Estimating the Number of Clusters in Logistic Regression Clustering by an Information Theoretic Criterion

Guoqi Qian[1], C. Radhakrishna Rao[2], Yuehua Wu[3] and Qing Shao[4]

[1] Department of Mathematics and Statistics, University of Melbourne, VIC 3010, Australia g.qian@ms.unimelb.edu.au
[2] Department of Statistics, Penn State University, University Park, PA 16802, U.S.A. crr1@psu.edu
[3] Department of Mathematics and Statistics, York University, 4700 Keele Street, Toronto, Ontario, M3J 1P3, Canada wuyh@yorku.ca
[4] Biostatistics and Statistical Reporting, One Health Plaza, Bldg. 435 - 4173, Novartis Pharmaceuticals Corporation, East Hanover, NJ 07936, U.S.A. qing.shao@novartis.com

1 Introduction

It is well-known that a logistic regression model aims at finding how a response variable Y is influenced by a set of explanatory variables $\{x_1, \ldots, x_p\}$ when Y is either binary with values 0 and 1 or a proportion of values between 0 and 1. A logistic regression model consists of three components (McCullagh and Nelder (1989)):

1. A random component Y that is either binary with values 0 and 1 or a proportion with values between 0 and 1. In the latter case, $Y = Z/m$ where Z is assumed to have a binomial distribution $B(m, \pi)$ with the probability of "success" π and the number of independent "experiments" m. We have binary data if $m \equiv 1$.
2. A systematic component (*linear predictor*) $\eta = \boldsymbol{x}'\boldsymbol{\beta}$, where $\boldsymbol{x} = (x_1, \ldots, x_p)'$ and $\boldsymbol{\beta}$ is the unknown p-vector parameter of interest.
3. A function $\pi = h(\eta) = e^{\eta}/(1 + e^{\eta})$ that relates the expectation π of Y with the linear predictor η. The inverse function $g(\pi)$ of $h(\eta)$ is named the *logistic link function*, where $g(\pi) \stackrel{\text{def}}{=} \log(\pi/(1-\pi)) = \eta$.

Logistic regression has been one of the most frequently used techniques in applications. Yet at times either the logistic curve does not describe the probability of success $\pi(\boldsymbol{x})$ adequately, or m is larger than 1 and Y

is more variable than the binomial distribution allows, which is termed *over-dispersion* in the literature. Over-dispersion relative to binomial distribution is possible if the m trials in a set are positively correlated, or an important covariate is omitted. A simple way to accommodate departures from a single logit link and over-dispersion is to introduce the logistic regression clustering model. Examples on the fitting of mixtures of logistic regression to biological and marketing data may be found in Farewell and Sprott (1988), Follmann and Lamber (1989, 1991), and Wedel and DeSarbo (1995), etc.

This paper studies the problem of estimating the number of clusters in the context of logistic regression clustering. The classification likelihood approach is employed to tackle this problem. An information theoretic criterion for selecting the number of logistic curves is proposed in the sequel and then its asymptotic property is considered.

The paper is arranged as follows: In Section 2, some notations are given and an information theoretic criterion is proposed for estimating the number of clusters. In Section 3, the small sample performance of the proposed criterion is studied by Monte Carlo simulation. In Section 4, the asymptotic property of the criterion proposed in Section 2 is investigated. Some lemmas needed in Section 4 are given in the appendix.

2 Notation and Preliminaries

Assume that we have n objects $\mathcal{O}^{(n)} = \{1, 2, \ldots, n\}$ with the associated data points $(\boldsymbol{x}_1, y_1)$, ..., $(\boldsymbol{x}_n, y_n)$, where $\boldsymbol{x}_j' = (x_{j1}, \ldots, x_{jp}) \in \mathbb{R}^p$ is a fixed explanatory p-vector and $y_j \in \mathbb{R}$ is a random dependent variable. The hidden true distributions of $y_1, \ldots, y_n$ are the binomial distributions $B(m_1, \pi_{01}), \ldots, B(m_n, \pi_{0n})$. The set of these n objects is a random sample coming from a structured population. Suppose that this population is composed of k_0 sub-populations, each of which has a distinct underlying linear predictor between the response variable and the explanatory variables. Then, there exists a hidden true partition of these n objects $\Pi_{k_0}^{(n)} = \{\mathcal{O}_1^{(n)}, \ldots, \mathcal{O}_{k_0}^{(n)}\}$, and each cluster $\mathcal{O}_i^{(n)} \triangleq \{i_1, \ldots, i_{n_i}\} \subseteq \mathcal{O}^{(n)}$ is characterized by a class-specific linear predictor

$$\eta_{j,\mathcal{O}_i} = \boldsymbol{x}_{j,\mathcal{O}_i}' \boldsymbol{\beta}_{0i}, \quad \eta_{j,\mathcal{O}_i} = \log\left(\frac{\pi_{0j,\mathcal{O}_i}}{1 - \pi_{0j,\mathcal{O}_i}}\right), \quad j \in \mathcal{O}_i^{(n)}, \tag{1}$$

where $\boldsymbol{x}_{j,\mathcal{O}_i}$ and $\pi_{0j,\mathcal{O}_i}$ are just relabeled $\boldsymbol{x}_j$ and π_{0j} which indicate that the associated object is the j-th object in the i-th cluster $\mathcal{O}_i^{(n)}$

$(i = 1, \ldots, k_0)$. We will use this double-index notation throughout this paper. Let $\boldsymbol{\beta}_{0i} \in \mathbb{R}^p$, $i = 1, \ldots, k_0$, be k_0 unknown class-specific true parameter vectors, which are assumed to be pairwise distinct. For convenience, we have suppressed the n in $\mathcal{O}_i^{(n)}$ in (1).

However the true partition Π_{k_0} and the associated model (1) are not observable. Hence, based on the observed data values $(\boldsymbol{x}_j, y_j), j = 1, \ldots, n$, we need to estimate the number of clusters first, and then the model (1).

Consider any possible partition of these n objects: $\Pi_k^{(n)} = \{\mathcal{C}_1^{(n)}, \ldots, \mathcal{C}_k^{(n)}\}$, where $k \leq K$ is a positive integer. Then under the clusterwise logistic regression model, the log-likelihood function for the k parameter vectors $\boldsymbol{\beta}_s$ is

$$\begin{aligned}
&l(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k | Y_n, X_n) \\
&= \sum_{s=1}^{k} \sum_{j \in \mathcal{C}_s} \{\log \binom{m_{j,\mathcal{C}_s}}{m_{j,\mathcal{C}_s} y_{j,\mathcal{C}_s}} + m_{j,\mathcal{C}_s} y_{j,\mathcal{C}_s} \log \pi_{j,\mathcal{C}_s} \\
&\qquad + m_{j,\mathcal{C}_s}(1 - y_{j,\mathcal{C}_s}) \log(1 - \pi_{j,\mathcal{C}_s})\} \\
&= \sum_{s=1}^{k} \sum_{j \in \mathcal{C}_s} \log \binom{m_{j,\mathcal{C}_s}}{m_{j,\mathcal{C}_s} y_{j,\mathcal{C}_s}} - \sum_{s=1}^{k} \sum_{j \in \mathcal{C}_s} \xi(\pi_{j,\mathcal{C}_s}; y_{j,\mathcal{C}_s}, m_{j,\mathcal{C}_s}) \\
&= \sum_{s=1}^{k} \sum_{j \in \mathcal{C}_s} \log \binom{m_{j,\mathcal{C}_s}}{m_{j,\mathcal{C}_s} y_{j,\mathcal{C}_s}} - \sum_{s=1}^{k} \sum_{j \in \mathcal{C}_s} \xi(h(\boldsymbol{x}'_{j,\mathcal{C}_s} \boldsymbol{\beta}_i); y_{j,\mathcal{C}_s}, m_{j,\mathcal{C}_s}),
\end{aligned}$$

