

Chapter 3

Regression Analysis for Imbalanced Binary Data: Multi-dimensional Case



Tomonari Sei

Abstract We consider regression models for binary response data and study their behavior when the response is highly imbalanced. Previous studies have shown that if the logistic regression model is adopted, the likelihood function tends to that of an exponential family under the imbalance limit. This phenomenon is closely related to extreme value theory. In this paper, we discuss a multi-dimensional analogue of these results. First, we examine quasi-linear logistic models, where the binary outcome is explained by the log-sum-exp function of several linear scores. Then, we define a generalized model called a detectable model, and derive its imbalance limit using multivariate extreme value theory. The max-stability of the copulas corresponds to an equivariant property of the predictors.

Keywords Copula · Detectable model · Extreme value theory · Imbalanced data · Log-sum-exp function · Logistic regression · Max-stability · Quasi-linear predictor · Semi-copula

3.1 Introduction

The logistic regression model is defined by

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

where Y is a binary response variable, X is a p -dimensional explanatory variable, and $a \in \mathbb{R}$ and $b \in \mathbb{R}^p$ are regression coefficients. The function $G(z) = e^z / (1 + e^z)$ is the logistic distribution function, and its inverse $G^{-1}(u) = \log(u / (1 - u))$ is the logit link function.

T. Sei (✉)

The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
e-mail: sei@mist.i.u-tokyo.ac.jp

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021
N. Hoshino et al. (eds.), *Pioneering Works on Extreme Value Theory*,
JSS Research Series in Statistics,
https://doi.org/10.1007/978-981-16-0768-4_3

55

Now, consider the imbalanced case; that is, the probability of $Y = 1$ is very small. In the same fashion as Poisson's law of rare events, we assume that the true parameters depend on the sample size n . Specifically, let the true parameters be $a_n = -\log n + \alpha$ and $b_n = \beta$. Then, we obtain

$$\begin{aligned} P(Y = 1 \mid X = x) &= \frac{\frac{1}{n} e^{\alpha + \beta^\top x}}{1 + \frac{1}{n} e^{\alpha + \beta^\top x}} \\ &= \frac{1}{n} e^{\alpha + \beta^\top x} + O(n^{-2}), \end{aligned}$$

as $n \rightarrow \infty$. If the marginal distribution $F(dx)$ of X does not depend on n and its support is compact, then the weak limit of the conditional distribution of X given $Y = 1$ is, by Bayes' theorem,

$$\lim_{n \rightarrow \infty} P(X \in dx \mid Y = 1) = \frac{e^{\beta^\top x} F(dx)}{\int e^{\beta^\top x} F(dx)}, \quad (3.1)$$

which is an exponential family [13]. Furthermore, the joint distribution of X and Y converges to an inhomogeneous Poisson point process with intensity measure $e^{\alpha + \beta^\top x} F(dx)$; see [17] for details. We call the limit of a regression model under the imbalance assumption the *imbalance limit*.

There are other binary regression models with the same imbalance limit. For example, the complementary log-log link, which corresponds to $G(z) = 1 - \exp(-e^z)$, has the same imbalance limit as the logit link. In this case, $G(z)$ is the negative Gumbel distribution function, one of the min-stable distributions.

Similarly, the limit of a binary regression model with a cumulative distribution function $G(z)$ is characterized by extreme value theory [15]. Models with distinct link functions have the same imbalance limit if the corresponding distribution functions belong to the same domain of attraction. Here, min-stability corresponds to stability with respect to a resolution change of the explanatory variables [2].

In this study, we develop a multivariate analogue of the above facts. The function G is generalized to include multi-dimensional functions. A practical class is the quasi-linear logistic regression model proposed by [12], which combines several linear predictors using the log-sum-exp function. See Sect. 3.2 for a precise definition. We define a generalized class, called a detectable model. The imbalance limit of the model is obtained using the multivariate extreme value theory (e.g., [4, 14, 16]). Here, the max-stability of the copulas corresponds to an equivariant property of the detectable predictors.

The rest of the paper is organized as follows. In Sect. 3.2, we review the quasi-linear logistic regression model. The model is further generalized in Sect. 3.3, and the imbalance limit is studied in Sect. 3.4. Examples of equivariant predictors are provided in Sect. 3.5. Finally, Sect. 3.6 concludes the paper.

3.2 Quasi-linear Logistic Regression Model and Its Imbalance Limit

In this section, we first define the quasi-linear logistic regression model, and then derive its imbalance limit, as in (3.1).

