情報論的学習理論テクニカルレポート 2009
Technical Report on Information-Based Induction Sciences 2009 (IBIS2009)

独立性の一般化に基づく統計モデルの拡張
An Extension of Some Statistical Models based on Generalized Independence

藤本 悠* 村田 昇†
Yu Fujimoto Noboru Murata

Abstract: In this paper, a generalized statistical independence is proposed from the viewpoint of a generalized multiplication motivated by the Bregman divergence, and it is implemented in some simple statistical models for joint probability distribution, such as independent models and naive Bayes models. Our models are deeply related to Archimedean copulas but slightly different. This paper also proposes an idea of their estimation method which directly uses empirical marginal distributions to retain simplicity of calculation. Our method is interpreted as an optimization of a rough approximation of the Bregman divergence and geometrically explained. Effectiveness of our proposed models is shown by numerical experiments on some benchmark data sets.

1 まえがき

Naive Bayes model(NB) や Bayesian network[10]，あるいは pLSA[8] など，何らかの形で独立性を仮定した統計モデルは，扱いやすさなどの点から様々な状況における統計的推論で広く用いられている．通常このような独立性の仮定はデータに基づいて導入されるべきものであるが，例えば NB などは独立性の仮定が崩れていても判別器としての性能をある程度維持できることが知られているため [4]，データから厳密な独立性が言えない場合であっても実用上の近似として用いられることが多い．また逆に変数間の依存関係を表現しようとすると，特に状態数を多く持つ離散変数を扱う場合には非常に多くのパラメタを導入しなければならず，データへのオーバーフィットの傾向が強くなってしまうという問題も出てくる．本稿ではこの独立性を一般化することで一種の弱い依存関係を表現することを試み，その一般化独立性を利用した独立モデル，及び NB の拡張を提案する．

統計的推論の場では確率値の乗除算が自然に用いられ，例えば変数間の独立性を論じる際には周辺分布の積が基本演算として重要な役割を果たしている．この乗除算は統計的な分布間の乖離尺度を表すのに広く用いられている Kullback-Leibler(KL) 情報量の意味での統計分布の一種の混合と解釈することができる．一般的な統計モデルの推定を考えた時，KL 情報量に基づく推定（いわゆる最尤推定）はサンプル数の増加に伴う漸近有効性などの好ましい性質を持っているが，外れ値には著しく弱いという一面もある．一方，KL 情報量の一種の一般化である Bregman 情報量 [11, 5] の意味での推定を考えることで最尤推定にはないロバスト性 [7] などの好ましい推定結果を獲得し得ることが知られている．本稿ではこのような背景をふまえ，Bregman 情報量に基づく乗除算の一般化を行い，それに伴うモデルの拡張とその性質について議論を行う．

まず第 2 章では Bregman 情報量を導入し，それに伴う乗除算，独立性の一般化を考える．第 3 章では独立性の一般化による統計モデルの拡張，及びその性質や推定方法について述べ，特に NB への実装に関する議論を行う．また，第 4 章では一般化独立性を導入した NB をベンチマークデータを使った数値実験によって評価する．最後に第 5 章で本稿のまとめと今後の展望を述べる．

*青山学院大学, 229-8558 , 神奈川県相模原市渕野辺 5-10-1 tel. 042-759-6322, e-mail yu.fujimoto@it.aoyama.ac.jp,
Aoyama Gakuin University, 5-10-1 Fuchinobe, Sagamihara, Kanagawa 229-8558
†早稲田大学, 169-8555, 東京都新宿区大久保 3-4-1, tel. 03-5292-4563, e-mail noboru.murata@eb.waseda.ac.jp,
Waseda University, 3-4-1 Ohkubo, Shinjuku, Tokyo 169-8555

2 Bregman情報量と独立性の一般化

本章では適当な関数$U(\cdot)$によって特徴づけられるBregman情報量を導入し，その関数Uによって記述される乗除算の一般化を行う．

2.1 Bregman情報量

離散変数空間$\mathcal{X}$が与えられた時に，次のような性質を持つ関数の集合として表現される空間$\mathcal{F}_\mathcal{X}$を考える．

$$\mathcal{F}_\mathcal{X} = \left\{ p_\mathcal{X}(x) \;\middle|\; p_\mathcal{X} : \mathcal{X} \to R_+, \sum_{x\in\mathcal{X}} p_\mathcal{X}(x) < \infty \right\}$$

ただし，ここでR_+は0を含む正の実数値の集合とする．すると一般に統計学で解析の対象とされる離散確率分布としての性質を持つ分布関数の空間$\mathcal{P}_\mathcal{X}$は次のように$\mathcal{F}_\mathcal{X}$の部分集合として表現されることになる．

$$\mathcal{P}_\mathcal{X} = \left\{ p_\mathcal{X}(x) \;\middle|\; p_\mathcal{X} : \mathcal{X} \to R_+, \sum_{x\in\mathcal{X}} p_\mathcal{X}(x) = 1 \right\} \subset \mathcal{F}_\mathcal{X}$$

これにより，ある離散分布$p_\mathcal{X}$は分布空間$\mathcal{P}_\mathcal{X}$，さらにより一般的な関数空間$\mathcal{F}_\mathcal{X}$上の1点に対応することになる．本稿では関数空間を$\mathcal{F}$，分布空間を$\mathcal{P}$と記述し，対象とする変数空間が自明でない場合は添字で明示する．

ここで2つの離散分布の間の乖離の度合いを測る1つの指標としてBregman情報量 [11] を導入する．

定義 1 (Bregman情報量) 離散変数Xに関する*2*つの分布関数$p, q \in \mathcal{P}_\mathcal{X}$間の統計的な乖離を測る*Bregman*情報量は凸関数$U(\cdot)$とその導関数$u(\cdot) = U'(\cdot)$の逆関数$\xi(\cdot) = u^{-1}(\cdot)$によって

$$\mathrm{D}_U(p,q) = \sum_{x\in X} \Big\{ U(\xi(q(x))) - U(\xi(p(x))) - p(x)(\xi(q(x)) - \xi(p(x))) \Big\}$$

