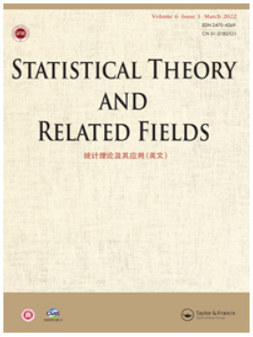


Statistical Theory and Related Fields

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20

Optimal model averaging estimator for multinomial logit models

Rongjie Jiang, Liming Wang & Yang Bai

To cite this article: Rongjie Jiang, Liming Wang & Yang Bai (2022): Optimal model averaging estimator for multinomial logit models, Statistical Theory and Related Fields, DOI: 10.1080/24754269.2022.2037204

To link to this article: https://doi.org/10.1080/24754269.2022.2037204

Published online: 17 Feb 2022.

STATISTICAL THEORY AND RELATED FIELDS
https://doi.org/10.1080/24754269.2022.2037204

Taylor & Francis
Taylor & Francis Group

Optimal model averaging estimator for multinomial logit models

Rongjie Jiang, Liming Wang and Yang Bai

School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, People's Republic of China

ABSTRACT

In this paper, we study optimal model averaging estimators of regression coefficients in a multinomial logit model, which is commonly used in many scientific fields. A Kullback–Leibler (KL) loss-based weight choice criterion is developed to determine averaging weights. Under some regularity conditions, we prove that the resulting model averaging estimators are asymptotically optimal. When the true model is one of the candidate models, the averaged estimators are consistent. Simulation studies suggest the superiority of the proposed method over commonly used model selection criterions, model averaging methods, as well as some other related methods in terms of the KL loss and mean squared forecast error. Finally, the website phishing data is used to illustrate the proposed method.

ARTICLE HISTORY
Received 29 March 2020
Revised 19 October 2021
Accepted 10 January 2022

KEYWORDS
Model averaging; multinomial logit model; Kullback–Leibler loss; asymptotically optimal

1. Introduction

Model selection is a traditional data analysis methodology. By minimizing a model selection criterion, such as Akaike information criterion (AIC) (Akaike, 1973), Bayesian information criterion (BIC) (Schwarz, 1978) and Mallow's C_p (Mallows, 1973), one model can be chosen from a number of candidate models. After that, one can make statistical inferences under the selected model. In this progress, we ignore the additional uncertainty or even bias introduced by the model selection produce, and thus often underreport the variance of inferences, as discussed in H. Wang et al. (2009). Instead of focusing on one model, the model averaging approach considers a series of candidate models and gives higher weights to the better models. It is an integrated progress that avoids ignoring the uncertainty introduced by the model selection procedure and reduces the risk in regression estimations.

Model averaging can be classified as Bayesian model averaging (BMA) and Frequentist model averaging (FMA). Compared with the FMA approach, there has been an enormous literatures about the BMA method, See Hoeting et al. (1999) for a comprehensive review. Unlike the BMA approach which considers the model uncertainty by giving a prior probability to each candidate model, FMA approach does not require priors and the corresponding estimators are totally determined by the data itself. Therefore, the current studies pay more attention to the FMA approach in statistics and econometrics.

In recent years, optimal model averaging methods have received a substantial amount of interests. Hansen (2007) proposed a Mallows model averaging (MMA) method for linear regression models with independent and homoscedastic errors. He developed its asymptotic optimality for a class of nested models by constraining the model weights in a special discrete set. A. T. Wan et al. (2010) provided a more flexible theoretical framework for the MMA which kept its asymptotic optimality for continuous weights and non-nested models. Hansen and Racine (2012), Liu and Okui (2013) developed a jackknife model averaging (JMA) method and heteroscedasticity-robust C_p model averaging (HRC_p) for the linear regression with independent and heteroscedastic errors, respectively. Zhang et al. (2013) broadened the JMA to the linear regression with dependent errors. Cheng et al. (2015) provided a feasible autocovariance-corrected MMA method to select weights across generalized least squares for the linear regression with time series errors. Zhu et al. (2018) proposed the MMA for multivariate multiple regression models.

Hansen's approach and the subsequent extensions listed above mainly focus on linear models. Recently, some optimal model averaging literatures for nonlinear models have also been developed, including optimal model averaging criterion for partially linear models (Zhang & Wang, 2019), quantile regressions (Lu & Su, 2015), generalized linear models and generalized linear mixed-effects models (Zhang et al., 2016), varying coefficient models (Li et al., 2018), varying-coefficients partially linear models (Zhu et al., 2019), and spatial autoregressive models (Zhang & Yu, 2018), and among others. All of these

CONTACT Yang Bai statbyang@mail.shufe.edu.cn

This article has been republished with minor changes. These changes do not impact the academic content of the article.

methods are asymptotically optimal in the sense of achieving the lowest loss in the large sample case. To the best of our knowledge, there are few optimal averaging estimations for a multinomial logit model that allows all candidate models to be possibly misspecified. The main contribution of this paper is to fill this gap.

The multinomial logit model is widely used in marketing research (Guadagni & Little, 1983), risk analysis (Bayaga, 2010), credit ratings (Ederington, 1985) and other fields including categorical data. A. T. Wan et al. (2014) developed the 'approximately optimal' (A-opt) method for the multinomial logit model under a local misspecification model assumption but did not establish asymptotic optimality of the resulting model averaging estimator. Besides, there have been many debates concerning the realism of the local misspecification assumption, e.g., Raftery and Zheng (2003). After that, S. Zhao et al. (2019) proposed M-fold cross-validation (MCV) criterion for the multinomial logit model and yielded forecasting superior to the strategy proposed by A. T. Wan et al. (2014). Then, its asymptotic optimality is proved for the dimension of covariates being fixed.

These two papers for multinomial logit models listed above both concerned a squared estimation error-based risk. Different from squared errors, the KL loss was produced to measure the closeness between the model and the true data generating process. Then, there are amounts of criterion are developed from the KL loss, such as Generalised information criteria (GIC) (Konishi & Kitagawa, 1996), Kullback–Leibler information criteria (KIC) (Cavanaugh, 1999) and an improved version of a criterion based on the Akaike information criterion (AIC_c) (Hurvich et al., 1998). In addition, Zhang et al. (2015) clarified that the model averaging methods based on the KL loss yield better forecasts than these model averaging approaches in terms of squared errors under linear regressions. Motivated by these facts, to propose a novel model averaging method based on the KL loss seems to be potentially interesting. Our simulation study demonstrates the model averaging method based on the KL loss has strong competitive advantages than the model averaging strategy by considering the squared estimation error for the multinomial logit model.

In order to develop an optimal model averaging method for the multinomial logit model, the weights are obtained through minimizing the KL loss. That is, we use a plug-in estimator of the KL loss plus a penalty term as the weight choice criterion, which is equivalent to penalizing the negative log-likelihood. It is interesting to note that this criterion reduces to the Mahalanobis Mallows criterion of Zhu et al. (2018) where they assume that the distribution of multiple responses is multivariate normal. The asymptotic optimality based on the KL loss of the proposed method is built on the consistency of estimators in misspecified models which is more flexible than the above-mentioned local misspecification assumption. Moreover, the asymptotic optimality will be established for the dimension of covariates being either fixed or diverging.

This article is the first study that proposes optimal model averaging estimation for multinomial logit models based on the KL loss. When the number of candidate models is small, the corresponding numerically solutions obtained are nearly instantaneous. If the number of candidate models is large, the computational burden of our model averaging procedure will be heavy. In this case, a model screening step prior to model averaging is desirable. That is, we use penalized regression with LASSO (Friedman et al., 2010) to prepare candidate models. Different tuning parameters may results in different models, which will be included in our resulting candidate models. Using the website phishing data, we demonstrate the superiority of our proposed method.

Our work is related to Zhang et al. (2016), which developed the model averaging method for univariate generalized linear models. We differ from this study by establishing the asymptotic optimality based on some original conditions, while they prove the asymptotic optimality by assuming some conclusions are valid. Moreover, we discuss the case when the true model is one of the candidate models and prove that the model averaging estimators based on our weight choice criterion are consistent.

The remainder of this article is organized as follows. In Section 2, we first describe the multinomial logit model. Then, we introduce the model averaging estimation for the multinomial logit model and propose a weight choice criterion by considering the KL loss. The asymptotic optimality of the proposed method and the estimation consistency are discussed in Sections 3 and 4, respectively. Sections 5 and 6 present the numerical results through various simulations and a real data example, respectively. Technical proofs of the main results are presented in the Appendix.

2. Model framework and weight choice

2.1. Multinomial logit model

Consider a general discrete choice model with n independent individuals and d nominal alternatives. And let $y_i = j$ means individual i selects alternative j. We use a multinomial logit regression to describe the discrete choice model (A. T. Wan et al., 2014). The corresponding assumption is that the *log* odds of category j relative to the reference category (without losing generality, we regard alternative d as reference) are determined by a linear combination of regressors. Thus, the choice probabilities for the ith individual can then be

expressed as

$$\begin{cases} f(y_i = j \mid \boldsymbol{X}_i) \\ \quad = \frac{\exp(\boldsymbol{X}_i\boldsymbol{\beta}_j)}{1+\sum_{j=1}^{d-1}\exp(\boldsymbol{X}_i\boldsymbol{\beta}_j)}, \quad \text{for } j = 1, \ldots, (d-1), \\ f(y_i = d \mid \boldsymbol{X}_i) \\ \quad = \frac{1}{1+\sum_{j=1}^{d-1}\exp(\boldsymbol{X}_i\boldsymbol{\beta}_j)}, \end{cases} \tag{1}$$

where $\boldsymbol{X}$ is an $n \times k$ covariate matrix with full column rank, $\boldsymbol{X}_i$ is constructed from the ith row of $\boldsymbol{X}$, $\boldsymbol{\beta}_j$ is an unknown parameter vector. We first assume k to be fixed and discuss the diverging situation in Section 3. Formula (1) can be rewritten as an exponential family form.

$$f(y_i \mid \boldsymbol{\theta}_i, \boldsymbol{X}_i) = \exp\left\{T(y_i)^T\boldsymbol{\theta}_i - b(\boldsymbol{\theta}_i)\right\}, \ i = 1, \ldots, n, \tag{2}$$

where $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{i(d-1)})^T$ is a parameter vector, with the canonical parameter θ_{ij} connecting the parameters $\boldsymbol{\beta}_j$ and the k-dimension covariate vector in the form $\theta_{ij} = \boldsymbol{X}_i\boldsymbol{\beta}_j$. And $\boldsymbol{\theta}_i = (I_{d-1} \otimes \boldsymbol{X}_i)\boldsymbol{\beta}$, where $\otimes$ is a Kronecker product, I_{d-1} is a $(d-1) \times (d-1)$ identity matrix, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_{d-1}^T)^T$, which is a $k(d-1) \times 1$ parameter vector. Besides, $b(\boldsymbol{\theta}_i) = \log(\sum_{j=1}^{d-1}\exp(\theta_{ij}) + 1)$ is a vector-valued function, $T(y) = (\boldsymbol{I}_{\{y=1\}}, \ldots, \boldsymbol{I}_{\{y=d-1\}})^T$, $\boldsymbol{I}_{\{\cdot\}}$ is an indicator function.

2.2. Model averaging estimator

We denote a set of S candidate models $M_1, \ldots, M_S$ by

$$\begin{aligned} M_s : & f(y_i \mid \boldsymbol{\theta}_i\{\boldsymbol{\beta}_s\}, \boldsymbol{X}_{(s),i}) \\ & = \exp\left\{T(y_i)^T(I_{d-1} \otimes \boldsymbol{X}_{(s),i})\boldsymbol{\beta}_s \right. \\ & \quad \left. - b\left((I_{d-1} \otimes \boldsymbol{X}_{(s),i})\boldsymbol{\beta}_s\right)\right\}, \end{aligned} \tag{3}$$

where S is fixed, and $\boldsymbol{X}_{(s)}$ is an $n \times k_s$ matrix containing k_s columns of $\boldsymbol{X}$ with full column rank, whose rows are $\boldsymbol{X}_{(s),1}, \ldots, \boldsymbol{X}_{(s),n}$. Under the sth candidate model, the maximum-likelihood estimator of the regression coefficients is $\hat{\boldsymbol{\beta}}_s$. And let $\hat{\boldsymbol{\beta}}_s \in R^{(d-1)\times k_s}$ be the subvector containing estimators in $\hat{\boldsymbol{\beta}}_{(s)} \in R^{(d-1)\times k}$. Note that the rest components of $\hat{\boldsymbol{\beta}}_{(s)}$ are restricted to be zeros. Let $\boldsymbol{\theta}_{0i}$ be the true value of $\boldsymbol{\theta}_i$. And $\boldsymbol{\theta}_{0i}$ is not required that there exists a $\boldsymbol{\beta}_0$ so that $\boldsymbol{\theta}_{0i} = (I_{d-1} \otimes \boldsymbol{X}_i)\boldsymbol{\beta}_0$. Thus, each of the candidate models can be misspecified. After their maximum-likelihood estimators are obtained, we need to determine the weight of each candidate model. Let $\boldsymbol{\omega} = (w_1, \ldots, w_S)$ be a weight vector in the unit simplex of $R^S : H = \{\boldsymbol{\omega} \in [0,1]^S, \sum_{s=1}^S w_s = 1\}$. Then the model averaging estimator of $\hat{\boldsymbol{\beta}}(\boldsymbol{\omega})$ is $\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}) = \sum_{s=1}^S w_s\hat{\boldsymbol{\beta}}_{(s)}$.