where $Y_n = (y_1, \ldots, y_n)'$, $X_n = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)'$. Again $y_{j,\mathcal{C}_s}$, $\boldsymbol{x}_{j,\mathcal{C}_s}$, $\pi_{j,\mathcal{C}_s}$ and $m_{j,\mathcal{C}_s}$ are just relabeled $y_j, \boldsymbol{x}_j, \pi_j$ and m_j $(j = 1, \ldots, n)$ to indicate the cluster to which the associated object belongs, and

$$\xi(\pi; y, m) = -my \log \pi - m(1 - y) \log(1 - \pi).$$

Note that by convention $\xi(0; y, m) = \xi(1; y, m) = 0$. The clusterwise maximum likelihood estimator (MLE) $\widehat{\boldsymbol{\beta}}_s$ based on the partition $\Pi_k^{(n)}$ is defined to be

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_s &= \arg\max_{\boldsymbol{\beta}_s} l(\boldsymbol{\beta}_s | Y_n, X_n) \\
&\equiv \arg\min_{\boldsymbol{\beta}_s} \sum_{j \in \mathcal{C}_s} \xi(h(\boldsymbol{x}'_{j,\mathcal{C}_s} \boldsymbol{\beta}_s); y_{j,\mathcal{C}_s}, m_{j,\mathcal{C}_s}), \quad s = 1, \ldots, k.
\end{aligned}$$

We then propose an information theoretic criterion for determining the number of clusters and subsequently classifying the data as follows:

Let $q(k)$ be a strictly increasing function of k, and A_n be a sequence of constants. We define

$$D_n(\Pi_k^{(n)}) \stackrel{\text{def}}{=} \sum_{s=1}^{k} \sum_{j \in \mathcal{C}_s} \xi(h(\boldsymbol{x}'_{j,\mathcal{C}_s} \widehat{\boldsymbol{\beta}}_s); y_{j,\mathcal{C}_s}, m_{j,\mathcal{C}_s}) + q(k) A_n, \tag{2}$$

and define $\widehat{k}_n$, the estimate of k_0, to satisfy the equation

$$D_n(\widehat{k}_n) = \min_{1 \le k \le M} \min_{\Pi_k^{(n)}} D_n(\Pi_k^{(n)}). \tag{3}$$

It is named Criterion LG-C, which stands for *clustering by logistic regression* in this paper. It can be seen that in (2), the first term is basically the negative maximum log-likelihood; the second term is the penalty term measuring the complexity of the underlying model. In addition, Criterion LG-C in (3) shows that we determine the optimal number of clusters and the corresponding partitioning of the data simultaneously.

3 Monte Carlo Simulation

We constructed three models in the simulation study: the two-cluster case; the three-cluster case with only one covariate; and the three-cluster case with two covariates. The parameter values used to build these models are listed in Table 1. We generate the covariates as follows: for the first two cases, the covariate x is generated from $N(0, 1)$, and the two covariates x_1, x_2 in case 3 are generated from a bivariate Normal distribution with the mean of 0, variance of 1 and the covariance being 0.3.

In this simulation study, $q(k) = 3k(p + 3)$, where p is the number of regression coefficients in the model and is a constant in our study; k is the unknown number of clusters that we are seeking, and, $A_n = A_n^{(i)}, i = 1, 2, 3, 4$, where $A_n^{(i)} = (1/\lambda)((\log n)^\lambda) - 1$, with $\lambda_1 = 1.5, \lambda_2 = 1.8, \lambda_3 = 2$ and $\lambda_4 = 2.3$.

For reducing the exhaustive computation needed by Criterion LG-C, we adopt the approach used in Shao and Wu (2005) here. The only change is that we fit a logistic regression other than a regression model within each cluster in every iteration. Then, we first do logistic regression clustering for each k (the choice of k being the same as the previous studies), and subsequently select the best k using Criterion LG-C. We run the simulation for each model 500 times and obtain the relative frequencies of selecting every k out of these 500 repetitions. The results

Table 1. Parameter values used in the simulation study of logistic regression clustering

Case	k_0	Regression coefficients	Number of observations
1	2	$\boldsymbol{\beta}_{01} = \begin{pmatrix} 1 \\ 6 \end{pmatrix}$, $\boldsymbol{\beta}_{02} = \begin{pmatrix} 1 \\ -6 \end{pmatrix}$	$n_1 = 70$, $n_2 = 50$
2	3	$\boldsymbol{\beta}_{01} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$, $\boldsymbol{\beta}_{02} = \begin{pmatrix} -2 \\ -1 \end{pmatrix}$, $\boldsymbol{\beta}_{03} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$	$n_1 = 35$, $n_2 = 35$, $n_3 = 50$
3	3	$\boldsymbol{\beta}_{01} = \begin{pmatrix} 1 \\ -1 \\ \frac{1}{2} \end{pmatrix}$, $\boldsymbol{\beta}_{02} = \begin{pmatrix} -2 \\ -1 \\ -\frac{1}{2} \end{pmatrix}$, $\boldsymbol{\beta}_{03} = \begin{pmatrix} -1 \\ 1 \\ \frac{1}{2} \end{pmatrix}$	$n_1 = 35$, $n_2 = 35$, $n_3 = 50$

are summarized in Table 2. It can be seen that Criterion LG-C does nearly perfect a job to detect the underlying number of groups for the models considered in this simulation study.

Table 2. Relative frequencies of selecting k based on 500 simulations of logistic regression clustering

Case	B1C2 ($k_0 = 2$)				B1C3 ($k_0 = 3$)				B2C3 ($k_0 = 3$)			
Model	$A_n^{(1)}$	$A_n^{(2)}$	$A_n^{(3)}$	$A_n^{(4)}$	$A_n^{(1)}$	$A_n^{(2)}$	$A_n^{(3)}$	$A_n^{(4)}$	$A_n^{(1)}$	$A_n^{(2)}$	$A_n^{(3)}$	$A_n^{(4)}$
$k = 1$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$k = 2$	1.000	1.000	1.000	1.000	0.000	0.000	0.000	0.012	0.000	0.000	0.000	0.006
$k = 3$	0.000	0.000	0.000	0.000	1.000	1.000	1.000	0.988	1.000	1.000	1.000	0.994
$k = 4$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$k = 5$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