と定義される[1]． ■

ここで$U(\cdot) = \exp(\cdot)$の時にはKL情報量

$$\mathrm{D}_{\mathrm{KL}}(p,q) = \sum_{x\in\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \tag{1}$$

と一致することから，Bregman情報量は関数UによってKLを一般化したものと考えることができる．表1に典型的なBregman情報量を構成するu関数，ξ関数の例を示す．

情報量を定義することにより関数空間の中での2つの関数間の位置関係を幾何学的に議論することが可能になる [11] が，その際，次の2種類の平坦性が重要となる．

[1] 本稿では離散変数の分布関数のみに着目しているが，Bregman情報量自体は任意の2つの関数$p_\mathcal{X}, q_\mathcal{X} \in \mathcal{F}_\mathcal{X}$に対して定めることができ，また同様に連続値変数に関しても定義することができる．

定義 2 (平坦性) *2*つの関数$f, g \in \mathcal{F}_\mathcal{X}$の次のような内分点$r(t)$の集合を$m$-平坦な空間と呼ぶ．

$$r(t) = (1-t)\cdot f + t\cdot g, \quad 0 \le t \le 1$$

同様に次のような内分点$r(t)$の集合をu-平坦な空間と呼ぶ．

$$r(t) = u((1-t)\cdot\xi(f) + t\cdot\xi(g)), \quad 0 \le t \le 1$$

この内分点$r(t)$のようにf, gの線形和（または$\xi(f), \xi(g)$の意味での線形和）をm-混合（u-混合）と呼ぶ． ■

ここでは一般にm-平坦な空間とu-平坦な空間は一致せず，特にu-平坦な空間は関数Uの形状に依存していることのみを述べておく．

2.2 U乗算と独立性の一般化

2変数X，Yの同時確率$p_{\mathcal{XY}}$から，周辺確率が次のように与えられているとする．

$$p_\mathcal{X}(x) = \sum_{y\in\mathcal{Y}} p_{\mathcal{XY}}(x,y), \quad p_\mathcal{Y}(y) = \sum_{x\in\mathcal{X}} p_{\mathcal{XY}}(x,y)$$

ここで同時確率が

$$\begin{aligned} p_{\mathcal{XY}}(x,y) &= p_\mathcal{X}(x) \times p_\mathcal{Y}(x) \\ &= \exp(\log(p_\mathcal{X}(x)) + \log(p_\mathcal{Y}(y))) \end{aligned} \tag{2}$$

のように周辺確率の積によって表される時，2変数は統計的に独立であるとみなされる．これはKL情報量を特徴づけるu関数$(\exp(\cdot))$内のξ関数$(\log(\cdot))$の意味での周辺確率の和で同時確率を定義したものと解釈できる．ここで式(2)を関数Uの意味で一般化すると次のような式が得られる．

$$f_{\mathcal{XY}}(x,y) = u(\xi(p_\mathcal{X}(x)) + \xi(p_\mathcal{Y}(y))) \in \mathcal{F}_{\mathcal{XY}} \tag{3}$$

ただし式(3)を用いると通常$f_{\mathcal{XY}} \notin \mathcal{P}_{\mathcal{XY}}$となるため，通常このような関数$f_{\mathcal{XY}}$を直接統計演算の中で用いることはできない．そこで

$$f_{\mathcal{XY}} \in \mathcal{F}_{\mathcal{XY}} \longmapsto p_{\mathcal{XY}} \in \mathcal{P}_{\mathcal{XY}} \tag{4}$$

という分布空間への射影(正規化)を考える．このような正規化としては例えば次のようなものが考えられる．

定義 3 *(m-正規化)* 任意の関数$f_{\mathcal{XY}} \in \mathcal{F}_{\mathcal{XY}}$から$\mathcal{P}_{\mathcal{XY}}$への射影として

$$p_{\mathcal{XY}}(x,y) = \frac{f_{\mathcal{XY}}(x,y)}{\sum_{x'\in\mathcal{X}, y'\in\mathcal{Y}} f_{\mathcal{XY}}(x',y')} \in \mathcal{P}_{\mathcal{XY}} \tag{5}$$

という正規化を考えることができる．式(5)によって実現される射影をm-正規化と呼ぶ． ■

表 1: Bregman 情報量の例.

	$u(z)$	$\mathrm{dom}(z)$	$\mathrm{range}(u)$	$\xi(z)=u^{-1}(z)$	$\mathrm{dom}(z)$	$\mathrm{range}(\xi)$	$\mathrm{dom}(\pi)$
KL	$\exp(z)$	$(-\infty,\infty)$	$(0,\infty)$	$\log(z)$	$(0,\infty)$	$(-\infty,\infty)$	-
Ex.1	$\exp(\mathrm{sgn}(z)\lvert z\rvert^{\frac{1}{\pi}})$	$(-\infty,\infty)$	$(0,\infty)$	$\mathrm{sgn}(\log(z))\lvert\log(z)\rvert^{\pi}$	$(0,\infty)$	$(-\infty,\infty)$	$(0,\infty)$
Ex.2	$(\pi z+1)^{\frac{1}{\pi}}$	$[-\frac{1}{\pi},\infty)$	$[0,\infty)$	$\frac{z^{\pi}-1}{\pi}$	$[0,\infty)$	$[-\frac{1}{\pi},\infty)$	$(0,\infty)$
		$(-\infty,-\frac{1}{\pi})$	$(0,\infty)$		$(0,\infty)$	$(-\infty,-\frac{1}{\pi})$	$(-\infty,0)$
Ex.3	$\exp(z)+\pi$	$(-\infty,\infty)$	(π,∞)	$\log(z-\pi)$	(π,∞)	$(-\infty,\infty)$	$(-\infty,\inf(z))$
Ex.4	$\exp\left(\frac{1-\exp(-z)}{\pi}\right)$	$(-\infty,\infty)$	$(0,\exp(\frac{1}{\pi}))$	$-\log(1-\pi\log(z))$	$(0,\exp(\frac{1}{\pi}))$	$(-\infty,\infty)$	$(0,\frac{1}{\log(\sup(z))})$

