbol{\theta})$ のとき，$E_p\{s_U(\boldsymbol{z};\boldsymbol{\theta})\} = 0$ となる．このように推定関数の一致性が保証される．最尤法($U = \exp$)では，特別に $b_U(\boldsymbol{\theta}) = 1$ であることに注意する．

推定量 $\hat{\theta}_U$ は推定方程式 $n^{-1}\sum_{i=1}^n s_U(\boldsymbol{z}_i;\boldsymbol{\theta}) = 0$ の解となるので，テイラー近似から導かれる漸近評価は

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_U - \boldsymbol{\theta}) = J_{\boldsymbol{\theta}}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_U(\boldsymbol{z}_i;\boldsymbol{\theta}) + o_P(1)$$

となる．これより，中心極限定理によって $\sqrt{n}(\hat{\theta}_U - \theta)$ の分布は平均 0，分散行列 $J_{\boldsymbol{\theta}}^{-1} V_\theta J_{\boldsymbol{\theta}}^{-1}$ の正規分布に収束することが証明される．ここで

$$J_{\boldsymbol{\theta}} = E_{p(\cdot;\boldsymbol{\theta})}\{(\partial/\partial\boldsymbol{\theta}) s_U^{\mathrm{T}}(\boldsymbol{z},\boldsymbol{\theta})\}, \quad V_{\boldsymbol{\theta}} = \mathrm{var}(s_U(\boldsymbol{z};\boldsymbol{\theta})) \tag{2.10}$$

とする．

2.5 Γ-ミニマックス

この小節では Γ-ミニマックスの議論が $\log$-ロスから U-ロス(2.7)へ拡張されることを紹介する．Γ-ミニマックスのベイズの考察については Good (1952)，一般的なゲーム論の観点からは Grünwald and Dawid (2004)を参照．これによって U-エントロピー最大化が U-ロスの下で自

然と意思決定者とのゲームのミニマックス性と等価になることが示される．簡単のため，自然が想定する分布は平均値制約の空間

$$\Gamma_\tau = \{p \in \mathcal{P} : E_p\{\boldsymbol{t}(\boldsymbol{z})\} = \boldsymbol{\tau}\}$$

にあると制限しよう．ここで $\boldsymbol{t}(\boldsymbol{z})$ は k 変量の統計量，$\boldsymbol{\tau}$ は $\boldsymbol{t}(\mathcal{Z})$ の凸包の内点とする．このとき，自然が与える分布 p が Γ_τ 上にあるとき(2.4)で定義された U-エントロピー $H_U(p)$ の最大化は以下のように与えられる．ラグランジュ関数

$$\mathcal{L}(p, \boldsymbol{\theta}, \kappa) = H_U(p) - \left\langle \boldsymbol{\theta}^{\mathrm{T}}\{\boldsymbol{t} - \boldsymbol{\tau}\}, p \right\rangle - \kappa\{\langle p \rangle - 1\}$$

の変分をとると平衡条件は，$\xi(p^*) = \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{t} - \kappa$，すなわち

$$p^*(\boldsymbol{z}) = u\big(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{t}(\boldsymbol{z}) - \kappa\big) \tag{2.11}$$

が得られる．ここで，θ と κ は制約条件

$$\left\langle u\big(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{t} - \kappa\big) \right\rangle = 1, \quad \left\langle \boldsymbol{t}, u\big(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{t} - \kappa\big) \right\rangle = \boldsymbol{\tau}$$

から決まる定数ベクトルと定数である．実際に，任意の $p \in \Gamma_\tau$ に対して p^* と p の U-エントロピーは

$$H_U(p^*) - H_U(p) = D_U(p, p^*)$$

を満たすので U-ダイバージェンス(2.2)の基本性質から p^* が U-エントロピー最大分布であることが分かる．この考察において 恒等化性

$$L_U(p, p^*) = H_U(p^*) \quad (\forall p \in \Gamma_\tau)$$

が本質的な役割を果たす．

次にミニマックス・ゲームについて考えよう．不等式(2.5)に注意すると

$$H_U(p^*) \le L_U(p^*, q) \le \max_{p \in \Gamma_\tau} L_U(p, q)$$

が得られる．このことから

$$H_U(p^*) \le \min_{q \in \mathcal{P}} L_U(p^*, q) \le \min_{q \in \mathcal{P}} \max_{p \in \Gamma_\tau} L_U(p, q)$$

であり，同様に恒等化性から $\max_{p \in \Gamma_\tau} L_U(p, p^*) = H_U(p^*)$ である．従って，

$$\max_{p \in \Gamma_\tau} \min_{q \in \mathcal{P}} L_U(p, q) = H_U(p^*) = \min_{q \in \mathcal{P}} \max_{p \in \Gamma_\tau} L_U(p, q)$$

であることが証明された．これが U-ロス・ゲームの Γ-ミニマックスである．このように任意の生成関数 U に対して Γ-ミニマックス性は成立する．これは，本質的には U-ロス関数(2.7)が自然に想起する p に関する線形性によって導かれる性質の一つである．U-ダイバージェンス(2.2)の連想する接続 $\nabla^{(U)}$ が常にミクスチュア接続になることと合わせて最小 U-ダイバージェンス法の基本的性質である．

2.6 U-モデル

前小節で議論されたように U-ロス・ゲームの Γ-ミニマックス解は U-エントロピー最大分布(2.11)で与えられることが分かった．よく知られているように指数モデルはボルツマン・シャノン エントロピーを最大にする．そして指数モデルの最尤法(経験 $\log$ ロス最小推定量)は最小

十分性，有効性，不偏性のエレガントな性質を発揮する．この小節では U-モデルを導入して最小 U-ダイバージェンス推定法の基本性質を調べよう．最初に葉層構造

$$\bigcup_{\boldsymbol{\tau}\in \mathrm{cvh}(\boldsymbol{t}(\mathcal{Z}))} \Gamma_{\boldsymbol{\tau}} \tag{2.12}$$

を考えよう．ここで $\mathrm{cvh}(A)$ は A の凸包を表す．このようにパラメータ $\boldsymbol{\tau}$ はリーフ $\Gamma_{\boldsymbol{\tau}}$ をつなぐ錘となっているが，同時に次のように U-エントロピー最大分布(2.11)をモデル化している．

$$M_U = \{p_U(\boldsymbol{z};\boldsymbol{\theta}) = u\big(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{t}(\boldsymbol{z}) - \kappa_{\boldsymbol{\theta}}\big) : \boldsymbol{\theta}\in\Theta\} \tag{2.13}$$

を U-モデルと呼ぶ．ここで $\kappa_{\boldsymbol{\theta}}$ は規格化条件 $\big\langle u\big(\boldsymbol{\theta}^{\mathrm{T}}t - \kappa_{\boldsymbol{\theta}}\big)\big\rangle = 1$ によって定まる定数．例えば，ベータエントロピー

$$H_\beta(p) = \frac{1}{\beta(1+\beta)}\big\langle p^{\beta+1}\big\rangle$$

の最大化分布は

$$p_\beta(\boldsymbol{z};\boldsymbol{\theta}) = \big[1 + \beta\{\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{t}(\boldsymbol{z}) - \kappa_{\beta\boldsymbol{\theta}}\}\big]^{\frac{1}{\beta}}$$

となる．

標準パラメータ θ と期待値パラメータ $\boldsymbol{\tau} = E_{p_U(\cdot;\boldsymbol{\theta})}\{\boldsymbol{t}(z)\}$ のパラメータ変換は 1:1 であると仮定する．これは変換のヤコビ行列

$$\frac{\partial\boldsymbol{\tau}}{\partial\boldsymbol{\theta}} = \Big\langle \Big\{\boldsymbol{t} - \frac{\partial\kappa_{\boldsymbol{\theta}}}{\partial\boldsymbol{\theta}}\Big\}\Big\{\boldsymbol{t} - \frac{\partial\kappa_{\boldsymbol{\theta}}}{\partial\boldsymbol{\theta}}\Big\}^{\mathrm{T}} u'\big(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{t} - \kappa_{\boldsymbol{\theta}}\big)\Big\rangle$$

が非特異であれば十分である．この仮定の下で，2 つのパラメータ θ と $\boldsymbol{\tau}$ は線形接続 ∇ と ${}^*\nabla^{(U)}$ に関してアフィンパラメータとなる．データ $z_1,\ldots,z_n$ に基づく最小 U-ダイバージェンス法の推定関数は

$$s_U(\boldsymbol{z};\boldsymbol{\theta}) = \boldsymbol{t} - \frac{\partial\kappa_{\boldsymbol{\theta}}}{\partial\boldsymbol{\theta}} - E_{p_U(\cdot;\boldsymbol{\theta})}\Big\{\boldsymbol{t} - \frac{\partial\kappa_{\boldsymbol{\theta}}}{\partial\boldsymbol{\theta}}\Big\} = \boldsymbol{t}(\boldsymbol{z}) - \boldsymbol{\tau}$$

となる．従ってパラメータ $\boldsymbol{\tau}$ の最小 U-ダイバージェンス推定量は $\hat{\boldsymbol{\tau}}_U = n^{-1}\sum_{i=1}^n \boldsymbol{t}(\boldsymbol{z}_i)$ となり $\hat{\boldsymbol{\tau}}_U$ の不偏性と統計量 $t(z)$ の十分性が示されるだろう．$\hat{\boldsymbol{\tau}}_U$ を統計汎関数と見ると $\hat{\boldsymbol{\tau}}_U(p) = E_p\{\boldsymbol{t}(\boldsymbol{z})\}$ と書けるので上の葉層構造(2.12)と一致する．任意の葉 $\Gamma_{\boldsymbol{\tau}}$ の像 $\hat{\boldsymbol{\tau}}_U(\Gamma_{\boldsymbol{\tau}})$ は真値 $\boldsymbol{\tau}$ からなる一点集合である．しかしながら，$\hat{\boldsymbol{\tau}}_U$ の有効性は成立しない．漸近的には一般論から最尤推定量が有効になる．

次に U-モデル M_U への U-ダイバージェンス(2.2)の射影を考えよう．空間 $\mathcal{P}$ から，かってな元 p を取ってくる．このとき，$\boldsymbol{\tau} = E_p\{\boldsymbol{t}(\boldsymbol{z})\}$ とおいて，

$$q^* = \operatorname*{argmin}_{q\in M_U} D_U(p,q)$$

を考えよう．このとき，ピタゴラスの定理

$$D_U(p,q) - \{D_U(p,q^*) + D_U(q^*,q)\} = 0 \tag{2.14}$$

が成立する．実際，(2.14)の左辺は内積 $\langle p - q^*, \xi(q^*) - \xi(q)\rangle$ になることから

$$(\boldsymbol{\theta}^* - \boldsymbol{\theta})^{\mathrm{T}}\big[E_p\{\boldsymbol{t}(\boldsymbol{z})\} - E_{q^*}\{\boldsymbol{t}(\boldsymbol{z})\}\big]$$

と表され，$E_p\{\boldsymbol{t}(\boldsymbol{z})\} = E_{q^*}\{\boldsymbol{t}(\boldsymbol{z})\} = \boldsymbol{\tau}$ に注意すると 0 になる．

次に節 2.3 で考察されたように親空間 $\mathcal{M}$ の中のモデル N 上でも最小 U-ダイバージェンス推定量は定義できる．(2.3)の U-クロスエントロピー $L(p,\nu)$ の ν に N の元 $\mu(\cdot;\theta)$ を代入すれば自然に U-ロス関数(2.7)が定義できる．今，$\mathcal{M}$ の中にシフトされた U-モデル

$$M_U^{\mathrm{shift}} = \{u\big(\boldsymbol{\theta}^{\mathrm{T}}\{\boldsymbol{t}(\boldsymbol{z}) - \boldsymbol{\tau}\}\big) : \boldsymbol{\theta} \in \Theta\} \tag{2.15}$$

