数理解析研究所講究録 第2124巻 2019年 87-100

A possible extension of regression analysis for imbalanced binary data

東京大学・情報理工学系研究科　清 智也

Tomonari Sei

Graduate School of Information Science and Technology,

The University of Tokyo.

概要

二値回帰モデルにおいて，応答変数の分布が限りなく不均衡になるときのモデルの極限を考えると，指数型分布族が現れることが知られている。この結果は極値理論と関係している。本研究では，Omae, Komori and Eguchi (2017, BMC Informatics) で提案された準線形ロジスティック回帰モデルにおける不均衡極限を考え，多次元の極値理論と対応することを示す。

1　導入：ロジスティック回帰モデルの不均衡極限

ロジスティック回帰モデルは

$$\mathrm{P}(Y=1 \mid X=x) = G(a+b^\top x), \quad G(z) = \frac{e^z}{1+e^z} \tag{1}$$

と表すことができる。ここで Y は応答変数，X は p 次元の説明変数ベクトル，a, b は回帰係数である。また G は ロジスティック分布である。

いま $Y=1$ となるケースが稀な場合 (imbalanced; 不均衡) を考える。ポアソンの少数の法則と同様に，真のパラメータがサンプルサイズ n に依存することを許容すると，$n\to\infty$ において非自明な漸近形が得られる。具体的には，回帰係数を $a_n = -\log n + \alpha$, $b_n = \beta$ (α, β は定数) とおけば

$$\begin{aligned}\mathrm{P}(Y=1 \mid X=x) &= \frac{\frac{1}{n}e^{\alpha+\beta^\top x}}{1+\frac{1}{n}e^{\alpha+\beta^\top x}} \\ &\simeq \frac{1}{n}e^{\alpha+\beta^\top x}, \quad n\to\infty.\end{aligned}$$

となる。もし X の周辺分布 F が n に依存しないならば，$Y=1$ のもとでの X の条件付き分布は，ベイズの定理から

$$\mathrm{P}(X\in dx \mid Y=1) \simeq \frac{e^{\beta^\top x}F(dx)}{\int e^{\beta^\top x}F(dx)}$$

という指数型分布に収束する [9]。また点過程として扱うと強度 $e^{\alpha+\beta^\top x}F(dx)$ の非斉次ポアソン点過程に収束することも示される [13]。このように，$Y = 1$ となる確率が n の増加とともに 0 に近づくときの極限を，本稿では不均衡極限 (imbalance limit) と呼ぶ。

同じ性質を持つ二値回帰モデルは他にも存在し，例えば complementary log-log link に対応する

$$\mathrm{P}(Y = 1 \mid X = x) = G(a + b^\top x), \quad G(z) = 1 - \exp(-e^z) \tag{2}$$

についても，$a_n = -\log n + \alpha$, $b_n = \beta$ とおけば

$$\mathrm{P}(Y = 1 \mid X = x) \simeq \frac{1}{n} e^{\alpha+\beta^\top x}$$

となる。ところで，式 (2) の $G(z)$ を累積分布関数として持つような確率変数 Z を考えたとき，その符号を変えた $W = -Z$ の累積分布関数は

$$H(w) := 1 - G(-w) = \exp(-e^{-w})$$

となる。これは Gumbel 分布であり，独立なサンプルの最大値をとる操作に関して閉じている。

以上の議論と同様にして，累積分布関数 $G(z)$ を使った二値回帰モデルの極限は極値理論によって特徴付けられる [10]。上で述べた logit リンクと complementary log-log リンクが同じ極限を持つことは，極値理論の言葉ではロジスティック分布が Gumbel 分布の吸引域に含まれることに対応する。また最大値安定性は，空間的な説明変数の解像度に対する安定性 [1] に対応する。

さて極値理論は多次元の場合にも深く研究されている [11, 4, 2]。したがって多次元の極値分布に対応する回帰モデルを形式的に得ることができる。しかしその応用上の意味は明らかではなかった。これに意味を与えてくれるのが，次節で述べる準線形ロジスティック回帰モデル [8] である。