また，同様に次のような射影も考えることができる．

定義 4 *(u-正規化)* 任意の関数 $f_{\mathcal{XY}}\in\mathcal{F}_{\mathcal{XY}}$ から $\mathcal{P}_{\mathcal{XY}}$ への射影として

$$p_{\mathcal{XY}}(x,y)=u(\xi(f_{\mathcal{XY}}(x,y))-c) \tag{6}$$

という正規化を考えることができる．ただしここで c は

$$\sum_{x\in\mathcal{X},y\in\mathcal{Y}}u(\xi(f_{\mathcal{XY}}(x,y))-c)=1 \tag{7}$$

を満たすための定数項の意である．式 (6) によって実現される射影を u-正規化と呼ぶ． ■

以下では後に述べる整合性の観点から u-正規化を採用し，同時分布が

$$\begin{aligned}p_{\mathcal{XY}}(x,y)&=u(\xi(p_{\mathcal{X}}(x))+\xi(p_{\mathcal{Y}}(y))-c)\in\mathcal{P}_{\mathcal{XY}}\\&=p_{\mathcal{X}}(x)\otimes p_{\mathcal{Y}}(y)\end{aligned} \tag{8}$$

のように周辺分布のみによって表現されることを U-独立と呼ぶ [6]．またこのような同時分布を得るための演算を U 乗算と呼び，記号 $\otimes$ で表す．

また式 (8) より，同時分布から周辺分布を得るような演算

$$\begin{aligned}p_{\mathcal{X}}(x)&=u(\xi(p_{\mathcal{XY}}(x,y))-\xi(p_{\mathcal{Y}}(y))-c')\\&=p_{\mathcal{XY}}(x,y)\oslash p_{\mathcal{Y}}(y)\end{aligned} \tag{9}$$

を U 除算と呼び，記号 $\oslash$ で表すことにする．なお式 (8) 及び式 (9) はそれぞれ関数 $u(\xi(p_{\mathcal{X}})+\xi(p_{\mathcal{Y}}))$，$u(\xi(p_{\mathcal{XY}})-\xi(p_{\mathcal{Y}}))$ を u-正規化したものとなっているが，m-正規化を用いた場合にはこのような整合性は現れないことを付記しておく．

U-独立性を論じる際には各関数の定義域と値域について把握しておく必要がある．表 1 に代表的な関数 u 及び ξ の定義域，値域，及びそれらを特徴づける関数のパラメタ π の定義域を示している．これらの定義域を外れると，前提としていた関数 $U(\cdot)$ の凸性が成り立たなくなったり，逆関数が定義できなかったりなどの不整合が生じる．U-独立性の議論を行うためにはこれに加えて

1.式 (8) の関数 u の中での各項の和 (正規化項 c も含む) が u 関数の定義域内にある．

2.「関数 u の値域 $\supseteq p_{\mathcal{XY}}$ の値域」である．

という 2 つの条件が満たされている必要がある[2]．そのため，以降 U-独立性を論じる際には上記の条件を満たすように u 関数，ξ 関数，及びこれらのパラメタ π を選ぶことにする．

ここで U-独立性と通常の独立性の違いについて少し触れておく．直感的に関数 u(及び ξ) と U 乗算の関係を調べる為に $z_1,z_2\in[0,1]$ に対する $u(\xi(z_1)+\xi(z_2))$ の形状を描いたのが図 1 である．なお，ここでは演算結果の概形にのみ興味があるため正規化項 c については考慮していない．図から分かるように U 乗算は関数 u (ξ) を変えることで，通常の乗算 (図 1(c)) と比べて周辺確率値の大小の違いを部分的に誇張したり，逆に鈍らせたりする効果がある．この結果より式 (8) で定義された U-独立性は周辺分布による弱い特殊な依存関係の表現とも解釈することができる．

2.3 Copula との関係

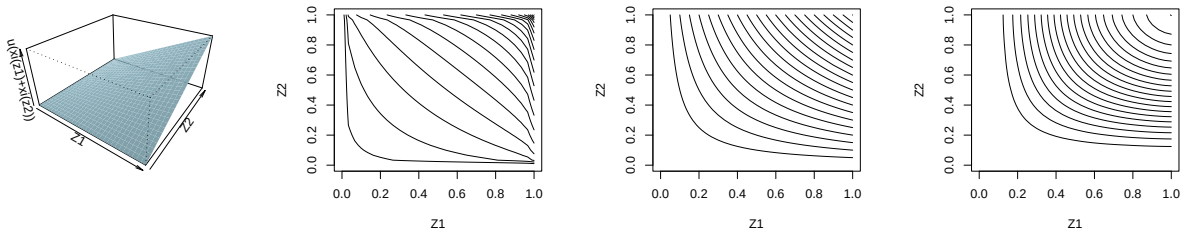
前節のような周辺分布に基づく独立性の議論は copula[12] と深い関係があり，特に逆関数の性質を用いて定義を行う点は Archimedean copula と呼ばれる copula と同じ形式になっている．本章では U-独立性と copula の比較を述べていく．

Copula は連続値変数を対象に，累積周辺分布によって累積同時分布を再現する手法のことで，次のように定義される．

定義 5 (copula) 全ての $u\in[0,1],v\in[0,1]$ に対して

$$C(u,0)=0=C(0,v)$$

[2]なお，条件 2 は厳密には KL の場合にも成り立っていないが，確率分布の値域を便宜的に $p_{\mathcal{XY}}(x,y)\in(0,1)$(連続値の場合は $p_{\mathcal{XY}}(x,y)\in(0,\infty)$) とすることで回避できる．



(a) 鳥瞰図 ($\pi = 1.0$). (b) 等高線図 ($\pi = 0.5$). (c) 等高線図 ($\pi = 1.0$). (d) 等高線図 ($\pi = 1.5$).