Let $\boldsymbol{Y} = (T(y_1), \ldots, T(y_n))^T$, $\boldsymbol{U} = E(\boldsymbol{Y})$ be $n \times (d-1)$ matrices, $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)^T$ be an $n \times (d-1)$ parameter matrix and $\boldsymbol{\Theta_0} = (\boldsymbol{\theta}_{01}, \ldots, \boldsymbol{\theta}_{0n})^T$ be the true value of $\boldsymbol{\Theta}$. We put the model estimator in vector form by using the vectoring operation $\mathrm{Vec}(\cdot)$, which creates a column vector by stacking the column vectors of below one another. Then, the model averaging estimator can be expressed as

$$\mathrm{Vec}(\boldsymbol{\Theta}^T\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega})\}) = \boldsymbol{Z}\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}), \tag{4}$$

where $\boldsymbol{Z} = ((I_{d-1} \otimes \boldsymbol{X}_1)^T, \ldots, (I_{d-1} \otimes \boldsymbol{X}_n)^T)^T$, which is an $n(d-1) \times k(d-1)$ matrix.

2.3. KL loss-based weight choice criterion

For linear models, the weight choice criterion is based on squared prediction error. In this paper, we use the KL divergence as a replacement for the squared prediction error to establish the asymptotic optimality. The KL loss of $\boldsymbol{\Theta}\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega})\}$ is

$$\begin{aligned} KL(\boldsymbol{\omega}) &= 2\sum_{i=1}^n E_{\boldsymbol{Y}^*}\{\log\{f(\boldsymbol{Y}^* \mid \boldsymbol{\Theta}_0)\} \\ &\quad - \log(f[\boldsymbol{Y}^* \mid \boldsymbol{\Theta}\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega})\}])\} \\ &= 2B\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega})\} - 2\,\mathrm{Vec}(\boldsymbol{U}^T)^T\,\mathrm{Vec}(\boldsymbol{\Theta}^T\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega})\}) \\ &\quad - 2B_0 + 2\,\mathrm{Vec}(\boldsymbol{U}^T)^T\,\mathrm{Vec}(\boldsymbol{\Theta}_0^T) \\ &= 2J(\boldsymbol{\omega}) - 2B_0 + 2\,\mathrm{Vec}(\boldsymbol{U}^T)^T\,\mathrm{Vec}(\boldsymbol{\Theta}_0^T), \end{aligned}$$

where $\boldsymbol{Y}^* = \boldsymbol{U} + \boldsymbol{\Xi}^*$ is another realization from $f(\boldsymbol{Y}^* \mid \boldsymbol{\Theta}_0)$, $\boldsymbol{\Xi}^*$ is independent of $\boldsymbol{\Xi}$, $\boldsymbol{\Xi} = \boldsymbol{Y} - \boldsymbol{U}$, $B_0 = \sum_{i=1}^n b(\boldsymbol{\theta}_{0i})$, $J(\boldsymbol{\omega}) = B\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega})\} - \mathrm{Vec}(\boldsymbol{U}^T)^T\,\mathrm{Vec}\,\boldsymbol{\Theta}^T\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega})\}$, and $B\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega})\} = \sum_{i=1}^n b(\boldsymbol{\theta}_i\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega})\})$.

Because of the relationship between $J(\boldsymbol{\omega})$ and $KL(\boldsymbol{\omega})$, we can obtain $\boldsymbol{\omega}$ to minimize $J(\boldsymbol{\omega})$ instead of $KL(\boldsymbol{\omega})$. In practice, minimizing $J(\boldsymbol{\omega})$ is infeasible owing to the value of $\boldsymbol{U}$ is unknown. A intuition idea is that we can plug $\boldsymbol{Y}$ into $J(\boldsymbol{\omega})$ instead of $\boldsymbol{U}$. That is, we can get $\boldsymbol{\omega}$ by minimizing $J^*(\boldsymbol{\omega}) = B\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega})\} - \mathrm{Vec}(\boldsymbol{Y}^T)^T\,\mathrm{Vec}(\boldsymbol{\Theta}^T\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega})\})$. But, this progress will lead to overfitting. Motivated by that $J^*(\boldsymbol{\omega})$ equals the corresponding negative log-likelihood of $\mathrm{Vec}(\boldsymbol{\Theta}^T\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega})\}) = \boldsymbol{Z}\hat{\boldsymbol{\beta}}(\boldsymbol{\omega})$. We add penalty term $\lambda_n(d-1)\boldsymbol{\omega}^T\boldsymbol{K}$ to $2J^*(\boldsymbol{\omega})$, where $\boldsymbol{K} = (k_1, \ldots, k_S)^T$, k_s is the number of columns of $\boldsymbol{X}$ used in the sth candidate model. And the weight choice criterion is introduced as

$$\begin{aligned} \wp(\boldsymbol{\omega}) &= 2B\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega})\} - 2\,\mathrm{Vec}(\boldsymbol{Y}^T)^T\,\mathrm{Vec}(\boldsymbol{\Theta}^T\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega})\}) \\ &\quad + \lambda_n(d-1)\boldsymbol{\omega}^T\boldsymbol{K}. \end{aligned}$$

The resultant weight vector is defined as $\hat{\boldsymbol{\omega}} = \arg\min_{\boldsymbol{\omega}\in H}\wp(\boldsymbol{\omega})$. Because $\wp(\boldsymbol{\omega})$ is convex in $\boldsymbol{\omega}$, the global optimization can be performed efficiently through constrained optimization programming. For example, the fmincon of MATLAB can be applied for this purpose. Note that when we restrict one element of $\boldsymbol{\omega}$ is 1 others 0, then $\wp(\boldsymbol{\omega})$ is equivalent to AIC and

BIC in the sense of choosing weights when $\lambda_n = 2$ and $\lambda_n = \log(n)$, respectively. In addition, when $\lambda_n = 2$, the criterion $\wp(\boldsymbol{\omega})$ can reduce to the Mahalanobis Mallows criterion of Zhu et al. (2018) where they assume that the distribution of multiple responses is multivariate normal.

3. Asymptotic optimality

This section presents the main theoretic results of this paper, which demonstrates the asymptotic optimality of the model averaging estimator $\boldsymbol{\Theta}\{\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}})\}$. We define the pseudo true regression parameter vector as $\boldsymbol{\beta}^*_{(s)}$ which minimizes the KL divergence between the true model and the sth candidate model. From Theorem 3.2 of White (1982), when the dimension of $\boldsymbol{\beta}^*_{(s)}$, k, is fixed, under some regularity conditions, we have

$$\left\| \hat{\boldsymbol{\beta}}_{(s)} - \boldsymbol{\beta}^*_{(s)} \right\| = O_p(n^{-1/2}). \tag{5}$$

Before we provide the relevant theorems, we first list the relevant notations in this paper. Let $\boldsymbol{\beta}^*(\boldsymbol{\omega}) = \sum_{s=1}^{S} w_s \boldsymbol{\beta}^*_{(s)}$, $KL^*(\boldsymbol{\omega}) = 2B\{\boldsymbol{\beta}^*(\boldsymbol{\omega})\} - 2\,\mathrm{Vec}(\boldsymbol{U}^T)^T \mathrm{Vec}(\boldsymbol{\Theta}^T\{\boldsymbol{\beta}^*(\boldsymbol{\omega})\}) - 2B_0 + 2\,\mathrm{Vec}(\boldsymbol{U}^T)^T \mathrm{Vec}\,(\boldsymbol{\Theta}_0^T)$, $\xi_n = \inf_{\boldsymbol{\omega}\in H} KL^*(\boldsymbol{\omega})$, $\boldsymbol{\Xi}_i$ be the ith row of $\boldsymbol{\Xi}$, $\bar{\lambda} = \max_{i\in\{1,\ldots,n\}} \lambda_{\max}\{\mathrm{Cov}(\boldsymbol{\Xi}_i)\}$, and $\underline{\lambda} = \min_{i\in\{1,\ldots,n\}} \lambda_{\min}\{\mathrm{Cov}(\boldsymbol{\Xi}_i)\}$, where $\mathrm{Cov}(\cdot)$, $\lambda_{\max}\{\cdot\}$ and $\lambda_{\min}\{\cdot\}$ denote the covariance, the maximum and minimum eigenvalues of a matrix, respectively. Note that all the limiting properties here and throughout the text hold under $n \to \infty$. The following conditions will be made:

R.1 There exist constants $\underline{C}$ and $\bar{C}$, such that $0 < \underline{C} < \lambda_{\min}\{\boldsymbol{X}^T\boldsymbol{X}/n\} < \lambda_{\max}\{\boldsymbol{X}^T\boldsymbol{X}/n\} < \bar{C}$.
R.2 $\max_{i\in\{1,\ldots,n\}} \|\boldsymbol{X}_i\|^2/n \to 0$, and there exist constants C_1 and C_2, such that $0 < C1 < \underline{\lambda} < \bar{\lambda} < C_2$.
R.3 $n\xi_n^{-2} = o(1)$.
R.4 $n^{-1/2}\lambda_n = O(1)$.

Remark 3.1: Conditions R.1 and R.2 are regular. Condition R.1 is the same as condition C.2 and the second part of condition C.3 of Zhang et al. (2020). The first part of Condition R.2 is the same as the first part of condition C.3 of Zhang et al. (2020). The second part of Condition R.2 assumes the covariance matrix of $\boldsymbol{\Xi}_i$ is positive definite. Condition R.3 requires ξ_n grows at a rate no slower than $n^{\frac{1}{2}}$. And both $\lambda_n = 2$ and $\lambda_n = \log(n)$ satisfy Condition R.4, which means that if one prefers AIC or BIC, this can be achieved by choosing $\lambda_n = 2$ or $\lambda_n = \log(n)$. Condition R.4 is also used by Theorem 2 of Zhang et al. (2020).

The following theorem illustrates the asymptotic optimality of the model averaging estimators for fixed k situation.

Theorem 3.1: *For fixed k, if Equation* (5) *and the regularity Conditions R.1–R.4 hold. Then $\hat{\boldsymbol{\omega}}$ is asymptotically optimal in the sense that*

$$\frac{KL(\hat{\boldsymbol{\omega}})}{\inf_{\boldsymbol{\omega}\in H} KL(\boldsymbol{\omega})} \to 1, \tag{6}$$

where the convergence is in probability.

Remark 3.2: Theorem 1 of S. Zhao et al. (2019) is based on the squared loss. The squared loss only concerns on the point distance. Different from them, the KL loss measures the closeness between the model and the true data generating process, which concerns on the full distribution. In addition, from $KL(\boldsymbol{\omega}) = \int_{-\infty}^{+\infty} f(Y^* \mid \boldsymbol{\Theta}_0) \log \frac{f(Y^* \mid \boldsymbol{\Theta}_0)}{f(Y^* \mid \boldsymbol{\Theta}\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega})\})}\, \mathrm{d}_{Y^*}$, we know that the KL loss pays more attention to these points with high probability. However, the squared loss considers all points are equally important.

Considering for diverging k, let $\boldsymbol{\beta}^*_s \in R^{(d-1)\times k_s}$ be the corresponding subvector of $\boldsymbol{\beta}^*_{(s)}$ and define

$$B_n(\boldsymbol{\beta}^*_s \mid \delta) = \left\{ \boldsymbol{\beta}_s \in R^{(d-1)\times k_s} : \left\| \frac{n^{1/2}}{\{(d-1)k\}^{1/2}} (\boldsymbol{\beta}_s - \boldsymbol{\beta}^*_s) \right\| \le \delta \right\}.$$

Let $b^{(2)} = \frac{\partial^2 b(\boldsymbol{x})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^T}$, $\boldsymbol{D}_s = \mathrm{diag}\{b^{(2)}[(I_{d-1} \otimes X_{(s),i})\boldsymbol{\beta}_s]\}_{i=1,\ldots,n}$ and $\boldsymbol{Z}_{(s)} = ((I_{d-1} \otimes X_{(s),1})^T, (I_{d-1} \otimes X_{(s),2})^T, \ldots, (I_{d-1} \otimes X_{(s),n})^T)^T$, which are $n(d-1) \times n(d-1)$ and $n(d-1) \times k_s(d-1)$ matrices, respectively. We list the following conditions required for the case with diverging k.

R.5 There exists a constant $C_3 > 0$ such that $\sum_{i=1}^{n} \|\boldsymbol{X}_i\|/(k^{1/2}n) \le C_3 < \infty$.
R.6 There exists a constant $C_0 > 0$ such that for any fixed $\delta > 0$, any $\boldsymbol{\beta}_s \in B_n(\boldsymbol{\beta}^*_s \mid \delta)$ and every $s = 1, \ldots, S$, the minimum eigenvalue of $\frac{1}{n}\boldsymbol{Z}^T_{(s)}\boldsymbol{D}_s\boldsymbol{Z}_{(s)}$ is bound below by C_0 for all sufficiently large n.
R.7 $k^2 n \xi_n^{-2} = o(1)$.