3.2.1 The Quasi-linear Logistic Regression Model

Omae et al. [12] define a *quasi-linear logistic regression model* as follows:

$$P(Y = 1 | X = x) = \frac{e^Q}{1 + e^Q}, \quad (3.2)$$

$$Q = \frac{1}{\tau} \log \left(\sum_{k=1}^K e^{\tau(a_k + b_k^\top x)} \right), \quad (3.3)$$

where X is a p -dimensional explanatory variable, $a_k \in \mathbb{R}$ and $b_k \in \mathbb{R}^p$ are regression coefficients for each $k = 1, \dots, K$, and $\tau > 0$ is a tuning parameter. It is also possible to define (3.3) for $\tau < 0$ (see [11]), but we restrict τ to be positive, owing to a property discussed later (Lemma 3.1 in Sect. 3.3). We assume $K \geq 2$, unless otherwise stated.

The model reduces to the logistic regression model if $K = 1$, but is not even identifiable with respect to the regression coefficients if $K \geq 2$. Therefore, some restrictions and regularizations are imposed in practice. For example, the explanatory variable X is partitioned into K subvectors $X_{(1)}, \dots, X_{(K)}$ using a clustering method such as the K -means method. Then, the coordinates of b_k , except for those corresponding to $X_{(k)}$, are set to zero for each k .

Denote the K linear predictors by $z_k = a_k + b_k^\top x$. Then, the right-hand side of (3.3) is written as

$$Q(z_1, \dots, z_K) = \frac{1}{\tau} \log \left(\sum_k e^{\tau z_k} \right),$$

which we call the *quasi-linear predictor* or the *log-sum-exp function* (refer to [3]). The log-sum-exp function tends to the simple sum $\sum_k z_k$ as $\tau \rightarrow 0$ up to a constant term, and tends to $\max(z_1, \dots, z_K)$ as $\tau \rightarrow \infty$ for fixed (z_1, \dots, z_K) .

Reference [12] proposed the following generalized class:

$$Q = \phi^{-1} \left(\sum_k \phi(z_k) \right), \quad (3.4)$$

where ϕ is an invertible function. The log-sum-exp function is a particular case of $\phi(z) = e^{\tau z}$. Further generalization is discussed in the next section. In what follows, we call (3.4) the *generalized quasi-linear predictor* with the *generator* ϕ .

Remark 3.1 In [11], a slightly different definition is used,

$$Q = \phi^{-1} \left(\frac{1}{K} \sum_k \phi(z_k) \right),$$

and is called the generalized average or the Kolmogorov–Nagumo average. The difference is the factor $1/K$. In this study, we adopt the form in (3.4) because we focus on a property shown in Lemma 3.1, later.

3.2.2 Imbalance Limit

We derive the imbalance limit of the quasi-linear logistic regression model. Suppose the true parameters a_k and b_k in (3.3) are given by

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

which depend on the sample size n . Then, we have

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

and obtain the asymptotic form

$$P(Y = 1 \mid X = x) = \frac{e^Q}{1 + e^Q} = \frac{1}{n} \left(\sum_k e^{\tau(\alpha_k + \beta_k^\top x)} \right)^{1/\tau} + O(n^{-2}).$$

The conditional distribution of X , given $Y = 1$, is

$$\begin{aligned} P(X \in dx \mid Y = 1) &= \frac{P(Y = 1 \mid X = x)F(dx)}{\int P(Y = 1 \mid X = x)F(dx)} \quad (\text{Bayes' theorem}) \\ &\rightarrow \frac{\{\sum_k e^{\tau(\alpha_k + \beta_k^\top x)}\}^{1/\tau} F(dx)}{\int \{\sum_k e^{\tau(\alpha_k + \beta_k^\top x)}\}^{1/\tau} F(dx)}, \end{aligned}$$

where $F(dx)$ is the marginal distribution of X . In particular, the distribution is reduced to a mixed exponential family if $\tau = 1$.

Remark 3.2 In [12], the authors note that the quasi-linear logistic model with $\tau = 1$ is Bayes optimal if the conditional distribution of X , given Y , is mixture normal.

Specifically, suppose that the ratio of the conditional distributions of X is a mixture exponential family

$$\frac{P(X \in dx \mid Y = 1)}{P(X \in dx \mid Y = 0)} = \frac{1}{Z} \sum_k e^{\alpha_k + \beta_k^\top x},$$

where α_k and β_k are parameters, and Z is a normalization constant. Then, the logit of the predictive distribution is

$$\log \frac{P(Y = 1 \mid X = x)}{P(Y = 0 \mid X = x)} = \log \left(\sum_k e^{\alpha_k^* + \beta_k^\top x} \right),$$

where $\alpha_k^* = \alpha_k - \log Z + \log(\pi_1/\pi_0)$ and $\pi_y = P(Y = y)$. This is the quasi-linear predictor.

3.3 Extension of the Model and Its Copula Representation

In this section, we extract several features of the quasi-linear logistic model, and use these to define a generalized class of regression models. We also discuss the relationship between this class of models and copula theory.

3.3.1 Detectable Model

We first focus on the following property of the generalized quasi-linear predictor (3.4), with generator ϕ .