4 Asymptotic Property of Criterion LG-C

Denote the eigenvalues of a symmetric matrix B of order p by $\lambda_1(B) \geq \ldots \geq \lambda_p(B)$. Let $\mathcal{O}_\ell = \{\ell_1, \ldots, \ell_{n_\ell}\}$ be any cluster or a subset of a cluster corresponding to the true partition $\Pi_{k_0}^{(n)}$ of $\mathcal{O}^{(n)}$, and $n_\ell = |\mathcal{O}_\ell|$. Let $X_{n_\ell} = (\boldsymbol{x}_{\ell_1,\mathcal{O}_\ell}, \ldots, \boldsymbol{x}_{\ell_{n_\ell},\mathcal{O}_\ell})'$ be the design matrix in $\mathcal{O}_\ell$. The Fisher information for the parameter $\boldsymbol{\beta}_{0\ell}$ is defined as

$$\begin{aligned}\mathcal{I}_{n_\ell}(\boldsymbol{\beta}_{0\ell}) &= -E\frac{\partial^2 l}{\partial\boldsymbol{\beta}_{0\ell}\partial\boldsymbol{\beta}_{0\ell}'} \\ &= X_{n_\ell}' M_{n_\ell} M_{\pi_\ell} X_{n_\ell},\end{aligned}$$

where

$$\begin{aligned}M_{n_\ell} &= \mathrm{diag}(m_{\ell_1,\mathcal{O}_\ell}, \ldots, m_{\ell_{n_\ell},\mathcal{O}_\ell}) \\ M_{\pi_\ell} &= \mathrm{diag}\{\pi_{0\ell_1,\mathcal{O}_\ell}(1-\pi_{0\ell_1,\mathcal{O}_\ell}), \ldots, \pi_{0\ell_{n_\ell},\mathcal{O}_\ell}(1-\pi_{0\ell_{n_\ell},\mathcal{O}_\ell})\}.\end{aligned}$$

The following assumptions are needed in the discussion on the asymptotic property of the criterion (3).

(A) For the true partition $\Pi_{k_0}^{(n)} = \{\mathcal{O}_1^{(n)}, \ldots, \mathcal{O}_{k_0}^{(n)}\}$, let $n_{0i} = |\mathcal{O}_i|$ be the number of objects in the cluster $\mathcal{O}_i^{(n)}$. Then there exists a fixed constant $a_0 > 0$ such that

$$a_0 n \leq n_{0i} \leq n, \quad \forall i = 1, \ldots, k_0. \tag{4}$$

(X1) $\lim_{n_\ell\to\infty} \lambda_\zeta\{\mathcal{I}_{n_\ell}(\boldsymbol{\beta}_{0\ell})\} = \infty,\ \zeta = 1, \ldots, p$. Also, there exists some constant $a_1 > 0$ such that $0 < \lambda_p\{\mathcal{I}_{n_\ell}(\boldsymbol{\beta}_{0\ell})\} \leq a_1\lambda_1\{\mathcal{I}_{n_\ell}(\boldsymbol{\beta}_{0\ell})\}$.

(X2) Let $\delta_{n_\ell} = \left(\max_{j\in\mathcal{O}_\ell} m_{j,\mathcal{O}_\ell}^2 \boldsymbol{x}_{j,\mathcal{O}_\ell}' \mathcal{I}_{n_\ell}(\boldsymbol{\beta}_{0\ell})^{-1} \boldsymbol{x}_{j,\mathcal{O}_\ell}\right)^{\frac{1}{2}}$, then

$$\delta_{n_\ell}(\log\log\lambda_p\{\mathcal{I}_{n_\ell}(\boldsymbol{\beta}_{0\ell})\})^{\frac{1}{2}} = o(1).$$

(X3) $a_2 n_\ell \leq \lambda_p\{\mathcal{I}_{n_\ell}(\boldsymbol{\beta}_{0\ell})\} \leq a_3 n_\ell$ holds for some positive constants a_2 and a_3.

(X4) $a_4 n_\ell \leq \lambda\{X_{n_\ell}' M_{n_\ell} X_{n_\ell}\} \leq a_5 n_\ell$ holds for some positive constants a_4 and a_5.

(X5) Let $d_0 = \frac{1}{4}\min_{1\leq i\neq\ell\leq k_0} |\boldsymbol{\beta}_{0i} - \boldsymbol{\beta}_{0\ell}|$. Also let

$$Q_{n_\ell} = \mathrm{diag}\{\upsilon_1, \ldots, \upsilon_{n_\ell}\},$$

where $\upsilon_i = m_{\ell_i,\mathcal{O}_\ell} e^{-d_0||\boldsymbol{x}_{\ell_i,\mathcal{O}_\ell}||}\pi_{0\ell_i,\mathcal{O}_\ell}(1-\pi_{0\ell_i,\mathcal{O}_\ell})$, $i = 1, \ldots, n_\ell$. Then there exists a constant $a_6 > 0$ such that $\lambda_1\{X_{n_\ell}' Q_{n_\ell} X_{n_\ell}\} \geq a_6 n_\ell$.

(Z) $n^{-1}A_n \to 0, \quad (\log\log n)^{-1}A_n \to \infty, \quad \text{as } n \to \infty.$

Remark 4.1 Assumption (A) implicitly implies that the population is comprised of k_0 sub-populations with proportions $p_1, \ldots, p_{k_0}$, where $0 < p_i \leq 1$, $i = 1, \ldots, k_0$, $\sum_{i=1}^{k_0} p_i = 1$, and $a_0 = \min_{1 \leq i \leq k_0} p_i$.

Remark 4.2 Assumptions (X1)–(X5) are essentially about the behaviour of the explanatory variables $\boldsymbol{x}$. Roughly speaking, they mean that most of the $\boldsymbol{x}$ observations should be finite and stay away from $\mathbf{0}$. In fact, as observed by Qian and Field (2002), if we assume $\boldsymbol{x}$ to be a random vector and $\boldsymbol{x} \in \mathcal{O}_i$ are i.i.d. observations within each cluster $\mathcal{O}_i$ of the true partitioning Π_{k_0}, for all $i = 1, \ldots, k_0$, then by applying the strong law of large numbers given in Chung (2001, p. 132, Theorem 5.4.1), it is easy to show that the following assumptions are sufficient for (X1) to (X5) to hold:

(S1) $P\{\boldsymbol{x}'\boldsymbol{t} \neq 0\} > 0$ for any $\boldsymbol{t} \neq 0$ in $\mathbb{R}^p$, which implies that $\mathrm{E}(\boldsymbol{x}\boldsymbol{x}')$ is positive definite.

(S2) $P\{h(\boldsymbol{x}'\boldsymbol{\beta}_{0i})(1 - h(\boldsymbol{x}'\boldsymbol{\beta}_{0i})) \neq |\boldsymbol{x}'\boldsymbol{t} \neq 0\} > 0$ for any $\boldsymbol{t} \neq 0$ in $\mathbb{R}^p$, which implies that both $\mathrm{E}(\pi_{0\mathcal{O}_i}(1-\pi_{0\mathcal{O}_i})\boldsymbol{x}\boldsymbol{x}')$ and $\mathrm{E}(e^{-d_0\|\boldsymbol{x}|}\pi_{0\mathcal{O}_i}(1-\pi_{0\mathcal{O}_i})\boldsymbol{x}\boldsymbol{x}')$ are positive definite, where $\pi_{0\mathcal{O}_i} = h(\boldsymbol{x}'\boldsymbol{\beta}_{0i})$, $\forall i = 1, \ldots, k_0$.

(S3) $\mathrm{E}\|\boldsymbol{x}\|^{2+\kappa} < \infty$ for some constant $\kappa > 0$.

(S4) $\sup_{1 \leq k \leq n} m_k < \infty$.

Since there is no essential complexity with random $\boldsymbol{x}$, we will treat the observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ as deterministic in the sequel for ease of notation throughout the rest of this paper.