を考える．ここで $\boldsymbol{\tau} = E_p\{\boldsymbol{t}(z)\}$．このモデルはデータ分布 p に依存していることに注意する．このとき，U-ロス関数(2.7)は $\langle p, \boldsymbol{\theta}^{\mathrm{T}}\{\boldsymbol{t} - \boldsymbol{\tau}\}\rangle = 0$ となるので

$$\ell_U(\boldsymbol{\theta}, p) = \big\langle U\big(\boldsymbol{\theta}^{\mathrm{T}}\{\boldsymbol{t}(\boldsymbol{z}) - \boldsymbol{\tau}\}\big)\big\rangle \tag{2.16}$$

と表現される．従って，パラメータ

$$\boldsymbol{\tau}^{\mathrm{shift}} = \frac{\big\langle \boldsymbol{t}, u\big(\boldsymbol{\theta}^{\mathrm{T}}\{\boldsymbol{t}(\boldsymbol{z}) - \boldsymbol{\tau}\}\big)\big\rangle}{\big\langle u\big(\boldsymbol{\theta}^{\mathrm{T}}\{\boldsymbol{t}(\boldsymbol{z}) - \boldsymbol{\tau}\}\big)\big\rangle}$$

の推定量は，再び，十分統計量 $n^{-1}\sum_{i=1}^n \boldsymbol{t}(\boldsymbol{z}_i)$ になる．あとで議論するようにアダブースト法はシフトされた U-モデルを定め，$U = \exp$ の(2.15)の指数ロスの最小化アルゴリズムで定式化されている．しかし，その統計的な一般的な理論や応用についてはこれからの問題である．

Ohara and Eguchi (2005)では，共分散行列全体の空間を情報幾何の方法によって考察している．行列式が一定の部分空間を葉とする葉層構造が明らかにされた．この構造によって共分散行列をパラメータに持つモデルとその統計推測の双対関係が特徴付けられた．無限次元指数モデルについて，Pistone and Sempi (1995)が興味深い数学的方法を導入した．このような枠組みで，無限次元 U-モデルについても考える必要があるだろう．

私たちは指数モデルと対数尤度の関係を凸性の関係(2.1)を経由してクルバック・ライブラーダイバージェンスによる理解を与えて，更に凸関数 U が導く関係

$$\xi(s) = \operatorname*{argmin}_{t \in \mathbf{R}}\{U(t) - ts\}$$

から U-モデル(2.13)と U-ロス関数(2.7)の関係を U-ダイバージェンス(2.2)による理解まで拡張した．次節では，このクラスの中からロバストネスの観点から有効な方法について考察する．

2.7 ロバストネス

最小ダイバージェンス法のクラスの中からロバストネスを持つ推定法を考察する．最尤法が少数のはずれ値によって大きく影響を受けるときにロバストである推定法を見つけたい．正規混合モデルの下でのロバストネスの考察は Fujisawa and Eguchi (2006)にある．再び，統計モデルを $M = \{p(z;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$，データの基礎分布を p としよう．このとき，尤度推定(スコアー)関数を単に $s(z;\boldsymbol{\theta})$ と書くと，最小ダイバージェンス法の推定関数(2.9)は

$$s_U(\boldsymbol{z};\boldsymbol{\theta}) = w(\boldsymbol{z};\boldsymbol{\theta})s(\boldsymbol{z};\boldsymbol{\theta}) - E_{p(\cdot;\boldsymbol{\theta})}\{w(\boldsymbol{z};\boldsymbol{\theta})s(\boldsymbol{z};\boldsymbol{\theta})\} \tag{2.17}$$

という関係をもつ．ここで $w(z;\boldsymbol{\theta})$ は $p(z;\boldsymbol{\theta})\xi'(p(z;\boldsymbol{\theta}))$ で定義されるスカラー関数とする．この関係から，$s_U(z;\boldsymbol{\theta})$ は重み関数 $w(z;\boldsymbol{\theta})$ による重み付き尤度推定(重み付きスコアー)関数とみられる．例えばベータダイバージェンスでは $w(z;\boldsymbol{\theta}) = p(z;\boldsymbol{\theta})^\beta$ となる．データに，はずれ値 z_{out} が混入され，最尤推定値が z_{out} によって大きなバイアスがもたらされたとしよう．このとき，文脈から z_{out} の尤度の貢献は小さいと考えて良いだろう．したがって，任意の $\beta > 0$ に対して $w(z_{\mathrm{out}};\boldsymbol{\theta})$ も十分小さいので(2.17)から，最小ベータダイバージェンス法の推定方程式からは z_{out} の影響は自動的に除かれることになる．これが最小ベータダイバージェンス法のロバストネスをもつ仕組みである．勿論 β が 0 に近づくと重み関数 $w(z;\theta)$ は恒等関数 1 に近づく

のでロバストネスの性能は落ちる．更に，特別なロバストネスの研究が Fujisawa and Eguchi (2005)で考察されている．

ロバストネスを計る尺度として影響関数がある．θ の推定量を経験分布関数 P_n の汎関数とみたとき，推定量は $\hat{\theta}(P_n)$ と書かれる．このとき，推定量の影響関数は

$$\mathrm{IF}(\hat{\boldsymbol{\theta}}, \boldsymbol{z}) = \frac{\partial}{\partial \varepsilon} \hat{\boldsymbol{\theta}}\big((1-\varepsilon)P_{\boldsymbol{\theta}} + \varepsilon \delta_{\boldsymbol{z}}\big)|_{\varepsilon=0}$$

で与えられる．ここで P_θ は密度関数 $p(\cdot, \theta)$ から作られる分布，δ_z は点 z で退化する一点分布を表す．これはデータの基礎分布 P がモデル分布 P_θ から外れて，確率 ε で，はずれ値 z が混入する状況を $P = (1-\varepsilon)P_\theta + \varepsilon\delta_z$ と記述し，そのモデルからの揺らぎを 0 へ近づけたときの推定量は $\hat{\theta}(P)$ の振る舞いを計っている．例えば，1 変量正規モデルの平均パラメータ θ の推定を考えると標本平均(最尤推定)の影響関数は $z-\theta$，一方で標本中央値の影響関数は $\mathrm{sgn}(z-\theta)$ に比例する．このように $|z| \to 0$ のとき最尤推定の影響関数は発散するが標本中央値の影響関数は有界である．では，最小ダイバージェンス推定量 $\hat{\theta}_U$ の影響関数は

$$\mathrm{IF}(\hat{\boldsymbol{\theta}}_U, \boldsymbol{z}) = J_{\boldsymbol{\theta}}^{-1} s_U(\boldsymbol{z}; \boldsymbol{\theta}) \tag{2.18}$$

という関係をもつ．ここで J_θ は(2.10)で定義される定数行列，$s_U(z;\theta)$ は(2.17)で定義されたベクトル値関数．あとの考察において具体的なモデルで計算される最小ベータダイバージェンス推定量の影響関数は有界であることが示される．要するに，(2.17)の関係から尤度推定(スコアー)関数 $s(z;\theta)$ のノルムが ∞ に発散するときに重み関数 $w(z;\theta)$ がそれよりも速いオーダーで 0 に収束することを確認すれば良いことが分かる．

2.8 局所学習の統合

この小節では最小 U-ダイバージェンス法のロバストネスを超越した性質に着目して，データ空間を局所的にモデルフィットを行い，それらを統合することを考える．その基本的なアイデアは推定関数(2.17)の $w(z,\theta)$ の性質を積極的に利用することにある．

異なる設定から観測されたデータセットが，潜在的な構造によって混合され，1 つのデータセットとしてまとめられた設定を考える．ミクスチュアーモデルを含み，より柔軟な局所モデリングを定式化して，そのモデルの下で有効に働く統計方法を提案する．ここで，局所モデルとは，厳密な定義はなく，局所線形モデル，折れ線回帰，変換点モデル，正規混合モデル，階層的エクスパート・ミクスチュアーなどが適用される場面を総称したものである．応用例として，主成分分析，独立成分分析，線形判別分析の文脈で上の潜在的な局在化を表現する局所モデリングの具体的な形を導出し，提案方法の妥当性について検討する．

主成分分析において主成分空間にデータ点が配置されたときに幾つかのクラスターに分解されることがある．このとき，オリジナルのデータをクラスター毎に主成分分析ができたなら，よりデータへの理解が深まるだろう．このように通常一般の多変量解析の局所版の提案のためモデルの共局在化を自動的に探索することによってそれぞれの局所モデルを反映したそれぞれの多変量解析を考える．

このために重要な点は，それぞれの局所モデルに対して適用するデータの範囲をデータから学習しなければならない点である．予め，データが，どの局所モデルに属すかのラベルが既知であれば簡単な作業となるが，現在の設定ではこのラベルは観測できない．キーとなる考えは，局所モデルの局在性をフィットさせるために異なる性能をもつ推定量のファミリーを用意することにする．

典型モデル $\{p(z;\theta) : \theta \in \Theta\}$ が選ばれたときに尤度方程式 $\sum_{i=1}^n s(z_i, \theta) = 0$ を解くことで本質的には尤度解析は完成する．一方で局所モデリングでは次の K ステップで完成される．

ステップ 1．**データ** $\{z_1,\ldots,z_n\}$ **の中から** 1 **つランダムに取り出し** $z_{(1)}$ **とおく．** $z=z_{(1)}$ **の周りで局所モデル** $\pi_z f(z,\theta_z)$ **を学習するために，次の推定方程式**

$$\sum_{i=1}^{n} w^*(z_i,z,\boldsymbol{\theta},\boldsymbol{\alpha})s(z_i,\boldsymbol{\theta}) = \mathrm{E}\{w^*(\boldsymbol{Z},z,\boldsymbol{\theta},\boldsymbol{\alpha})s(\boldsymbol{Z},\boldsymbol{\theta})\} \tag{2.19}$$

によって定義される推定量のファミリー $\{\hat{\theta}_{z\alpha}:\alpha\in\mathcal{A}\}$ **を求める．ここで** $w^*(z_i,\theta,\alpha)$ **は重み関数で，例えば，カーネル関数** K_h **と尤度に比例するように**

$$w^*(z_i,z,\boldsymbol{\theta},\boldsymbol{\alpha}) \propto K_h(z-z_i)f(z_i,\boldsymbol{\theta})^{\beta} \tag{2.20}$$

と定める．このとき $\alpha=(h,\beta)$ **とする．適切な** α **の選択のために，ロバストな経験ロス関数** $L(\theta)$ **の交互検証法(**CV**)によるものを使って，次のように定める**: $\hat{\alpha}_z=\mathrm{argmin}_{\alpha\in\mathcal{A}} L_{\mathrm{CV}}(\hat{\boldsymbol{\theta}}_{z\alpha})$. **これより** $z=z_{(1)}$ **の周りで局所モデルで使われたデータは次で与えられる：**

$$D_\epsilon(z)=\{z_i : w^*(z_i,\hat{\boldsymbol{\theta}}_{z\hat{\alpha}_z},\hat{\boldsymbol{\alpha}}_z)\geq\epsilon\}.$$

ステップ 2．$D_\epsilon(z_{(1)})$ **以外の観測値** $z_{(2)}$ **を取り出し，** $z=z_{(2)}$ **とおく．可能な全ての** $\alpha\in\mathcal{A}$ **に対して推定方程式(**2.19**)を解き，同様に** CV **によって** $\hat{\alpha}_z$ **を求める．学習に使われたデータセットを** $D_\epsilon(z_{(2)})$ **と置く．**

ステップ K**．このような手続きを繰り返し，ステップ** K **の終了時点でデータセット** $\{D_\epsilon(z_{(j)}):j=1,\ldots,K\}$ **の和集合の元の個数が** $100a$**%を超えるときアルゴリズムを停止させる．**