Lemma 3.1 *Suppose $\phi : \mathbb{R} \rightarrow (0, \infty)$ is continuous, strictly increasing, and has boundary values $\phi(-\infty) = 0$ and $\phi(\infty) = \infty$. Then, the generalized quasi-linear predictor (3.4) satisfies*

$$Q(z_1, \dots, z_k, \dots, z_K) \text{ is increasing in } z_k, \quad (3.5)$$

$$Q(-\infty, \dots, z_k, \dots, -\infty) = z_k, \quad (3.6)$$

for each k and $(z_1, \dots, z_K) \in \mathbb{R}^K$.

Properties (3.5) and (3.6) are also satisfied by $Q(z_1, \dots, z_K) = \max(z_1, \dots, z_K)$, where the increasing property in (3.5) is interpreted as nondecreasing. In a sense, property (3.6) respects the maximum of the K linear scores (p. 4 of [12]).

In general, we call a function $Q : \mathbb{R}^K \rightarrow \mathbb{R}$ a *detectable predictor* if it satisfies (3.5) and (3.6). Note that the quantity $Q(-\infty, \dots, z_k, \dots, -\infty)$ does not

depend on the choice of the diverging sequence of $(z_1, \dots, z_{k-1}, z_{k+1}, \dots, z_K)$ to $(-\infty, \dots, -\infty)$ under the monotonicity condition (3.5).

Remark 3.3 The term “detectable” is borrowed from the neural network literature (e.g., Chaps. 6 and 9 of [8]), where a number of compositions of one-dimensional nonlinear functions and multi-dimensional linear functions are applied. In contrast, we focus on the properties of the multi-dimensional nonlinear function Q using the copula theory.

Then, we define a model class, as follows.

Definition 3.1 (*Detectable model*) Let Q be a detectable predictor, and let G_1 be a strictly increasing continuous distribution function. Then, a *detectable model* with G_1 and Q is defined by

$$P(Y = 1 | X = x) = G_1(Q), \quad (3.7)$$

$$Q = Q(a_1 + b_1^\top x, \dots, a_K + b_K^\top x). \quad (3.8)$$

We call G_1 the *inverse link function*.

For example, the quasi-linear logistic model is a detectable model with $G_1(Q) = e^Q / (1 + e^Q)$ and $Q(z_1, \dots, z_K) = \tau^{-1} \log(\sum_k e^{\tau z_k})$. Similarly to the quasi-linear model, the detectable model aggregates K linear predictors into a quantity Q .

We give two properties of detectable predictors. The proofs are easy, and thus are omitted.

Lemma 3.2 *Any detectable predictor Q satisfies an inequality*

$$Q(z_1, \dots, z_K) \geq \max(z_1, \dots, z_K).$$

Lemma 3.3 *Let Q_1 and Q_2 be detectable predictors. Then, $(Q_1 + Q_2)/2$, $\max(Q_1, Q_2)$ and $\min(Q_1, Q_2)$ are also detectable. More generally, if a function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ is increasing in each argument and satisfies $f(x, x) = x$, for all $x \in \mathbb{R}$, then $f(Q_1, Q_2)$ is detectable.*

The generalized average mentioned in Remark 3.1 is an example of f in which $f(x, x) = x$.

3.3.2 Copula Representation

The detectable model has a copula representation. Consider a detectable model with an inverse link function G_1 , and a detectable predictor Q . Denote the composite map of G_1 and Q by

$$G(z_1, \dots, z_K) = G_1(Q(z_1, \dots, z_K)).$$

Then, G is increasing in each variable and satisfies

$$G(-\infty, \dots, z_k, \dots, -\infty) = G_1(z_k).$$

Next, define a dual of G by

$$H(w_1, \dots, w_K) = 1 - G(-w_1, \dots, -w_K), \quad (w_1, \dots, w_K) \in \mathbb{R}^K,$$

and

$$H_1(w) = 1 - G_1(-w), \quad w \in \mathbb{R}. \quad (3.9)$$

Then, H is increasing in each variable and satisfies

$$H(\infty, \dots, w_k, \dots, \infty) = H_1(w_k).$$

Thus, H_1 is considered the k th marginal distribution function of H . Note that H itself may not be a multivariate distribution function because the K -increasing property may fail. Recall that a function H is said to be K -increasing if $\Delta_1 \cdots \Delta_K H \geq 0$, where Δ_k is the difference operator with respect to the k th argument.

Finally, as with Sklar's theorem, we define

$$C(u_1, \dots, u_K) = H(H_1^{-1}(u_1), \dots, H_1^{-1}(u_K)). \quad (3.10)$$

Then, C satisfies the following conditions:

$$\begin{aligned} C(1, \dots, u_k, \dots, 1) &= u_k, \\ C(u_1, \dots, u_K) &\text{ is increasing in } u_k, \end{aligned}$$

for each k . A function $C : [0, 1]^K \rightarrow [0, 1]$ satisfying the two conditions is called a *semi-copula* (see Chap. 8 of [5]). Any copula is a semi-copula, but the converse is not true. The K th-order difference $\Delta_1 \cdots \Delta_K C$ of a semi-copula, which measures a rectangular region, may be negative.