本論文の構成は以下の通りである。まず 2 節で準線形ロジスティック回帰モデルの定義を述べ，その不均衡極限を導く。3 節ではモデルをさらに拡張する。その不均衡極限を 4 節で考察する。最後に 5 節でまとめを述べる。

2　準線形ロジスティック回帰モデル

大前ら [8] は，次の準線形ロジスティック回帰モデル (quasi-linear logistic regression model) を考えた：

$$\mathrm{P}(Y=1 \mid X) = \frac{e^{Q}}{1+e^{Q}}, \tag{3}$$

$$Q = Q(X) = \frac{1}{\tau}\log\left(\sum_{k=1}^{K}\exp(\tau(a_k + b_k^{\top}X_{(k)}))\right). \tag{4}$$

ここで $X=(X_{(1)},\dots,X_{(K)})$ は事前にクラスタリングによってグループ分けされた説明変数ベクトルであり，各 $X_{(k)}$ は部分ベクトルである。また a_k, b_k は回帰係数であり，$\tau>0$ はチューニングパラメータである。$K=1$ の場合は普通のロジスティック回帰モデルに帰着される。事前にクラスタリングしない場合については [7] で議論されている。

式 (4) の右辺は準線形予測子 (quasi-linear predictor) と呼ばれる。準線形予測子は $\tau\to 0$ のとき（定数項を除いて）単純和 $\sum_k(a_k+b_k^{\top}X_{(k)})$ に近づき，$\tau\to\infty$ のときは最大値 $\max_k(a_k+b_k^{\top}X_{(k)})$ に近づく。準線形予測子を用いる意図は,「複数の線形予測子のうち，どれか一つでも値が大きくなれば生起確率が上がるようにしたい」というものである。

[8] はさらに，式 (4) を一般化した

$$Q = \phi^{-1}\left(\sum_{k}\phi(a_k + b_k^{\top}X_{(k)})\right) \tag{5}$$

という形の予測子も提案している。ただし関数 $\phi:\mathbb{R}\to(0,\infty)$ は連続で狭義単調増加かつ $\phi(-\infty)=0, \phi(\infty)=\infty$ と仮定する。特に $\phi(z)=e^{\tau z}$ とおけば式 (4) が得られる。さらなる一般化については次節で改めて議論することにする。以降，式 (5) も準線形予測子と呼ぶことにし，式 (4) のことは log-sum-exp と呼ぶことにする。

log-sum-exp の場合の不均衡極限を求めてみよう。真のパラメータ a_k および b_k がサンプルサイズ n に依存して

$$a_{k,n} = -\log n + \alpha_k, \quad b_{k,n} = \beta_k$$

と書けると仮定すれば

$$Q = -\log n + \frac{1}{\tau}\log\left(\sum_{k=1}^{K} e^{\tau(\alpha_k+\beta_k^{\top}X_{(k)})}\right)$$

となる。よって

$$\mathrm{P}(Y=1 \mid X) = \frac{e^{Q}}{1+e^{Q}} \simeq \frac{1}{n}\left(\sum_{k} e^{\tau(\alpha_k+\beta_k^{\top}X_{(k)})}\right)^{1/\tau} \tag{6}$$

という漸近形が得られる。また $Y=1$ のもとでの X の条件付き分布は

$$\mathrm{P}(X \in dx \mid Y=1) \simeq \frac{\{\sum_k e^{\tau(\alpha_k+\beta_k^{\top}X_{(k)})}\}^{1/\tau} F(dx)}{\int\{\sum_k e^{\tau(\alpha_k+\beta_k^{\top}X_{(k)})}\}^{1/\tau} F(dx)}$$

という分布に収束する。ただし F は X の周辺分布である。特に $\tau=1$ の場合は混合指数型分布族となる。関連して，$X|Y$ の条件付き分布を混合正規分布としたときに，$Y|X$ の分布が準線形ロジスティック回帰モデルになることが [8] で指摘されている。