重み関数(2.20**)から観測値** z_i **がデータ点** z **に遠ければ，あるいは尤度の貢献が小さければ** z_i **はほとんど効かない．ここで** α **は局所モデルの局在性を計るため重要である．特に** $\alpha=(\infty,0)$ **のときは(**2.19**)は尤度方程式に帰着される．上のアルゴリズムは各々のステップで，データ点は単一にしか与えてないが，並列的に幾つかのデータ点を発生させ，最も成績の良かったデータ点を採択する方法がより適切であろう．**

データの局所学習については，初めは局所尤度のアイディアから研究が進められた．局所尤度法はデータに関わる非線形性を解析するために考案された方法論の一つである．Hjort and Jones **(**1996**)，**Eguchi and Copas **(**1998**)，**Eguchi et al. **(**2003**)，**Park et al. **(**2005**)などの研究がある．局所モデルの上での尤度法は必ずしも適切ではない場合がある．従来の局所尤度法は，重み関数(**2.20**)がカーネル関数** K_h **だけで与えられている．このためカーネルに対する詳細な理論的な要請と実際の適用に相容れない点が指摘されていた(次元の呪い問題)．異なるさまざまな推定法の統合によって局所モデルとそれに適合するデータを同時に求めることを可能にした．局所モデルに適合しないデータは，全てはずれ値とするロバスト推定法が選択された．この手続きを繰り返し，全体の適合度を測りながら停止規則が与えられたので，局所モデルの個数は自動的に決めることができた．このような** U**-ロス関数(**2.7**)の局所化による方法が** Kawakita and Eguchi **(**2006**)，**Mollah et al. **(**2006**)で考察されている．**

3. 主成分分析と独立成分分析

主成分分析と独立成分分析は，教師なしデータの解析のために様々な社会科学，自然科学分野において広く利用されている方法である．目的は高次元データからの特徴抽出，次元縮小などである．主成分分析と独立成分分析は互いに相補うものとして定着している．バイオインフォマティクスにおいても遺伝子発現データやプロテオームデータの解析などにも有力な方法として注目されている．

p**-次元観測ベクトル** $x=(x_1,\ldots,x_p)^{\mathrm{T}}$ **が与えられたとする．このとき主成分分析とは，** $y=Wx$ **の成分が無相関となる** $k\times p$ **行列** W **を求める解析法であり，独立成分分析は** $y=Wx$ **の成分**

が独立となる $k \times p$ 行列 W を求める解析法である．一般に確率ベクトル y の成分が独立であれば無相関となるが，必ずしも逆は成立しない．このことに主成分分析と独立分析の本質的な違いがある．しかし，特徴抽出の目的のためにはどちらが有効であるかは解析を行う以前から分かることは少ない．寧ろ，同時に並列して主成分分析と独立分析を行い，比較を通してそれぞれの結果を補うことが重要である．

3.1 主成分分析

主成分分析において，あとの独立成分分析との対応からデータ次元の削減に焦点を当てよう．これより，p-次元の空間から k 次元空間への線形写像を表す k-主成分行列 W は，制約条件 $WW^{\mathrm{T}} = I_k$（k 次単位行列）の下での最小化問題

$$\min\left\{E\{\|\boldsymbol{x}-\boldsymbol{\mu}\|^2 - \|\boldsymbol{W}(\boldsymbol{x}-\boldsymbol{\mu})\|^2\}\right\} \tag{3.1}$$

で定義される．ニューラルネットとの関連は Bishop (1995) を参照．このとき x の分散行列を $V(\boldsymbol{x})$ とすると，

$$E\{\|\boldsymbol{x}-\boldsymbol{\mu}\|^2 - \|\boldsymbol{W}(\boldsymbol{x}-\boldsymbol{\mu})\|^2\} = \mathrm{trace}\{V(\boldsymbol{x})\} - \sum_{i=1}^{k} \|\boldsymbol{w}_k V(\boldsymbol{x}) \boldsymbol{w}_k^{\mathrm{T}}\|^2 \tag{3.2}$$

となる．ここで $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$ は W の行ベクトルを表す．このことに注意すると，最小化問題 (3.1) の解 W は，分散行列 $V(\boldsymbol{x})$ の最大固有値から k 番目までの固有値に対応する固有ベクトル $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$ を行ベクトルとして並べた行列で与えられる．このように k-主成分ベクトルでは $\boldsymbol{y} = W\boldsymbol{x} = (\boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x}, \ldots, \boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x})^{\mathrm{T}}$ の成分は無相関となる．実際にデータ集合 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ が与えられたとき，(3.1) の経験ロス関数は次で与えられる．

$$\frac{1}{n}\sum_{i=1}^{n}\{\|\boldsymbol{W}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\|^2\}, \quad \bar{\boldsymbol{x}} = \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{x}_j.$$

この問題に関連して，p-次元の空間のデータ点を逐次的に 1 次元空間から k 次元空間へ射影する方法も同様である．初めに

$$\hat{\boldsymbol{w}} = \underset{\|\boldsymbol{w}\|=1}{\mathrm{argmin}}\{E(\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{x}-\bar{\boldsymbol{x}}))^2\} \tag{3.3}$$

を求め，次に $\hat{w}$ に直交する超平面に x を射影した x_1 に対して (3.3) を適用する．このように適当な次元 k まで求めることが可能である．この標準的な主成分分析は多変量正規モデルの下で最尤法によるものである．

節 2 で考察された最小 U-ダイバージェンス法は，制約条件 $WW^{\mathrm{T}} = I_k$（k 次単位行列）の下での最小化問題

$$\ell_U(\boldsymbol{W}, \boldsymbol{\mu}) = -E\{\psi\big(\|\boldsymbol{x}-\boldsymbol{\mu}\|^2 - \|\boldsymbol{W}(\boldsymbol{x}-\boldsymbol{\mu})\|^2\big)\}$$

で定義される．ここで $\psi(t) = -\xi(\exp(t))$ を表す．関数 ξ は生成関数 U の導関数 u の逆関数である．一般の U-ロス関数 (2.7) の第 2 項は W に関して定数なので無視されていることに注意する．この最小化問題は次の反復アルゴリズムで解ける．

$$\boldsymbol{\mu}^* = \sum_{i=1}^{n} w(\boldsymbol{x}_i; \boldsymbol{\mu}, \boldsymbol{W})\boldsymbol{x}_i, \quad \boldsymbol{W}^* = \mathrm{Eigen}_k(S(\boldsymbol{\mu}, \boldsymbol{W})).$$

ここで $\mathrm{Eigen}_k(A)$ は行列 A の k 番目までの固有値に対応する固有ベクトルを順番に並べてできる行列を表し，重み付き確率関数 p と重み付き行列 S は

$$p(\boldsymbol{x}_i; \boldsymbol{\mu}, \boldsymbol{W}) = \frac{w\big(\|\boldsymbol{x}_i-\boldsymbol{\mu}\|^2 - \|\boldsymbol{W}(\boldsymbol{x}_i-\boldsymbol{\mu})\|^2\big)}{\sum_j^n w\big(\|\boldsymbol{x}_j-\boldsymbol{\mu}\|^2 - \|\boldsymbol{W}(\boldsymbol{x}_j-\boldsymbol{\mu})\|^2\big)}$$

$$S(\boldsymbol{\mu},\boldsymbol{W})=\sum_{i=1}^{n}p(\boldsymbol{x}_i;\boldsymbol{\mu},\boldsymbol{W})(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^{\mathrm{T}}$$

と表す．ここで $w(t)=-\psi'(t)$ とする．このアルゴリズムでは一様に目的関数が減少

$$L_U(\boldsymbol{W}^*,\boldsymbol{\mu}^*)<L_U(\boldsymbol{W},\boldsymbol{\mu})$$

していることが示される．

影響関数については一般公式(2.18)の主成分分析における具体的な考察は Higuchi and Eguchi (1998)，Kamiya and Eguchi (2001)で与えられている．また，ベータダイバージェンスの β のようなチューニングパラメータについてデータに応じた決定方法は Higuchi and Eguchi (2004) にある．

3.2 独立成分分析

次に独立成分分析について見てみよう．初めにその基本的な設定であるインスタント・ミクスチュアーモデル(instant mixture)について説明しよう．p-次元入力ベクトル $\boldsymbol{x}$ が $p\times p$ の非退化行列 $\boldsymbol{A}$ に対して

$$\boldsymbol{x}=\boldsymbol{A}\boldsymbol{s} \tag{3.4}$$

と書かれると仮定しよう．ここで p-次元確率ベクトル s は独立な成分を持つと仮定する．このようにモデル(3.4)は，観測ベクトル x が A によって観測不能な s の成分の線形な混合によって得られることを意味する．解析の目的は Wx が独立成分を持つような $p\times p$ 行列 W をデータ集合 $x_1,\ldots,x_n$ から推定することである．例えば $W=A^{-1}$ とすれば良いが，必ずしも正確に A^{-1} を推定する必要はなく，WA が対角であれば十分である．独立性の仮定より復元ベクトル Wx の密度関数は

$$q(\boldsymbol{y})=q_1(y_1)\cdots q_m(y_m)$$

と書けることより，入力ベクトルの密度関数は

$$p(\boldsymbol{x};\boldsymbol{W},\boldsymbol{\mu})=|\det(W)|q_1(\boldsymbol{w}_1\boldsymbol{x}-\mu_1)\cdots q_m(\boldsymbol{w}_m\boldsymbol{x}-\mu_m)$$

となる．ここで $w_1,\ldots,w_m$ は W の行ベクトルを表す．標準的な方法では $q_1,\ldots,q_m$ を適当に決めて，データ集合 $x_1,\ldots,x_n$ から求まる擬似尤度関数

$$L(\boldsymbol{W},\boldsymbol{\mu})=\frac{1}{n}\sum_{i=1}^{n}\log p(\boldsymbol{x}_i;\boldsymbol{W},\boldsymbol{\mu}) \tag{3.5}$$

の最大化によって W，μ の推定量を得る．ここで U-ダイバージェンス(2.2)を採用して最小ダイバージェンス法を適用しよう．U-経験ロス関数は，一般公式(2.8)から

$$\ell_U^{\mathrm{emp}}(\boldsymbol{W},\boldsymbol{\mu})=-\frac{1}{n}\sum_{i=1}^{n}\xi\big(p(\boldsymbol{x}_i;\boldsymbol{W},\boldsymbol{\mu})\big)+\int U\big(\xi(p(\boldsymbol{z};\boldsymbol{W},\boldsymbol{\mu}))\big)d\boldsymbol{z} \tag{3.6}$$

で与えられる．一般に(3.6)の右辺の第 2 項は $\det(W)$ だけの関数となるが特に β-ダイバージェンスでは $c_\beta|\det(\boldsymbol{W})|^\beta$ となる．ただし c_β は W に関して定数とする．詳しくは Minami and Eguchi (2002)を参照.(3.5)の推定関数は,

$$s(\boldsymbol{x};\boldsymbol{W},\boldsymbol{\mu})=\begin{bmatrix}\frac{\partial}{\partial\boldsymbol{W}}\log p(\boldsymbol{x};\boldsymbol{W},\boldsymbol{\mu})\\ \frac{\partial}{\partial\boldsymbol{\mu}}\log p(\boldsymbol{x};\boldsymbol{W},\boldsymbol{\mu})\end{bmatrix}$$

$$= \begin{bmatrix} \{I_m - \boldsymbol{h}(\boldsymbol{W}\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{W}\boldsymbol{x})^{\mathrm{T}}\}\boldsymbol{W}^{-\mathrm{T}} \\ \boldsymbol{h}(\boldsymbol{W}\boldsymbol{x}-\boldsymbol{\mu}) \end{bmatrix} \tag{3.7}$$