We summarize this result as follows.

Theorem 3.1 (Copula representation) *A detectable model specified by an inverse link function G_1 and a detectable predictor Q is represented as*

$$\begin{aligned} G_1(Q(z_1, \dots, z_K)) &= G(z_1, \dots, z_K) \\ &= 1 - H(-z_1, \dots, -z_K) \\ &= 1 - C(H_1(-z_1), \dots, H_1(-z_K)), \end{aligned}$$

where C is a semi-copula, and H_1 is a univariate continuous distribution function. The correspondence

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

is one-to-one.

Proof It is sufficient to prove the one-to-one correspondence. Indeed, if G_1 and Q are given, then H_1 and C are determined by (3.9) and (3.10), respectively. Conversely, if H_1 and C are given, then we have $G_1(z) = 1 - H_1(-z)$ by (3.9), and

$$Q(z_1, \dots, z_K) = -H_1^{-1}(C(H_1(-z_1), \dots, H_1(-z_K)))$$

holds. □

Consider again the quasi-linear logistic regression model with the log-sum-exp predictor, which corresponds to (3.2) and (3.3). Then, the functions H_1 and C in Theorem 3.1 are the logistic distribution function and

$$C(u_1, \dots, u_K) = \frac{1}{1 + (\sum_{k=1}^K (\frac{1-u_k}{u_k})^\tau)^{1/\tau}}, \quad (3.11)$$

respectively. The function C is a copula if $\tau \geq 1$, as shown in Example 4.26 of [10]. In particular, if $\tau = 1$, then

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

which belongs to the Clayton copula family [10]. If $\tau \rightarrow \infty$, then C converges to $\min_k u_k$, the upper Fréchet–Hoeffding bound. If $0 < \tau < 1$, C is not a copula, in the strict sense, because it is not K -increasing.

We say that a semi-copula C is Archimedean if it is written as

$$C(u_1, \dots, u_K) = \psi^{-1}(\psi(u_1) + \dots + \psi(u_K)),$$

with a decreasing function $\psi : (0, 1) \rightarrow (0, \infty)$ called the generator (e.g., [10]). For example, the semi-copula in (3.11) is Archimedean with the generator $\psi(u) = (\frac{1-u}{u})^\tau$.

Archimedean semi-copulas characterize the generalized quasi-linear models as stated in the following theorem. The proof is straightforward.

Theorem 3.2 (Archimedean case) *Let $\{G_1, Q\}$ be a detectable model and $\{H_1, C\}$ be the corresponding pair determined by Theorem 3.1. Then, Q is a generalized quasi-linear predictor (3.4) with a generator ϕ if and only if C is an Archimedean semi-copula with a generator ψ . The relation between the generators ϕ and ψ is given by $\phi(z) = \psi(H_1(-z))$.*

Note that ψ depends not just on ϕ , but also on the inverse link G_1 .

Remark 3.4 Here, we briefly discuss the merit of having a genuine copula in the copula representation of a detectable model, where a genuine copula means a semi-copula with the K -increasing property. If C is a genuine copula, then the detectable model $P(Y = 1 \mid X = x) = G_1(Q(z_1, \dots, z_K))$ has the following latent variable representation. Take a random vector $U = (U_1, \dots, U_K)$, distributed according to the copula C . Then, $G_1(Q(z_1, \dots, z_K)) = 1 - C(H_1(-z_1), \dots, H_1(-z_K))$ coincides with the probability of an event, such that $U_k > H_1(-z_k)$ for at least one k . Now, the response variable Y can be assumed to be the indicator function of that event. The random vector U is seen as a latent variable. Once a latent variable representation is obtained, we can also consider a state-space model for time-dependent data. This is left to future research.

3.4 The Imbalance Limit of Detectable Models

In this section, we characterize the imbalance limit of detectable models using the copula representation in Theorem 3.1 and the multivariate extreme value theory [4, 14, 16].

Recall that detectable models are specified by a univariate distribution function H_1 and a semi-copula C . Throughout this section, we fix H_1 as the Gumbel distribution function

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

and focus on the semi-copulas C . In this case, the inverse link function is $G_1(z) = 1 - \exp(-e^z)$, which corresponds to the complementary log-log link function. The relation between C and the detectable predictor Q is given by

$$C(u_1, \dots, u_K) = \exp(-e^{Q(z_1, \dots, z_K)}), \quad u_k = \exp(-e^{z_k}), \quad (3.12)$$

by Theorem 3.1.