3　モデルの一般化とコピュラ表現

この節では準線形ロジスティック回帰モデルの特徴を抽出し，その特徴に基づいた一般化を行う。またコピュラとの関連を指摘する。

K 個の線形予測子を $z_k = a_k + b_k^{\top}X_{(k)}$ と表すとき，式 (5) は

$$Q(z_1,\ldots,z_K) = \phi^{-1}\left(\sum_{k=1}^{K}\phi(z_k)\right) \tag{7}$$

と表される。ϕ に課された条件 $\phi(-\infty)=0$ と $\phi(\infty)=\infty$ から，Q は次の 2 つの条件を満たす：

$$Q(z_1,\ldots,\infty,\ldots,z_K) = \infty, \tag{8}$$

$$Q(-\infty,\ldots,z_k,\ldots,-\infty) = z_k. \tag{9}$$

式 (8) より，K 個の予測子のうち一つでも大きな値を取るものがあれば予測分布 $\mathrm{P}(Y=1 \mid X=x)$ は 1 に近づく。また式 (9) より，一つの予測子を除いて全て小さい値を取るならば，その予測子にもとづく通常の回帰モデルに帰着される。

以上の考察をもとに，次のようなモデルのクラスを考える。

定義 1 (検出可能モデル)**.** 式 (8), (9) を満たす $Q(z_1, \ldots, z_K)$ のことを検出可能 (detectable) な予測子と呼ぶ。検出可能な予測子を用いて以下のように表される二値回帰モデルを検出可能モデルという：

$$\mathrm{P}(Y = 1 \mid X = x) = G(a_1 + b_1^\top X_{(1)}, \ldots, a_K + b_K^\top X_{(K)}), \tag{10}$$

$$G(z_1, \ldots, z_K) = G_1(Q(z_1, \ldots, z_K)). \tag{11}$$

ただし $G_1 : \mathbb{R} \to [0, 1]$ は 1 次元の連続な累積分布関数である。

準線形モデルと同様に，検出可能モデルにおいても複数の線形予測子が 1 つの予測子 Q に集約され，Q に対する二値回帰モデルが当てはめられる。

さて，検出可能モデルはコピュラと密接に関係する。コピュラとは以下の 3 条件を満たす関数 $C : [0, 1]^K \to [0, 1]$ のことである：

$$C(u_1, \ldots, 0, \ldots, u_K) = 0, \tag{12}$$

$$C(1, \ldots, u_k, \ldots, 1) = u_k, \tag{13}$$

$$\Delta_1 \cdots \Delta_K C(u_1, \ldots, u_K) \geq 0 \quad (K\text{-increasing}). \tag{14}$$

ここで Δ_k は差分演算子を表す。Sklar の定理により，任意の累積分布関数 H は周辺分布 $H_1, \ldots, H_K$ とコピュラ C を用いて

$$H(w_1, \ldots, w_K) = C(H_1(w_1), \ldots, H_K(w_K))$$

と一意的に分解される。コピュラの詳細については [12] を参照されたい。

ここでは式 (12), (13) と式 (8), (9) の類似性に注目する。コピュラの条件を緩め，式 (12) と (13) を満たす $C : [0, 1]^K \to [0, 1]$ のことを符号付きコピュラ (signed copula) と呼ぶことにする。符号付きと言っているのは式 (14) の値が負になることも許容しているためである。

定理 1 (コピュラ表現)**.** 式 (11) の検出可能モデルは次のように表される：

$$G(z_1, \ldots, z_K) = 1 - C(H_1(-z_1), \ldots, H_1(-z_K)). \tag{15}$$

ここで H_1 は 1 次元分布，C は符号付きコピュラであり，それぞれ

$$H_1(w) = 1 - G_1(-w), \tag{16}$$

$$C(u_1, \ldots, u_K) = H_1(-Q(-H_1^{-1}(u_1), \ldots, -H_1^{-1}(u_K))) \tag{17}$$