となる．ここで $\boldsymbol{h}(\boldsymbol{y}) = \left((\partial/\partial y_i)\log q_i(y_i)\right)_{i=1}^{m}$ を表す．一方で最小 β-ダイバージェンス法の推定関数は，

$$s_\beta(\boldsymbol{x};\boldsymbol{W},\boldsymbol{\mu}) = p^\beta(\boldsymbol{x};\boldsymbol{W},\boldsymbol{\mu})s(\boldsymbol{x};\boldsymbol{W},\boldsymbol{\mu}) + \begin{bmatrix} \beta c_\beta |\det(\boldsymbol{W})|^\beta \boldsymbol{W}^{-\mathrm{T}} \\ 0 \end{bmatrix}$$

となる．このように，$\beta=0$ のとき推定関数 $s_\beta(\boldsymbol{x};\boldsymbol{W},\boldsymbol{\mu})$ は，$s(\boldsymbol{x};\boldsymbol{W},\boldsymbol{\mu})$ になる．任意の $\beta>0$ に対して，推定方程式 $\sum_{i=1}^n s_\beta(\boldsymbol{x}_i;\boldsymbol{W},\boldsymbol{\mu})=0$ は本質的には第 i 番目の重みが $p^\beta(\boldsymbol{x}_i;\boldsymbol{W},\boldsymbol{\mu})$ である重み付き尤度方程式とみられる．任意の $\beta>0$ に対して $\boldsymbol{x}_i$ が小さな密度 $p(\boldsymbol{x}_i;\boldsymbol{W},\boldsymbol{\mu})$ を持てば推定方程式に対して貢献度は小さい．このように，この推定方程式は，はずれ値に対してロバストネスがあるといえる．厳密には，任意の i に対して

$$\sup_{z\in\mathbf{R}} q_i(z)\exp(|z|) < \infty$$

が成立すれば，任意の $\beta>0$ に対して推定関数 $s_\beta(\boldsymbol{x};\boldsymbol{W},\boldsymbol{\mu})$ は有界となる．β が 0 に近づくにつれて，この有界性は徐々に破綻していき，$\beta=0$ で非有界となる．それと対照的に β が 0 に近づくにつれて最小 β-ダイバージェンス推定量の漸近分散は小さくなる．

遺伝子発現データに見られるように，データ次元 p に対してサンプルサイズ n が小さいときは，しばしば復元行列 W を求めることは不能となる．このとき，一般にデータ次元 p が大きいときは Hyvarinen et al. (2001) の Fast-ICA が有効である．この方法は，独立成分を逐次的に求める方法である．前処理として主成分分析を行い，$\boldsymbol{x}^*$ の分散行列 $V(\boldsymbol{x}^*)$ が単位行列となるように $\boldsymbol{x}$ を変換する．必要であれば，特異値分解アルゴリズムを使う．これから

$$\hat{\boldsymbol{w}} = \underset{\|\boldsymbol{w}\|=1}{\operatorname{argmin}}\{E\{G(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}^*)\}\} \tag{3.8}$$

を求める．ここで，コントラスト関数 G は非 2 次関数で，典型的な例は独立成分の裾が軽いときは 尖度(kurtosis)$G(y)=y^4$，独立成分の裾が重いときは $G(y)=\log\{\exp(y)+\exp(-y)\}$ が挙げられる．もしコントラスト関数 G が 2 次関数であれば(3.8)は w の定数関数になることに注意する．このことは前処理によって主成分分析(3.3)の手続きが完了していることを意味する．次に $\hat{w}$ に直交する超平面に $\boldsymbol{x}^*$ を射影した $\boldsymbol{x}_1^*$ に対して(3.8)を適用する．このように適当な次元 k まで求めることが可能である．

主成分分析と独立成分分析の内容で最小ベータダイバージェンスの局所学習については Mollah et al. (2005a, 2005b, 2006) で与えられている．更にこの理論的な考察についての研究を継続中である．最近，カーネル主成分分析についても最小ベータダイバージェンス法の拡張が検討されている．

4. 統計的パタン認識

パタン認識は，特徴ベクトル x が与えられたとき，そのクラスラベル y を決定する問題である．ここで x は p 次元の特徴空間 $\mathcal{X}$ の元，y はラベル集合 $\mathcal{Y}$ の元とする．一般的な議論は McLachlan (2004) を参照．認知科学は我々が持つ生物脳は‘学習’という過程を通して合理的な判別ルールを創ることが研究されている．例えば，ある生物集団の中で最適な配偶子を見つけるために多くの場合，メスは交配可能なオスの特徴をベクトル化した x を観測して 2 値ラベルを決定することによって最適な配偶子を選び出す．このようにパタン認識は生物脳の基本的な

機能の一つである．このような文脈で複数の判別子をより巧妙に連結させて新しい判別子を構成するアルゴリズムを目指すブースト法が活発に議論されている．多くの文献があるが，例えば，Schapire (1990)，Freund and Schapire (1997)，Schapire et al. (1998)，Eguchi and Copas (2001)，江口(2004)，Eguchi (2005a)が参照される．

特徴空間 $\mathcal{X}$ からラベル集合 $\mathcal{Y}$ への写像 h を一つ与えれば $y=h(\boldsymbol{x})$ によってパタン認識の方法が得られることになる．以後，$h(\boldsymbol{x})$ を判別子と呼ぶ．しばしば判別子は空間 $\mathcal{X}\times\mathcal{Y}$ 上の関数 $F(\boldsymbol{x},y)$ を使って

$$h_F(\boldsymbol{x})=\operatorname{argmax}\{F(\boldsymbol{x},y):y\in\mathcal{Y}\} \tag{4.1}$$

によって構成される．この $F(\boldsymbol{x},y)$ を判別関数とよぶ．データ空間 $\mathcal{Z}$ は空間 $\mathcal{X}\times\mathcal{Y}$ で，$\boldsymbol{z}=(\boldsymbol{x},y)$ と分解されている．統計的にはトレーニングデータ $E_n=\{(\boldsymbol{x}_1,y_1),\ldots,(\boldsymbol{x}_n,y_n)\}$ から判別関数 F を学習して(4.1)のルールによるパタン認識の良い方法をつくることがゴールである．多くの場合，判別子 $h:\mathcal{X}\to\mathcal{Y}$ の性能を計るためにトレーニングデータとは独立にテストデータ $E_m^*=\{(\boldsymbol{x}_1^*,y_1^*),\ldots,(\boldsymbol{x}_m^*,y_m^*)\}$ が用意される．テストエラー

$$\operatorname{testerr}(h)=\frac{1}{m}\sum_{i=1}^{m}I(h(\boldsymbol{x}_i^*)\neq y_i^*) \tag{4.2}$$

によって評価される．高性能の判別子 h を構成するときに，利用できるものはトレーニングデータのみでテストデータは使えない．これより，トレーニングエラーをいくら小さくしても，テストエラーが小さくなるとは限らない．ここに予測問題の本質がある．

可能な全ての判別関数の空間を $\mathcal{F}$，可能な全ての判別子の空間を $\mathcal{H}$ と表す．判別関数の選択には冗長性がある．集合

$$\mathcal{F}_h=\{F\in\mathcal{F}:h_F=h\}$$

の任意の元 F_1,F_2 を考えると，正数 c_1,c_2 に対して $c_1F_1+c_2F_2$ は，また $\mathcal{F}_h$ の元である．一般に，$\mathcal{F}$ の元 F_1,F_2 が共通の判別子を作るとき同値であると定めると $\mathcal{H}$ は $\mathcal{F}$ をこの同値関係で割ったものに等しくなる．一つの判別子 h を構成するためには，0 と 1 の値しか取らない判別関数 $F(\boldsymbol{x},y)=I(h_F(\boldsymbol{x})=y)$ を求めれば十分である．ここで I は定義関数とする．しかし，データから豊かな学習をするためには，実はこの冗長性が役に立つ．

以後の議論において，特徴ベクトル $\boldsymbol{x}$，クラスラベル y に対して，節 2 で考察されたデータ空間 $\mathcal{Z}$ とデータベクトル $\boldsymbol{z}$ を $\mathcal{Z}=\mathcal{X}\times\mathcal{Y}$，$\boldsymbol{z}=(\boldsymbol{x},y)$ とおいて，節 2 で準備された結果を利用しよう．統計的判別子のファミリィ $\mathcal{H}_1$ が用意されていたとする．ブースト法とはこのファミリィの中から判別子をうまく組み合わせて，より良い判別子を作ることを考える．ブースト法のキーとなる考えは判別子のファミリィ $\mathcal{H}_1$ のかってな元 $h_1,\ldots,h_d$ を判別関数全体の空間 $\mathcal{F}$ に次のように線型に埋め込むことにある：

$$\mathcal{F}=\left\{F(\boldsymbol{x},y,\boldsymbol{\alpha})=\sum_{j=1}^{d}\alpha_jI(h_j(\boldsymbol{x})=y):\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_d)\in A\right\}. \tag{4.3}$$

線型な結合係数たち $\{\alpha_j\}$ をより合理的に与えることができれば $h_1,\ldots,h_d$ の中の最良の判別子を超えるより強力な判別子が構成できる．あとの議論で明らかになるよう，実際には逐次的に

$$F(\boldsymbol{x},y,\alpha_1,\ldots,\alpha_{t+1})=F(\boldsymbol{x},y,\alpha_1,\ldots,\alpha_t)+\alpha_{t+1}I(h_{t+1}(\boldsymbol{x})=y)$$

と最適化される．このようにブースト法は各々のステップ t で最適な判別子 h_{t+1} と結合係数 α_{t+1} を与えるアルゴリズムによって定義される．

4.1 判別関数の U-ロス関数

節 2 の一般の考察をこの判別問題に適用しよう．詳細な考察が Murata et al. (2004)で与えられている．特徴空間 $\mathcal{X}$ とラベル集合 $\mathcal{Y}$ の直積空間上の確率分布は

$$p(\boldsymbol{x},y)=P(y|\boldsymbol{x})q(\boldsymbol{x}) \tag{4.4}$$

と書ける．ここで $P(y|x)$ は x が与えられたときの y の条件付分布，$q(x)$ は x の周辺分布の密度関数を表す．このとき(4.3)で与えられるような判別子を線型結合した判別関数を

$$F(\boldsymbol{x},y)=\boldsymbol{\alpha}^{\mathrm{T}}f(\boldsymbol{x},y)$$

と書く．ここで $f(\boldsymbol{x},y)=(I(h_1(\boldsymbol{x})=y),\ldots,I(h_d(\boldsymbol{x})=y))$．これに対して節 2 で考察されたシフトした U-モデル

$$\tilde{\mu}_{\boldsymbol{\alpha}}(y|\boldsymbol{x})=u(\boldsymbol{\alpha}^{\mathrm{T}}f(\boldsymbol{x},y)-b(\boldsymbol{x},\boldsymbol{\alpha}))$$

を考えよう．$b(\boldsymbol{x},\boldsymbol{\alpha})=\sum_{y'\in\mathcal{Y}}\boldsymbol{\alpha}^{\mathrm{T}}f(\boldsymbol{x},y')p(y'|\boldsymbol{x})$ とする．これは，小節 2.6 の(2.15)式で一般に定義されたモデルを今の文脈に書いたものである．これより，U-ロス関数は一般の(2.7)式から

$$\ell_U(\boldsymbol{\alpha})=\int_{\mathcal{X}}\sum_{y\in\mathcal{Y}}U(\boldsymbol{\alpha}^{\mathrm{T}}f(\boldsymbol{x},y)-b(\boldsymbol{x},\boldsymbol{\alpha}))q(\boldsymbol{x})d\boldsymbol{x}$$

になる．節 2.4 と同様な議論から

$$\ell_U(\boldsymbol{\alpha})-\ell_U(\boldsymbol{\alpha}^*)=D_U(\tilde{\mu}_{\boldsymbol{\alpha}^*},\tilde{\mu}_{\boldsymbol{\alpha}})$$