図 1: Ex.1 の場合の $u(\xi(z_1) + \xi(z_2))$ の形状（$\pi = 1.0$ において通常の乗算と一致）.

及び,

$$C(u, 1) = u \quad and \quad C(1, v) = v$$

が成り立ち，$u_1 \leq u_2$，$v_1 \leq v_2$ であるような $u_1, u_2, v_1, v_2 \in [0, 1]$ に対して

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0$$

が成り立つような関数 $C : [0, 1]^2 \to [0, 1]$ を *(2次元)copula* と呼ぶ． ■

中でも，特に次の形を持つ copula のことを Archimedean copula と呼ぶ.

定義 6 (Archimedean copula) $\phi(1) = 0$ であるような単調減少関数 $\phi : [0, 1] \to [0, \infty]$ によって次のように定義された C を *Archimedean copula* と呼ぶ.

$$C(u, v) = \phi^{[-1]}(\phi(u) + \phi(v)) \tag{10}$$

ただし，ここで $\phi^{[-1]}$ は ϕ の厳密な逆関数 ϕ^{-1} ではなく，

$$\phi^{[-1]}(t) = \begin{cases} \phi^{-1}(t) & (0 \leq t \leq \phi(0)) \\ 0 & (\phi(0) \leq t \leq \infty) \end{cases} \tag{11}$$

の意である． ■

ここで Archimedean copula を構成する $\phi(z)$ を $\xi(z)$ とすると，定数項の存在を除いて前述の U-独立性の式 (8) に対応することが分かる．例えば関数 $\phi(z) = (-\log z)^\pi$ を用いて構成した Archimedean copula[9] は表 1 に示した Ex.1 によって構築される U-独立性と発想としては同じものになるが，連続値の累積周辺分布を対象としている点などが異なる[3]．両者の比較をまとめると表 2 のようになる.

[3]U-独立性は周辺分布が連続値に対するものであっても定義でき，その場合には $[0, \infty]^2 \to [0, \infty]$ という射影を考えることになる．そのため U-独立性は単純に「copula の離散版」に対応している訳ではない.

表 2: U-独立性と Archimedean copula の比較

	U-独立性	Archimedean copula
式	$u(\xi(p_\mathcal{X}(x)) + \xi(p_\mathcal{Y}(y)) - c)$	$\phi^{[-1]}(\phi(u) + \phi(v))$
対象	離散 (連続) の分布関数	連続の累積分布関数
特徴	• 定数項 c による正規化で同時分布が得られる. • 周辺確率値の高低差の同時確率への影響を $U(\cdot)$ の形状によって制御する.	• 定義 5 より累積同時分布が得られる. • 同時確率分布の裾の広がり方を関数 ϕ の形状によって制御する.

3 U-独立性に基づくモデルの拡張

Log linear model(LLM)[1] や Bayesian network[10] など多くの統計モデルにおける基本的な考え方として変数間の（条件付き）独立性が重要視される．しかし，実世界のデータを考えると変数間に弱い特殊な関係があるために厳密には独立性が崩れていることもあり，そのような場合には積極的に弱い相関関係をモデル化することが解析上効果的だと考えられる．本章では U 乗除算の 1 つの応用として，弱い相関関係のモデリングを目的とした独立モデル，及び NB の拡張を試みる.

3.1 独立モデルの拡張とその幾何学的解釈

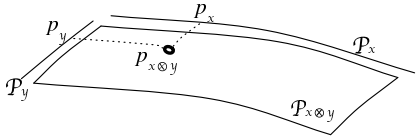
ある変数 X，Y に対して変数間の独立性を仮定した統計モデルの集合を考える.

$$\mathcal{P}_{\mathcal{X}\times\mathcal{Y}} = \{p(x, y; \theta) = p_\mathcal{X}(x) \times p_\mathcal{Y}(y), \theta = \{p_\mathcal{X}, p_\mathcal{Y}\}\}$$

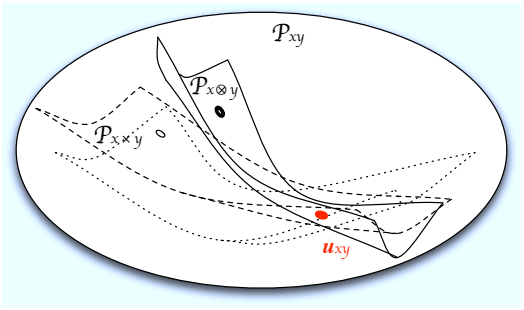
これに対して，適当な関数 U の下で次のような U-独立性を仮定したモデルの集合を考える.

$$\mathcal{P}_{\mathcal{X}\otimes\mathcal{Y}} = \{p(x, y; \theta, U) = p_\mathcal{X}(x) \otimes p_\mathcal{Y}(y), \theta = \{p_\mathcal{X}, p_\mathcal{Y}\}\}$$