Suppose that the assumptions (A), (X1)–(X5), (Z) hold, and that $\Pi_{k_0}^{(n)} = \{\mathcal{O}_1^{(n)}, \ldots, \mathcal{O}_{k_0}^{(n)}\}$ is the underlying true classification of the n objects in $\mathcal{O}^{(n)}$. Observe that the true partition $\Pi_{k_0}^{(n)}$ is a sequence of naturally nested classifications as n increases, i.e.,

$$\mathcal{O}_i^{(n)} \subseteq \mathcal{O}_i^{(n+1)}, \quad i = 1, \ldots, k_0, \quad \text{for large } n.$$

Consider any given sequence of classifications with k clusters $\Pi_k^{(n)} = \{\mathcal{C}_1^{(n)}, \ldots, \mathcal{C}_k^{(n)}\}$ of $\mathcal{O}^{(n)}$ such that

$$\mathcal{C}_s^{(n)} \subseteq \mathcal{C}_s^{(n+1)}, \quad s = 1, \ldots, k, \quad \text{for large } n,$$

when n increases. For simplicity, when no confusion appears, n will be suppressed in $\Pi_{k_0}^{(n)}$, $\Pi_k^{(n)}$, $\mathcal{O}_i^{(n)}, 1 \leq i \leq k_0$, and $\mathcal{C}_s^{(n)}, 1 \leq s \leq k$.

Case 1: When $k_0 < k \le K$, where $K < \infty$ is a fixed constant

First we have

$$\begin{aligned}
&D_n(\Pi_k) - D_n(\Pi_{k_0}) \\
&= \sum_{s=1}^{k} \sum_{j \in \mathcal{C}_s} \xi(h(\boldsymbol{x}'_{j,\mathcal{C}_s} \widehat{\boldsymbol{\beta}}_s); y_{j,\mathcal{C}_s}, m_{j,\mathcal{C}_s}) \\
&\quad - \sum_{i=1}^{k_0} \sum_{j \in \mathcal{O}_i} \xi(h(\boldsymbol{x}'_{j,\mathcal{O}_i} \widehat{\boldsymbol{\beta}}_{0i}); y_{j,\mathcal{O}_i}, m_{j,\mathcal{O}_i}) + (q(k) - q(k_0))A_n,
\end{aligned}$$

where

$$\widehat{\boldsymbol{\beta}}_s = \arg\min_{\boldsymbol{\beta}} \sum_{j \in \mathcal{C}_s} \xi(h(\boldsymbol{x}'_{j,\mathcal{C}_s} \boldsymbol{\beta}); y_{j,\mathcal{C}_s}, m_{j,\mathcal{C}_s}), \quad s = 1, \ldots, k, \tag{5}$$

$$\widehat{\boldsymbol{\beta}}_{0i} = \arg\min_{\boldsymbol{\beta}} \sum_{j \in \mathcal{O}_i} \xi(h(\boldsymbol{x}'_{j,\mathcal{O}_i} \boldsymbol{\beta}); y_{j,\mathcal{O}_i}, m_{j,\mathcal{O}_i}), \quad i = 1, \ldots, k_0. \tag{6}$$

Note that

$$\mathcal{O}^{(n)} = \bigcup_{i=1}^{k_0} \mathcal{O}_i = \bigcup_{s=1}^{k} \mathcal{C}_s = \bigcup_{s=1}^{k} \bigcup_{i=1}^{k_0} (\mathcal{C}_s \cap \mathcal{O}_i).$$

Then

$$\begin{aligned}
&D_n(\Pi_k) - D_n(\Pi_{k_0}) \\
&= \sum_{s=1}^{k} \sum_{i=1}^{k_0} \sum_{j \in \mathcal{C}_s \cap \mathcal{O}_i} \Big[\xi(h(\boldsymbol{x}'_{j,\mathcal{C}_s \cap \mathcal{O}_i} \widehat{\boldsymbol{\beta}}_s); y_{j,\mathcal{C}_s \cap \mathcal{O}_i}, m_{j,\mathcal{C}_s \cap \mathcal{O}_i}) \\
&\quad - \xi(h(\boldsymbol{x}'_{j,\mathcal{C}_s \cap \mathcal{O}_i} \widehat{\boldsymbol{\beta}}_{0i}); y_{j,\mathcal{C}_s \cap \mathcal{O}_i}, m_{j,\mathcal{C}_s \cap \mathcal{O}_i}) \Big] + (q(k) - q(k_0))A_n \\
&= \sum_{s=1}^{k} \sum_{i=1}^{k_0} \sum_{j \in \mathcal{C}_s \cap \mathcal{O}_i} \Big[\xi(h(\boldsymbol{x}'_{j,\mathcal{C}_s \cap \mathcal{O}_i} \widehat{\boldsymbol{\beta}}_s); y_{j,\mathcal{C}_s \cap \mathcal{O}_i}, m_{j,\mathcal{C}_s \cap \mathcal{O}_i}) \\
&\quad - \xi(h(\boldsymbol{x}'_{j,\mathcal{C}_s \cap \mathcal{O}_i} \widehat{\boldsymbol{\beta}}_{0si}); y_{j,\mathcal{C}_s \cap \mathcal{O}_i}, m_{j,\mathcal{C}_s \cap \mathcal{O}_i}) \Big] \\
&\quad + \sum_{s=1}^{k} \sum_{i=1}^{k_0} \sum_{j \in \mathcal{C}_s \cap \mathcal{O}_i} \Big[\xi(h(\boldsymbol{x}'_{j,\mathcal{C}_s \cap \mathcal{O}_i} \widehat{\boldsymbol{\beta}}_{0si}); y_{j,\mathcal{C}_s \cap \mathcal{O}_i}, m_{j,\mathcal{C}_s \cap \mathcal{O}_i}) \\
&\quad - \xi(h(\boldsymbol{x}'_{j,\mathcal{C}_s \cap \mathcal{O}_i} \widehat{\boldsymbol{\beta}}_{0i}); y_{j,\mathcal{C}_s \cap \mathcal{O}_i}, m_{j,\mathcal{C}_s \cap \mathcal{O}_i}) \Big] + (q(k) - q(k_0))A_n,
\end{aligned}$$

where $\widehat{\boldsymbol{\beta}}_{0si}$ is the MLE of $\boldsymbol{\beta}$ defined by

$$\widehat{\boldsymbol{\beta}}_{0si} = \arg\min_{\boldsymbol{\beta}} \sum_{j \in \mathcal{C}_s \cap \mathcal{O}_i} \xi(h(\boldsymbol{x}'_{j,\mathcal{C}_s \cap \mathcal{O}_i}\boldsymbol{\beta}); y_{j,\mathcal{C}_s \cap \mathcal{O}_i}, m_{j,\mathcal{C}_s \cap \mathcal{O}_i}). \quad (7)$$

By (5) and (7), we have

$$\sum_{j \in \mathcal{C}_s \cap \mathcal{O}_i} \Big[\xi(h(\boldsymbol{x}'_{j,\mathcal{C}_s \cap \mathcal{O}_i}\widehat{\boldsymbol{\beta}}_s); y_{j,\mathcal{C}_s \cap \mathcal{O}_i}, m_{j,\mathcal{C}_s \cap \mathcal{O}_i}) - \xi(h(\boldsymbol{x}'_{j,\mathcal{C}_s \cap \mathcal{O}_i}\widehat{\boldsymbol{\beta}}_{0si}); y_{j,\mathcal{C}_s \cap \mathcal{O}_i}, m_{j,\mathcal{C}_s \cap \mathcal{O}_i})\Big] \geq 0.$$