と表される。逆に H_1 と C を指定すれば検出可能モデルが得られる。

Proof. 式 (11) を仮定する。関数 H を

$$\begin{aligned} H(w_1,\dots,w_K) &= 1 - G(-w_1,\dots,-w_K) \\ &= 1 - G_1(Q(-w_1,\dots,-w_K)) \end{aligned}$$

で定義する。また $H_1(w) = 1 - G_1(-w)$ とおく。このとき

$$H(w_1,\dots,w_K) = H_1(-Q(-w_1,\dots,-w_K))$$

となる。式 (8), (9) より，H は性質

$$H(w_1,\dots,-\infty,\dots,w_K) = 0, \quad H(\infty,\dots,w_k,\dots,\infty) = H_1(w_k)$$

を満たすことが確かめられる。そこで，Sklar の定理にならって

$$C(u_1,\dots,u_K) = H(H_1^{-1}(u_1),\dots,H_1^{-1}(u_K))$$

とおけば，C は符号付きコピュラの性質を満たす。また式 (15) も成り立つ。逆に H_1 と C が与えられれば G_1 と Q が定まる。 □

この定理から，模式的に

$$\{G_1, Q\} \leftrightarrow \{H_1, C\}$$

という対応が得られた。検出可能モデルを指定するにはどちらの組を用いてもよい。

元々の準線形ロジスティック回帰モデルの場合を確認しておく。

例 1. 式 (3), (4) の準線形ロジスティック回帰モデルの場合，G_1 はロジスティック分布，Q は log-sum-exp の形である。定理 1 によって定まる 1 次元分布 H_1 はロジスティック分布であり，また符号付きコピュラは

$$C(u_1,\dots,u_K) = \frac{1}{1 + (\sum_{k=1}^{K} (\frac{1-u_k}{u_k})^{\tau})^{1/\tau}} \tag{18}$$

となる。この C はアルキメデス型（後述）であり，[6] の Table 4.1 にある 12 番目のコピュラと一致する。特に $\tau = 1$ の場合は

$$C(u_1,\dots,u_K) = \frac{1}{1 + \sum_{k=1}^{K} \frac{1-u_k}{u_k}}. \tag{19}$$

となり，Clayton コピュラと呼ばれる [6, 12]。$0 < \tau < 1$ の場合，C は厳密な意味でのコピュラにはならない（K-increasing でない）。 □

符号付きコピュラがアルキメデス型 (Archimedean) であるとは，

$$C(u_1, \ldots, u_K) = \psi^{-1}(\psi(u_1) + \cdots + \psi(u_K)) \tag{20}$$

の形で書けることと定義する。ψ を生成素 (generator) と呼ぶ。式 (18) の場合，生成素は $\psi(u) = (\frac{1-u}{u})^\tau$ で与えられる。

一般化平均とアルキメデス型の関係は次の定理で与えられる。これら 2 つの類似性については [3] でも指摘されている。

定理 2. 検出可能モデルにおいて，予測子 Q が式 (7) の一般化平均の形に書けるための必要十分条件は，対応する符号付きコピュラ C がアルキメデス型となることである。このときの生成素の対応は $\psi(u) = \phi(H_1^{-1}(u))$ で与えられる。

Proof. 式 (17) より，$Q = \phi^{-1}(\sum_k \phi(z_k))$ と $C = \psi^{-1}(\sum_k \psi(u_k))$ は $\psi(u) = \phi(H_1^{-1}(u))$ のもとで同値である。 □

4 主結果

検出可能モデルは 1 次元分布 H_1 と符号付きコピュラ C で指定された (定理 1)。本節では H_1 を Gumbel 分布

$$H_1(w) = \exp(-e^{-w})$$

に固定した上で，不均衡極限において C の違いがどう反映されるかを明らかにしたい。これは多変量極値理論 [11, 4, 2] と密接に関係する。Gumbel 分布以外の場合については付録 B で補足する。