が云える．ここで

$$\boldsymbol{\alpha}^*=\operatorname*{argmin}_{\boldsymbol{\alpha}\in A}D_U(p,\tilde{\mu}_{\boldsymbol{\alpha}}).$$

経験 U-ロス関数は

$$\ell_U^{\mathrm{emp}}(\boldsymbol{\alpha})=\frac{1}{n}\sum_{i=1}^{n}\sum_{y\in\mathcal{Y}}U(\boldsymbol{\alpha}^{\mathrm{T}}\{f(\boldsymbol{x}_i,y)-f(\boldsymbol{x}_i,y_i)\}) \tag{4.5}$$

で与えられる．ここで

$$b(\boldsymbol{x}_i,\boldsymbol{\alpha})=\boldsymbol{\alpha}^{\mathrm{T}}f(\boldsymbol{x}_i,y_i)$$

と成ることに注意する．

一方で確率制約のある場合はどうなるのだろうか．U-確率モデルは

$$\bar{\mu}_{\boldsymbol{\alpha}}(y|\boldsymbol{x})=u(\boldsymbol{\alpha}^{\mathrm{T}}f(\boldsymbol{x},y)-\kappa(\boldsymbol{x},\boldsymbol{\alpha})) \tag{4.6}$$

で与えられる．ここで $\kappa(\alpha)$ は規格化条件

$$\sum_{y\in\mathcal{Y}}u(\boldsymbol{\alpha}^{\mathrm{T}}f(\boldsymbol{x},y)-\kappa(\boldsymbol{x},\boldsymbol{\alpha}))=1$$

を満たすとする．この下での U-ロス関数は

$$\bar{\ell}_U(\boldsymbol{\alpha})=\int_{\mathcal{X}}\sum_{y\in\mathcal{Y}}\left[U(\boldsymbol{\alpha}^{\mathrm{T}}f(\boldsymbol{x},y)-\kappa(\boldsymbol{x},\boldsymbol{\alpha}))-P(y|\boldsymbol{x})\{\boldsymbol{\alpha}^{\mathrm{T}}f(\boldsymbol{x},y)-\kappa(\boldsymbol{x},\boldsymbol{\alpha})\}\right]q(\boldsymbol{x})d\boldsymbol{x}$$

になる．ピタゴラス関係

$$\bar{\ell}_U(\boldsymbol{\alpha})-\bar{\ell}_U(\boldsymbol{\alpha}^*)=D_U(\bar{\mu}_{\boldsymbol{\alpha}^*},\bar{\mu}_{\tilde{\boldsymbol{\alpha}}})$$

が云える．経験 U-ロス関数は

$$\bar{\ell}_U^{\mathrm{emp}}(\boldsymbol{\alpha}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{y\in\mathcal{Y}} U(\boldsymbol{\alpha}^{\mathrm{T}} f(\boldsymbol{x}_i, y) - \kappa(\boldsymbol{x}_i, \boldsymbol{\alpha})) + \kappa(\boldsymbol{x}_i, \boldsymbol{\alpha}) - \boldsymbol{\alpha}^{\mathrm{T}} f(\boldsymbol{x}_i, y_i) \tag{4.7}$$

で与えられる．

では典型例 $U = \exp$ の場合の 2 つのロス関数を見てみよう．

$$\ell_{\exp}^{\mathrm{emp}}(\boldsymbol{\alpha}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{y\in\mathcal{Y}} \exp\{\boldsymbol{\alpha}^{\mathrm{T}}\{f(\boldsymbol{x}_i, y) - f(\boldsymbol{x}_i, y_i)\}\},$$

$$\bar{\ell}_{\exp}^{\mathrm{emp}}(\boldsymbol{\alpha}) = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{\exp\{\boldsymbol{\alpha}^{\mathrm{T}} f(\boldsymbol{x}_i, y_i)\}}{\sum_{y\in\mathcal{Y}} \exp\{\boldsymbol{\alpha}^{\mathrm{T}} f(\boldsymbol{x}_i, y)\}} \tag{4.8}$$

である．それぞれ指数ロス，対数ロスと呼ばれアダブーストとロジットブーストを定めるロス関数である．この興味深い理解は Lebanon and Lafferty (2002) で与えられている．特に統計学の分野では対数ロスはロジスティック回帰モデルの負の対数尤度関数として確立されているものである．指数ロスは，機械学習の分野から Freund and Schapire (1997) で提案されるまで知られていなかった．ロジスティック判別関数との関連は Eguchi and Copas (2002) にある．加法ロジスティックモデルとの関連は Friedman et al. (2000) を参照．

4.2 U-ブースト

これより，節 4.1 で得られた判別のロス関数の最小化アルゴリズムを考えよう．その基本的なアイデアは m-射影を繰り返し利用することにある．現在ある判別関数 $F(\boldsymbol{x}, y)$ から新たな判別子 $h(\boldsymbol{x})$ を埋め込むためのワンステップ

$$F(\boldsymbol{x}, y) \mapsto F^*(\boldsymbol{x}, y) = F(\boldsymbol{x}, y) + \alpha I(h(\boldsymbol{x}) = y)$$

を次のような更新

$$(\alpha^*, h^*) = \operatorname*{argmin}_{(\alpha,h)\in\mathbb{R}\times\mathcal{H}} \ell_U^{\mathrm{emp}}(F(\boldsymbol{x}, y) + \alpha I(h(\boldsymbol{x}) = y))$$

で考える．これはワンパラメータ α のモデルへの m-射影によって求められ，ここにピタゴラス関係が成立する直角三角形が得られる．この操作を繰り返すことによって次々と，直角三角形が連想され判別の U-ロス関数の逐次最適化が計られる．

実際の確率制約を課さない U-ブーストアルゴリズムは判別子のクラス $\mathcal{H}_1$ と例題の集合 E_n が与えられたとき次のように提案される．

A. 初期分布を $w_1(i, y) = \frac{1}{n(g-1)} I(y \neq y_i)$ とする．ここで $g = \mathrm{card}(\mathcal{Y})$．

B. 反復回数 $t = 1, \ldots, T$ に対して，重み付きエラーレイトを

$$\epsilon_t(h) = \frac{1}{2}\sum_{i=1}^{n}\sum_{y\in\mathcal{Y}} w_t(i, y) I(y \neq y_i)\{f(\boldsymbol{x}_i, y) - f(\boldsymbol{x}_i, y_i) + 1\} \tag{4.9}$$

と定める．次に

(B-1) $h_*^{(t)} = \mathrm{argmin}_{h\in\mathcal{H}_1} \epsilon_t(h)$ と選ぶ．

(B-2) $\alpha_t^* = \mathrm{argmin}_\alpha \ell_U^{\mathrm{emp}}(F_{t-1} + \alpha f_*^{(t)})$ を求める．ここで ℓ_U^{emp} は(4.5)で定義したもの．

(B-3) これより $F_t(\boldsymbol{x},y)=\sum_{j=1}^{t}\alpha_j^* I(h_*^{(j)}(\boldsymbol{x})=y)$ とおき，

$$w_{t+1}(i,y)\propto u\{F_t(\boldsymbol{x}_i,y)-F_t(\boldsymbol{x}_i,y_i)\}$$

によって重み付きエラーレイト(4.9)を更新する．

C. 最後に $h_{\mathrm{final}}(\boldsymbol{x})=\mathrm{argmax}_{y\in\mathcal{Y}}F_T(\boldsymbol{x},y)$ を得る．ここで $F_T(\boldsymbol{x},y)=\sum_{t=1}^{\mathrm{T}}\alpha_t^* I(h_*^{(t)}(\boldsymbol{x})=y)$ である．

確率制約の下での U-ブーストアルゴリズムはステップ(B-2)の ℓ_U^{emp} を $\bar{\ell}_U^{\mathrm{emp}}$ に変更するだけである．このように U-ブーストアルゴリズムは共に非常に単純な反復アルゴリズムによって判別子のクラス $\mathcal{H}_1$ の中で異なる識別能力をもつ判別子を統合することに成功している．その特徴は例題に対する重み $w_t(i,y)$ のステップ t に関するダイナミックな変動である．U-ブーストアルゴリズムの共通の性質として「例題の重み $w_t(i,y)$ に関して最小のエラーレイトを持つ $h_*^{(t)}$ が，ステップ(B-2)で決められた結合係数 α_t^* を加入させ更新した重み $w_{t+1}(i,y)$ では最悪のエラーレイトを持つ」ことが示される．即ち全ての U-ブーストアルゴリズムにおいて，共通に任意の t に対して

$$\epsilon_{t+1}(h_t)=\frac{1}{2}$$

が成立する．証明はステップ(B-2)の結合係数 α_t^* において α に関する $\ell_U{}^{\mathrm{emp}}(F_{t-1}+\alpha f_*^{(t)})$ の勾配が 0 と成ることから示される．詳細は Murata et al. (2004)の定理 3 の証明を参照されたい．

再び典型のケース $U(t)=\exp(t)$ に戻るとステップ(B-2)は陽に解けて

$$\alpha_t^*=\frac{1}{2}\log\frac{1-\epsilon_t(h_*^{(t)})}{\epsilon_t(h_*^{(t)})}$$

に置き換えられて，アダブースト M2 と呼ばれるものになる．

U-ブーストの統計的性質を考察しよう．よく知られているようにベイズルールは

$$h_B(\boldsymbol{x})=\mathrm{argmax}\{P(y|\boldsymbol{x}):y\in\mathcal{Y}\} \tag{4.10}$$

で与えられ，例題の事後分布 $P(y|x)$ の時，すべての判別子の中でエラーレイトの下限を与える．提案されているほとんどの統計的判別子は例題から事後分布 $P(y|x)$ を推定する問題に帰着している（Eguchi and Copas, 2001; Eguchi and Copas, 2002 を参照）．では U-ロス関数の逐次最適化で定義された U-ブーストはベイズルールとどんな関連をもっているか調べよう．ベイズルールと同値な判別関数の全体，

$$\mathcal{F}_B=\{F(\boldsymbol{x},y):h_F=h_B\}$$

を考えよう．判別関数のノンパラメトリックなシフトされた U-モデルは

$$\mathcal{M}_U=\{\mu_F(y|\boldsymbol{x})=u(F(\boldsymbol{x},y)-b_F(\boldsymbol{x})):F\in\mathcal{F}\}$$

で与えられる．ここでシフト定数 $b_F(x)$ はパラメトリックの場合と同様に

$$b_F(\boldsymbol{x})=\sum_{y'\in\mathcal{Y}}F(\boldsymbol{x},y')p(y'|\boldsymbol{x}) \tag{4.11}$$

となる．このとき，ある判別関数のクラス $\mathcal{F}$ の中に F^* が存在して

$$u(F^*(\boldsymbol{x},y)-b_{F^*}(\boldsymbol{x}))=c(\boldsymbol{x})P(y|\boldsymbol{x}) \tag{4.12}$$

を満たすとする．ここで $c(\boldsymbol{x})$ は正の関数．この仮定は，$F^*(\boldsymbol{x},y)$ を判別関数とする判別ルールがベイズルールと等価になることを意味する．このとき，

$$F^* = \mathrm{argmin}\{\ell_U(F) : F \in \mathcal{F}\} \tag{4.13}$$

が成立する．ただし，

$$\ell_U(F) = \int_{\mathcal{X}} \sum_{y\in\mathcal{Y}} U(F(\boldsymbol{x},y) - b_F(\boldsymbol{x}))q(\boldsymbol{x})d\boldsymbol{x}.$$