この U-独立モデルは図 2 のように幾何学的に解釈することができる．まず，ある U-独立な分布 $p_{\mathcal{X}\otimes\mathcal{Y}} \in \mathcal{P}_{\mathcal{X}\otimes\mathcal{Y}}$ は周辺分布 $p_\mathcal{X} \in \mathcal{P}_\mathcal{X}$，および $p_\mathcal{Y} \in \mathcal{P}_\mathcal{Y}$ によって記述される空間内の 1 点で表現される (図 2(a) 参照)．直感的には U-独立な制約を持つ空間 $\mathcal{P}_{\mathcal{X}\otimes\mathcal{Y}}$ は同時分布の空間 $\mathcal{P}_{\mathcal{X}\mathcal{Y}}$ の中で図 2(b) のように存在していることになる．また，考えている関数 U が変わると U 乗算の結果表現



(a) $\mathcal{P}_{\mathcal{X}\otimes\mathcal{Y}}$ の空間



(b) $\mathcal{P}_{\mathcal{X}\mathcal{Y}}$ 内の $\mathcal{P}_{\mathcal{X}\otimes\mathcal{Y}}$ のイメージ

図 2: U-独立性の幾何学的イメージ．

される同時分布の形状も変わる（図 1 を参照）ことから，$\mathcal{P}_{\mathcal{X}\otimes\mathcal{Y}}$ は $\mathcal{P}_{\mathcal{X}\mathcal{Y}}$ の中で関数 U に応じて異なる曲がり方をしていることも分かる．

また，次のような一様分布を考えてみる．

$$u_{\mathcal{X}}(x) = \frac{1}{|\mathcal{X}|}, \quad u_{\mathcal{Y}}(y) = \frac{1}{|\mathcal{Y}|}, \quad u_{\mathcal{X}\mathcal{Y}}(x,y) = \frac{1}{|\mathcal{X}|\times|\mathcal{Y}|}$$

すると独立性を構成する関数 U の種類に依らずに

$$u_{\mathcal{X}\mathcal{Y}}(x,y) = u_{\mathcal{X}}(x) \otimes u_{\mathcal{Y}}(y)$$

が必ず成り立つため，図 2(b) のように，あらゆる $\mathcal{P}_{\mathcal{X}\otimes\mathcal{Y}}$ は同時分布の空間 $\mathcal{P}_{\mathcal{X}\mathcal{Y}}$ の中の 1 点 $u_{\mathcal{X}\mathcal{Y}}$ を必ず含むことになる．

3.2 拡張独立モデルの推定

さて，ここで経験同時分布 $\tilde{p}_{\mathcal{X}\mathcal{Y}}$ がデータから与えられている時に U-独立モデルの最尤推定を行うことを考える．まず通常の独立モデル $q \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ の場合，最尤推定は

$$\begin{aligned}\hat{q} &= \operatorname*{argmin}_{q\in\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}} \mathrm{D}_{\mathrm{KL}}(\tilde{p}_{\mathcal{X}\mathcal{Y}}, q) \\ &= \tilde{p}_{\mathcal{X}} \times \tilde{p}_{\mathcal{Y}} \end{aligned} \tag{12}$$

のようにそれぞれの経験周辺分布のみによって実現することが出来る (図 3(a) 参照)．一方，U-独立モデル $q \in \mathcal{P}_{\mathcal{X}\otimes\mathcal{Y}}$ について考えてみると

$$\hat{q} = \operatorname*{argmin}_{q\in\mathcal{P}_{\mathcal{X}\otimes\mathcal{Y}}} \mathrm{D}_{\mathrm{KL}}(\tilde{p}_{\mathcal{X}\mathcal{Y}}, q) \tag{13}$$

を得るには一般にモデルのパラメタ θ，つまり周辺分布 $p_{\mathcal{X}} \in \mathcal{P}_{\mathcal{X}}$，$p_{\mathcal{Y}} \in \mathcal{P}_{\mathcal{Y}}$，及び乗算を定める関数 U に関する非線形最適化問題を解く必要がある (図 3(b) 参照)．

これを避けるために，周辺分布は経験周辺分布で固定し，関数 U のみに依存するようなモデルの集合

$$\tilde{\mathcal{P}}_{\mathcal{X}\otimes\mathcal{Y}} = \{p(x,y;U) = \tilde{p}_{\mathcal{X}}(x) \otimes \tilde{p}_{\mathcal{Y}}(y)\}$$

を考え，その下で最適な U-独立モデルを適当な関数 U のクラスの中から探索することを考える．

$$\hat{q} = \operatorname*{argmin}_{q\in\tilde{\mathcal{P}}_{\mathcal{X}\otimes\mathcal{Y}}} \mathrm{D}_{\mathrm{KL}}(\tilde{p}_{\mathcal{X}\mathcal{Y}}, q) \tag{14}$$

このように経験周辺分布で構成した U-独立モデルの推定は直感的には図 3(c) のように解釈することができ，以下では式 (14) によって得られたモデルを経験 U-独立モデルと呼ぶ[4]．経験 U-独立モデルの最適化は関数 U のチューニングによって行われるため，実際には cross-validation などによって実現することができ，式 (13) の最適化と比べると著しく簡単になる．

前述の議論より，U 乗算を構成する $U(\cdot)$ の形状が変わると周辺確率が高い部分と低い部分の差の影響が様々に変化することが分かっているが，周辺分布として一様分布 $u_{\mathcal{X}}$，$u_{\mathcal{Y}}$ が与えられている時には U-独立なモデルは関数 U に依らずに $u_{\mathcal{X}\mathcal{Y}}$ となる．このことは経験 U-独立モデルの空間 $\tilde{P}_{\mathcal{X}\otimes Y}$ (図 3(c) の太線で表される空間) は経験周辺分布 $\tilde{p}_{\mathcal{X}} \simeq u_{\mathcal{X}}$ かつ $\tilde{p}_{\mathcal{Y}} \simeq u_{\mathcal{Y}}$ の時には，関数 U をどのように動かしても，ほとんど $u_{\mathcal{X}\mathcal{Y}}$ に一致してしまうことを示唆している．逆に言うと $\tilde{p}_{\mathcal{X}}$，及び $\tilde{p}_{\mathcal{Y}}$ が局所的に高い（低い）確率値をピークとして持つような分布の時 (一様分布とは対照に空間の端の方の点で表現される) には，$\tilde{P}_{\mathcal{X}\otimes\mathcal{Y}}$ は関数 U の形状によって様々な同時分布を表すことができる．このため