符号付きコピュラ $\bar{C}$ が極値 (extreme) であるとは，

$$\bar{C}(u_1, \ldots, u_K) = \lim_{n\to\infty} C^n(u_1^{1/n}, \ldots, u_K^{1/n}), \quad u \in [0,1]^K \tag{21}$$

を満たす符号付きコピュラ C が存在することとする。またこのとき C は $\bar{C}$ の吸引域 (domain of attraction) に属すという。

符号付きコピュラ C が最大値安定 (max-stable) であるとは，全ての $n \geq 1$ に対して

$$C(u_1, \ldots, u_K) = C^n(u_1^{1/n}, \ldots, u_K^{1/n}), \quad u \in [0,1]^K$$

が成り立つこととする。次の補題はよく知られている。

補題 1. 符号付きコピュラ C に対し，それが極値であることと最大値安定であることは同値である。

Proof. 最大値安定ならば極値なのは明らかである。逆に $\bar{C}$ を極値符号付きコピュラとすると，任意の $m \geq 1$ に対して

$$\begin{aligned}\bar{C}^m(u_1^{1/m}, \ldots, u_K^{1/m}) &= \lim_{n\to\infty} C^{nm}(u_1^{1/nm}, \ldots, u_K^{1/nm}) \\ &= \bar{C}(u_1, \ldots, u_K)\end{aligned}$$

となるので最大値安定である。 □

検出可能モデルにおける最大値安定性の役割は次の補題で与えられる。

補題 2. Gumbel 分布 H_1 と符号付きコピュラ C で指定される検出可能モデルを考える。このとき C が最大値安定であるための必要十分条件は

$$Q(z_1 + \alpha, \ldots, z_K + \alpha) = Q(z_1, \ldots, z_K) + \alpha, \quad \alpha \in \mathbb{R}, \tag{22}$$

つまり Q が位置共変 (location equivariant) になることである。

Proof. C を最大値安定とする。このとき式 (17) より

$$\begin{aligned}Q(z_1, \ldots, z_K) &= -H_1^{-1}(C(H_1(-z_1), \ldots, H_1(-z_K))) \\ &= \log(-\log C(\exp(-e^{z_1}), \ldots, \exp(-e^{z_K}))) \\ &= \log(-\log C^n(\exp(-\frac{1}{n}e^{z_1}), \ldots, \exp(-\frac{1}{n}e^{z_K}))) \\ &= \log n + Q(-\log n + z_1, \ldots, -\log n + z_K)\end{aligned}$$

が任意の $n \geq 1$ に対して成り立つ。あとは Q の連続性による。逆も同様である。 □

検出可能モデルの不均衡極限は次のように特徴付けられる。これは極値理論で知られている事実（[2], Corollary 6.1.3 など）の類推である。

定理 3. Gumbel 分布 H_1 と符号付きコピュラ C によって指定される検出可能モデルを考え，式 (15) の G が

$$\lim_{n\to\infty} n\, G(-\log n + z_1, \ldots, -\log n + z_K) = g(z_1, \ldots, z_K) \tag{23}$$

という極限を持つとする。ただし g は連続関数とする。このとき，C は次の極値符号付きコピュラ $\bar{C}$ の吸引域に属す：

$$\bar{C}(u_1, \ldots, u_K) = \exp(-g(z_1, \ldots, z_K)), \quad z_k = \log(-\log u_k).$$

また H_1 と $\bar{C}$ で指定される検出可能モデルも式 (23) を満たす。

Proof. 式 (15) より，定理の仮定は

$$g(z_1,\dots,z_K)=\lim_{n\to\infty} n\{1-C(H_1(\log n-z_1),\dots,H_K(\log n-z_K))\}$$

と書ける。$u_k=\exp(-\exp(z_k))$ とおけば，$H_1(\log n-z_k)=u_k^{1/n}$ となるので

$$\begin{aligned} g(z_1,\dots,z_K) &= \lim_{n\to\infty} n\{1-C(u_1^{1/n},\dots,u_K^{1/n})\} \\ &= \lim_{n\to\infty}\{-\log C^n(u_1^{1/n},\dots,u_K^{1/n})\} \end{aligned}$$