By Assumptions (X1)-(X4), (25) in Lemma 3, (6), (7) and again the fact that $\mathcal{C}_s \cap \mathcal{O}_i$ is a subset of the cluster $\mathcal{O}_i$ corresponding to the true partition Π_{k_0}, we have

$$\sum_{s=1}^{k}\sum_{i=1}^{k_0}\sum_{j \in \mathcal{C}_s \cap \mathcal{O}_i} \Big[\xi(h(\boldsymbol{x}'_{j,\mathcal{C}_s \cap \mathcal{O}_i}\widehat{\boldsymbol{\beta}}_{0si}); y_{j,\mathcal{C}_s \cap \mathcal{O}_i}, m_{j,\mathcal{C}_s \cap \mathcal{O}_i}) - \xi(h(\boldsymbol{x}'_{j,\mathcal{C}_s \cap \mathcal{O}_i}\boldsymbol{\beta}_{0i}); y_{j,\mathcal{C}_s \cap \mathcal{O}_i}, m_{j,\mathcal{C}_s \cap \mathcal{O}_i})\Big] = O(\log\log n),$$

and

$$\sum_{i=1}^{k_0}\sum_{j \in \mathcal{O}_i} \Big[\xi(h(\boldsymbol{x}'_{j,\mathcal{O}_i}\widehat{\boldsymbol{\beta}}_{0i}); y_{j,\mathcal{O}_i}, m_{j,\mathcal{O}_i}) - \xi(h(\boldsymbol{x}'_{j,\mathcal{O}_i}\boldsymbol{\beta}_{0i}); y_{j,\mathcal{O}_i}, m_{j,\mathcal{O}_i})\Big] = O(\log\log n).$$

Using the fact that

$$\sum_{s=1}^{k}\sum_{i=1}^{k_0}\sum_{j \in \mathcal{C}_s \cap \mathcal{O}_i} \xi(h(\boldsymbol{x}'_{j,\mathcal{C}_s \cap \mathcal{O}_i}\boldsymbol{\beta}_{0i}); y_{j,\mathcal{C}_s \cap \mathcal{O}_i}, m_{j,\mathcal{C}_s \cap \mathcal{O}_i}) \equiv \sum_{i=1}^{k_0}\sum_{j \in \mathcal{O}_i} \xi(h(\boldsymbol{x}'_{j,\mathcal{O}_i}\boldsymbol{\beta}_{0i}); y_{j,\mathcal{O}_i}, m_{j,\mathcal{O}_i}),$$

where

$$\xi(h(\boldsymbol{x}'_{j,\mathcal{O}_i}\boldsymbol{\beta}_{0i}); y_{j,\mathcal{O}_i}, m_{j,\mathcal{O}_i}) = -m_{j,\mathcal{O}_i} y_{j,\mathcal{O}_i} \log \pi_{0j,\mathcal{O}_i} - m_{j,\mathcal{O}_i}(1 - y_{j,\mathcal{O}_i}) \log(1 - \pi_{0j,\mathcal{O}_i}), \quad (8)$$

and $\xi(h(\boldsymbol{x}'_{j,\mathcal{C}_s \cap \mathcal{O}_i}\boldsymbol{\beta}_{0i}); y_{j,\mathcal{C}_s \cap \mathcal{O}_i}, m_{j,\mathcal{C}_s \cap \mathcal{O}_i})$ is similarly defined, we obtain

$$
\begin{aligned}
&\sum_{s=1}^{k}\sum_{i=1}^{k_0}\sum_{j\in\mathcal{C}_s\cap\mathcal{O}_i}\Big[\xi(h(\boldsymbol{x}'_{j,\mathcal{C}_s\cap\mathcal{O}_i}\widehat{\boldsymbol{\beta}}_{0si});y_{j,\mathcal{C}_s\cap\mathcal{O}_i},m_{j,\mathcal{C}_s\cap\mathcal{O}_i})\\
&\quad-\xi(h(\boldsymbol{x}'_{j,\mathcal{C}_s\cap\mathcal{O}_i}\widehat{\boldsymbol{\beta}}_{0i});y_{j,\mathcal{C}_s\cap\mathcal{O}_i},m_{j,\mathcal{C}_s\cap\mathcal{O}_i})\Big]\\
&=\sum_{s=1}^{k}\sum_{i=1}^{k_0}\sum_{j\in\mathcal{C}_s\cap\mathcal{O}_i}\Big[\xi(h(\boldsymbol{x}'_{j,\mathcal{C}_s\cap\mathcal{O}_i}\widehat{\boldsymbol{\beta}}_{0si});y_{j,\mathcal{C}_s\cap\mathcal{O}_i},m_{j,\mathcal{C}_s\cap\mathcal{O}_i})\\
&\quad-\xi(h(\boldsymbol{x}'_{j,\mathcal{C}_s\cap\mathcal{O}_i}\boldsymbol{\beta}_{0i});y_{j,\mathcal{C}_s\cap\mathcal{O}_i},m_{j,\mathcal{C}_s\cap\mathcal{O}_i})\Big]\\
&\quad-\sum_{i=1}^{k_0}\sum_{j\in\mathcal{O}_i}\Big[\xi(h(\boldsymbol{x}'_{j,\mathcal{O}_i}\widehat{\boldsymbol{\beta}}_{0i});y_{j,\mathcal{O}_i},m_{j,\mathcal{O}_i})\\
&\quad-\xi(h(\boldsymbol{x}'_{j,\mathcal{O}_i}\boldsymbol{\beta}_{0i});y_{j,\mathcal{O}_i},m_{j,\mathcal{O}_i})\Big]\\
&=O(\log\log n). \qquad (9)
\end{aligned}
$$

Hence by (8), (9) and Assumption (Z) and the fact that $q(k)-q(k_0)>0$, we have that for large n,

$$D_n(\Pi_k)-D_n(\Pi_{k_0})\geq O(\log\log n)+(q(k)-q(k_0))A_n>0. \qquad (10)$$

Case 2: When $k<k_0$

By Lemma 1, for any partition $\Pi_k^{(n)}=\{\mathcal{C}_1^{(n)},\ldots,\mathcal{C}_k^{(n)}\}$, there exist one cluster in $\Pi_k^{(n)}$ and two distinct clusters in the true partition $\Pi_{k_0}^{(n)}$, say $\mathcal{C}_1\in\Pi_k^{(n)}$ and $\mathcal{O}_1,\mathcal{O}_2\in\Pi_{k_0}^{(n)}$, such that

$$b_0n<|\mathcal{C}_1\cap\mathcal{O}_1|<n \quad\text{and}\quad b_0n<|\mathcal{C}_1\cap\mathcal{O}_2|<n, \qquad (11)$$

where $b_0=a_0/k_0>0$ is a constant.