定義から仮定(4.12)に注意すると

$$\ell_U(F) - \ell_U(F^*) = D_U(\mu_{F^*}, \mu_F)$$

となり，従って，U-ダイバージェンス(2.2)の基本性質から(4.13)が結論される．詳細な証明は，Eguchi (2005a)を参照．

このように最小 U-ダイバージェンス推定 F^* はベイズルールと同値になる．このようにアブストラクトなロス関数 $\ell_U(F)$ の最適化はベイズルールと同値な判別関数の探索と一致する．

実際の統計的パタン認識では，(4.5)で与えられるよう例題の集合から経験的な期待値を取り，有限次元のパラメータベクトル α の逐次最適化によって U-ブーストアルゴリズムで実行されている．n 個の例題に対して分布の仮定(4.10)が更に漸近的に成立すると仮定すると上の命題の経験的な命題も成立するだろう．ここでは，上の形式的な考察に留める．理論的な整合性のために幾つかの強い仮定を課すとき，この経験的な命題が成立することが VC 次元の考察から Lugosi and Vayatis (2004)で証明されている．アダブーストの正則化については Rätsch et al. (2001)の方法がある．

4.3 イータブースト

特に統計的パタン認識においてミスラベルの問題に焦点を当てる．そのため次の U 関数を調べよう．

$$U_\eta(t) = (1-\eta)\exp(t) + \eta t.$$

ここで η は $0 \le \eta < 1$ の定数とする．これより

$$u_\eta(t) = (1-\eta)\exp(t) + \eta, \quad \xi_\eta(u) = \log\frac{u-\eta}{1-\eta}$$

となるので，生成されるダイバージェンスは

$$D_\eta(\mu,\nu) = \int_{\mathcal{X}} \sum_{y\in\mathcal{Y}} \left[\nu(\boldsymbol{x},y) - \mu(\boldsymbol{x},y) - \{\mu(\boldsymbol{x},y)-\eta\}\log\frac{\nu(\boldsymbol{x},y)-\eta}{\mu(\boldsymbol{x},y)-\eta}\right] q(\boldsymbol{x})d\boldsymbol{x}$$

となり，イータダイバージェンスと呼ぼう．これより，公式(4.6)から判別関数 $F(\boldsymbol{x},y) = \boldsymbol{\alpha}^{\mathrm{T}} f(\boldsymbol{x},y)$ の U_η-確率モデルは

$$\bar{\mu}_\eta(y|\boldsymbol{x},\boldsymbol{\alpha}) = (1-\eta)\exp\{\boldsymbol{\alpha}^{\mathrm{T}} f(\boldsymbol{x},y) - \kappa(\boldsymbol{x},\boldsymbol{\alpha})\} + \eta$$

となる．ここで，規格化定数は

$$\kappa(\boldsymbol{x},\boldsymbol{\alpha}) = \log\frac{1-\eta}{1-g\eta} + \log\left[\sum_{y'\in\mathcal{Y}} \exp\{\boldsymbol{\alpha}^{\mathrm{T}} f(\boldsymbol{x},y')\}\right]$$

で与えられる．ただし $g = \mathrm{card}(\mathcal{Y})$ とする．これを書き直すと

$$\bar{\mu}_\eta(y|\boldsymbol{x},\boldsymbol{\alpha}) = \{1-\eta(g-1)\}P_L(y|\boldsymbol{x},\boldsymbol{\alpha}) + \eta\sum_{y'\neq y} P_L(y|\boldsymbol{x},\boldsymbol{\alpha}) \tag{4.14}$$

と表される．ここで

$$P_L(y|\boldsymbol{x},\boldsymbol{\alpha}) = \frac{\exp\{\boldsymbol{\alpha}^{\mathrm{T}} f(\boldsymbol{x},y)\}}{\sum_{y'\in\mathcal{Y}} \exp\{\boldsymbol{\alpha}^{\mathrm{T}} f(\boldsymbol{x},y')\}}.$$

この確率モデル(4.14)は，次の解釈を与える．理想的な仮定では，例題の分布(4.4)に対してラベル y が観測される条件付確率 $P(y|\boldsymbol{x})$ がロジステックモデル $P_L(y|\boldsymbol{x},\alpha)$ で与えられるとしよう．しかし，実際には何らかの理由で理想仮定が崩れて，確率 η で誤って y 以外のラベルが観測されたとする．このとき，帰結される条件付確率は $\bar{\mu}_\eta(y|\boldsymbol{x},\alpha)$ となる．このようにイータダイバージェンス D_η は，ミスラベルの生成的な確率モデルを連想することが分かる．

U_η から作られる U-ブーストをイータブーストと呼ぼう．上のミスラベルの確率モデルの解釈から，イータブーストは自然にロバストな方法となっている．実際，確率制約を入れたイータブーストは，Copas (1988)にあるようにノイズの多いデータに対して 2 値回帰分析で提案された方法と等価である．アダブーストのロバスト化については Takenouchi and Eguchi (2004)を参照．一般の特徴空間のはずれ値に対するロバストネスについては Kanamori et al. (2004, 2006)を参照．

4.4 スペースブースト

空間パタン認識の問題を考える．与えられたある領域のピクセルの番地(添え字)の集合を $\mathcal{D}$ とする．番地 i が指定するピクセルの上で観測される特徴ベクトルを $\boldsymbol{x}_i$，クラスラベルを y_i とする．ここで考えられている空間パタン認識の問題はリモートセンシングに動機付けられている (Nishii and Eguchi, 2005a, 2005b, 2006a)．このとき，$\boldsymbol{x}_i$ はその地点の反射光を p 個のバンドで観測されたベクトルで，y_i はその地点の g 個に区分された土地利用のカテゴリーを表す．添え字集合 $\mathcal{D}$ が記述する風景全体の特徴量を $\boldsymbol{x}=\{\boldsymbol{x}_i : i\in\mathcal{D}\}$，クラスラベルを $\boldsymbol{y}=\{\boldsymbol{y}_i : i\in\mathcal{D}\}$ と書く．通常のパタン認識の問題に空間情報が加わっていることに注意する．このため，$\mathcal{D}$ の配置の情報を $\mathcal{D}$ 上の離散値距離 d を使って，次のように‘ 近傍 ’を定義しよう．番地 i の距離 r の近傍を

$$\mathcal{U}_r(i) = \{j\in\mathcal{D} : d(i,j) = r\}$$

と定める．これより，r-近傍は $\mathcal{N}_r(i) = \cup_{\gamma=1}^{r}\mathcal{U}_\gamma(i)$ と定義される．適切な r に対して，任意の $j\in\mathcal{N}_r(i)$ に対して y_j は y_i と等しい確率が高いことが想定される．例えば，通常は，広大な水域に孤立した針葉樹林帯などは現れないはずだ．標準的な仮定はこのような直感を次のように定式化する．

仮定 1. (条件独立性)$\mathcal{D}$ 上のクラスラベル y が与えられたとき，全ての $\boldsymbol{x}_i$ は統計独立である．

仮定 2. (マルコフ・ランダムフィールド性)$\mathcal{D}$ の任意の番地 i に対してクラスラベル y を y_i と残りの $\boldsymbol{y}_{-i}$ に分割する．このとき正数 r が存在して，条件付分布 $P(y_i|\boldsymbol{y}_{-i})$ はラベル集合 $\{y_j : j\in\mathcal{N}_r(i)\}$ にのみに依存する．

この仮定に基づいて Besag (1986)の ICM (反復的条件付モード)法は

$$P(y_i|\boldsymbol{x}_i,\boldsymbol{y}_{-i}) = \frac{\exp\{-\beta\Delta_i(y_i)\}p(\boldsymbol{x}_i|y_i)}{\sum_{\ell=1}^{g}\exp\{-\beta\Delta_i(\ell)\}p(\boldsymbol{x}_i|\ell)}$$

から擬似尤度関数 $\prod_{i\in\mathcal{D}} P(y_i|\boldsymbol{x}_i,\boldsymbol{y}_{-i})$ を導き，この関数の最大化アルゴリズムからなる(詳細は McLachlan, 2004 を参照)．

私たちは，空間の特徴抽出のためにマルコフ・ランダムフィールドのモデルから作られた学習機をブーストする方法論を考察している．最初に，仮定 1 を利用して空間情報を無視したパ

タン識別を学習データ $\{(\boldsymbol{x}_i, y_i) : i \in \mathcal{D}\}$ を使って行う．この結果，番地 i の得られた判別関数を $f_0(\boldsymbol{x}_i, y)$ と書く．次に，仮定 2 から r-近傍 $\mathcal{N}_r(i)$ を $\{\mathcal{U}_\gamma(i) : \gamma = 1, \ldots, r\}$ に分解して，各成分 $\mathcal{U}_\gamma$ 毎に判別関数を

$$f_\gamma(i, \boldsymbol{x}, y) = \begin{cases} \frac{1}{|\mathcal{U}_\gamma(i)|} \sum_{j \in \mathcal{U}_\gamma(i)} \log p(y|\boldsymbol{x}_j) & \text{if } |\mathcal{U}_\gamma(i)| \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

と構成する．ここで $\boldsymbol{x} = \{\boldsymbol{x}_i : i \in \mathcal{D}\}$．このように f_γ は $\mathcal{U}_\gamma$ だけの空間情報を組み込んだ判別関数である．従って，これらを空間情報を使ってない判別関数 $f_0(\boldsymbol{x}_i, y)$ と線形に結合させて作られた

$$F_r(i, \boldsymbol{x}, y) = \alpha_0 f_0(\boldsymbol{x}_i, y) + \alpha_1 f_1(i, \boldsymbol{x}, y) + \cdots + \alpha_r f_r(i, \boldsymbol{x}, y)$$

の係数 $\alpha_\gamma, \gamma = 0, \ldots, r$ をブーストアルゴリズムによって学習する人工的な例題や実際のリモートセンシングデータでこの方法を実行した．ICM などと比較して，高速な計算時間で有効な解析結果が得られることが示された（Nishii and Eguchi，2005a, 2006a, 2006b）．さらに空間の特徴抽出において，学習データの中に特定のラベル間のミスラベルが発生することがある．この問題に対してロバストな方法が提案されている（Nishii and Eguchi，2005c）．また，異なる解像度のリモートセンシングデータの融合のためのパラレルブーストの方法が考察されている．

海洋資源に関する問題で，特にマグロ漁獲における混獲問題について Kawakita et al.（2005）の研究も空間パタン認識の問題と密接な関連があるだろう．

5. ゲノムデータ解析

ゲノム科学は現在，急速な規模でゲノムデータを産出している．従来の想定を逸脱するデータの次元 p を持つことが特徴である．例えば，遺伝子発現のためのマイクロアレイ実験において一回の実験で計測される発現遺伝子数 p は，かつては数例だったものが現在では数万例に拡大されている．次元数 p が設定内であれば，何の問題もなく有効に働く統計方法であっても，その性能が保証される p と観測数 n との関係は相対的である．この意味で，現代のバイオテクノロジーで生産されるマイクロアレイ，SNP，プロテオームのデータ解析などは，従来のデータ解析方法では全くうまく働かない．ここに統計学の新しい課題が始まった．

この課題は，「$p \gg n$ 問題」と呼ばれ，広く認識されるようになってきた．このような形式を持つデータに対する適切なデータ解析法が，一部の発見的な方法を除いては整備されていないのが現状である．しかもこの問題は一過性のものでなく，バイオテクノロジーが進展すると共により深刻化するものである．実際，バイオテクノロジーの急速な発展に伴い p は膨大となる一方で，観測数 n は被験者のインフォームドコンセントの問題や，遺伝的背景と関連しない共変量の調整が必要となるので，一定数より大きくすることは困難である．この $p \gg n$ 形式のデータの解析の最大の問題点は見せかけの発見である．高次元 p の発現ベクトルが過剰な情報を与え，少数の例題 n では適切な検証性が得られないことから，本質的な困難さが生じる．多数の遺伝子に渡って調べると，全く遺伝子の情報を持たない発現量であっても見せかけの関連が出てしまう．