A semi-copula C is said to be *extreme* if there exists a semi-copula C_0 such that

$$C(u_1, \dots, u_K) = \lim_{n \rightarrow \infty} C_0^n(u_1^{1/n}, \dots, u_K^{1/n}), \quad u \in [0, 1]^K. \quad (3.13)$$

In this case, we say that C_0 belongs to the domain of attraction of C . A semi-copula C is said to be *max-stable* if, for all $n \geq 1$,

$$C(u_1, \dots, u_K) = C^n(u_1^{1/n}, \dots, u_K^{1/n}), \quad u \in [0, 1]^K. \quad (3.14)$$

The following lemma is widely known for copulas.

Lemma 3.4 *A semi-copula C is extreme if and only if it is max-stable.*

Proof It is obvious that any max-stable semi-copula is also extreme. Conversely, assume that C is extreme. Let C_0 be a semi-copula that satisfies (3.13). Then, we have

$$\begin{aligned} C^m(u_1^{1/m}, \dots, u_K^{1/m}) &= \lim_{n \rightarrow \infty} C_0^{nm}(u_1^{1/nm}, \dots, u_K^{1/nm}) \\ &= C(u_1, \dots, u_K), \end{aligned}$$

for all $m \geq 1$, which means C is max-stable. \square

Max-stability is reflected in detectable models as follows.

Lemma 3.5 *Consider a detectable model specified by the Gumbel distribution function H_1 and a semi-copula C . Then, C is max-stable if and only if the detectable predictor Q is equivariant with respect to location; that is,*

$$Q(z_1 + \alpha, \dots, z_K + \alpha) = Q(z_1, \dots, z_K) + \alpha, \quad \alpha \in \mathbb{R}. \quad (3.15)$$

Proof Let C be max-stable. Then, by (3.12) and (3.14), we have

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

for all $n \geq 1$. Then, (3.15) is proved for $\alpha = \log x$, with positive rational numbers x . The result follows from the monotonicity of Q . The converse is proved in a similar manner. \square

Remark 3.5 According to extreme value theory, the stable tail dependence function corresponding to a max-stable copula C is defined by

$$l(x_1, \dots, x_K) = -\log C(e^{-x_1}, \dots, e^{-x_K}), \quad (x_1, \dots, x_K) \in [0, \infty)^K,$$

which satisfies a homogeneous property $l(tx_1, \dots, tx_K) = tl(x_1, \dots, x_K)$ (see [6]). The equivariance of Q in Lemma 3.5 is interpreted as another representation of the max-stable property. Note that l is not suitable for constructing predictors because its domain is not the whole space.

The imbalance limit of detectable models is characterized as follows. The result is an analogue of that in extreme value theory (e.g., Corollary 6.1.3 of [4]).

Theorem 3.3 (Imbalance limit) *Consider a detectable model specified by the Gumbel distribution function H_1 and a semi-copula C . Let G_1 , Q , and G be the functions determined by Theorem 3.1. Then, the following three conditions are equivalent, where \tilde{Q} denotes an equivariant predictor:*

1. The predictor Q admits a limit

$$\lim_{n \rightarrow \infty} \{n Q(z_1 - \log n, \dots, z_K - \log n) + \log n\} = \bar{Q}(z_1, \dots, z_K).$$

2. The function G admits a limit

$$\lim_{n \rightarrow \infty} \{n G(z_1 - \log n, \dots, z_K - \log n)\} = e^{\bar{Q}(z_1, \dots, z_K)}.$$

3. The semi-copula C belongs to the domain of attraction of

$$\bar{C}(u_1, \dots, u_K) = \exp(-e^{\bar{Q}(z_1, \dots, z_K)}), \quad u_k = \exp(-e^{z_k}).$$

Under these conditions, if the true regression coefficients are $a_{k,n} = -\log n + \alpha_k$ and $b_{k,n} = \beta_k$, then the weak limit of the conditional distribution of X is

$$\lim_{n \rightarrow \infty} P(X \in dx \mid Y = 1) = \frac{e^{\bar{Q}(\alpha_1 + \beta_1^\top x, \dots, \alpha_K + \beta_K^\top x)} F(dx)}{\int e^{\bar{Q}(\alpha_1 + \beta_1^\top x, \dots, \alpha_K + \beta_K^\top x)} F(dx)}$$

whenever the support of $F(dx) = P(X \in dx)$ is compact.