Remark 3.3: Condition R.5 is implied by condition A.1(iii) of Lu and Su (2015). Condition R.6 guarantees that $\|\hat{\boldsymbol{\beta}}_{(s)} - \boldsymbol{\beta}^*_{(s)}\| = O_p(n^{-1/2}k_s^{1/2})$, which is an extension of the first part of condition C.4 of Zhang et al. (2016). Condition R.7 is an extension of Condition R.3 under the diverging k situation. Condition R.7 allows k to increase with n, but restricts its rate. Obviously, as k increases, ξ_n decreases. Therefore, the smaller k is easier to satisfy Condition R.7. In practice, we can exclude redundant variables from the candidate set prior to model averaging to control k.

Theorem 3.2: *For diverging k, if Conditions R.1–R.2 and R.4–R.7 are satisfied, then* (6) *remains valid as $n \to \infty$.*

Remark 3.4: Note that both $\lambda_n = 2$ and $\lambda_n = \log(n)$ satisfy Condition R.4. In Section 4, the numerical analysis shows that both of them outperform alternative model selection methods (AIC and BIC) and model averaging methods (Smoothed AIC and Smoothed BIC), respectively. And, when the sample size is small, the optimal value of λ_n increases as the level of the model misspecification improves.

4. Estimation consistency

Here we would like to comment on the case when the true model is included in the candidate models. That is,

$$\boldsymbol{\theta}_{0i} = (I_{d-1} \otimes \boldsymbol{X}_i)\boldsymbol{\beta}_0,$$

where $\boldsymbol{\beta}_0 \in R^{(d-1)\times k}$ is the true value of $\boldsymbol{\beta}$ and the number of non-zero coefficients of $\boldsymbol{\beta}_0$ is k_{true}. Let $\boldsymbol{\omega}_{true}$ be a weight vector in which the element corresponding to the true model is 1, and all others are 0. When k is fixed, from chapter 3.4.1 of Fahrmeir and Tutz (2013), under some regularity conditions, we have

$$\left\| \hat{\boldsymbol{\beta}}(\boldsymbol{\omega}_{true}) - \boldsymbol{\beta}_0 \right\| = O_p(n^{-1/2}). \tag{7}$$

For diverging k, from Theorem 2.1 of Portnoy (1988), under some regularity conditions, we can obtain

$$\left\| \hat{\boldsymbol{\beta}}(\boldsymbol{\omega}_{true}) - \boldsymbol{\beta}_0 \right\| = O_p(k_{true}^{1/2} n^{-1/2}). \tag{8}$$

Denote $\boldsymbol{D}_i(\boldsymbol{\beta}) = b^{(2)}[(\boldsymbol{I}_{d-1} \otimes \boldsymbol{X}_i)\boldsymbol{\beta}]$. In order to prove the estimation consistency, we further impose the following condition.

R.8 There exists $\delta(r) \geq \underline{d} > 0$, such that uniformly for $\boldsymbol{\omega} \in H$ and $r \in (0, 1)$ and for almost all $i \in \{1, \ldots, n\}$,

$$\left\| \boldsymbol{D}_i^{1/2}[\boldsymbol{\beta}_0 + r(\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}) - \boldsymbol{\beta}_0)](\boldsymbol{I}_{d-1} \otimes \boldsymbol{X}_i) \times (\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}) - \boldsymbol{\beta}_0) \right\|^2 / \left\| \hat{\boldsymbol{\beta}}(\boldsymbol{\omega}) - \boldsymbol{\beta}_0 \right\|^2 > \delta(r)$$

Remark 4.1: Condition R.8 states that most $\boldsymbol{D}_i^{1/2}[\boldsymbol{\beta}_0 + r(\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}) - \boldsymbol{\beta}_0)](\boldsymbol{I}_{d-1} \otimes \boldsymbol{X}_i)s$ do not degenerate, in the sense that their inner products with $\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}) - \boldsymbol{\beta}_0$ do not approach zero. Which is implied by $\lambda_{\min}\{\text{diag}\{\boldsymbol{D}_i[\boldsymbol{\beta}_0 + r(\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}) - \boldsymbol{\beta}_0)]\}_{i=1,\ldots,n}\} > 0$, uniformly for $\boldsymbol{\omega} \in H$ and $r \in (0, 1)$, and the first part of condition R.1. Our asymptotic study mainly requires Condition R.8 so that the positive definition of $\text{diag}\{\boldsymbol{D}_i[\boldsymbol{\beta}_0 + r(\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}) - \boldsymbol{\beta}_0)]\}_{i=1,\ldots,n}$ is not necessary.

We now describe the performance of the weighted estimator when the true model is among the candidate models.

Theorem 4.1: *When k is fixed and the true model is one of the candidate models, if Conditions R.1, R.2, R.4, R.8 and Equation* (7) *are satisfied, then the weighted estimator satisfies*

$$\left\| \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}) - \boldsymbol{\beta}_0 \right\| = O_p(n^{-1/2}\lambda_n^{1/2}) = o_p(1). \tag{9}$$

Remark 4.2: Theorem 4.1 states that $\|\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}) - \boldsymbol{\beta}_0\| = o_p(1)$. I conjecture that it may be possible to extend the converge rate of the weighted estimator to $n^{1/2}$, similar to Theorem 2 of Zhang and Liu (2019).

Theorem 4.2: *For diverging k, if $k = o(n^{1/2})$, Conditions R.1, R.2, R.4, R.8 and Equation* (8) *are satisfied, then the weighted estimator satisfies*

$$\left\| \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}) - \boldsymbol{\beta}_0 \right\| = O_p(k^{1/2}n^{-1/2}\lambda_n^{1/2}) = o_p(1). \tag{10}$$

5. Monte Carlo simulations

In this section, we evaluate the empirical performance of our proposed model averaging strategy for the multinomial logit model. We use two versions of our proposed model averaging method named OPT1-KL with $\lambda_n = 2$ and OPT2-KL with $\lambda_n = \log(n)$ to compare with some alternative FMA methods and model selection strategies. Model selection methods include AIC, BIC, and LASSO proposed by Friedman et al. (2010), where the tuning parameter, $\hat{\zeta}$, is selected by cross-validation. Model averaging strategies include A-OPT (A. T. Wan et al., 2014), MCV (S. Zhao et al., 2019)), Smoothed AIC (SAIC) and Smoothed BIC (SBIC) (Buckland et al., 1997). The SAIC strategy assigns the weight

$$\exp(-\text{AIC}_s/2) \Big/ \sum_{s=1}^{S} \exp(-\text{AIC}_s/2)$$

to the sth model and SBIC allocates weights similarly.

We use the KL loss for assessment and generate 1000 simulated data. For the convenient comparison, we normalize all KL losses by dividing the KL loss corresponding to the best method. The sample size varies at 100, 200. Note that MCV and A-OPT are the model averaging methods to average estimate of the probability $y_i = j$. Which leads to computing the KL loss is infeasible. Therefore, we compare our methods with MCV and A-OPT in terms of the mean squared forecast error (MSFE) instead of the KL loss. Without loss of generality, we also normalize all MSFEs by dividing the MSFE corresponding to the best method.

Two situations of simulations are used. In the first situation, when the candidate models do not contain the true model, we examine the effect of the changing magnitude of coefficients and the changing level of the model misspecification. Moreover, we consider the case when the number of covariates is diverging with the sample size. Note that all candidate models are misspecified in this situation, so there does not exist the full

model. It implies that the A-OPT method is not feasible for this situation. In the second situation, when the candidate models include the true model, we compare our methods with other competitive methods and validate the estimation consistency.

Setting 1. We generate y_i from the setup of the multinomial logit model (2) with the following specifications: $d = 3, \boldsymbol{X}_{i1} = 1, \boldsymbol{X}_{ij}, j = 2, \ldots, 6$ follow normal distributions with mean zeros, variance ones and the correlation between different components of $\boldsymbol{X}_i$ is $\rho = 0.75$, and $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2})^T = (I_2 \otimes \boldsymbol{X}_i)(\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$, where

$$\boldsymbol{\beta}_1 = \gamma_1(1, 1, 0.2, -1.2, -0.5, 0.7)^T;$$
$$\boldsymbol{\beta}_2 = \gamma_1(0.7, 0.9, 0.3, -1.1, -0.6, 0.7)^T.$$

In order to imitate that all candidate models are misspecified, we pretend the last covariate missed. Let $\boldsymbol{X}_{i1}$ contain in all candidate models. So there are $2^4 - 1 = 15$ candidate models to combine. The parameter γ_1 is used to control the magnitude of coefficients, and we let it vary in the set $\{0.5, 1, 2\}$.

Simulation results are shown in Table 1. One remarkable aspect of the results is that OPT1-KL and OPT2-KL yield smaller mean and standard deviance (SD) values than the other four competitions (SAIC, SBIC, AIC and BIC) in different magnitudes of coefficients. In the majority of cases, FMA approaches are superior to model selection methods. This pattern appears to be more obvious when γ_1 is small than when it is large. For example, when $\gamma_1 = 0.5$, all model averaging methods deliver smaller mean values than model selection strategies. When γ_1 increases to 2, for $n = 200$, AIC has marginal advantages than SBIC regarding the mean values. This result is reasonable, because when γ_1 is small, and the non-zero coefficients in the true model are all close to zero, which makes it difficult to distinguish the non-zero coefficients from a false model that contains many zeros. As a result, model selection criterion scores can be quite similar for different candidate models and the choice of models becomes unstable. On the other hand, when the absolute values of the non-zero coefficients are large, and a model selection criterion can identify a non-zero coefficient more readily.

Table 1. Simulations results of the KL loss for setting 1.

γ_1	n		OPT1-KL	OPT2-KL	SAIC	SBIC	AIC	BIC
0.5	100	Mean	1.2854	1.0000	2.6779	2.2431	3.8806	3.0228
		SD	2.1785	1.0000	5.2774	4.5048	4.5793	4.5313
	200	Mean	1.0181	1.0000	1.8427	1.7991	2.4709	2.2732
		SD	1.6082	1.0000	3.3939	3.0055	2.9032	2.8244
1	100	Mean	1.0359	1.0000	1.9150	1.8519	2.4402	2.3318
		SD	1.5007	1.0000	3.0423	2.7406	2.7229	2.5896
	200	Mean	1.0000	1.0373	1.5212	1.6712	1.8182	1.9670
		SD	1.1837	1.0000	2.2872	2.3921	2.3530	2.4111
2	100	Mean	1.0260	1.0000	1.7058	1.7868	1.9758	2.0719
		SD	1.3210	1.0000	2.6279	2.6401	2.7112	2.6516
	200	Mean	1.0000	1.0173	1.3756	1.5121	1.5009	1.6475
		SD	1.1638	1.0000	2.3457	2.8700	2.6478	3.2289

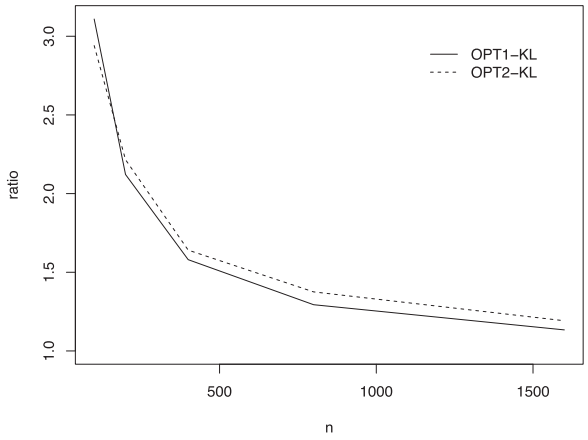


Figure 1. The means of ratio by methods of OPT1-KL and OPT2-KL with $\gamma_1 = 1$.

In addition, we calculate the means of $KL(\hat{\boldsymbol{\omega}})/inf_{\boldsymbol{\omega}\in H}KL(\boldsymbol{\omega})$(ratio) by methods of OPT1-KL and OPT2-KL with $\gamma_1 = 1$. The values of mean, presented in Figure 1, decrease and approach to 1 as the sample size n increases. This feature verifies asymptotic optimality numerically. Then, we compare our strategy with MCV in terms of MSFE. We generate 100 observations as the training sample and 10 observations as the test sample under setting 1 with $\gamma_1 = 1$. And the simulation results are based on 1000 replications.

$$\text{MSFE} = \frac{1}{10,000}\sum_{r=1}^{1000}\sum_{v=1}^{10}\sum_{j=1}^{3}(\hat{p}_{vj}^{[r]} - p_{vj}^{[r]})^2,$$

where $\hat{p}_{vj}^{[r]}$ is the forecast of $p_{vj}^{[r]}$, which represents the probability that the vth test observation selects alternative j for the rth replication.

Table 2 shows that our proposed approaches outperform other strategies. Then, SAIC and SBIC perform better than MCV. Note that MCV is the model averaging method based on the squared loss, and our strategy is based on the KL loss. It implies that the approach based on the KL loss has a strong competitive advantage than this approach based on the square loss for a multinomial logit model.