5.1 SNP 解析

ゲノムの DNA 塩基配列の中には，種が同じであっても個体によって異なるサイトがある．このように同一種内の多様性の 1 つに 1 塩基置換（SNP）がある．ヒトゲノムの場合，数百万の箇所に SNP があると報告されている．現在，SNP の特定のパターンと疾病，薬剤感受性などとの関連性について精力的な研究がされている．近い将来，個人の全ての SNP タイプを計測

して，それに基づいて医療がデザインされることが目標にされている．しかし，膨大な SNP タイプの中から有効なパタンを見つけることは，$p \gg n$ の典型的な例題となり解決しなければならない問題が山積している．

5.2 SNP タイピング

各被験者の SNP のタイプを計測することは SNP 解析の基礎となる．1 つの SNP サイトで可能なタイプは 3 である．例えば DNA 塩基を A と T とすると，可能なタイプは AA，AT，TT となる．SNP タイピングの一つにインベイダー法がある．この方法によって各被験者に 2 つの塩基 A と T に関係して 2 種類の蛍光色強度 x_A と x_T が観測される．このように 2 次元の観測ベクトル $\boldsymbol{x} = (x_A, x_T)$ に対して，高々3 グループのクラスタリングを行う統計問題に帰着される．このデータに次の付加的な情報が着目された．それは，一回の実験に 300 程度の被験者のデータに必ず，無反応な物質を複数個加えて，蛍光反応が観測されている．これによってバックグランドノイズが推定できて，原点の補正が可能である．私たちは幾つかの試験的な試みの後に，たどり着いた結論は，直接 $\boldsymbol{x} = (x_A, x_T)$ を使うのではなく，補正後の角度

$$y = \tan^{-1}\left(\frac{x_T}{x_A}\right)$$

だけを使うことにする．詳細な解析の方法については Fujisawa et al. (2004a) を参照．これにより x の長さの情報を無視することになる．このデータの縮約は，実験方法の性質上から蛍光反応の総量がタイピングの情報と無関係なことから支持される．これにより 3 タイプのクラスタリングは 1 変量の正規混合モデル

$$\pi_{\mathrm{AA}}\varphi(y;\mu_{\mathrm{AA}},\sigma^2_{\mathrm{AA}}) + \pi_{\mathrm{AT}}\varphi(y;\mu_{\mathrm{AT}},\sigma^2_{\mathrm{AT}}) + \pi_{\mathrm{TT}}\varphi(y;\mu_{\mathrm{TT}},\sigma^2_{\mathrm{TT}})$$

によって実行された．ここで $\varphi(y;\mu,\sigma^2)$ は平均 μ 分散 σ^2 の正規密度関数を表す．このモデルでの推測法は節 2 で紹介された β-ダイバージェンス最小法や更に罰則付き法も加えた方法で幾つかの困難な点を克服した．特に，データセットには，3 つのタイプが揃ってなくて 2 つのタイプや 1 つのタイプだけから構成されている場合があり，この場合は従来の最尤法はうまく働かないが，ここで開発された方法は適切な SNP タイピングが可能である．

5.3 SNP ハプロタイプブロック

前小節で個人の SNP タイピングが議論された．この小節では全ての SNP タイプが同定された被験者グループのデータ解析について考察する．このように個人の全ての SNP サイトでは塩基の一組からなる遺伝子型が特定され，SNP の遺伝子型の集合が得られているとする．このとき，SNP の遺伝子型のペアの塩基はどちらの親から受け継いだか分かれば，SNP の遺伝子型の集合を母方と父方で 2 本の長い DNA 配列のハプロタイプによって分けることができる．そのハプロタイプの可能な組み合わせは膨大な自然数である．目的はある疾患と関連するハプロタイプの発見である．ここにゲノムデータ解析の基本課題 $p \gg n$ が立ちはだかる．しかし，集団遺伝学の知見から現実のハプロタイプの分布は可能な全ての組み合わせに一様に分布しているのでなく，かなり限定的な組み合わせのみに分布していることが分かってきた．少数の‘祖先ハプロタイプ’が存在して，その組み換えによって現在の集団のハプロタイプ全体が構成されているという説から支持されている．この理解によってハプロタイプの全体はブロック毎に祖先ハプロタイプの一部が組み合わさって，これをハプロタイプ・ブロックと呼ぶ．ここでブロックとは隣接する SNP サイトの 1 かたまりを表す．このような考察から先祖由来モデルが提案され，癌研究所情報解析グループと共同研究によって特定のハプロタイプ・ブロックと疾患の関連性について共同研究がされている (Fujisawa et al., 2004b を参照)．このため SNP のハプロタイプ・ブロックの特定のためのアルゴリズムを提案し，癌研究所の SNP タイピング

データと臨床の情報を結合させて，このブロックアルゴリズムを適用して，特定のハプロタイプと薬剤感受性の高い関連を示す事例についての研究が進行中である．

5.4 プロテオーム解析

質量分光分析によって被験者の血清からタンパク・プロファイルを生成する方法が急速に普及している．このプロテオーム・データの解析について考察する．Fushiki et al.（2006）を参照．目的はプロテオームのマスの関数データから検体の正常細胞とガン細胞を識別するピークパタンの発見である．言うまでもなくこのような関数データには原理的には無限自由度のピークパタンが埋め込み可能である．これを少数例から探索することはゲノムデータの基本課題と本質的に関わっている．しかも実験の方法の性質上，タンパクの高分子量の領域は観測精度が劣っている点，質量に対応する x 軸も観測誤差が顕在化する点など特別な注意をすべき点も多い．標準的な前処理の方法によって被験者グループ全体に渡って TOF マス分布のピーク時点の補正が成された．

次に目的のパタンの抽出のために，特徴量としてのピークの数量化について 2 つ方法を考えた．連続値のピーク値そのものを使う方法とピークの有無の判定によって 0-1 に縮約して使う方法である．このように，被験者に p 個のピーク値の 2 通りの数値化を考え，それぞれの場合において特徴ベクトル x から正常細胞であるかガン細胞かのクラスラベル y を予測する問題に帰着される．パタン抽出のために採用された方法はアダブーストである．両方の特徴ベクトルから異なる傾向のパタンが選ばれる傾向があることが分った．この傾向について更に，実解析と理論の方向から詳細な考察が成されている．

この研究のキーは“ 情報の持つ特徴は共通ピークに現れる ”という考えである．ある被験者のプロファイルに明確なピークが検出されてもグループとしてマイナーであると認定されると特徴空間から除外する．このアイデアによって，対象にするピーク集合は現実的な規模に収めることができ，ピークパタン予測に有効であることが米国の国立癌研究所（NIC）において公開された卵巣ガンのデータ ‘Ovarian Dataset 8-7-02’ の解析によって示されている．

5.5 疾病と遺伝子発現パタン

ゲノムデータの中核を成すのは，やはりマイクロアレイデータであろう．遺伝子の発現を見るために大規模なアレイ実験が急速に進歩して飛躍的な普及が遂げられている．この節ではマイクロアレイデータから疾病との関連の特徴抽出を考察する．標準的な前処理が施された後に対象にする p 個の遺伝子の発現量からなる p 次元特徴ベクトルから疾患のランクを表すラベル y を予測する問題を考えよう．簡単のため，疾病の有無だけを表す 2 クラスの問題を扱う（解説は江口, 2005 参照）．癌研究所から提供されるデータを中心にデータ解析を行っている．そのための節 4 で紹介された方法論の開発が主要な目的である．

5.6 グループブースト

節 4 で紹介された U-ブーストの標準法であるアダブースト法がマイクロアレイデータのパタン認識の問題に試みられたが，実は良い結果は得られなかった．様々な理由が考えられるが，第 1 の理由は“ $p \gg n$ ”問題であろう．実際には $(p,n) = (1000, 20)$ 程度であったが，やはり，n が相対的に小さすぎる．学習に使う判別子セットは単一の遺伝子の発現量が，あらかじめ決められた閾値を超えるかどうかで疾病の有無を決める判別子の全てを採った．要するに

$$\mathcal{H}_1 = \{\mathrm{sgn}(x_j - b) : j = 1, \ldots, p, b \in \mathbf{R}\} \cup \{-\mathrm{sgn}(x_j - b) : j = 1, \ldots, p, b \in \mathbf{R}\}$$

に限定した．これより，複数個の遺伝子の交互作用による特徴量は考えていない．しかし，この制限された $\mathcal{H}_1$ でさえ特徴量の情報の過剰から適切な解析結果は得られない．典型的なアダブーストの挙動は T が 3，4 でトレーニングエラーが 0 となり，選ばれた T 個の遺伝子を除去

して，残りの $p-T$ 個の遺伝子で，再度，繰り返しても，やはりほぼ，たった T 回でアルゴリズムが終了してしまう．結局，この繰り返しによって，適切な結果が得られない．過大な特徴量から本質的な働きをする遺伝子発現が見せかけの関連性を示す遺伝子発現の中に埋没してしまう．

この困難な問題を解決するためにグループブーストという方法を提案する（Takenouchi et al., 2005 を参照）．このキーとなる考えは小節 4.2 の U-ブーストアルゴリズムのステップ(B-1)の代わりに判別子 h の重み付きエラーレイト $\epsilon_t(h)$ に関して $\mathcal{H}_1$ の中から上位 k まで選ぶ．選ばれた判別子 $h_t^{(1)},\ldots,h_t^{(k)}$ を一つの判別関数

$$\bar{h}_t = \frac{1}{k}\sum_{j=1}^{k} \alpha_t^{(j)} h_t^{(j)}$$

に集約して次のステップ(B-3)に渡すというものである．ここで，結合係数 $\alpha_t^{(1)},\ldots,\alpha_t^{(k)}$ は(B-2)と同様に

$$\alpha_t^{(j)} = \operatorname*{argmin}_{\alpha} \ell_U^{\mathrm{emp}}\left(F_{t-1} + \alpha h_t^{(j)}\right)$$

と定める．この方法では，ブーストの学習の過程で使われる判別子セット $\mathcal{H}_1$ から各ステップでベストワンを選ぶのではなく，ベスト k を選んでいる．これによって，あるステップで意味のない遺伝子が選ばれたときでも，残りの $k-1$ の遺伝子によってその影響を緩和している．通常 k は想定される候補遺伝子の数と同程度に選択される．これによって，ここでの $p \gg n$ 問題の解決が図られた．

$p \gg n$ 問題の別のアプローチは同じ目的によって異なる施設で得られたマイクロアレイデータを結合させる方法がある．要するに過大な p に対して，複数の n を併合しようというわけだ．しかしながら，マイクロアレイ解析は施設ごとに異なる実験条件で行われて，時にはプラットホーム自体が，異なる方式だったりする．これらの相違性から生じるバイアスを考慮しながら，複数のデータセットを融合させて高次元データ・小サンプル数の問題を検討している．アダブーストの並列化アルゴリズムを考案して幾つかの適用を試行中である．

新領域融合研究の一環で イネゲノムと系統マウスの遺伝子発現の研究を行っている．従来のマイクロアレイのデータと異なり非常に厳密に管理された系統の発現データであることが特徴的である．ただし Affymetrix 社のプローブは 1 つの系統しか用意されてない．従って系統間の SNP による偽の発現差が生じる現象の特定がこの研究プロジェクトのキーとなる．現在，イネゲノム，系統マウス，双方の発現データに対して適切な統計方法について開発の途上にある．

6. 予測と発見の先に

予測と発見という大きなパラダイムからゲノムデータ解析のための統計的方法を目指して最小ダイバージェンス法のクラス適用について検討した．特に，具体的なデータの形式，SNP，マイクロアレイ，プロテオームについてゲノムデータ解析のための統計的方法の基本課題から問題点が考察された．まだまだ，これからの課題が山積され，より精力的に多くの努力と時間とが必要であることは言うまでもない．