となる。これは $\bar{C}=e^{-g(z_1,\dots,z_K)}$ $(z_k=\log(-\log u_k))$ が極値符号付きコピュラであることを意味する。また C がもともと最大値安定ならば $\bar{C}=C$ となる。よって定理が示された。 □

たとえば式 (3), (4) の準線形ロジスティック回帰モデルの場合，符号付きコピュラ C は式 (18) で与えられた。このとき定理 3 の $\bar{C}$ は

$$\begin{aligned} \bar{C}(u_1,\dots,u_K) &= \lim_{n\to\infty} C^n(u_1^{1/n},\dots,u_K^{1/n}) \\ &= \lim_{n\to\infty}\left(\frac{1}{1+(\sum_{k=1}^K(u_k^{-1/n}-1)^\tau)^{1/\tau}}\right)^n \\ &= \lim_{n\to\infty}\left(\frac{1}{1+(\sum_{k=1}^K(-\frac{1}{n}\log u_k)^\tau)^{1/\tau}}\right)^n \\ &= \lim_{n\to\infty}\left(\frac{1}{1+\frac{1}{n}(\sum_{k=1}^K(-\log u_k)^\tau)^{1/\tau}}\right)^n \\ &= \exp\left(-\left(\sum_{k=1}^K(-\log u_k)^\tau\right)^{1/\tau}\right) \end{aligned}$$

となる。これは（$\tau\geq 1$ のとき）Gumbel–Hougaard コピュラと呼ばれ，生成素 $\psi(u)=(-\log u)^\tau$ を持つアルキメデス型コピュラとなる。実は極値アルキメデス型コピュラは Gumbel–Hougaard コピュラに限られることが知られている [5]。

Gumbel–Hougaard コピュラに対応する予測子は，式 (17) より

$$\begin{aligned} Q(z_1,\dots,z_K) &= -H_1^{-1}(C(H_1(-z_1),\dots,H_K(-z_K))) \\ &= \frac{1}{\tau}\log\left(\sum_{k=1}^K e^{\tau z_k}\right) \end{aligned}$$

となる。結果的に，位置共変性を満たす準線形予測子は log-sum-exp の形に限られることが分かった。この事実は直接的に確かめることもできる（付録 A）。

図 1 は検出可能モデルを分類したものである。

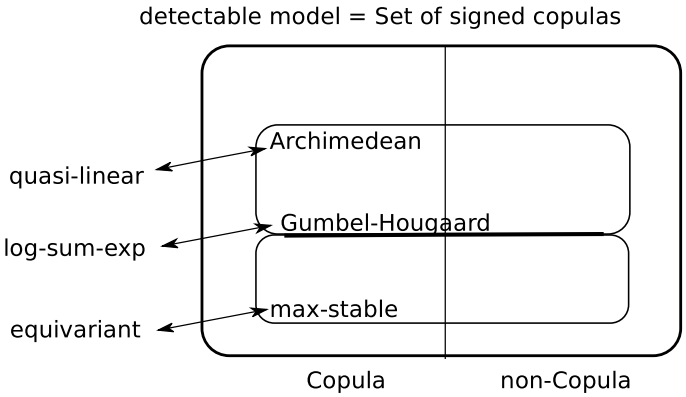

図 1: 検出可能モデルの分類。H_1 は Gumbel 分布に固定するものとする。

5　まとめと今後の課題

本論文では準線形ロジスティック回帰モデルを検出可能モデルに拡張した上で，その不均衡極限を特徴付けた（定理 3）。結論として，準線形ロジスティック回帰モデルは極値アルキメデス型コピュラによって特徴付けられることが分かった。

しかしながら準線形以外の検出可能予測子 Q の具体例は挙げなかった。定理 1 より，コピュラを一つ与えればそのような予測子が一つ定まる。リスク理論では多くのコピュラが提案され，利用されている。それらに対応する予測子がどのような性質を満たすかは今後の課題とする。