Proof The equivalence of conditions 1 and 3 follows immediately from the relation (3.12). We prove the equivalence of conditions 1 and 2. Because

$$G(z_1, \dots, z_K) = 1 - \exp(-e^{Q(z_1, \dots, z_K)}),$$

condition 2 is written as

$$\lim_{n \rightarrow \infty} n \{1 - \exp(-e^{Q(z_1 - \log n, \dots, z_K - \log n)})\} = e^{\bar{Q}(z_1, \dots, z_K)},$$

which is also equivalent to

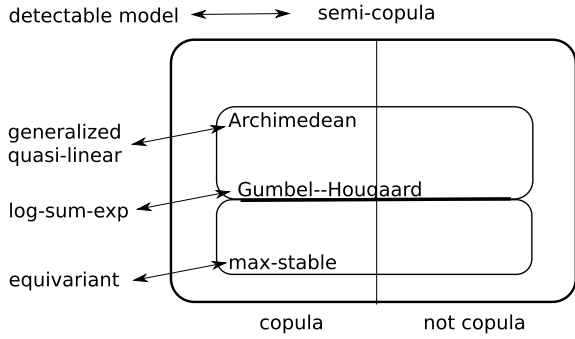
$$\lim_{n \rightarrow \infty} n e^{Q(z_1 - \log n, \dots, z_K - \log n)} = e^{\bar{Q}(z_1, \dots, z_K)}.$$

The logarithm of both sides yields condition 1.

Next, we show the convergence of the conditional distribution. Note that the convergence of G in condition 2 is locally uniform with respect to (z_1, \dots, z_K) because G is monotone in each argument. Then, Bayes' theorem and the compactness of the support of F imply

$$\begin{aligned} P(X \in dx \mid Y = 1) &= \frac{G(-\log n + \alpha_1 + \beta_1^\top x, \dots, -\log n + \alpha_K + \beta_K^\top x) F(dx)}{\int G(-\log n + \alpha_1 + \beta_1^\top x, \dots, -\log n + \alpha_K + \beta_K^\top x) F(dx)} \\ &\rightarrow \frac{e^{\bar{Q}(\alpha_1 + \beta_1^\top x, \dots, \alpha_K + \beta_K^\top x)} F(dx)}{\int e^{\bar{Q}(\alpha_1 + \beta_1^\top x, \dots, \alpha_K + \beta_K^\top x)} F(dx)}, \end{aligned}$$

Fig. 3.1 Classification of detectable models, where H_1 is fixed to be Gumbel



as stated. □

For example, consider the semi-copula in (3.11), which is derived from the log-sum-exp logistic model. Here, the extreme semi-copula \bar{C} in Theorem 3.3 is

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

which is called the Gumbel–Hougaard copula [10] if $\tau \geq 1$. In particular, it reduces to the independent copula if $\tau = 1$. The Gumbel–Hougaard copula is an Archimedean copula with generator $\psi(u) = (-\log u)^\tau$. In fact, this class is characterized by the max-stable Archimedean property [7].

The detectable predictor Q corresponding to the Gumbel–Hougaard copula when H_1 is Gumbel is the log-sum-exp

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

As a result, the generalized quasi-linear predictor with the equivariant property (3.15) is limited to be the log-sum-exp predictor. This fact is directly confirmed in Lemma 3.6 in Sect. 3.5. Note too that the independent copula corresponds to $\tau = 1$.

Figure 3.1 classifies the detectable models.

If H_1 is not Gumbel, the imbalance limit depends on the domain of attraction to which H_1 belongs. For example, the logistic distribution belongs to the domain of attraction of the Gumbel. For such a case, the statements in Theorem 3.3 still hold.

3.5 Examples of Equivariant Predictors

In this section, we provide examples of equivariant predictors, where the equivariance is defined by (3.15). Recall that equivariant predictors correspond to max-stable semi-copulas if H_1 is Gumbel (Lemma 3.5). In the following, we construct the predictors directly and do not use the copula representations (except for Lemma 3.7).

It is obvious that, by definition, the log-sum-exp predictor is equivariant. Conversely, the log-sum-exp predictor is characterized as follows.

Lemma 3.6 *Let Q be a generalized quasi-linear predictor with a generator ϕ , where $\phi : \mathbb{R} \rightarrow (0, \infty)$ is continuous and strictly increasing, $\phi(-\infty) = 0$, and $\phi(\infty) = \infty$. Then, Q is equivariant if and only if it is the log-sum-exp predictor for some $\tau > 0$.*

Proof We prove the “only if” part. It is enough to consider the case $K = 2$, because we can set $\phi(z_k) = 0$ for $3 \leq k \leq K$ by letting $z_k \rightarrow -\infty$. Because Q is equivariant, we have

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

for any $\alpha \in \mathbb{R}$. Applying ϕ to the both sides and putting $z_k = \phi^{-1}(x_k)$, we obtain

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

This is Cauchy’s functional equation (Theorem 2.1.1 of [1]) on $\eta(x) := \phi(\phi^{-1}(x) + \alpha)$. Because η is increasing, the solution has to be $\eta(x) = \phi(\phi^{-1}(x) + \alpha) = \sigma_\alpha x$, for some $\sigma_\alpha > 0$. Put $z = \phi^{-1}(x)$ to obtain $\phi(z + \alpha) = \sigma_\alpha \phi(z)$. By letting $z = 0$, we have $\sigma_\alpha = \phi(\alpha)/\phi(0)$ and, therefore,