Consider

$$\sum_{j\in\mathcal{C}_1\cap\mathcal{O}_1}\xi(h(\boldsymbol{x}_{j,\mathcal{C}_1\cap\mathcal{O}_1})'\widehat{\boldsymbol{\beta}}_1;y_{j,\mathcal{C}_1\cap\mathcal{O}_1},m_{j,\mathcal{C}_1\cap\mathcal{O}_1})$$

and

$$\sum_{j\in\mathcal{C}_1\cap\mathcal{O}_2}\xi(h(\boldsymbol{x}_{j,\mathcal{C}_1\cap\mathcal{O}_2})'\widehat{\boldsymbol{\beta}}_1;y_{j,\mathcal{C}_1\cap\mathcal{O}_1},m_{j,\mathcal{C}_1\cap\mathcal{O}_2}),$$

where $\widehat{\boldsymbol{\beta}}_1$ is defined in (5) with $s=1$. Then in view of the convexity of $\xi(\cdot)$ and (5), (11) and the fact that $\boldsymbol{\beta}_{01},\boldsymbol{\beta}_{02}$ are two distinct true parameter vectors, at least one of the below two inequalities hold:

$$\begin{aligned}
&\sum_{j\in\mathcal{C}_1\cap\mathcal{O}_1}\xi(h(\boldsymbol{x}_{j,\mathcal{C}_1\cap\mathcal{O}_1})'\widehat{\boldsymbol{\beta}}_1;y_{j,\mathcal{C}_1\cap\mathcal{O}_1},m_{j,\mathcal{C}_1\cap\mathcal{O}_1})\\
&>\sum_{j\in\mathcal{C}_1\cap\mathcal{O}_1}\xi(h(\boldsymbol{x}_{j,\mathcal{C}_1\cap\mathcal{O}_1})'\boldsymbol{\beta};y_{j,\mathcal{C}_1\cap\mathcal{O}_1},m_{j,\mathcal{C}_1\cap\mathcal{O}_1}),\quad\forall\boldsymbol{\beta}:|\boldsymbol{\beta}-\boldsymbol{\beta}_{01}|\le d_0,\\
&\sum_{j\in\mathcal{C}_1\cap\mathcal{O}_2}\xi(h(\boldsymbol{x}_{j,\mathcal{C}_1\cap\mathcal{O}_2})'\widehat{\boldsymbol{\beta}}_1;y_{j,\mathcal{C}_1\cap\mathcal{O}_2},m_{j,\mathcal{C}_1\cap\mathcal{O}_2})\\
&>\sum_{j\in\mathcal{C}_1\cap\mathcal{O}_2}\xi(h(\boldsymbol{x}_{j,\mathcal{C}_1\cap\mathcal{O}_2})'\boldsymbol{\beta};y_{j,\mathcal{C}_1\cap\mathcal{O}_2},m_{j,\mathcal{C}_1\cap\mathcal{O}_2}),\quad\forall\boldsymbol{\beta}:|\boldsymbol{\beta}-\boldsymbol{\beta}_{02}|\le d_0,
\end{aligned}\tag{12}$$

where d_0 is defined in Assumption (X5). Without loss of generality, we assume that (12) holds. Now let us focus our discussion on the set $\mathcal{C}_1\cap\mathcal{O}_1$ first. Let $n_{11}=|\mathcal{C}_1\cap\mathcal{O}_1|$. We want to find out the order of

$$\begin{aligned}
\textstyle\sum_{j\in\mathcal{C}_1\cap\mathcal{O}_1}&\Big[\xi(h(\boldsymbol{x}'_{j,\mathcal{C}_1\cap\mathcal{O}_1}\widehat{\boldsymbol{\beta}}_1);y_{j,\mathcal{C}_1\cap\mathcal{O}_1},m_{j,\mathcal{C}_1\cap\mathcal{O}_1})\\
&-\xi(h(\boldsymbol{x}'_{j,\mathcal{C}_1\cap\mathcal{O}_1}\widehat{\boldsymbol{\beta}}_{011});y_{j,\mathcal{C}_1\cap\mathcal{O}_1},m_{j,\mathcal{C}_1\cap\mathcal{O}_1})\Big]\overset{\text{def}}{=}T
\end{aligned}$$

as n increases to infinity, where $\widehat{\boldsymbol{\beta}}_{011}$ is defined in (7). For simplicity, we will use single indices exclusively for observations in the set $\mathcal{C}_1\cap\mathcal{O}_1$, i.e., $\boldsymbol{x}_j$, y_j, m_j and π_{0j} will respectively represent $\boldsymbol{x}_{j,\mathcal{C}_1\cap\mathcal{O}_1}$, $y_{j,\mathcal{C}_1\cap\mathcal{O}_1}$, $m_{j,\mathcal{C}_1\cap\mathcal{O}_1}$ and $\pi_{0j,\mathcal{C}_1\cap\mathcal{O}_1}$ until the equation (18).

First note that

$$\begin{aligned}
T&=\sum_{j\in\mathcal{C}_1\cap\mathcal{O}_1}\Big[\xi(h(\boldsymbol{x}'_j\widehat{\boldsymbol{\beta}}_1);y_j,m_j)-\xi(h(\boldsymbol{x}'_j\widehat{\boldsymbol{\beta}}_{011});y_j,m_j)\Big]\\
&=\sum_{j\in\mathcal{C}_1\cap\mathcal{O}_1}\Big[\xi(h(\boldsymbol{x}'_j\widehat{\boldsymbol{\beta}}_1);y_j,m_j)-\xi(h(\boldsymbol{x}'_j\boldsymbol{\beta}_{01});y_j,m_j)\Big]\\
&\quad-\sum_{j\in\mathcal{C}_1\cap\mathcal{O}_1}\Big[\xi(h(\boldsymbol{x}'_j\widehat{\boldsymbol{\beta}}_{011});y_j,m_j)-\xi(h(\boldsymbol{x}'_j\boldsymbol{\beta}_{01});y_j,m_j)\Big]\\
&\overset{\text{def}}{=}T_1+T_2.
\end{aligned}$$

By Lemma 3 and (7), we have that for large n,

$$\begin{aligned}
T_2&=\sum_{j\in\mathcal{C}_1\cap\mathcal{O}_1}\Big[\xi(h(\boldsymbol{x}'_j\widehat{\boldsymbol{\beta}}_{011});y_j,m_j)-\xi(h(\boldsymbol{x}'_j\boldsymbol{\beta}_{01});y_j,m_j)\Big]\\
&=\log\log n_{11}=o(n_{11}).
\end{aligned}\tag{13}$$

Now let us consider the order of T_1. For any $\boldsymbol{\beta}$, define

$$H(\boldsymbol{\beta}) = \sum_{j \in \mathcal{C}_1 \cap \mathcal{O}_1} \left\{ \xi(h(\boldsymbol{x}_j'\boldsymbol{\beta}); y_j, m_j) - \xi(h(\boldsymbol{x}_j'\boldsymbol{\beta}_{01}); y_j, m_j) \right\}.$$

From the definitions of $\xi(\pi; y, m)$ and $w(u, v)$, it follows that

$$\begin{aligned} H(\boldsymbol{\beta}) &= \sum_{j \in \mathcal{C}_1 \cap \mathcal{O}_1} \left\{ -m_j y_j \boldsymbol{x}_j'(\boldsymbol{\beta} - \boldsymbol{\beta}_{01}) - m_j \log \frac{1 - h(\boldsymbol{x}_j'\boldsymbol{\beta})}{1 - h(\boldsymbol{x}_j'\boldsymbol{\beta}_{01})} \right\} \\ &= - \sum_{j \in \mathcal{C}_1 \cap \mathcal{O}_1} m_j (y_j - \pi_{0j}) \boldsymbol{x}_j'(\boldsymbol{\beta} - \boldsymbol{\beta}_{01}) \\ &\quad + \sum_{j \in \mathcal{C}_1 \cap \mathcal{O}_1} m_j w(\boldsymbol{x}_j'\boldsymbol{\beta}, \boldsymbol{x}_j'\boldsymbol{\beta}_{01}) \stackrel{\text{def}}{=} H_1(\boldsymbol{\beta}) + H_2(\boldsymbol{\beta}). \end{aligned} \tag{14}$$