Setting 2. In order to examine the effects of the changing level of the model misspecification. We set $\theta_{i1} = \boldsymbol{U}_i\boldsymbol{\beta}_1 + \gamma_2 \exp(0.5X_{i6})$, $\theta_{i2} = \boldsymbol{U}_i\boldsymbol{\beta}_2 + \gamma_2 \exp(0.6X_{i6})$, and

$$\boldsymbol{\beta}_1 = (1, 1, 0.2, -1.2, -0.5)^T;$$
$$\boldsymbol{\beta}_2 = (0.7, 0.9, 0.3, -1.1, -0.6)^T,$$

where $\boldsymbol{U}_i = (1, X_{i2}, \ldots, X_{i5})$, $X_{i2}, \ldots, X_{i6}$ have the same specification as the previous design, γ_2 controls

Table 2. Simulation results of MSFE.

	OPT1-KL	OPT2-KL	SAIC	SBIC	AIC	BIC	MCV
Mean	1.0000	1.0195	1.0703	1.0728	1.2243	1.2424	1.1675
SD	1.0858	1.0906	1.1869	1.1656	1.3430	1.2729	1.0000

Table 3. Simulations results of the KL loss for setting 2.

γ_2	n		OPT1-KL	OPT2-KL	SAIC	SBIC	AIC	BIC
0.25	100	Mean	1.0000	1.0302	1.0886	1.0955	1.3113	1.2951
		SD	1.1114	1.0000	1.2915	1.0422	1.4790	1.1611
	200	Mean	1.0000	1.1340	1.1028	1.2919	1.2384	1.4722
		SD	1.0000	1.0499	1.1158	1.0986	1.2548	1.2265
0.50	100	Mean	1.0000	1.0189	1.0736	1.0566	1.2941	1.2694
		SD	1.1225	1.0000	1.3122	1.0794	1.4152	1.1858
	200	Mean	1.0000	1.1282	1.1018	1.2484	1.2400	1.4465
		SD	1.0000	1.0259	1.1005	1.0071	1.2501	1.1436
0.75	100	Mean	1.0084	1.0000	1.0614	1.0103	1.2938	1.2220
		SD	1.1939	1.0000	1.3298	1.0451	1.6517	1.1342
	200	Mean	1.0000	1.1333	1.0764	1.2175	1.2037	1.4217
		SD	1.0324	1.0108	1.1369	1.0000	1.3275	1.1295

the level of the model misspecification, we study it in the set $\{0.25, 0.5, 0.75\}$. We take the multinomial logit model (3) to fit the data. We still omit the last covariate X_{i6} and consider $S = 2^4 - 1$ candidate models.

The simulation results under different levels of the model misspecification are shown in Table 3. It is seen that OPT1-KL and OPT2-KL always deliver better performances than their competitors SAIC/AIC and SBIC/BIC in terms of mean values, respectively. Focusing on SD values, OPT1-KL always performs much better than SAIC and AIC, and OPT2-KL outperforms SBIC and BIC in most cases. It demonstrates the superiority of our methods.

In addition, we explore our strategies with other values of λ_n differing from 2 and $\log(n)$. That is, we vary λ_n from 0.5 to $n^{0.4}$. The simulation results are presented in Figure 2. For cases of $\gamma_2 = 0.25, 0.5$ and 0.75, when $n = 100$, the means of KL loss are minimized at $\lambda_n = 2.5$, $\lambda_n = 2.75$ and $\lambda_n = 3.25$, respectively. It states that when the level of the model misspecification improves, the optimal value of λ_n increases slightly for the small sample size. For a larger sample size $n = 200$, the optimal values of λ_n are same for all cases with $\lambda_n = 2$.

Setting 3. This setup discusses the case when the number of covariates is diverging. The data generate progress is the same as those in setting 1. Except that we adapt the covariance matrix to $R = (r_{ij})$ with $r_{ij} = 0.40^{|i-j|}$, for that the model screening method is not suitable for the case when the covariances have strong dependence which is implied by the first part of Lemma 3.2 in Ando and Li (2014). Then, β_1 and β_2 are chosen according to the following cases:

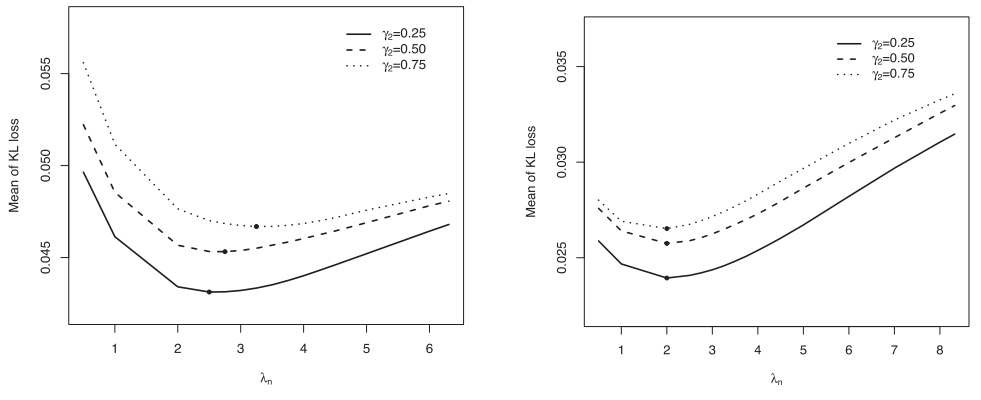
$$\begin{aligned}\beta_1 &= (1, 1.2, 0, 0, 1.5, 0, 0, 1.1, 0, 0, 0.1,\\ &\quad \ldots, 0.1, 0.9)^T_{[3n^{1/3}]\times 1};\\ \beta_2 &= (1, 1.3, 0, 0, 2, 0, 0, 1.2, 0, 0, 0.1,\\ &\quad \ldots, 0.1, 0.8)^T_{[3n^{1/3}]\times 1}.\end{aligned}$$

Similar to setting 1, we also pretend the last covariate was missed. Then, there are $2^{[3n^{1/3}]-2} - 1$ candidate models. The computation burden will be heavy. Therefore, a screening method to screen candidate models is desirable. That is, we use penalized regression with LASSO (Friedman et al., 2010) to prepare candidate models. Different tuning parameters may result in different models, which will be included in our resulting candidate models. Obviously, the resulting candidate model contains lots of redundant variables when the tuning parameter is very small. In order to avoid the generated candidate models including a lot of redundant variables, we use tuning parameters larger than $\hat{\zeta}$ to prepare candidate models.

Simulation results are provided in Table 4. Focusing on the mean values, Table 4 shows that OPT1-KL always performs better than SAIC and AIC, and OPT2-KL still outperforms SBIC and BIC, respectively. In addition, OPT2-KL always outperforms LASSO. It implies the advantages of our proposed method comparing with other competitive methods.

Setting 4. This setup verifies the estimation consistency. The data generate progress is the same as those in setting 3. Except that we choose $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ as follows:

$$\begin{aligned}\text{case1} \ : \boldsymbol{\beta}_1 &= (1, 1.5, 3, 0, 0)^T;\\ \boldsymbol{\beta}_2 &= (1, 1.7, 4, 0, 0)^T;\end{aligned}$$



(a) $n = 100$ (b) $n = 200$

Figure 2. The relationship between the mean of KL loss and λ_n. The points with the smallest losses are indicated by the filled circle •. (a) $n = 100$ and (b) $n = 200$.

Table 4. Simulations results of the KL loss for setting 3.

n		OPT1-KL	OPT2-KL	SAIC	SBIC	AIC	BIC	LASSO
100	Mean	1.0863	1.0000	1.3113	1.1209	1.4165	1.2136	1.0172
	SD	1.7691	1.3568	2.7897	1.6756	3.1682	1.9177	1.0000
200	Mean	1.0322	1.0000	1.1918	1.0406	1.2664	1.0821	1.1053
	SD	1.1986	1.1635	1.7025	1.4451	1.8834	1.5681	1.0000

and

$$\text{case2} \ : \boldsymbol{\beta}_1 = (1, 1.5, 3, 1.2, 0)^T;$$
$$\boldsymbol{\beta}_2 = (1, 1.7, 4, 1.3, 0)^T.$$

All candidate models include X_{i1}. Thus, we consider a total of $2^4 - 1 = 15$ candidate models. Note that the true model is included in the candidate models. We compare our proposed method with other competitive methods based on the KL loss and MSFE. The results are presented in Tables 5 and 6, respectively. These results show that OPT2-KL always obtain the smallest KL loss and MSFE among these methods, which validates the superiority of our method.

Also, we calculate the mean squared error (MSE) by methods of OPT1-KL and OPT2-KL.

$$\text{MSE} = \frac{1}{1000}\sum_{r=1}^{1000} \|\boldsymbol{\beta}^{(r)}(\hat{\boldsymbol{\omega}}) - \boldsymbol{\beta}\|^2,$$

where $\boldsymbol{\beta}^{(r)}(\hat{\boldsymbol{\omega}})$ represents the estimator of $\boldsymbol{\beta}$ for the rth replication and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$. The MSE curves, shown in Figure 3, decrease and approach zero with the increase of sample size n. The feature confirms estimation consistency numerically.

Table 5. Simulations results of the KL loss for setting 4.

Case	n		OPT1-KL	OPT2-KL	SAIC	SBIC	AIC	BIC
1	100	Mean	1.2060	1.0000	1.4672	1.1862	1.2667	1.2462
		SD	1.5345	1.1268	2.1304	1.4188	1.4251	1.0000
	200	Mean	1.1020	1.0000	1.2307	1.0853	1.0256	1.1022
		SD	1.2962	1.0000	1.7283	1.3664	1.0145	1.1259
2	100	Mean	1.1659	1.0000	1.5795	1.3179	1.3301	1.3501
		SD	1.7378	1.0000	2.9602	1.9637	1.8268	1.4928
	200	Mean	1.0311	1.0000	1.2623	1.2498	1.0708	1.3389
		SD	1.5307	1.1471	2.0867	1.6240	1.0798	1.0000

6. An empirical application

In this part, we apply the proposed method to the website phishing data, which was previously used by Abdelhamid et al. (2014). This data set contains three types of website (702 phishing websites, 548 legitimate websites and 103 suspicious websites). The dependent variables consist of Server Form Handler, Using Pop-Up Window, Fake HTTPs protocol, Request URL, URL of Anchor, Website Traffic, URL Length, Age of Domain and Having IP address. These variables are categorical (or binary). We transform this information into indicator variables. After this operation, the total number of predictors is 16. After the screening method, we analyse this dataset using candidate multinomial logit models. We randomly select 677 observations as the training sample and predict the remaining instances. We repeat this progress 500 times. We use the following KL-type prediction loss L_{KL} to measure the prediction performance.

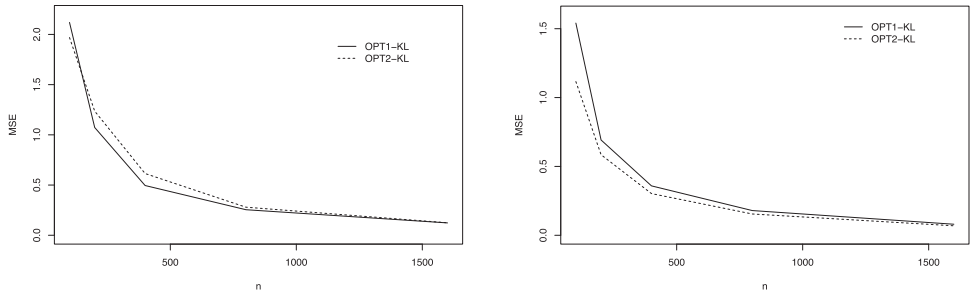
$$L_{KL} = -\frac{2}{n_0}\sum_{v=1}^{n_0}\sum_{j=1}^{d} I_{\{y_{test,v}=j\}} \log\{\hat{p}(y_{test,v} = j)\},$$

where $\{y_{test,1}, \ldots, y_{test,n_0}\}$ are testing observations, $\hat{p}(y_{test,v} = j)$ is the predicted probability of the vth test observation taking on j.

Figure 4 shows the box plot of all KL-type prediction losses by seven methods. It is observed that our proposed methods of OPT1-KL and OPT2-KL produce better performances than their competitions SAIC/AIC and SBIC/BIC, respectively. In addition, OPT2-KL outperforms LASSO in terms of the KL loss.

Table 6. Simulations results of MSFE for setting 4.

Case		OPT1-KL	OPT2-KL	SAIC	SBIC	AIC	BIC	A-OPT	MCV
1	Mean	1.0901	1.0000	1.2252	1.1081	1.2523	1.2838	2.0084	1.0225
	SD	1.1847	1.0764	1.3439	1.1847	1.4777	1.4331	14.1238	1.0000
2	Mean	1.0694	1.0000	1.2367	1.1510	1.4000	1.4735	1.8465	1.0653
	SD	1.1278	1.0278	1.2833	1.1722	1.4611	1.5611	10.6498	1.0000



(a) case 1 (b) case 2

Figure 3. Assessing the estimation consistency of OPT1-KL and OPT2-KL. (a) case 1 and (b) case 2.