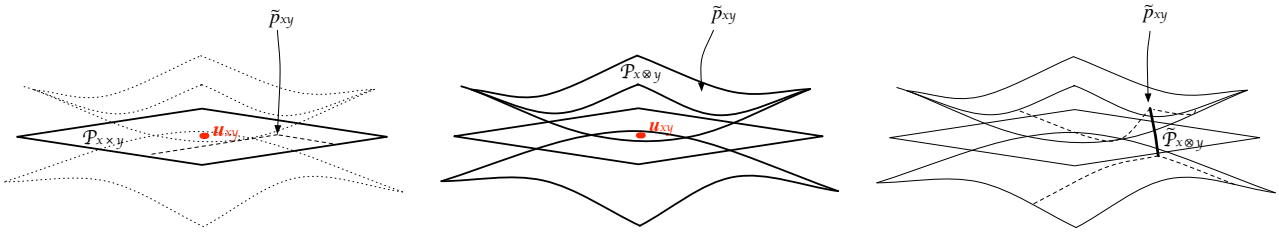
- サンプルの数が少ない
- サンプルが少数の外れ値を含んでいる

などの理由によって，扱う分布が一様分布から遠く離れている場合には，経験 U- 独立モデルは特に柔軟な表現力を持っていると考えられる．

経験 U-独立モデルの推定では経験周辺分布を用いているが，これは U-独立モデル $p_{\mathcal{X}\otimes\mathcal{Y}}(p_{\mathcal{X}}, p_{\mathcal{Y}}, U) \in \mathcal{P}_{\mathcal{X}\otimes\mathcal{Y}}$ の $\{p_{\mathcal{X}}, p_{\mathcal{Y}}\}$ に関する推定を

$$\begin{aligned}&\operatorname*{argmin}_{\{p_{\mathcal{X}},p_{\mathcal{Y}}\}} \mathrm{D}_U(\tilde{p}_{\mathcal{X}\mathcal{Y}}, p_{\mathcal{X}\otimes\mathcal{Y}}(\theta,U)) \\ &= \operatorname*{argmin}_{\{p_{\mathcal{X}},p_{\mathcal{Y}}\}} \sum_{x\in\mathcal{X},y\in\mathcal{Y}} U(\xi(p_{\mathcal{X}}(x)) + \xi(p_{\mathcal{Y}}(y)) - c) \\ &\qquad - \sum_{x\in\mathcal{X},y\in\mathcal{Y}} \tilde{p}_{\mathcal{X}\mathcal{Y}}(x,y)(\xi(p_{\mathcal{X}}(x)) + \xi(p_{\mathcal{Y}}(y)) - c) \\ &\simeq \operatorname*{argmin}_{\{p_{\mathcal{X}},p_{\mathcal{Y}}\}} \sum_{x\in\mathcal{X}} U(\xi(p_{\mathcal{X}}(x))) + \sum_{y\in\mathcal{Y}} U(\xi(p_{\mathcal{Y}}(y)))\end{aligned}$$

[4] copula の文脈でも，経験累積周辺分布から累積同時分布を作るという話題があり，経験 copula と呼ばれる．



(a) 独立モデルの最尤推定　(b) U-独立モデルの最尤推定　(c) 経験 U-独立モデルの最尤推定

図 3: モデルの最尤推定の幾何学的イメージ.

$$
\begin{aligned}
&- \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} \tilde{p}_{\mathcal{XY}}(x,y)(\xi(p_{\mathcal{X}}(x)) + \xi(p_{\mathcal{Y}}(y))) \\
&= \{\tilde{p}_{\mathcal{X}}, \tilde{p}_{\mathcal{Y}}\}
\end{aligned}
$$

のような近似によって得ているのと等価となる．このように経験 U-独立モデルの推定は Bregman 情報量の近似の観点からも解釈することができ，関数 U を変えることでロバストな統計的性質が実現されることが実験的に確認されている [6].

3.3 NB の拡張

変数 Y と変数集合 $X = (X_1, \ldots, X_I)$ が与えられている時，$I+1$ 個の変数集合 (X, Y) を考える．これらの同時確率 $p(X, Y)$ を

$$
\begin{aligned}
p_{KL}(X, Y) &= p(Y)p(X|Y) \\
&= p(Y)\prod_{i=1}^{I} p(X_i|Y) \qquad (15)
\end{aligned}
$$

のように表したものは naive Bayes model(NB) と呼ばれ，次のような特徴を持つ.

1. 条件付き独立性を仮定することでモデルを単純化
2. パラメタの推定が簡単
3. 大データへの拡張性 (Scalability)

また，条件付き独立性の仮定が崩れていても判別器としては比較的頑強に動作することが知られているため [4]，様々な状況においてプリミティブな判別器として用いられている．なお，式 (15) における $p(Y)$ の最尤推定量はデータ中で y が観測された頻度 $n(y)$ を用いて

$$
p(y) = \frac{n(y)}{\sum_{y'\in Y} n(y')} \qquad (16)
$$

によって得られるが，$p(X_i|Y)$ に関してはサンプリングゼロ等の問題を回避するために本稿では次のような Laplace smoother $\alpha \in [0, 1]$ を導入する.

$$
p(x_i|y) = \frac{n(x_i, y) + \alpha}{\sum_{x_i'\in X_i}(n(x_i', y) + \alpha)} \qquad (17)
$$

ただしここで $n(x_i, y)$ はデータの中で x_i と y が同時に観測された頻度を表す.