謝辞

本研究は科研費（課題番号：26108003, 17K00044）の助成を受けたものである。また本論文の執筆にあたり，大前勝弘博士との議論が大変参考になった。感謝を申し上げたい。

付録

A　log-sum-exp の特徴付け

式 (7) の準線形予測子のうち，位置共変なものは log-sum-exp に限られることを示す。

補題 3. 関数 $\phi : \mathbb{R} \to (0, \infty)$ は狭義単調増加とし，$\phi(-\infty) = 0, \phi(\infty) = \infty$ とする。もし $Q(z_1, \ldots, z_K) = \phi^{-1}(\sum_{k=1}^{K} \phi(z_k))$ が位置共変性 (22) を満たすならば，ある $\tau > 0$ が存在して $\phi(z) = \phi(0)e^{\tau z}$ となる。つまり Q は log-sum-exp となる。

Proof. $z_3, \ldots, z_K \to -\infty$ とおくことにより，最初から $K = 2$ の場合を考えればよい。位置共変性から任意の $\alpha \in \mathbb{R}$ に対して

$$\phi^{-1}(\phi(z_1 + \alpha) + \phi(z_2 + \alpha)) = \phi^{-1}(\phi(z_1) + \phi(z_2)) + \alpha$$

となる。したがって，$z_1 = \phi^{-1}(x_1), z_2 = \phi^{-1}(x_2)$ とおけば

$$\phi(\phi^{-1}(x_1) + \alpha) + \phi(\phi^{-1}(x_2) + \alpha) = \phi(\phi^{-1}(x_1 + x_2) + \alpha)$$

となる。これは関数 $\eta_\alpha(x) := \phi(\phi^{-1}(x) + \alpha)$ がコーシーの関数方程式

$$\eta_\alpha(x_1 + x_2) = \eta_\alpha(x_1) + \eta_\alpha(x_2)$$

を満たすことを意味する。η_α は単調増加だから，解は $\eta_\alpha(x) = \sigma_\alpha x$ $(\sigma_\alpha > 0)$ しかない。よって

$$\phi(\phi^{-1}(x) + \alpha) = \sigma_\alpha x$$

が成り立つ。さらに $x = \phi(z)$ とおけば

$$\phi(z + \alpha) = \sigma_\alpha \phi(z)$$

となる。$z = 0$ とおいて $\sigma_\alpha = \phi(\alpha)/\phi(0)$，よって

$$\phi(z + \alpha) = \frac{\phi(\alpha)\phi(z)}{\phi(0)}$$

を得る。そこで $\psi(z) = \log \phi(z) - \log \phi(0)$ とおけば

$$\psi(z + \alpha) = \psi(z) + \psi(\alpha)$$

となる。ψ は単調増加だから $\psi(z) = \tau z$ $(\tau > 0)$ と書ける。よって $\phi(z) = \phi(0)e^{\tau z}$ である。 □

B　H_1 が Gumbel 分布でない場合

定理 3 の拡張として次の補題が成り立つ。

補題 4. 1 次元分布 $H_1, \dots, H_K$ と符号付きコピュラ C で定義される関数

$$G(z_1, \dots, z_K) = 1 - C(H_1(-z_1), \dots, H_K(-z_K))$$

を考え，ある定数列 $c_{kn} \in \mathbb{R}, d_{kn} > 0$ および連続関数 $g(z_1, \dots, z_K)$ が存在して

$$\lim_{n\to\infty} nG(c_{1n} + d_{1n}z_1, \dots, c_{Kn} + d_{Kn}z_K) = g(z_1, \dots, z_K)$$

が成り立つと仮定する。このとき H_k は極値分布 $\bar{H}_k(w) = e^{-g(-\infty,\dots,-w,\dots,-\infty)}$ の吸引域に属し，また C は極値符号付きコピュラ

$$\bar{C}(u_1, \dots, u_K) = \exp(-g(\bar{H}_1^{-1}(u_1), \dots, \bar{H}_K^{-1}(u_K)))$$