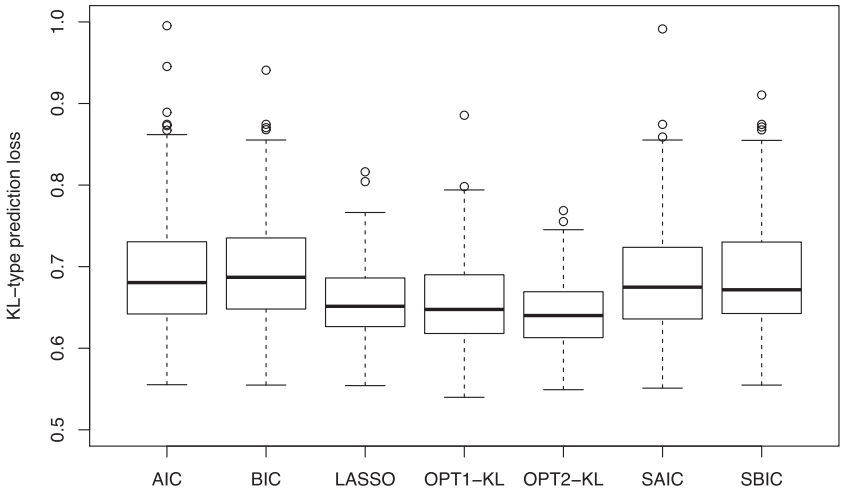


Figure 4. Boxplots of KL-type prediction losses by seven methods in the website phishing data.

In addition, we evaluate the prediction performance based on the hit-rate, which is computed by dividing the number of correct predictions by the size of the test sample. The prediction value of an observation is j ($1 \leq j \leq 3$) if the predicted probability of this observation taking on j has the largest value among the three predicted probability values. In addition, we also calculate the optimal rate (OPR) and the worst rate (WOR) of each method, which is the proportion of times with the largest hite-rate and the smallest hit-rate. Table 7 presents mean values of hit-rate (HRV), OPR and WOR corresponding to these methods. Which shows that OPT1-KL and OPT2-KL methods obtain the larger HRV, OPR and smaller WOR than other competitions SAIC, SBIC, AIC, BIC and LASSO, demonstrating the superiority of our proposed strategies.

Table 8 reports the Diebold–Mariano test (Diebold & Mariano, 2002) results for the differences in hit-rate. Note that a positive Diebold–Mariano statistic indicates that the estimator in the numerator produces a larger hit-rate than the estimator in the denominator. The test statistics and p-values show that the differences in hit-rate between our methods and other strategies are all statistically significant.

Table 7. Out-of-sample performances in the website phishing data.

	OPT1-KL	OPT2-KL	SAIC	SBIC	AIC	BIC	LASSO
HRV	0.8879	0.8921	0.8811	0.8812	0.8799	0.8800	0.8841
OPR	0.2200	0.6100	0.0200	0.0400	0.0100	0.0100	0.0900
WOR	0.0200	0.0200	0.2100	0.2600	0.2600	0.1600	0.0700

7. Discussion

In the context of multinomial logit model, we proposed model averaging estimator and weight choice criterion based on KL loss with a penalty term. And we proved the asymptotic optimality of the resulting model averaging estimator under some regularity conditions. When the true model is one of the candidate models, the averaged estimators are consistent. Also, in order to reduce the computational burden, we applied a model screening step before averaging. Numerical experiments were performed to demonstrate the superiority of the proposed methods over other commonly used model selection strategies, model averaging methods, MCV, Lasso in terms of KL loss and MSFE.

While we consider the multinomial logit model, the extension to other models, such as ordered logit model, warrants further investigation. And the data structure of the regressors further complicates this issue. Another interesting question is how to choose an optimal λ_n. Shen et al. (2004) have proposed an adaptive method to choose λ_n for model selection criterion in generalized

Table 8. Diebold–Mariano statistics of hit-rate in the website phishing data.

	$\frac{\text{OPT1}-\text{KL}}{\text{OPT2}-\text{KL}}$	$\frac{\text{OPT1}-\text{KL}}{\text{SAIC}}$	$\frac{\text{OPT1}-\text{KL}}{\text{SBIC}}$	$\frac{\text{OPT1}-\text{KL}}{\text{AIC}}$	$\frac{\text{OPT1}-\text{KL}}{\text{BIC}}$	$\frac{\text{OPT1}-\text{KL}}{\text{LASSO}}$	$\frac{\text{OPT2}-\text{KL}}{\text{SAIC}}$
DM	−7.6502	16.6320	9.7040	14.0080	13.1790	4.9572	16.8250
p-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	$\frac{\text{OPT2}-\text{KL}}{\text{SBIC}}$	$\frac{\text{OPT2}-\text{KL}}{\text{AIC}}$	$\frac{\text{OPT2}-\text{KL}}{\text{BIC}}$	$\frac{\text{OPT2}-\text{KL}}{\text{LASSO}}$	$\frac{\text{SAIC}}{\text{SBIC}}$	$\frac{\text{SAIC}}{\text{AIC}}$	$\frac{\text{SAIC}}{\text{BIC}}$
DM	22.0980	14.1900	16.1360	8.4057	−0.0124	5.7799	4.1535
p-value	0.0000	0.0000	0.0000	0.0000	0.9901	0.0000	0.0000
	$\frac{\text{SAIC}}{\text{LASSO}}$	$\frac{\text{SBIC}}{\text{AIC}}$	$\frac{\text{SBIC}}{\text{BIC}}$	$\frac{\text{SBIC}}{\text{LASSO}}$	$\frac{\text{AIC}}{\text{BIC}}$	$\frac{\text{AIC}}{\text{LASSO}}$	$\frac{\text{BIC}}{\text{LASSO}}$
DM	−5.0119	1.9948	4.0979	−4.7816	−0.0955	−5.7478	−6.4553
p-value	0.0000	0.0474	0.0000	0.0000	0.9240	0.0000	0.0000

linear models. Building similar methods for our proposed model averaging method to choose an optimal λ_n warrants future researches.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The research is supported by Natural Science Foundation of China (No. 11771268) and a center named Shanghai Research Center for Data Science and Decision Technology.

References

Abdelhamid, N., Ayesh, A., & Thabtah, F. (2014). Phishing detection based associative classification data mining. *Expert Systems with Applications*, *41*(13), 5948–5959. https://doi.org/10.1016/j.eswa.2014.03.019

Akaike, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika*, *60*(2), 255–265. https://doi.org/10.1093/biomet/60.2.255

Ando, T., & Li, K. C. (2014). A model-averaging approach for high-dimensional regression. *Journal of the American Statistical Association*, *109*, 254–265. https://doi.org/10.1080/01621459.2013.838168

Bayaga, A. (2010). Multinomial logistic regression: Usage and application in risk analysis. *Journal of Applied Quantitative Methods*, *5*, 288–297.

Buckland, S. T., Burnham, K. P., & Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics*, *53*(2), 603–618. https://doi.org/10.2307/2533961

Cavanaugh, J. E. (1999). A large-sample model selection criterion based on Kullback's symmetric divergence. *Statistics & Probability Letters*, *42*(4), 333–343. https://doi.org/10.1016/S0167-7152(98)00200-4

Cheng, T. C. F., Ing, C. K., & Yu, S. H. (2015). Toward optimal model averaging in regression models with time series errors. *Journal of Econometrics*, *189*(2), 321–334. https://doi.org/10.1016/j.jeconom.2015.03.026

Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *20*, 134–144. https://doi.org/10.1198/073500102753410444

Ederington, L. H. (1985). Classification models and bond ratings. *Financial Review*, *20*, 237–262. https://doi.org/10.1111/fire.1985.20.issue-4

Fahrmeir, L., & Tutz, G. (2013). *Multivariate statistical modelling based on generalized linear models*. Springer Science & Business Media.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*, 1–22. https://doi.org/10.18637/jss.v033.i01

Guadagni, P. M., & Little, J. D. C. (1983). A logit model of brand choice calibrated on scanner data. *Marketing Science*, *2*, 203–238. https://doi.org/10.1287/mksc.2.3.203

Hansen, B. E. (2007). Least squares model averaging. *Econometrica*, *75*, 1175–1189. https://doi.org/10.1111/ecta.2007.75.issue-4

Hansen, B. E., & Racine, J. S. (2012). Jackknife model averaging. *Journal of Econometrics*, *167*, 38–46. https://doi.org/10.1016/j.jeconom.2011.06.019

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*, 382–417. https://doi.org/10.1214/ss/1009212519

Hurvich, C. M., Simonoff, J. S., & Tsai, C. L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *60*, 271–293. https://doi.org/10.1111/1467-9868.00125

Konishi, S., & Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika*, *83*, 875–890. https://doi.org/10.1093/biomet/83.4.875

Li, C., Li, Q., Racine, J., & Zhang, D. Q. (2018). Optimal model averaging of varying coefficient models. *Statistica Sinica*, *28*, 2795–2809. https://doi.org/10.5705/ss.202017.0034

Liu, Q., & Okui, R. (2013). Heteroskedasticity-Robust C_p model averaging. *Econometrics Journal*, *16*(3), 463–472. https://doi.org/10.1111/ectj.12009

Lu, X., & Su, L. (2015). Jackknife model averaging for quantile regressions. *Journal of Econometrics*, *188*, 40–58. https://doi.org/10.1016/j.jeconom.2014.11.005

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, *15*, 661–675. https://doi.org/10.1080/00401706.1973.10489103

Portnoy, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *The Annals of Statistics*, *16*(1), 356–366. https://doi.org/10.1214/aos/1176350710

Raftery, A. E., & Zheng, Y. (2003). Discussion: Performance of Bayesian model averaging. *Journal of the American Statistical Association*, *98*(464), 931–938. https://doi.org/10.1198/016214503000000891

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Shen, X., Huang, H. C., & Ye, J. (2004). Adaptive model selection and assessment for exponential family distributions. *Technometrics*, *46*(3), 306–317. https://doi.org/10.1198/004017004000000338

Wan, A. T., Zhang, X., & Wang, S. (2014). Frequentist model averaging for multinomial and ordered logit models. *International Journal of Forecasting*, *30*(1), 118–128. https://doi.org/10.1016/j.ijforecast.2013.07.013

Wan, A. T., Zhang, X., & Zou, G. (2010). Least squares model averaging by Mallows criterion. *Journal of Econometrics*, *156*(2), 277–283. https://doi.org/10.1016/j.jeconom.2009.10.030

Wang, H., Zhang, X., & Zou, G. (2009). Frequentist model averaging estimation: A review. *Journal of Systems Science and Complexity*, *22*(4), 732–748. https://doi.org/10.1007/s11424-009-9198-y

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*(1), 1–25. https://doi.org/10.2307/1912526

Zhang, X., & Liu, C. A. (2019). Inference after model averaging in linear regression models. *Econometric Theory*, *35*(4), 816–841. https://doi.org/10.1017/S0266466618000269

Zhang, X., Wan, A. T., & Zou, G. (2013). Model averaging by jackknife criterion in models with dependent data. *Journal of Econometrics*, *174*(2), 82–94. https://doi.org/10.1016/j.jeconom.2013.01.004

Zhang, X., & Wang, W. (2019). Optimal model averaging estimation for partially linear models. *Statistica Sinica*, *29*, 693–718. https://doi.org/10.5705/ss.202015.0392

Zhang, X., Yu, D., Zou, G., & Liang, H. (2016). Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models. *Journal of*

the American Statistical Association, *111*(516), 1775–1790. https://doi.org/10.1080/01621459.2015.1115762

Zhang, X., & Yu, J. (2018). Spatial weights matrix selection and model averaging for spatial autoregressive models. *Journal of Econometrics*, *203*(1), 1–18. https://doi.org/10.1016/j.jeconom.2017.05.021

Zhang, X., Zou, G., & Carroll, R. J. (2015). Model averaging based on Kullback-Leibler distance. *Statistica Sinica*, *25*, 1583–1598. https://doi.org/10.5705/ss.2013.326

Zhang, X., Zou, G., Liang, H., & Carroll, R. J. (2020). Parsimonious model averaging with a diverging number of parameters. *Journal of the American Statistical Association*, *115*(530), 972–984. https://doi.org/10.1080/01621459.2019.1604363

Zhao, P., & Li, Z. (2008). Central limit theorem for weighted sum of multivariate random vector sequences. *Journal of Mathematics*, *28*, 171–176. https://doi.org/10.1007/s12033-007-0073-6

Zhao, S., Zhou, J., & Yang, G. (2019). Averaging estimators for discrete choice by M-fold cross-validation. *Economics Letters*, *174*, 65–69. https://doi.org/10.1016/j.econlet.2018.10.014

Zhu, R., Wan, A. T., Zhang, X., & Zou, G. (2019). A mallows-type model averaging estimator for the varying-coefficient partially linear model. *Journal of the American Statistical Association*, *114*(526), 882–892. https://doi.org/10.1080/01621459.2018.1456936

Zhu, R., Zou, G., & Zhang, X. (2018). Model averaging for multivariate multiple regression models. *Statistics*, *52*(1), 205–227. https://doi.org/10.1080/02331888.2017.1367794