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

By putting $\psi(z) = \log \phi(z) - \log \phi(0)$, we have $\psi(z + \alpha) = \psi(z) + \psi(\alpha)$. Again, because ψ is increasing, we have $\psi(z) = \tau z$, for some $\tau > 0$, which means $\phi(z) = \phi(0)e^{\tau z}$. Hence, ϕ is the generator of the log-sum-exp predictor. \square

For other examples, consider

$$Q(z_1, z_2) = \frac{z_1 + z_2 + \sqrt{(z_1 - z_2)^2 + 4\varepsilon^2}}{2}, \quad (3.16)$$

where $\varepsilon > 0$ is a fixed constant. This is actually an equivariant detectable predictor. Indeed, it satisfies the conditions $\partial Q/\partial z_k > 0$, $Q(z, -\infty) = Q(-\infty, z) = z$, and $Q(z_1 + \alpha, z_2 + \alpha) = Q(z_1, z_2) + \alpha$. The function Q in (3.16) is quite different from the log-sum-exp function when $|z_1 - z_2|$ is large. Indeed, if $z_1 > z_2$ and z_1 is fixed, then

$$Q(z_1, z_2) = z_1 + O((z_1 - z_2)^{-1}),$$

as $z_1 - z_2 \rightarrow \infty$, whereas

$$\frac{1}{\tau} \log(e^{\tau z_1} + e^{\tau z_2}) = z_1 + O(e^{-\tau(z_1 - z_2)}).$$

The case of $z_1 < z_2$ is derived in a similar manner. The behavior for large $|z_1 - z_2|$ may affect the numerical stability of the parameter estimation. This is left to future work.

A multivariate extension of (3.16) is the unique solution of

$$\prod_{k=1}^K (Q - z_k) = \varepsilon^K, \quad Q > \max(z_1, \dots, z_K), \quad (3.17)$$

which we call an *algebraic predictor*. The tail behavior is given by

$$Q = z_{(1)} + O((z_{(1)} - z_{(2)})^{-(K-1)}),$$

as $z_{(1)} - z_{(2)} \rightarrow \infty$, where $z_{(1)} \geq \dots \geq z_{(K)}$ is the order statistic of z_1, \dots, z_K .

We can construct a broad class of equivariant predictors using a direct consequence of extreme value theory, as follows.

Lemma 3.7 *Let μ be a (nonnegative) measure on the simplex $\Delta = \{s \mid \sum_{k=1}^K s_k = 1, s_1, \dots, s_K \geq 0\}$, such that $\int s_k \mu(ds) = 1$ for all k . Then,*

$$Q(z_1, \dots, z_K) = \log \int \max(s_1 e^{z_1}, \dots, s_K e^{z_K}) \mu(ds) \quad (3.18)$$

is an equivariant detectable predictor. Conversely, if Q is equivariant and the semi-copula C determined by Theorem 3.1 with the Gumbel marginal H_1 is a (genuine) copula, then there exists such a unique measure μ .

Proof It is easy to see that Q in (3.18) is actually an equivariant predictor. To prove the converse, suppose that Q is equivariant and C determined by Theorem 3.1, with the Gumbel marginal H_1 , is a copula. Lemma 3.5 implies that C is a max-stable copula and, therefore, H in Theorem 3.1 is a max-stable distribution function with the Gumbel marginal H_1 . Then, by Proposition 5.11' of [14], H has the spectral representation

$$H(x_1, \dots, x_K) = \exp \left\{ - \int_{\Delta} \max(s_1 e^{-x_1}, \dots, s_K e^{-x_K}) \mu(dx) \right\},$$

with a measure μ on Δ such that $\int s_k \mu(dx) = 1$, for all k . Equation (3.18) follows from the representation $Q(z_1, \dots, z_K) = \log(-\log H(-z_1, \dots, -z_K))$. \square

The measure μ is called the *spectral measure*. For example, let $K = 3$ and $\mu = (\delta_{(1/2, 1/2, 0)} + \delta_{(1/2, 0, 1/2)} + \delta_{(0, 1/2, 1/2)})/2$, where δ denotes the Dirac measure. Then,

$$Q(z_1, z_2, z_3) = \log \left(\frac{e^{\max(z_1, z_2)} + e^{\max(z_1, z_3)} + e^{\max(z_2, z_3)}}{2} \right).$$

Using the order statistic $z_{(1)} \geq z_{(2)} \geq z_{(3)}$ of (z_1, z_2, z_3) , we have

$$Q(z_1, z_2, z_3) = \log \left(e^{z_{(1)}} + \frac{e^{z_{(2)}}}{2} \right),$$

which, in particular, depends only on the top two scores $(z_{(1)}, z_{(2)})$. More generally,

$$Q(z_1, \dots, z_K) = \frac{1}{\tau} \log \left(e^{\tau z_{(1)}} + \sum_{k=2}^K \lambda_k e^{\tau z_{(k)}} \right)$$

is equivariant for any positive τ and nonnegative λ_k . The log-sum-exp function is the special case $\lambda_2 = \dots = \lambda_K = 1$.