ここでゲノムデータ解析から統計研究者に求められている別の面についても考えたい．この観点に立てば，予測と発見の先に‘ 検証 ’という別の大きなパラダイムがある．予測と発見という探索的な研究のスタンスから予測され，発見された仮説や命題の‘ 確からしさ ’を正確に計る必要性が生じる．

現在，いまだかつて経験のない規模のゲノムデータが産出されているが，それぞれのゲノム

データは，それぞれの背景から，それぞれの目的を持って獲得されている．そして，それぞれの研究視点から，予測と発見がなされている．最近，情報公開の流れから，多くの研究施設ではこれらのゲノムデータが公開されている．例えば，マイクロアレイデータについては世界規模のデータベースの構築が進められている．これらの公表されたデータに基づいて発見された仮説の検証はできないのだろうか．実は，このような解析は一般的にはメタアナライシス，またはシステマティクレビューと呼ばれている．この解析から適切な結果を得るためには，いつも観察研究の統計解析に問題となるいわゆる‘ 選択バイアス ’が問題となる．一般的な定式化が Copas and Eguchi (2001, 2005) にある．公表データの解析に重大な問題となる公表バイアス (出版バイアス) について研究を進める必要がある．Henmi et al. (2006) では，結果変数を条件付きにした出版確率によって定義される選択確率の及ぼす最悪評価を考察している．

多くのゲノムデータは，競争的な審査プロセスを持つ学術雑誌から採択された論文から公表される．したがって，大きな割合を占めるであろう公表されない (未出版) データはミッシングである．また，日進月歩のバイオテクノロジーの進展に伴い，ゲノムデータの精度の変化についても考察されるべきである．マイクロアレイ一つとっても，データとしての不均一性はかなり大きいといわざるを得ない．このような蓄積されるゲノムデータからの予測と発見の検証も私たちの大きな課題であると考えている．

謝　辞

この論文に対して建設的なコメントをいただいた査読者に感謝します．また原稿の段階でコメントいただいた藤澤 洋徳 氏，伏木 忠義 氏にも感謝します．

参 考 文 献

Amari, S. (1985). *Differential-Geometrical Methods in Statistics*, Lecture Notes in Statistics, Vol. 28, Springer-Verlag, New York.

Amari, S. and Nagaoka, H. (2000). *Methods of Information Geometry*, Translations of Mathematical Monographs, Vol. 191, Oxford University Press, Oxford.

Barndorff-Nielsen, O. E. (1978). *Information and Exponential Families in Statistical Theory*, Wiley, Chichester.

Besag, J. (1986). On the statistical analysis of dirty pictures, *Journal Royal Statistical Society B*, **48**, 259–302.

Bishop, C. (1995). *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford.

Copas, J. (1988). Binary Regression Models for Contaminated Data, *Journal Royal Statistical Society B*, **50**, 225–265.

Copas, J. and Eguchi, S. (2001). Local sensitivity approximation for selectivity bias, *Journal Royal Statistical Society B*, **63**, 871–895.

Copas, J. and Eguchi, S. (2005). Local model uncertainty and incomplete data bias (with discussion), *Journal Royal Statistical Society B*, **67**, 459–512.

Eguchi, S. (1983). Second order efficiency of minimum contrast estimators in a curved exponential family, *Annals of Statistics*, **11**, 793–803.

Eguchi, S. (1992). Geometry of minimum contrast, *Hiroshima Mathematical Journal*, **22**, 631–647.

江口真透. (2004). 統計的パタン認識の情報幾何—*U*-ブースト学習アルゴリズム—，数理科学，**489**, 53–59.

江口真透. (2005). DNA チップデータ解析において統計学の役割は何か?，バイオテクノロジージャー

ナル, **5**, 430–435.

Eguchi, S. (2005). Information geometry and statistical pattern recognition, *Sugaku Exposition*, American Mathematical Society (to appear).

Eguchi, S. and Copas, J. (1998). A class of local likelihood methods and near-parametric asymptotics, *Journal of Royal Statistical Society B*, **60**, 709–724.

Eguchi, S. and Copas, J. B. (2001). Recent developments in discriminant analysis from an information geometric point of view, *Journal Korean Statistical Society*, **30**, 247–264.

Eguchi, S. and Copas, J. B. (2002). A class of logistic type discriminant functions, *Biometrika*, **89**, 1–22.

Eguchi, S., Kim, T.-Y. and Park, B. U. (2003). Local likelihood method: A bridge over parametric and nonparametric regression, *Journal of Nonparametric Statistics*, **15**, 665–683.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences*, **55**, 119–139.

Friedman, J. H., Hastie, T. and Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting, *Annals of Statistics*, **28**, 337–407.

Fujisawa, H. and Eguchi, S. (2005). A new approach to robust parameter estimation against heavy contamination, Research Memo., No. 947, The Institute of Statistical Mathematics, Tokyo.

Fujisawa, H. and Eguchi, S. (2006). Robust estimation in the normal mixture model, *Journal Statistical Planning and Inference*, **136**, 3989–4011.

Fujisawa, H., Eguchi, S., Ushijima, M., Miyata, S., Miki, Y., Muto, T. and Matsuura, M. (2004a). Genotyping of single nucleotide polymorphism using model-based clustering, *Bioinformatics*, **20**, 718–726.

Fujisawa, H., Isomura, M., Eguchi, S., Ushijima, M., Miyata, S., Miki, Y. and Matsuura, M. (2004b). Identifying haplotype block structure by using ancestor-derived model and MDL principle, Research Memo., No. 928, The Institute of Statistical Mathematics, Tokyo.

Fushiki, T., Fujisawa, H. and Eguchi, S. (2006). Identification of biomarkers from mass spectrometry data, *BMC Bioinformatics*, **7**, 358.

Good, I. J. (1952). Rational decisions, *Journal Royal Statistical Society B*, **14**, 107–114.

Grünwald, P. D. and Dawid, A. P. (2004). Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory, *Annals of Statistics*, **32**, 1367–1433.

Hastie, T., Tibishirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning*, Springer, New York.

Henmi, M., Copas, J. and Eguchi, S. (2006). Confidence intervals and P-values for meta analysis with publication bias (to be submitted).

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

Higuchi, I. and Eguchi, S. (2004). Robust principal component analysis with adaptive selection for tuning parameters, *Journal of Machine Learning Research*, **5**, 453–471.

Hjort, N. L. and Jones, M. C. (1996). Locally parametric nonparametric density estimation, *Annals of Statistics*, **24**, 1619–1647.

Huber, P. J. (1981). *Robust Statistics*, John Wiley & Sons, New York.

Hyvarinen, Karhunen, A. and Oja, K. (2001). *Independent Component Analysis*, Wiley, New York.

Kamiya, H. and Eguchi, S. (2001). A class of robust principal component vectors, *Journal of Multivariate Analysis*, **77**, 239–269.

Kanamori, T., Takenouchi, T., Murata, N. and Eguchi, S. (2004). The most robust loss function for boosting, *Neural Information Processing*, 11th International Conference, ICONIP 2004, Calcutta, Lecture Notes in Computer Science, Vol. 3316, 496–501, Springer, Berlin.

Kanamori, T., Takenouchi, T., Murata, N. and Eguchi, S. (2006). Robust loss functions for boosting, *Neural Computation* (to appear).

Kawakita, M. and Eguchi, S. (2006). Boosting method for local learning in statistical pattern recognition (to be submitted).

Kawakita, M., Minami, M., Eguchi, S. and Lennert-Cody, C. E. (2005). An introduction to the predictive technique AdaBoost with a comparison to generalized additive models, *Fisheries Research*, **76**, 328–343.

Lebanon, G. and Lafferty, J. (2002). Boosting and maximum likelihood for exponential models, *Advances in Neural Information Processing Systems*, **14**.

Lugosi, G. and Vayatis, N. (2004). On the Bayes-risk consistency of regularized boosting methods, *Annals of Statistics*, **32**, 30–55.

McLachlan, G. J. (2004). *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, New York.

Minami, M. and Eguchi, S. (2002). Robust blind source separation by beta-divergence, *Neural Computation*, **14**, 1859–1886.

Mollah, N. H., Minami, M. and Eguchi, S. (2005a). Robust prewhitening for ICA by minimizing beta-divergence and its application to FastICA (in revision).

Mollah, N. H., Sultana, N., Minami, M. and Eguchi, S. (2005b). Exploring local PCA structure for dimensionality reduction by minimizing β-divergence, Research Memo., No. 956, The Institute of Statistical Mathematics, Tokyo.

Mollah, N. H., Minami, M. and Eguchi, S. (2006). Exploring latent structure of mixture ICA models by the minimum beta-divergence method, *Neural Computation*, **18**, 166–190.

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

Nishii, R. and Eguchi, S. (2005a). Supervised image classification by contextual AdaBoost based on posteriors in neighborhoods, *IEEE Transaction on Geoscience and Remote Sensing*, **43**, 2547–2554.

Nishii, R. and Eguchi, S. (2005b). Spatio-temporal contextual image classification based on spatial AdaBoost, IEEE International Geoscience and Remote Sensing Symposium, Seoul.

Nishii, R. and Eguchi, S. (2005c). Robust supervised image classifiers by spatial AdaBoost based on robust loss functions, SPIE Remote Sensing Europe, Bruges.

Nishii, R. and Eguchi, S. (2006a). Supervised image classification of multispectral images based on statistical machine learning, *Signal and Image Processing for Remote Sensing* (ed. C. H. Chen), Taylor and Francis Books (to appear).

Nishii, R. and Eguchi, S. (2006b). Image classification based on Markov random field models with Jeffreys divergence (in revision).

Ohara, A. and Eguchi, S. (2005). Geometry on positive definite matrices and V-potential function, Research Memo., No. 950, The Institute of Statistical Mathematics, Tokyo.

Park, B. U., Lee, Y. K., Kim, T. Y., Park, C. and Eguchi, S. (2005). Local likelihood density estimation when the bandwidth is large, *Journal of Statistical Planning and Inference*, **136**, 839–859.

Pistone, P. and Sempi, C. (1995). An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one, *Annals of Statistics*, **23**, 1543–1561.

Rao, C. R. (1945). Information and accuracy attainable in the estimation of statistical parameters, *Bulletin Calcutta Mathematical Society*, **37**, 81–91.

Rätsch, G., Onoda, T. and Müller, K.-R. (2001). Soft margins for AdaBoost, *Machine Learning*, **42**, 287–320.

Schapire, R. E. (1990). The strength of weak learnability, *Machine Learning*, **5**, 197–227.

Schapire, R. E., Freund, Y., Bartlett, P. and Lee, W. S. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods, *Annals of Statistics*, **26**, 1651–1686.

Scott, D. W. (2001). Parametric statistical modeling by minimum integrated square error, *Technometrics*, **43**, 274–285.

Takenouchi, T. and Eguchi, S. (2004). Robustifying AdaBoost by adding the naive error rate, *Neural Computation*, **16**, 767–787.

Takenouchi, T., Ushijima, M. and Eguchi, S. (2005). GroupAdaBoost for selecting important genes, IEEE 5th Symposium on Bioinformatics and Bioengineering, Mineapolis.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*, Springer-Verlag, New York.

Proceedings of the Institute of Statistical Mathematics Vol. 54, No. 2, 375–403 (2006) 403

Prediction and Discovery: Towards Novel Methodology for Genome Data Analysis

Shinto Eguchi

The Institute of Statistical Mathematics

A class of minimum divergence methods is proposed to improve the defect of the maximum likelihood method in terms of statistical discussion including applications of PCA, ICA and pattern recognition. A challenging problem in genome data analyses is discussed, and minimum divergence methods are applied to genome data including SNPs, proteome, and microarray as an approach to solving the problem.

Key words: U-divergence, U-model, U-loss function, gene expression, robust, information geometry.