Appendix

Proof of Theorem 3.1: Let $\tilde{\wp}(\boldsymbol{\omega}) = \wp(\boldsymbol{\omega}) - 2B_0 + 2\,\mathrm{Vec}(\boldsymbol{U}^T)^T\,\mathrm{Vec}(\boldsymbol{\Theta}_0^T)$. It's obvious that $\tilde{\wp}$ and $\wp$ are equivalent in the sense of choosing weights. From the proof of Theorem 1 in Zhang et al. (2016), Theorem 3.1 is valid if the following holds:

$$\sup_{\boldsymbol{\omega}\in H}\frac{|KL(\boldsymbol{\omega}) - KL^*(\boldsymbol{\omega})|}{KL^*(\boldsymbol{\omega})} \to o_p(1) \tag{A1}$$

and

$$\sup_{\boldsymbol{\omega}\in H}\frac{|\tilde{\wp}(\boldsymbol{\omega}) - KL^*(\boldsymbol{\omega})|}{KL^*(\boldsymbol{\omega})} \to o_p(1). \tag{A2}$$

By Equation (5), we can know that uniformly for $\boldsymbol{\omega} \in H$

$$\left\|\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}) - \boldsymbol{\beta}^*(\boldsymbol{\omega})\right\| = \left\|\sum_{s=1}^{S}\omega_s(\hat{\boldsymbol{\beta}}_{(s)} - \boldsymbol{\beta}^*_{(s)})\right\| = O_p(n^{-1/2}). \tag{A3}$$

From (A3), $\lambda_{\max}(\boldsymbol{Z}^T\boldsymbol{Z}) = \lambda_{\max}(\boldsymbol{X}^T\boldsymbol{X})$, Condition R.1 and Taylor expansion that uniformly for $\boldsymbol{\omega} \in H$

$$\begin{aligned}
&\left|B\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega})\} - B\left\{\boldsymbol{\beta}^*(\boldsymbol{\omega})\right\}\right| \\
&= \left|\sum_{i=1}^{n}\left[\left(\frac{\exp(\boldsymbol{X_i}\tilde{\boldsymbol{\beta}}(\boldsymbol{\omega})_1)\boldsymbol{X}_i}{\sum}{}_{j=1}^{d-1}(\exp(\boldsymbol{X_i}\tilde{\boldsymbol{\beta}}(\boldsymbol{\omega})_j)+1,\right.\right.\right. \\
&\left.\left.\left.\cdots, \frac{\exp(\boldsymbol{X_i}\tilde{\boldsymbol{\beta}}(\boldsymbol{\omega})_{d-1})\boldsymbol{X}_i}{\sum_{j=1}^{d-1}(\exp(\boldsymbol{X_i}\tilde{\boldsymbol{\beta}}(\boldsymbol{\omega})_j)+1}\right)(\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}) - \boldsymbol{\beta}^*(\boldsymbol{\omega}))\right]\right| \\
&\le \|\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}) - \boldsymbol{\beta}^*(\boldsymbol{\omega})\|\sum_{i=1}^{n}\left\|\left(\frac{\exp(\boldsymbol{X_i}\tilde{\boldsymbol{\beta}}(\boldsymbol{\omega})_1)\boldsymbol{X}_i}{\sum}{}_{j=1}^{d-1}\right.\right. \\
&\times(\exp(\boldsymbol{X_i}\tilde{\boldsymbol{\beta}}(\boldsymbol{\omega})_j)+1, \quad \ldots, \left.\left.\frac{\exp(\boldsymbol{X_i}\tilde{\boldsymbol{\beta}}(\boldsymbol{\omega})_{d-1})\boldsymbol{X}_i}{\sum_{j=1}^{d-1}(\exp(\boldsymbol{X_i}\tilde{\boldsymbol{\beta}}(\boldsymbol{\omega})_j)+1}\right)\right\| \\
&\le \|\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}) - \boldsymbol{\beta}^*(\boldsymbol{\omega})\|\sum_{l=1}^{n(d-1)}\|\boldsymbol{Z}_l\| \\
&\le \|\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}) - \boldsymbol{\beta}^*(\boldsymbol{\omega})\|\sum_{l=1}^{n(d-1)}(1+\|\boldsymbol{Z}_l\|^2) \\
&= \|\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}) - \boldsymbol{\beta}^*(\boldsymbol{\omega})\|[\mathrm{trace}(\boldsymbol{Z}^T\boldsymbol{Z}) + n(d-1)] \\
&\le \|\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}) - \boldsymbol{\beta}^*(\boldsymbol{\omega})\|[\lambda_{\max}(\boldsymbol{Z}^T\boldsymbol{Z})(k(d-1)) + n(d-1)] \\
&= O_p(n^{1/2}),
\end{aligned} \tag{A4}$$

where $\tilde{\boldsymbol{\beta}}(\boldsymbol{\omega})$ between $\hat{\boldsymbol{\beta}}(\boldsymbol{\omega})$ and $\boldsymbol{\beta}^*(\boldsymbol{\omega})$. From $\boldsymbol{U}_{ij} < 1, i = 1,\ldots,n, j = 1,\ldots,(d-1)$, we can obtain $\|\,\mathrm{Vec}(\boldsymbol{U}^T)\|^2 = O(n)$, which along with (A3), $\lambda_{\max}(\boldsymbol{Z}^T\boldsymbol{Z}) = \lambda_{\max}(\boldsymbol{X}^T\boldsymbol{X})$, Condition R.1, we have

$$\begin{aligned}
&\mathrm{Vec}(\boldsymbol{U}^T)^T\left[\mathrm{Vec}(\boldsymbol{\Theta}^T\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega})\}) - \mathrm{Vec}(\boldsymbol{\Theta}^T\left\{\boldsymbol{\beta}^*(\boldsymbol{\omega})\right\})\right] \\
&= \mathrm{Vec}(\boldsymbol{U}^T)^T\{\boldsymbol{Z}\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}) - \boldsymbol{Z}\boldsymbol{\beta}^*(\boldsymbol{\omega})\} \\
&\le \|\,\mathrm{Vec}(\boldsymbol{U}^T)^T\boldsymbol{Z}\|\left\|\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}) - \boldsymbol{\beta}^*(\boldsymbol{\omega})\right\| \\
&= [(\mathrm{Vec}(\boldsymbol{U}^T)^T\boldsymbol{Z}^{(1)})^2 \\
&+ \ldots + (\mathrm{Vec}(\boldsymbol{U}^T)^T\boldsymbol{Z}^{(k(d-1))})^2]^{1/2}\left\|\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}) - \boldsymbol{\beta}^*(\boldsymbol{\omega})\right\| \\
&\le [\|\,\mathrm{Vec}(\boldsymbol{U}^T)\|^2(\|\boldsymbol{Z}^{(1)}\|^2 \\
&+ \ldots + \|\boldsymbol{Z}^{(k(d-1))}\|^2)]^{1/2}\left\|\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}) - \boldsymbol{\beta}^*(\boldsymbol{\omega})\right\| \\
&\le [\|\,\mathrm{Vec}(\boldsymbol{U}^T)\|^2 trace(\boldsymbol{Z}^T\boldsymbol{Z})]^{1/2}\left\|\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}) - \boldsymbol{\beta}^*(\boldsymbol{\omega})\right\| \\
&\le [\lambda_{\max}(\boldsymbol{Z}^T\boldsymbol{Z})k(d-1)\|\,\mathrm{Vec}(\boldsymbol{U}^T)\|^2]^{1/2}\left\|\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}) - \boldsymbol{\beta}^*(\boldsymbol{\omega})\right\| \\
&= O_p(n^{1/2}),
\end{aligned} \tag{A5}$$

where $\boldsymbol{Z}^{(j)}$ is the jth column of $\boldsymbol{Z}$. Note that $\sum_{l=1}^{n(d-1)}\|\boldsymbol{Z}_l\|^2 = \mathrm{trace}(\boldsymbol{Z}^T\boldsymbol{Z}) \le \lambda_{\max}(\boldsymbol{Z}^T\boldsymbol{Z})k(d-1)$, which combines with central limit theorem, Condition R.1, and the second part of Condition R.2, we obtain $\|\,\mathrm{Vec}(\boldsymbol{\Xi}^T)^T\boldsymbol{Z}\| = O_p(n^{1/2})$. From $\|\,\mathrm{Vec}(\boldsymbol{\Xi}^T)^T\boldsymbol{Z}\| = O_p(n^{1/2})$ and (A3), we have

$$\begin{aligned}
&\mathrm{Vec}(\boldsymbol{\Xi}^T)^T\left[\mathrm{Vec}(\boldsymbol{\Theta}^T\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega})\}) - \mathrm{Vec}(\boldsymbol{\Theta}^T\left\{\boldsymbol{\beta}^*(\boldsymbol{\omega})\right\})\right] \\
&= \mathrm{Vec}(\boldsymbol{\Xi}^T)^T\{\boldsymbol{Z}\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}) - \boldsymbol{Z}\boldsymbol{\beta}^*(\boldsymbol{\omega})\} \\
&\le \left\|\mathrm{Vec}(\boldsymbol{\Xi}^T)^T\boldsymbol{Z}\right\|\left\|\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}) - \boldsymbol{\beta}^*(\boldsymbol{\omega})\right\| = O_p(1).
\end{aligned} \tag{A6}$$

From Condition R.1, the first part of Condition R.2, we have

$$\begin{aligned}
&\sum_{i=1}^{n}\boldsymbol{\theta}_i^T(\boldsymbol{\beta}^*_{(s)})\,\mathrm{Cov}(\boldsymbol{\Xi}_i)\boldsymbol{\theta}_i(\boldsymbol{\beta}^*_{(s)}) \\
&< C_2\sum_{i=1}^{n}\|\boldsymbol{\theta}_i(\boldsymbol{\beta}^*_{(s)})\|^2 = C_2{\boldsymbol{\beta}^*_{(s)}}^T\boldsymbol{Z}^T\boldsymbol{Z}\boldsymbol{\beta}^*_{(s)} \\
&\le C_2\lambda_{\max}(\boldsymbol{Z}^T\boldsymbol{Z})\|\boldsymbol{\beta}^*_{(s)}\|^2 = O(n),
\end{aligned}$$

and

$$\max_{i\in\{1,\ldots,n\}}\|\boldsymbol{\theta}_i(\boldsymbol{\beta}^*_{(s)})\|^2 \Big/ \sum_{i=1}^{n}\|\boldsymbol{\theta}_i(\boldsymbol{\beta}^*_{(s)})\|^2$$

$$
\begin{aligned}
&= \max_{i\in\{1,\ldots,n\}} \|(\boldsymbol{X}_i\boldsymbol{\beta}^*_{(s)1},\ldots,\boldsymbol{X}_i\boldsymbol{\beta}^*_{(s)(d-1)})\|^2 \Big/ \sum_{i=1}^{n}\|\boldsymbol{\theta}_i(\boldsymbol{\beta}^*_{(s)})\|^2 \\
&\le \max_{i\in\{1,\ldots,n\}} (d-1)\|\boldsymbol{X}_i\|^2\|\boldsymbol{\beta}^*_{(s)}\|^2 \Big/ \sum_{i=1}^{n}\|\boldsymbol{\theta}_i(\boldsymbol{\beta}^*_{(s)})\|^2 \\
&\le \max_{i\in\{1,\ldots,n\}} (d-1)\|\boldsymbol{X}_i\|^2\|\boldsymbol{\beta}^*_{(s)}\|^2/(\|\boldsymbol{\beta}^*_{(s)}\|^2\lambda_{\min}(\boldsymbol{Z}^T\boldsymbol{Z})) \\
&= o(1),
\end{aligned}
$$

these along with Theorem 1 in P. Zhao and Li (2008), the second part of Condition R.2, we know that uniformly for $\boldsymbol{\omega}\in H$

$$
\begin{aligned}
&\mathrm{Vec}(\Xi^T)^T\,\mathrm{Vec}(\boldsymbol{\Theta}^T\{\boldsymbol{\beta}^*(\omega)\}) \\
&\quad= \sum_{i=1}^{n}\Xi_i\boldsymbol{\theta}_i\{\boldsymbol{\beta}^*(\boldsymbol{\omega})\} \\
&\quad= \sum_{s=1}^{S} w_s\sum_{i=1}^{n}\Xi_i\boldsymbol{\theta}_i(\boldsymbol{\beta}^*_{(s)}) = O_p(n^{1/2}).
\end{aligned}
\tag{A7}
$$