Note that Q defined by (3.18) must satisfy

$$Q(z_1, \dots, z_K) \leq \log(e^{z_1} + \dots + e^{z_K}),$$

which follows from $\max_k (s_k e^{z_k}) \leq \sum_k s_k e^{z_k}$. In particular, employing the lower bound in Lemma 3.2, we can prove that the tail behavior of Q is restricted to

$$Q(z_1, \dots, z_K) = z_{(1)} + O(e^{-(z_{(1)} - z_{(2)})}),$$

as $z_{(1)} - z_{(2)} \rightarrow \infty$. Thus, the algebraic predictor (3.17) cannot be expressed as (3.18) with a (nonnegative) spectral measure μ .

3.6 Conclusion

In this paper, we introduced detectable models as generalizations of the quasi-linear logistic models, and then derived the imbalance limit (Theorem 3.3). A key property is that of equivariance (3.15). The log-sum-exp function is characterized as a unique equivariant quasi-linear predictor (Lemma 3.6); see Sect. 3.5 for examples of other equivariant predictors.

We have not conducted any simulation results. Thus, future work should investigate the numerical stability of the maximum likelihood estimator when an equivariant predictor such as the algebraic predictor (3.17) is adopted.

The generalized average of the form in Remark 3.1 can be extended to functions with the property $Q(z, \dots, z) = z$ instead of (3.6). Regression models with such a property may exhibit different behaviors.

Lastly, we have implicitly assumed that the conditional probability $P(Y = 1 \mid X = x)$ ranges from zero to one. However, this assumption may be relaxed. In fact,

[9] suggests an asymmetric logistic regression model that uses $G(z) = (e^z + \kappa)/(1 + e^z + \kappa)$, for $\kappa > 0$, as the inverse link function. This function is not even a distribution function because $G(-\infty) > 0$. Therefore, it would be interesting to investigate what happens if $\kappa_n \rightarrow 0$ as $n \rightarrow \infty$ under the imbalance limit.

Acknowledgements The author thanks the reviewer for his/her insightful comments and references. In particular, the author was not previously aware of the term “semi-copula.” This research was motivated by Prof. Masaaki Sibuya’s questions during a presentation at Keio University in 2013, and his subsequent comments during the workshop at the Institute of Statistical Mathematics in 2019. The author thanks Katsuhiko Omae, Osamu Komori, and Shinto Eguchi for providing helpful discussions and information. This work was supported by JSPS KAKENHI Grant Numbers 26108003 and 17K00044.

References

1. Aczél J (1966) Lectures on functional equations and their applications. Academic, New York
2. Baddeley A, Berman M, Fisher NI, Hardegen A, Milne RK, Schuhmacher D, Shah R, Turner R (2010) Spatial logistic regression and change-of-support in Poisson point processes. *Electron J Statist* 4:1151–1201
3. Boyd S, Vandenberghe L (2004) Convex optimization. Cambridge University Press, Cambridge
4. de Haan L, Ferreira A (2006) Extreme value theory – an introduction. Springer, Berlin
5. Durante F, Sempi C (2016) Principles of copula theory. CRC Press, Boca Raton
6. Genest C, Nešlehová J (2012) Copula modeling for extremes. In: El-Shaarawi AH, Piegorsch WW (eds) Encyclopedia of environmetrics, 2nd ed. Wiley, Hoboken
7. Genest C, Rivest L-P (1989) A characterization of Gumbel’s family of extreme value distributions. *Stat Probab Lett* 8:207–211
8. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. The MIT Press, Cambridge
9. Komori O, Eguchi S, Ikeda S, Okamura H, Ichinokawa M, Nakayama S (2016) An asymmetric logistic regression model for ecological data. *Methods Ecol Evol* 7:249–260
10. Nelsen RB (1999) An introduction to copula, 2nd edn. Springer, Berlin
11. Omae K (2017) Statistical learning by quasi-linear predictor. SOKENDAI, Ph. D. Thesis
12. Omae K, Komori O, Eguchi S (2017) Quasi-linear score for capturing heterogeneous structure in biomarkers. *BMC Bioinform* 18(308):1–15
13. Owen AB (2007) Infinitely imbalanced logistic regression. *J Mach Learn Res* 8:761–773
14. Resnick SI (1987) Extreme values, regular variation, and point processes. Springer, New York
15. Sei T (2014) Infinitely imbalanced binomial regression and deformed exponential families. *J Stat Plann Inference* 149:116–124
16. Sibuya M (1960) Bivariate extreme statistics. *I Ann Inst Stat Math* 11(3):195–210
17. Warton DI, Shepherd LC (2010) Poisson point process models solve the “pseudo-absence problem” for presence only data in ecology. *Ann Appl Stat* 4:1383–1402