の吸引域に属す。

Proof. z_k 以外を $-\infty$ とすることで

$$\lim_{n\to\infty} n\{1 - H_k(-c_{kn} - d_{kn}z_k)\} = g(-\infty, \dots, z_k, \dots, -\infty) =: g_k(z_k)$$

となる。よって $H_k(w)$ は極値分布 $\bar{H}_k(w) = e^{-g_k(-w)}$ の吸引域に属する。実際,

$$\begin{aligned} H_k(-c_{kn} + d_{kn}w)^n &= \exp(n \log H_k(-c_{kn} + d_{kn}w)) \\ &= \exp(n \log(1 - (1 - H_k(-c_{kn} + d_{kn}w)))) \\ &= \exp(-n(1 - H_k(-c_{kn} + d_{kn}w)) \\ &\to \exp(-g_k(-w)) \end{aligned}$$

となる。あとは定理 3 と同様にして $\bar{C}$ を求める。まず

$$\begin{aligned} & g(z_1, \dots, z_K) \\ &= \lim_{n\to\infty} n\{1 - C(H_1(-c_{1n} - d_{1n}z_1), \dots, H_K(-c_{Kn} - d_{Kn}z_K))\} \\ &\underset{(*)}{=} \lim_{n\to\infty} n\{1 - C(u_1^{1/n}, \dots, u_K^{1/n})\} \quad (u_k = \exp(-g_k(z_k))) \\ &= \lim_{n\to\infty} -\log C^n(u_1^{1/n}, \dots, u_K^{1/n}) \end{aligned}$$

と書ける。ただし $(*)$ の等号は議論を要するが略す。すると $\bar{C}$ が存在し，$\bar{C}(u_1, \dots, u_K) = e^{-g(z_1,\dots,z_K)}$ となる。 □

参考文献

[1] Baddeley, A., Berman, M., Fisher, N. I., Hardegen, A. Milne, R. K., Schuhmacher, D., Shah, R. and Turner, R. (2010). Spatial logistic regression and change-of-support in Poisson point processes, *Electron. J. Statist.*, **4**, 1151–1201.

[2] de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory – An Introduction*, Springer.

[3] 江口 真透 (2018). 統計モデリングのための一般化平均, ミニワークショップ「統計多様体の幾何学とその周辺 (10)」講演資料, 2018 年 11 月 17 日.

[4] Galambos, J. (1987). *The Asymptotic Theory of Extreme Order Statistics*, 2nd ed., Krieger.

[5] Genest, C. and Rivest, L.-P. (1989). A characterization of Gumbel's family of extreme value distributions, *Statist. Probab. Let.*, **8**, 207–211.

[6] Nelsen, R. B. (1999). *An Introduction to Copula*, 2nd ed., Springer.

[7] Omae, K. (2017). Statistical Learning by Quasi-linear Predictor, SOKENDAI, Ph. D. Thesis.

[8] Omae, K., Komori, O. and Eguchi, S. (2017). Quasi-linear score for capturing heterogeneous structure in biomarkers, *BMC Bioinformatics*, 18 (308), 1–15.

[9] Owen, A. B. (2007). Infinitely imbalanced logistic regression. *J. Mach. Learn. Res.*, **8**, 761–773.

[10] Sei, T. (2014). Infinitely imbalanced binomial regression and deformed exponential families, *J. Statist. Plann. Inference*, **149**, 116–124.

[11] Sibuya, M. (1960). Bivariate extreme statistics, I, *Ann. Inst. Statist. Math.*, **11** (3), 195–210.

[12] 塚原 英敦 (2011). 接合分布関数（コピュラ）の理論と応用, 「21 世紀の統計科学」（北川 源四郎，竹村 彰通 著）, 第 3 巻, 101–140, 日本統計学会, HP 版.

[13] Warton, D. I. and Shepherd, L. C. (2010). Poisson point process models solve the "pseudo-absence problem" for presence only data in ecology. *Ann. Applied Statist.*, **4**, 1383–1402.