Therefore, (A4), (A5) indicate that

$$
\begin{aligned}
&\sup_{\boldsymbol{\omega}\in H}|KL(\boldsymbol{\omega}) - KL^*\boldsymbol{\omega}| \\
&\le 2\sup_{\boldsymbol{\omega}\in H}\left|B\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega})\} - B\left\{\boldsymbol{\beta}^*(\boldsymbol{\omega})\right\}\right| \\
&\quad + 2\left|\mathrm{Vec}(\boldsymbol{U}^T)^T\left[\mathrm{Vec}(\boldsymbol{\Theta}^T\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega})\}) - \mathrm{Vec}(\boldsymbol{\Theta}^T\left\{\boldsymbol{\beta}^*(\boldsymbol{\omega})\right\})\right]\right| \\
&= O_p(n^{1/2}).
\end{aligned}
\tag{A8}
$$

and (A4), (A5), (A6), (A7) indicate that

$$
\begin{aligned}
&\sup_{\boldsymbol{\omega}\in H}|\tilde{\wp}(\boldsymbol{\omega}) - KL^*(\boldsymbol{\omega})| \le 2\sup_{\boldsymbol{\omega}\in H}\left|B\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega})\} - B\left\{\boldsymbol{\beta}^*(\boldsymbol{\omega})\right\}\right| \\
&\quad + 2\sup_{\boldsymbol{\omega}\in H}\Big|\mathrm{Vec}(\boldsymbol{Y}^T)^T\,\mathrm{Vec}(\boldsymbol{\Theta}^T\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega})\}) \\
&\quad - \mathrm{Vec}(\boldsymbol{U}^T)^T\,\mathrm{Vec}(\boldsymbol{\Theta}^T\left\{\boldsymbol{\beta}^*(\boldsymbol{\omega})\right\})\Big| + \lambda_n(d-1)\boldsymbol{\omega}^T\boldsymbol{K} \\
&\quad \le 2\sup_{\boldsymbol{\omega}\in H}\left|B\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega})\} - B\left\{\boldsymbol{\beta}^*(\boldsymbol{\omega})\right\}\right| \\
&\quad + 2\sup_{\boldsymbol{\omega}\in H}\Big|\mathrm{Vec}(U^T)(\mathrm{Vec}(\boldsymbol{\Theta}^T\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega})\}) \\
&\quad - 2\,\mathrm{Vec}(\boldsymbol{\Theta}^T\left\{\boldsymbol{\beta}^*(\boldsymbol{\omega})\right\}))\Big| \\
&\quad + 2\sup_{\boldsymbol{\omega}\in H}\Big|\mathrm{Vec}(\Xi^T)^T\,\mathrm{Vec}(\boldsymbol{\Theta}^T\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega})\}) \\
&\quad - \mathrm{Vec}(\boldsymbol{\Theta}^T\left\{\boldsymbol{\beta}^*(\boldsymbol{\omega})\right\})\Big| \\
&\quad + 2\sup_{\boldsymbol{\omega}\in H}\left|\mathrm{Vec}(\Xi^T)^T\,\mathrm{Vec}(\boldsymbol{\Theta}^T\{\boldsymbol{\beta}^*(\boldsymbol{\omega})\})\right| \\
&\quad + \lambda_n(d-1)\boldsymbol{\omega}^T\boldsymbol{K} = O_p(n^{1/2}) + \lambda_n(d-1)\boldsymbol{\omega}^T\boldsymbol{K}.
\end{aligned}
\tag{A9}
$$

Now, from (A8), (A9) and Conditions R.3 and R.4, we can get (A1) and (A2). This completes the proof of Theorem 3.1. ■

Proof of Theorem 3.2: From the result of Theorem 3.1, it suffices to prove that (A10)–(A14), as $n\to\infty$

$$
\sup_{\boldsymbol{\omega}\in H}\frac{\left|B\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega})\} - B\left\{\boldsymbol{\beta}^*(\boldsymbol{\omega})\right\}\right|}{KL^*(\boldsymbol{\omega})} = o_p(1), \tag{A10}
$$

$$
\begin{aligned}
&\sup_{\boldsymbol{\omega}\in H}\frac{\left|\mathrm{Vec}(\boldsymbol{U}^T)^T\left(\mathrm{Vec}(\boldsymbol{\Theta}^T\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega})\}) - \mathrm{Vec}(\boldsymbol{\Theta}^T\left\{\boldsymbol{\beta}^*(\boldsymbol{\omega})\right\})\right)\right|}{KL^*(\boldsymbol{\omega})} \\
&\quad = o_p(1),
\end{aligned}
\tag{A11}
$$

$$
\sup_{\boldsymbol{\omega}\in H}\frac{\left|\mathrm{Vec}(\Xi^T)^T\,\mathrm{Vec}(\boldsymbol{\Theta}^T\left\{\boldsymbol{\beta}^*(\omega)\right\})\right|}{KL^*} = o_p(1), \tag{A12}
$$

$$
\begin{aligned}
&\sup_{\boldsymbol{\omega}\in H}\frac{\left|\mathrm{Vec}(\Xi^T)^T\left(\mathrm{Vec}(\boldsymbol{\Theta}^T\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega})\}) - \mathrm{Vec}(\boldsymbol{\Theta}^T\left\{\boldsymbol{\beta}^*(\boldsymbol{\omega})\right\})\right)\right|}{KL^*} \\
&\quad = o_p(1),
\end{aligned}
\tag{A13}
$$

$$
\frac{\lambda_n(d-1)\boldsymbol{\omega}^T\boldsymbol{K}}{KL^*} = o(1). \tag{A14}
$$

First of all, we show that for any fixed $\varepsilon>0$, there exists $\delta_\varepsilon>0$ such that for all sufficiently large n

$$
P\left(\left\|\frac{n^{1/2}}{\{(d-1)k\}^{1/2}}(\hat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}^*_s)\right\| \le \delta_\varepsilon\right) \ge 1-\varepsilon.
$$

Write $\boldsymbol{u}^*_s = (b^{(1)}[(I_{d-1}\otimes X_{(s),i})\boldsymbol{\beta}^*_s]^T,\ldots,b^{(1)}[I_{d-1}\otimes X_{(s),n})\boldsymbol{\beta}^*_s]^T)^T$, where $\boldsymbol{u}^*_s$ is a $n(d-1)\times 1$ vector. The quasi true value $\boldsymbol{\beta}^*_s$ minimizes the KL divergence so that

$$
\begin{aligned}
&\partial\{B(\boldsymbol{\beta}_s) - \mathrm{Vec}(\boldsymbol{U}^T)^T\,\mathrm{Vec}(\boldsymbol{\Theta}^T(\boldsymbol{\beta}_s))\}/\partial\boldsymbol{\beta}_s|_{\boldsymbol{\beta}_s=\boldsymbol{\beta}^*_s} \\
&\quad = \mathbf{0}_{k_s(d-1)\times 1},
\end{aligned}
$$

which implies that $\boldsymbol{Z}^T_{(s)}\boldsymbol{u}^*_s = \boldsymbol{Z}^T_{(s)}\,\mathrm{Vec}(\boldsymbol{U}^T)$. Then, by using first-order Taylor expansion of $\partial\log f(\boldsymbol{Y}|\boldsymbol{Z}_{(s)}\ \boldsymbol{\beta}_s)/\partial\boldsymbol{\beta}_s = \mathbf{0}_{k_s(d-1)\times 1}$ at $\boldsymbol{\beta}^*_{(s)}$, and we can get

$$
\begin{aligned}
\mathbf{0}_{k_s(d-1)\times 1} &= -\boldsymbol{Z}^T_{(s)}\{\mathrm{Vec}(\boldsymbol{Y}^T) - \mathrm{Vec}(\boldsymbol{U}^T)\} \\
&\quad + \boldsymbol{Z}^T_{(s)}\boldsymbol{D}_s\boldsymbol{Z}_{(s)}(\hat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}^*_s),
\end{aligned}
$$

which implies $(\boldsymbol{Z}^T_{(s)}\boldsymbol{D}_s\boldsymbol{Z}_{(s)})^{-1}\boldsymbol{Z}^T_{(s)}\{\mathrm{Vec}(\boldsymbol{Y}^T) - \mathrm{Vec}(\mathbf{U}^T)\} = (\hat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}^*_s)$, then

$$
\begin{aligned}
&\frac{n^{1/2}}{\{(d-1)k\}^{1/2}}(\hat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}^*_s) \\
&\quad = \left(\frac{1}{n}\boldsymbol{Z}^T_{(s)}\boldsymbol{D}_s\boldsymbol{Z}_{(s)}|_{\boldsymbol{\beta}_s=\tilde{\boldsymbol{\beta}}_s}\right)^{-1}\frac{\boldsymbol{Z}^T_{(s)}\{\mathrm{Vec}(\boldsymbol{Y}^T) - \mathrm{Vec}(\boldsymbol{U}^T)\}}{\{(d-1)k\}^{1/2}n^{1/2}},
\end{aligned}
$$

where $\tilde{\boldsymbol{\beta}}_s$ between $\hat{\boldsymbol{\beta}}_s$ and $\boldsymbol{\beta}^*_s$. It follows Condition R.6 and sufficiently large n

$$
\begin{aligned}
&P\left(\left\|\frac{n^{1/2}}{\{(d-1)k\}^{1/2}}(\hat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}^*_s)\right\| \le \delta\right) \\
&\quad \ge P\left(C_0^{-1}\left\|\frac{\boldsymbol{Z}^T_{(s)}\{\mathrm{Vec}(\boldsymbol{Y}^T) - \mathrm{Vec}(\boldsymbol{U}^T)\}}{\{(d-1)k\}^{1/2}n^{1/2}}\right\| \le \delta\right) \\
&\quad \ge 1 - \frac{\sum_{i=1}^{n}\|\boldsymbol{X}_i\|^2\,\bar{\lambda}^2(d-1)}{C_0^2\delta^2(d-1)kn} \\
&\quad \ge 1 - \frac{C_1}{C_0^2\delta^2},
\end{aligned}
$$

By taking $\delta_\varepsilon = C_1^{1/2}/(\varepsilon^{1/2}C_0)$, we can obtain $\|\hat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}^*_s\| = \|\hat{\boldsymbol{\beta}}_{(s)} - \boldsymbol{\beta}^*_{(s)}\| = O_p(\{k(d-1)\}^{1/2}n^{-1/2})$, and thus

$$
\begin{aligned}
&\left\|\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}) - \boldsymbol{\beta}^*(\boldsymbol{\omega})\right\| \\
&\quad \le \sum_{s=1}^{S}\omega_s\left\|\hat{\boldsymbol{\beta}}_{(s)} - \boldsymbol{\beta}^*_{(s)}\right\| \\
&\quad = O_p(\{k(d-1)\}^{1/2}n^{-1/2}).
\end{aligned}
\tag{A15}
$$

From (A15) and Condition R.5, we can show that uniformly for $\boldsymbol{\omega} \in H$

$$\begin{aligned}\left|B\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega})\} - B\left\{\boldsymbol{\beta}^*(\boldsymbol{\omega})\right\}\right| &\le \|\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}) - \boldsymbol{\beta}^*(\boldsymbol{\omega}\| \sum_{l=1}^{n(d-1)} \|\boldsymbol{Z}_l\| \\ &= \|\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}) - \boldsymbol{\beta}^*(\boldsymbol{\omega}\|(d-1)\sum_{i=1}^{n}\|\boldsymbol{X}_i\| \\ &= O_p(k(d-1)n^{1/2}), \end{aligned} \tag{A16}$$

and similar to the proof of (A5), we can obtain

$$\begin{aligned}&\mathrm{Vec}(\boldsymbol{U}^T)^T\left[\mathrm{Vec}(\boldsymbol{\Theta}^T\{\hat{\boldsymbol{\beta}}(w)\}) - \mathrm{Vec}(\boldsymbol{\Theta}^T\left\{\boldsymbol{\beta}^*(\boldsymbol{\omega})\right\})\right] \\ &= \mathrm{Vec}(\boldsymbol{U}^T)^T\left(\boldsymbol{Z}\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}) - \boldsymbol{Z}\boldsymbol{\beta}^*(\boldsymbol{\omega})\right) \\ &\le [\lambda_{\max}(\boldsymbol{Z}^T\boldsymbol{Z})k(d-1)\|\,\mathrm{Vec}(\boldsymbol{U}^T)\|^2]^{1/2}\left\|\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}) - \boldsymbol{\beta}^*(\boldsymbol{\omega})\right\| \\ &= O_p(kn^{1/2}). \end{aligned} \tag{A17}$$

By combining (A16), (A17) and Condition R.7, we obtain (A10) and (A11). By using $\sum_{l=1}^{n(d-1)}\|\boldsymbol{Z}_l\|^2 = \mathrm{trace}(\boldsymbol{Z}^T\boldsymbol{Z}) \le \lambda_{\max}(\boldsymbol{Z}^T\boldsymbol{Z})k(d-1)$, central limit theorem, Condition R.1, and the second part of Condition R.2, we obtain $\|\mathrm{Vec}(\Xi^T)^T\boldsymbol{Z}\| = O_p(\{k(d-1)\}^{1/2}n^{1/2})$, which combines (A15) imply

$$\begin{aligned}&\mathrm{Vec}(\boldsymbol{\Xi}^T)^T\left(\mathrm{Vec}(\boldsymbol{\Theta}^T\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega})\}) - \mathrm{Vec}(\boldsymbol{\Theta}^T\left\{\boldsymbol{\beta}^*(\boldsymbol{\omega})\right\})\right) \\ &= O_p(k(d-1)). \end{aligned} \tag{A18}$$