Let $A_0 = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_{01}\| \le d_0\}$. Then by Lemma 3 it can be shown that

$$\begin{aligned} \inf_{\boldsymbol{\beta} \in \partial A_0} H_1(\boldsymbol{\beta}) &= O(\sqrt{n_{11} \log\log n_{11}}) \inf_{\boldsymbol{\beta} \in \partial A_0} \|\boldsymbol{\beta} - \boldsymbol{\beta}_{01}\| \\ &= O(\sqrt{n_{11} \log\log n_{11}}) \quad \text{a.s.} \end{aligned} \tag{15}$$

By (23) of Lemma 2 and Assumption (X5), we derive that

$$\begin{aligned} &\inf_{\boldsymbol{\beta} \in \partial A_0} H_2(\boldsymbol{\beta}) \\ &\ge \inf_{\boldsymbol{\beta} \in \partial A_0} \frac{1}{4} \sum_{j \in \mathcal{C}_1 \cap \mathcal{O}_1} m_j e^{-|\boldsymbol{x}_j'(\boldsymbol{\beta} - \boldsymbol{\beta}_{01})|} h(\boldsymbol{x}_j'\boldsymbol{\beta}_{01})(1 - h(\boldsymbol{x}_j'\boldsymbol{\beta}_{01})) \\ &\quad \times (\boldsymbol{x}_j'\boldsymbol{\beta} - \boldsymbol{x}_j'\boldsymbol{\beta}_{01})^2 \\ &= \frac{1}{4} \inf_{\boldsymbol{\beta} \in \partial A_0} (\boldsymbol{\beta} - \boldsymbol{\beta}_{01})' X_{\mathcal{C}_1 \cap \mathcal{O}_1}' Q_{n_{11}} X_{\mathcal{C}_1 \cap \mathcal{O}_1} (\boldsymbol{\beta} - \boldsymbol{\beta}_{01}) \\ &\ge \frac{1}{4} a_6 n_{11} \inf_{\boldsymbol{\beta} \in \partial A_0} \|\boldsymbol{\beta} - \boldsymbol{\beta}_{01}\| = \frac{1}{4} d_0 a_6 n_{11}. \end{aligned} \tag{16}$$

From (14), (15) and (16) it follows that there exists a constant $\tau > 0$ such that for large n,

$$\inf_{\boldsymbol{\beta} \in \partial A_0} H(\boldsymbol{\beta}) \ge \tau n_{11}. \tag{17}$$

By (12) and (17), we have that

$$\begin{aligned} T_1 &= \sum_{j \in \mathcal{C}_1 \cap \mathcal{O}_1} \left[\xi(h(\boldsymbol{x}_j'\widehat{\boldsymbol{\beta}}_1); y_j, m_j) - \xi(h(\boldsymbol{x}_j'\boldsymbol{\beta}_{01}); y_j, m_j) \right] \\ &\ge \inf_{\boldsymbol{\beta} \in \partial A_0} H(\boldsymbol{\beta}) \ge \tau n_{11}. \end{aligned} \tag{18}$$

Hence by combining results from (13) and (18), we have

$$\sum_{j\in\mathcal{C}_1\cap\mathcal{O}_1}\Big[\xi(h(\boldsymbol{x}'_{j,\mathcal{C}_1\cap\mathcal{O}_1}\widehat{\boldsymbol{\beta}}_1);y_{j,\mathcal{C}_1\cap\mathcal{O}_1},m_{j,\mathcal{C}_1\cap\mathcal{O}_1}) - \xi(h(\boldsymbol{x}'_{j,\mathcal{C}_1\cap\mathcal{O}_1}\widehat{\boldsymbol{\beta}}_{011});y_{j,\mathcal{C}_1\cap\mathcal{O}_1},m_{j,\mathcal{C}_1\cap\mathcal{O}_1})\Big] \geq \tau n_{11}. \tag{19}$$

Note that $D_n(\Pi_k) - D_n(\Pi_{k_0})$ can be partitioned as follows:

$$\begin{aligned}
&D_n(\Pi_k) - D_n(\Pi_{k_0})\\
&= \sum_{s=1}^{k}\sum_{j\in\mathcal{C}_s}\xi(h(\boldsymbol{x}'_{j,\mathcal{C}_s}\widehat{\boldsymbol{\beta}}_s);y_{j,\mathcal{C}_s},m_{j,\mathcal{C}_s})\\
&\quad -\sum_{i=1}^{k_0}\sum_{j\in\mathcal{O}_i}\xi(h(\boldsymbol{x}'_{j,\mathcal{O}_i}\widehat{\boldsymbol{\beta}}_{0i});y_{j,\mathcal{O}_i},m_{j,\mathcal{O}_i}) + (q(k)-q(k_0))A_n\\
&= \sum_{j\in\mathcal{C}_1\cap\mathcal{O}_1}\Big[\xi(h(\boldsymbol{x}'_{j,\mathcal{C}_1\cap\mathcal{O}_1}\widehat{\boldsymbol{\beta}}_1);y_{j,\mathcal{C}_1\cap\mathcal{O}_1},m_{j,\mathcal{C}_1\cap\mathcal{O}_1})\\
&\quad -\xi(h(\boldsymbol{x}'_{j,\mathcal{C}_1\cap\mathcal{O}_1}\widehat{\boldsymbol{\beta}}_{011});y_{j,\mathcal{C}_1\cap\mathcal{O}_1},m_{j,\mathcal{C}_1\cap\mathcal{O}_1})\Big]\\
&\quad +\sum_{\mathcal{J}_{is}}\sum_{j\in\mathcal{C}_s\cap\mathcal{O}_i}\Big[\xi(h(\boldsymbol{x}'_{j,\mathcal{C}_s\cap\mathcal{O}_i}\widehat{\boldsymbol{\beta}}_s);y_{j,\mathcal{C}_s\cap\mathcal{O}_i},m_{j,\mathcal{C}_s\cap\mathcal{O}_i})\\
&\quad -\xi(h(\boldsymbol{x}'_{j,\mathcal{C}_s\cap\mathcal{O}_i}\widehat{\boldsymbol{\beta}}_{0si});y_{j,\mathcal{C}_s\cap\mathcal{O}_i},m_{j,\mathcal{C}_s\cap\mathcal{O}_i})\Big]\\
&\quad +\sum_{s=1}^{k}\sum_{i=1}^{k_0}\sum_{j\in\mathcal{C}_s\cap\mathcal{O}_i}\Big[\xi(h(\boldsymbol{x}'_{j,\mathcal{C}_s\cap\mathcal{O}_i}\widehat{\boldsymbol{\beta}}_{0si});y_{j,\mathcal{C}_s\cap\mathcal{O}_i},m_{j,\mathcal{C}_s\cap\mathcal{O}_i})\\
&\quad -\xi(h(\boldsymbol{x}'_{j,\mathcal{C}_s\cap\mathcal{O}_i}\widehat{\boldsymbol{\beta}}_{0i});y_{j,\mathcal{C}_s\cap\mathcal{O}_i},m_{j,\mathcal{C}_s\cap\mathcal{O}_i})\Big] + (q(k)-q(k_0))A_n,
\end{aligned}$$