さて，条件付き独立性の一般化という観点からの最も単純な NB の拡張としては次のようなものが考えられる.

$$
p_U(X, Y) = p(Y)\left(\bigotimes_{i=1}^{I} p(X_i|Y)\right) \qquad (18)
$$

ただし，

$$
\bigotimes_{i=1}^{I} p(X_i|Y) = p(X_1|Y) \otimes \cdots \otimes p(X_I|Y)
$$

の意味とする.

また，$X_1, \ldots, X_I$ のうちのいくつかの変数間にのみ弱い相関がある場合には，

$$
p_U(X, Y) = p(Y)\left(\bigotimes_{i\in\bar{S}_I} p(X_i|Y)\right)\left(\prod_{j\in S_I} p(X_j|Y)\right) \qquad (19)
$$

のような一般化も可能である．ただし $\bar{S}_I$ は X の中の弱い依存関係のある変数番号の集合を表し，S_I はそれ以外の変数番号の集合の意味である.

X, Y の経験同時分布 $\tilde{p}$ が与えられている時，上式 (18) や (19) で与えられる $p_U(X, Y)$ の最尤推定は

$$
\hat{p}_U = \operatorname*{argmin}_{p_U} \mathrm{D}_{\mathrm{KL}}(\tilde{p}, p_U) \qquad (20)
$$

によって実現されるが，前節での議論と同様に U 乗算を用いて定義されたモデルは厳密解を得るには非線形最適化問題を解く必要がある．これは NB の特徴の中の「パラメタ推定の簡単さ」と「大データへの拡張性」という観点からは望ましく無い．そこで経験 U-独立モデルの推定と同様に，必要な周辺分布等の推定に関しては式 (16)，(17) をそのまま用い，乗算を構成する関数 U の方を変えることで経験 U-NB を得ることを考える．なお，実際にはどの変数間の乗算をどの関数 U によって一般化するか，という判断も行う必要があるが，本稿では

表 3: MONK's problem のデータ内容

データ名	I	訓練データ数	テストデータ数	$Y=1$ となる条件
MONKS1	6	124	432	$(X_1 = X_2)$ or $X_5 = 1$
MONKS2	6	169	432	X のうち 2 個の変数が 1

表 4: 実験で用いたモデル

モデル名	式	注釈
model1	式 (18)	$\pi = 1$ で通常の NB と一致
model2	式 (19) $(\bar{S}_I = \{1, 2\})$	$\pi = 1$ で通常の NB と一致
model3	式 (19) $(\bar{S}_I = \{3, 4, 5, 6\})$	$\pi = 1$ で通常の NB と一致

モデル中の全ての一般化乗算は共通の関数 U によって計算されることとし，関数 U は表 1 に示すようなパラメタ π によって特徴づけられるようなものに限定する．すると，経験 U-NB の推定はある関数 U のクラスの下でのパラメタ π に関する最適化問題になる．次章では実際のデータセットに対して経験 U-NB の推定を行い，その精度等について議論する．

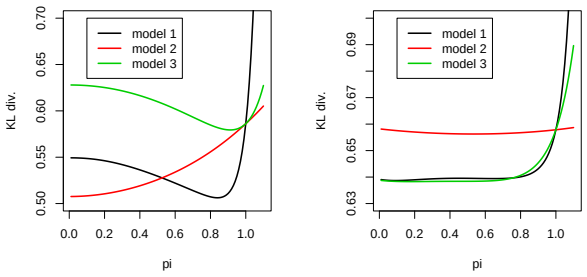
4 数値実験

本章では一般化独立性を導入した経験 U-NB の推定精度の良さを測る為に，いくつかのベンチマークデータセットを用いた実験を行う．まず実験 1 では U 乗算を用いた NB の拡張方法について比較，考察を行う．また経験 U-NB を推定する際には乗算を特徴づける U のチューニングが必要となるため，実験 2 では訓練データに基づく U のチューニング，及び推定結果の評価を行う．なお，以下の実験で用いる経験 U-NB は全て表 1 の Ex.1 で与えられる関数 U を用いて行う．

4.1 実験 1:モデルの拡張方法の比較

先に述べたように U 乗算を用いた NB の拡張方法にはいくつか考えられる．ここでは UCI の ML レポジトリ [3] にある"MONK's problem" より 2 つのデータ (以下 MONKS1，MONKS2) を用いてモデルの拡張方法に関する実験，考察を行った．これらは共に表 3 に示すようなクラス変数 Y と離散変数集合 $X = (X_1, \ldots, X_6)$ から成り立ち，Y の値の決定に際して X の中に一種の依存関係が存在するような人工データである．比較する NB の拡張としては表 4 に示す 3 つのモデルを使った．実験では訓練データの経験分布 $\tilde{p}$ からこれらのモデル $\hat{p}_U$ を構築し，テストデータから構築した p^* を用いて KL 情報量 $\mathrm{D}_{\mathrm{KL}}(p^*, \hat{p}_U)$ を指標とした評価を行った．なお，本実験ではモデル推定時の式 (17) の α はゼロとする．

MONKS1 に関する実験結果を，関数 U を特徴づける π の値に応じてプロットしたのが図 4(a) である．X_1 と X_2 の関係表現に U-乗算を用いた model 1，2 に関しては，本来データが持っている依存関係をよりうまく表現し得ることが分かる．一方，X_1 と X_2 間の乗算を一般化していない model 3 ではほとんど改善が見られない．同様に MONKS2 に関する結果を図 4(b) に示す．MONKS2 は変数間全てに依存関係があるため，X_1 と X_2 の間でのみ U-乗算を用いた model 2 ではあまり改善が見られない．一方で半数以上の変数間で U-乗算を用いた model 3，及び全変数間で U-乗算を用いた model 1 では改善が見られる．これより変数間の依存関係に関する知見が無い状況であっても model 1 のような一般化をすることで NB が改善される可能性があることが確かめられる．



(a) MONKS1 の場合　　(b) MONKS2 の場合

図 4: 実験 1 の結果.