From Condition R.1 and the first part of Condition R.2, we obtain

$$\begin{aligned}&\sum_{i=1}^{n}\boldsymbol{\theta}_i^T(\boldsymbol{\beta}_{(s)}^*)\,\mathrm{Cov}(\Xi_i)\boldsymbol{\theta}_i(\boldsymbol{\beta}_{(s)}^*) \\ &< C_2\sum_{i=1}^{n}\|\boldsymbol{\theta}_i(\boldsymbol{\beta}_{(s)}^*)\|^2 = C_2{\boldsymbol{\beta}_{(s)}^*}^T\boldsymbol{Z}^T\boldsymbol{Z}\boldsymbol{\beta}_{(s)}^* \\ &\le C_2\lambda_{\max}(\boldsymbol{Z}^T\boldsymbol{Z})\|\boldsymbol{\beta}_{(s)}^*\|^2 = O(nk), \end{aligned}$$

and

$$\max_{i\in\{1,\ldots,n\}}\|\boldsymbol{\theta}_i(\boldsymbol{\beta}_{(s)}^*)\|^2\Big/\sum_{i=1}^{n}\|\boldsymbol{\theta}_i(\boldsymbol{\beta}_{(s)}^*)\|^2 = o(1),$$

these along with Theorem 1 in P. Zhao and Li (2008), the second part of Condition R.2, we know that uniformly for $\boldsymbol{\omega} \in H$

$$\begin{aligned}&\mathrm{Vec}(\Xi^T)^T\,\mathrm{Vec}(\boldsymbol{\Theta}^T\{\boldsymbol{\beta}^*(\omega)\}) \\ &= \sum_{i=1}^{n}\Xi_i\boldsymbol{\theta}_i\{\boldsymbol{\beta}^*(\boldsymbol{\omega})\} \\ &= \sum_{s=1}^{S}w_s\sum_{i=1}^{n}\Xi_i\boldsymbol{\theta}_i(\boldsymbol{\beta}_{(s)}^*) = O_p((nk)^{1/2}). \end{aligned} \tag{A19}$$

Using (A18), (A19) and Conditions R.4, R.7, the claims (A12)–(A14) are obtained. This completes the proof of Theorem 3.2. ■

Proof of Theorem 4.1: Note that the true value $\boldsymbol{\beta}_0$ minimizes the KL divergence so that

$$\partial\{B(\boldsymbol{\beta}) - \mathrm{Vec}(\boldsymbol{U}^T)^T\,\mathrm{Vec}(\boldsymbol{\Theta}^T(\boldsymbol{\beta}))\}/\partial\boldsymbol{\beta}|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} = \boldsymbol{0}_{k(d-1)\times 1},$$

which implies that

$$\boldsymbol{Z}^T\boldsymbol{u}_0 = \boldsymbol{Z}^T\,\mathrm{Vec}(\boldsymbol{U}^T), \tag{A20}$$

where $\boldsymbol{u}_0 = (b^{(1)}[(I_{d-1}\otimes X_i)\boldsymbol{\beta}_0]^T, \ldots, b^{(1)}[I_{d-1}\otimes X_n)\boldsymbol{\beta}_0]^T)^T$, which is a $n(d-1)\times 1$ vector. Then by using second order Taylor expansion of $B\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}_{true})\}$ at $\boldsymbol{\beta}_0$, we have

$$\begin{aligned}&B\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}_{true})\} \\ &= B_0 + (\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}_{true}) - \boldsymbol{\beta}_0)^T\boldsymbol{Z}^T\boldsymbol{u}_0 \\ &\quad + \frac{1}{2}(\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}_{true}) - \boldsymbol{\beta}_0)^T\boldsymbol{Z}^T\tilde{\boldsymbol{D}}(\boldsymbol{\omega}_{true})\boldsymbol{Z}(\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}_{true}) - \boldsymbol{\beta}_0). \end{aligned} \tag{A21}$$

where $\tilde{\boldsymbol{D}}(\boldsymbol{\omega}_{true}) = \mathrm{diag}\{\boldsymbol{D}_i[\boldsymbol{\beta}_0 + r(\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}_{true}) - \boldsymbol{\beta}_0)]\}_{i=1,\ldots,n}$. From every element of symmetric matrix $\boldsymbol{D}_i[\boldsymbol{\beta}_0 + r(\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}_{true}) - \boldsymbol{\beta}_0)]$ is bound, (A20), (A21), (7) and Condition R.1, we have

$$\begin{aligned}KL(\boldsymbol{\omega}_{true}) &= 2(B\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}_{true})\} - B_0) \\ &\quad - 2[\mathrm{Vec}(U^T)(\mathrm{Vec}(\boldsymbol{\Theta}^T\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}_{true})\}) - \mathrm{Vec}(\boldsymbol{\Theta}_0^T)] \\ &= 2[\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}_{true}) - \boldsymbol{\beta}_0]^T\boldsymbol{Z}^T\boldsymbol{u}_0 \\ &\quad + [\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}_{true}) - \boldsymbol{\beta}_0]^T\boldsymbol{Z}^T\tilde{\boldsymbol{D}}(\boldsymbol{\omega}_{true})\boldsymbol{Z}[\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}_{true}) - \boldsymbol{\beta}_0] \\ &\quad - 2[(\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}_{true}) - \boldsymbol{\beta}_0)^T\boldsymbol{Z}^T\,\mathrm{Vec}(\boldsymbol{U}^T)] \\ &= [\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}_{true}) - \boldsymbol{\beta}_0]^T\boldsymbol{Z}^T\tilde{\boldsymbol{D}}(\boldsymbol{\omega}_{true})\boldsymbol{Z}[\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}_{true}) - \boldsymbol{\beta}_0] \\ &\le \lambda_{\max}(\tilde{\boldsymbol{D}}(\boldsymbol{\omega}_{true}))\lambda_{\max}(\boldsymbol{Z}^T\boldsymbol{Z})\|\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}_{true}) - \boldsymbol{\beta}_0\|^2 \\ &\le \max_{i=1,\ldots,n}\mathrm{trace}(b^{(2)}[(I_{d-1}\otimes X_i)\tilde{\boldsymbol{\beta}}(\boldsymbol{\omega})])\lambda_{\max} \\ &\quad \times (\boldsymbol{X}^T\boldsymbol{X})\|\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}_{true}) - \boldsymbol{\beta}_0\|^2 \\ &\le (d-1)\lambda_{\max}(\boldsymbol{X}^T\boldsymbol{X})\|\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}_{true}) - \boldsymbol{\beta}_0\|^2 \\ &\le O_p(n)\|\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}_{true}) - \boldsymbol{\beta}_0\|^2 \\ &= O_p(1). \end{aligned} \tag{A22}$$

In addition, let $\aleph$ be the set of i such that the inequality in Condition R.8 holds. From Condition R.8, and the second-order Taylor expansion of $B\{\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}})\}$ at $\boldsymbol{\beta}_0$, we have

$$\begin{aligned}KL(\hat{\boldsymbol{\omega}}) &= 2(B\{\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}})\} - B_0) \\ &\quad - 2[\mathrm{Vec}(U^T)(\mathrm{Vec}(\boldsymbol{\Theta}^T\{\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}})\}) - \mathrm{Vec}(\boldsymbol{\Theta}_0^T)] \\ &= 2(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}) - \boldsymbol{\beta}_0)^T\boldsymbol{Z}^T\boldsymbol{u}_0 \\ &\quad - 2(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}) - \boldsymbol{\beta}_0)^T\boldsymbol{Z}^T\,\mathrm{Vec}(\boldsymbol{U}^T) \\ &\quad + \sum_{i=1}^{n}\left\|\boldsymbol{D}_i^{1/2}[\boldsymbol{\beta}_0 + r(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}) - \boldsymbol{\beta}_0)]\right. \\ &\quad \left.\times\ (\boldsymbol{I}_{d-1}\otimes\boldsymbol{X}_i)(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}) - \boldsymbol{\beta}_0)\right\|^2 \\ &= \sum_{i=1}^{n}\left\|\boldsymbol{D}_i^{1/2}[\boldsymbol{\beta}_0 + r(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}) - \boldsymbol{\beta}_0)]\right. \\ &\quad \left.\times\ (\boldsymbol{I}_{d-1}\otimes\boldsymbol{X}_i)(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}) - \boldsymbol{\beta}_0)\right\|^2 \\ &\ge \sum_{i\in\aleph}\underline{d}\left\|\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}) - \boldsymbol{\beta}_0)\right\|^2 \\ &\ge \underline{d}n^*\left\|\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}) - \boldsymbol{\beta}_0)\right\|^2, \end{aligned} \tag{A23}$$

where n^* is the number of elements in $\aleph$, and from Condition R.8 we know that n^* has the same order as n. Note that

$$\tilde{\wp}(\boldsymbol{\omega}_{true}) = KL(\boldsymbol{\omega}_{true}) + 2\,\mathrm{Vec}(\Xi^T)^T\,\mathrm{Vec}(\boldsymbol{\Theta}^T\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}_{true})\})$$

$$+ \lambda_n(d-1)k_{true},$$

and

$$\tilde{\wp}(\hat{\boldsymbol{\omega}}) = KL(\hat{\boldsymbol{\omega}}) + 2\,\mathrm{Vec}(\Xi^T)^T\,\mathrm{Vec}(\boldsymbol{\Theta}^T\{\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}})\}) + \lambda_n(d-1)\hat{\boldsymbol{\omega}}^T\boldsymbol{K}. \tag{A24}$$

These along (A22), (A24) and $\tilde{\wp}(\boldsymbol{\omega}_{true}) \geq \tilde{\wp}(\hat{\boldsymbol{\omega}})$, we have

$$\begin{aligned} &KL(\boldsymbol{\omega}_{true}) + 2\,\mathrm{Vec}(\Xi^T)^T\,\mathrm{Vec}(\boldsymbol{\Theta}^T\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}_{true})\}) \\ &\quad + \lambda_n(d-1)k_{true} \geq KL(\hat{\boldsymbol{\omega}}) \\ &\quad + 2\,\mathrm{Vec}(\Xi^T)^T\,\mathrm{Vec}(\boldsymbol{\Theta}^T\{\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}})\}) + \lambda_n(d-1)\hat{\boldsymbol{\omega}}^T\boldsymbol{K}, \end{aligned}$$

which follows that

$$\begin{aligned} &KL(\boldsymbol{\omega}_{true}) + 2\,\mathrm{Vec}(\Xi^T)^T[\mathrm{Vec}(\boldsymbol{\Theta}^T\{\hat{\boldsymbol{\beta}}(\boldsymbol{\omega}_{true})\}) \\ &\quad - \mathrm{Vec}(\boldsymbol{\Theta}_0^T)] + \lambda_n(d-1)k_{true} \\ &\quad - 2\,\mathrm{Vec}(\Xi^T)^T[\mathrm{Vec}(\boldsymbol{\Theta}^T\{\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}})\}) - \mathrm{Vec}(\boldsymbol{\Theta}_0^T)]] \\ &\quad - \lambda_n(d-1)\hat{\boldsymbol{\omega}}^T\boldsymbol{K} \geq KL(\hat{\boldsymbol{\omega}}). \end{aligned}$$

Note that $\sum_{l=1}^{n(d-1)} \|\boldsymbol{Z}_l\|^2 = \mathrm{trace}(\boldsymbol{Z}^T\boldsymbol{Z}) \leq \lambda_{\max}(\boldsymbol{Z}^T\boldsymbol{Z})k(d-1)$, which along with central limit theorem, Condition R.1, and the second part of Condition R.2, we obtain $\|\mathrm{Vec}(\Xi^T)^T\boldsymbol{Z}\| = O_p(n^{1/2})$. From $\|\mathrm{Vec}(\Xi^T)^T\boldsymbol{Z}\| = O_p(n^{1/2})$, (A23) and (7), we can know

$$\begin{aligned} &O_p(1) + \lambda_n(d-1)k_{true} + O_p(n^{1/2})\|\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}) - \boldsymbol{\beta}_0\| \\ &\quad + \lambda_n(d-1)\hat{\boldsymbol{\omega}}^T\boldsymbol{K} \geq \underline{d}n^* \left\|\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}) - \boldsymbol{\beta}_0)\right\|^2. \end{aligned}$$

Thus, there exists $\tilde{a}_n = O_p(n)$, $\tilde{c_n} = O_p(n^{1/2})$, such that

$$\tilde{a}_n\|\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}) - \boldsymbol{\beta}_0\|^2 + \tilde{c_n}\|\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}) - \boldsymbol{\beta}_0\| \leq O_p(\lambda_n).$$

This lead to

$$\|\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}) - \boldsymbol{\beta}_0\|^2 + \frac{\tilde{c_n}}{\tilde{a}_n}\|\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}) - \boldsymbol{\beta}_0\| \leq O_p\left(\frac{\lambda_n}{\tilde{a}_n}\right),$$

and thus,

$$\left(\|\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}) - \boldsymbol{\beta}_0\| + \frac{\tilde{c_n}/2}{\tilde{a}_n}\right)^2 \leq O_p\left(\frac{\lambda_n}{\tilde{a}_n}\right) + \frac{\tilde{c_n}^2/4}{\tilde{a}_n^2},$$

which implies that

$$\|\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}}) - \boldsymbol{\beta}_0\| = O_p(n^{-1/2}\lambda_n^{1/2}).$$

This completes the proof of Theorem 4.1. ■

Proof of Theorem 4.2: The proof of Theorem 4.2 can be treated analogously to the proof of Theorem 4.1. ■