where $\mathcal{J}_{is} = \{i,s : i = 1,\ldots,k; s = 1,\ldots,k_0; i \text{ and } s \text{ can not be 1 simultaneously}\}$ and hence $\mathcal{J}_{is}$ corresponds to all possible intersection sets of Π_k and Π_{k_0} excluding $\mathcal{C}_1\cap\mathcal{O}_1$; $\widehat{\boldsymbol{\beta}}_i$, $\widehat{\boldsymbol{\beta}}_{0i}$ and $\widehat{\boldsymbol{\beta}}_{0si}$ are defined in (5), (6), and (7), respectively. By (8), we obtain

$$\sum_{\mathcal{J}_{is}}\sum_{j\in\mathcal{C}_s\cap\mathcal{O}_i}\Big[\xi(h(\boldsymbol{x}'_{j,\mathcal{C}_s\cap\mathcal{O}_i}\widehat{\boldsymbol{\beta}}_s);y_{j,\mathcal{C}_s\cap\mathcal{O}_i},m_{j,\mathcal{C}_s\cap\mathcal{O}_i}) - \xi(h(\boldsymbol{x}'_{j,\mathcal{C}_s\cap\mathcal{O}_i}\widehat{\boldsymbol{\beta}}_{0si});y_{j,\mathcal{C}_s\cap\mathcal{O}_i},m_{j,\mathcal{C}_s\cap\mathcal{O}_i})\Big] \geq 0. \tag{20}$$

By following the same line of argument as in proving (9), we can show that

$$\sum_{s=1}^{k}\sum_{i=1}^{k_0}\sum_{j\in\mathcal{C}_s\cap\mathcal{O}_i}\Big[\xi(h(\boldsymbol{x}'_{j,\mathcal{C}_s\cap\mathcal{O}_i}\widehat{\boldsymbol{\beta}}_{0si});y_{j,\mathcal{C}_s\cap\mathcal{O}_i},m_{j,\mathcal{C}_s\cap\mathcal{O}_i})$$
$$-\xi(h(\boldsymbol{x}'_{j,\mathcal{C}_s\cap\mathcal{O}_i}\widehat{\boldsymbol{\beta}}_{0i});y_{j,\mathcal{C}_s\cap\mathcal{O}_i},m_{j,\mathcal{C}_s\cap\mathcal{O}_i})\Big]=O(\log\log n)=o(n). \quad (21)$$

Hence in terms of (11), (19), (20) and (21) and Assumption (Z), we obtain that for large n,

$$D_n(\Pi_k)-D_n(\Pi_{k_0})\geq\tau b_0 n+o(n)+(q(k)-q(k_0))A_n>0. \quad (22)$$

Therefore combining the results from (10) in Case 1 and (22) in Case 2, we have showed that the true classification is preferable when n increases to infinity.

Appendix

Lemma 1. *Suppose that Assumption (A) holds, for any possible partition $\Pi_k^{(n)}$ of $\mathcal{O}^{(n)}$, if $k<k_0$, where k is the number of clusters for $\Pi_k^{(n)}$ and k_0 is the true number of clusters in $\mathcal{O}^{(n)}$, there exist $\mathcal{C}_s\in\Pi_k^{(n)}$ and $\mathcal{O}_i,\mathcal{O}_l\in\Pi_{k_0}^{(n)}$ such that*

$$|\mathcal{C}_s\cap\mathcal{O}_i|>b_0 n \quad \textit{and} \quad |\mathcal{C}_s\cap\mathcal{O}_l|>b_0 n,$$

where $b_0=a_0/k_0>0$ is a fixed constant.

The proof can be found in Shao and Wu (2005).

Lemma 2. *Define $w(u,v)=-\log(1-h(u))/(1-h(v))-h(v)(u-v)$, where $h(u)=e^u/(1+e^u)$. Then $w(u,v)$ is strictly convex with respect to u. Further, we have*

$$w(u,v)\geq\frac{1}{4}e^{-\zeta}h(v)(1-h(v))(u-v)^2 \quad \textit{if} \quad |u-v|\leq\zeta,\ \forall\zeta>0. \quad (23)$$

The proof can be found in Qian and Field (2002).

Lemma 3. *Suppose that Assumptions (X1)–(X4) hold. Then we have that for large n,*

$$\left.\frac{\partial l}{\partial\boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_{0\ell}}=\sum_{j\in\mathcal{O}_\ell}m_{j,\mathcal{O}_\ell}(y_{j,\mathcal{O}_\ell}-\pi_{0j,\mathcal{O}_\ell})\boldsymbol{x}_{j,\mathcal{O}_\ell}$$
$$=X'_{n_\ell}M_{n_\ell}(Y_{n_\ell}-\Pi_{0n_\ell})=O(\sqrt{n_\ell\log\log n_\ell}), \quad (24)$$

and

$$0 \leq \sum_{j \in \mathcal{O}_\ell} \{\xi(h(\boldsymbol{x}'_{j,\mathcal{O}_\ell}\widehat{\boldsymbol{\beta}}_{n_\ell}); y_{j,\mathcal{O}_\ell}, m_{j,\mathcal{O}_\ell}) - \xi(h(\boldsymbol{x}'_{j,\mathcal{O}_\ell}\boldsymbol{\beta}_{0\ell}); y_{j,\mathcal{O}_\ell}, m_{j,\mathcal{O}_\ell})\}$$
$$= O(\log\log n_\ell), \tag{25}$$

where $Y_{n_\ell} = (y_{\ell_1}, \ldots, y_{\ell_{n_\ell}})'$ and $\Pi_{0n_\ell} = \mathrm{diag}\{\pi_{\ell_1}, \ldots, \pi_{\ell_{n_\ell}}\}$.
See Qian and Field (2002) for the proof. In fact, (24) and (25) are respectively the results of Lemma 2 and Theorem 2 in that paper.

Acknowledgement

The research was partially supported by the Natural Sciences and Engineering Research Council of Canada.

References

Chung KL (2001) A Course in Probability Theory (3rd edition). Academic Press

Farewell BT, Sprott D (1988) The use of a mixture model in the analysis of count data. Biometrics 44:1191–1194

Follmann DA, Lambert D (1989) Generalizing logistic regression by nonparametric mixing. Journal of the American Statistical Association 84:295–300

Follmann DA, Lambert D (1991) Identifiability for nonparametric mixtures of logistic regressions. Journal of Statistical Planning and Inference 27:375–381

McCullagh P, Nelder JA (1989) Generalized Linear Models (2nd edition). Chapman and Hall

Qian G, Field C (2002) Law of iterated logarithm and consistent model selection criterion in logistic regression. Statistics & Probability Letters 56:101–112

Shao Q, Wu Y (2005) A consistent procedure for determining the number of clusters in regression clustering. Journal of Statistical Planning and Inference 135:461–476

Wedel M, DeSarbo WS (1995) A mixture likelihood approach for generalized linear models. Journal of Classification 12:21–55