表 5: 実験 2 で用いたデータセット

データ名	I	訓練データ数	テストデータ数
CAR	6	300	1728
NUR	8	300	12960

4.2 実験 2:パラメタ π のチューニング

実験 2 では訓練データが比較的少数の時に，Laplace smoother を導入した通常の NB をパラメタ π のチューニングによってさらに改善することが可能かどうかを確認した．なお，この実験では前述の MONKS1，MONKS2 に加え，UCI の ML レポジトリにある"car evaluation"(以下 CAR)，"nursery"(以下 NUR) の 2 つのデータセット (表 5 参照) を用い，モデルとしては表 4 の model 1 を用いている．

この実験ではまず α を 0.01 刻みで動かしながら訓練データを用いた 10-fold クロスバリデーション (CV) によって通常の NB における $\hat{\alpha}$ を求め，さらにその $\hat{\alpha}$ の下での最適なパラメタ $\hat{\pi}$ を選んだ．その後，訓練データ全てを用いて推定したモデルのテストデータに対する良さ $\mathrm{D}_{\mathrm{KL}}(p^*, \hat{p}_U)$ を測った．表 6 に示す実験の結果より，CV によって妥当な $\hat{\pi}$ を得ることが可能であり，またどのデータセットに対しても $\hat{\alpha}$ を用いた通常の NB をさらに改善していることが確認できる．通常の NB では Laplace smoother の導入によって少数の訓練データに対するモデルのオーバーフィットを緩和することが可能だが，経験 U-NB ではそれに加えてモデルの背後にある Bregman 情報量のロバストな性質がさらなる緩和を

表 6: 実験 2 の結果

	$D_{\mathrm{KL}}(p^*, \hat{p}_U)$	
データ名	NB	U-NB
MONKS1	0.5796	**0.5340** ($\hat{\pi} = 0.92$)
MONKS2	0.6520	**0.6386** ($\hat{\pi} = 0.01$)
CAR	0.6307	**0.6242** ($\hat{\pi} = 0.96$)
NUR	0.4321	**0.4141** ($\hat{\pi} = 0.98$)

可能にすることがこの結果より示唆される．

5 まとめ

本稿では Bregman 情報量から発想を得た乗除算の一般化を試み，そこから導きだされた U-独立性に基づく統計モデル (独立モデル，NB) の拡張について議論を行った．また U-独立性をモデル化することで厳密な推定を行うための計算コストが増すため，これを避けるために経験周辺分布を用いることを提案した．経験周辺分布を用いた U-独立モデルの推定は Bregman 情報量の近似の観点からも解釈することができ，モデルが一種のロバスト性を実現できる可能性を示唆していると考えられる．

本稿では述べなかったが U-独立モデルを一般化線形モデルの一種と見ると，関数 $\xi(\cdot)$ の意味での線形モデルを構築していることになり，log linear model の拡張となる．また NB の場合と同様の議論によって Bayesian network などの一般化も行うことができる．一般的な log linear model における交互作用項や Bayesian network におけるリンクによる依存関係の表現と比べるとパラメタ数，オーバーフィットの度合いなどの点で本手法には利点が見込めるため，例えば弱い相関を表すリンクを省いた単純なネットワーク構造と U-独立性の議論を組み合わせることで，より簡潔なモデルの記述が期待される．

また，本稿では統計モデルの精度の観点から U-NB の改善の様子を実験的に示したが，NB の応用先としてしばしば用いられる判別器としての観点からどのような性質を持つのかを調べるのは興味深い．特に判別器として有効性が謳われている complement naive Bayes[13] のような手法と組み合わせることも考えられるため，今後判別器としての評価等も行う予定である．また，同じ周辺分布であっても，用いる関数 U によって様々な同時分布の表現が可能となるため，$U(\cdot)$ の形状とモデルの統計的な性質に関しても議論を行う必要があると考えている．

参考文献

[1] A. Agresti. *Categorical Data Analysis*. Wiley Inc., 2 edition, 2002.

[2] S. Amari. Integration of stochastic models by minimizing α-divergence. *Neural Computation*, 19:2780–2796, 2007.

[3] A. Asuncion and D. Newman. UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences, 2007.

[4] P. Domingos and M. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29:103–130, 1997.

[5] Y. Fujimoto and N. Murata. A modified EM algorithm for mixture models based on Bregman divergence. *Annals of the Institute of Statistical Mathematics*, 59(1):3–25, 2007.

[6] Y. Fujimoto and N. Murata. A generalized product rule and weak independence based on Bregman divergence. In *Proc. 12th World Multi-Conference on Systemics, Cybernetics and Informatics*, 2008.

[7] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics*. Wiley Inc., 1986.

[8] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42:177–196, 2001.

[9] P. Hougaard. A class of multivariate failure time distributions. *Biometrika*, 73(3):671–678, Dec. 1986.

[10] F. V. Jensen. *Bayesian Networks and Decision Graphs*. Statistics for Engineering and Information Science. Springer, 2001.

[11] N. Murata, T. Takenouchi, T. Kanamori, and S. Eguchi. Information geometry of $\mathcal{U}$-boost and Bregman divergence. *Neural Computation*, 16:1437–1481, 2004.

[12] R. B. Nelsen. *An Introduction to Copulas*. Springer Series in Statistics. Springer, second edition, 2006.

[13] J. D. Rennie, L. Shih, J. Teevan, and D. Karger. Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML2003)*, 2003.