ℓ_1-penalized Multinomial Regression: Estimation, inference, and prediction, with an application to risk factor identification for different dementia subtypes

Ye Tian[1] | **Henry Rusinek[2]** | **Arjun V. Masurkar[2]** | **Yang Feng[3]**

[1]Department of Statistics, Columbia University, New York, USA

[2]Grossman School of Medicine, New York University, New York, USA

[3]Department of Biostatistics, School of Global Public Health, New York University, New York, USA

Correspondence
Yang Feng, Department of Biostatistics, School of Global Public Health, New York University, USA
Email: yang.feng@nyu.edu

Abstract

High-dimensional multinomial regression models are very useful in practice but have received less research attention than logistic regression models, especially from the perspective of statistical inference. In this work, we analyze the estimation and prediction error of the contrast-based ℓ_1-penalized multinomial regression model and extend the debiasing method to the multinomial case, providing a valid confidence interval for each coefficient and p-value of the individual hypothesis test. We also examine cases of model misspecification and non-identically distributed data to demonstrate the robustness of our method when some assumptions are violated. We apply the debiasing method to identify important predictors in the progression into dementia of different subtypes. Results from extensive simulations show the superiority of the debiasing method compared to other inference methods.

1 | INTRODUCTION

High-dimensional data, where the number of variables is comparable to or larger than the sample size, are ubiquitous in many applications, including health care and genomics. As a result, the topic has generated intensive recent research, such as the Lasso[1], the non-concave penalty including SCAD[2] and MCP[3], the general ℓ_p penalty ($0 < p \leq 1$)[4], elastic net[5], adaptive Lasso[6], group Lasso[7], variable selection consistency[8,9], and the general penalized M-estimators[10,11].

Most previous works focus on estimating the sparse parameter and predicting new observations in high-dimensional classification or regression models. In practice, a valid statistical inference procedure is also essential. However, for penalized high-dimensional models like the Lasso, inference can be very challenging because the limiting distribution of the estimator is untraceable[12]. The most popular inference method is based on random sampling or bootstrapping[13,14,15], such as the residual bootstrapping (usually for fixed design) and the vector bootstrapping (usually for random design). Another approach is the multiple sample splitting[16,17], which can be used to control both the false discovery rate (FDR) and the family-wise error rate (FWER). Post-selection inference, conducted after model selection, is another popular approach[18]. As pointed out in[19], the inference procedure needs to be adjusted to accommodate the model selection process.[20] proposed an inference method by exactly characterizing the distribution of the statistic conditioned on the model selection result. Another approach is the debiasing or desparsifying method[21,22,23], where confidence intervals and p-values are obtained by adding a correction term (derived from KKT conditions) to the Lasso estimator to eliminate the bias caused by penalization.

Many of the aforementioned approaches can be extended to generalized linear models or general M-estimators. However, the theoretical analysis of the logistic regression model (a *single-response* model) may not carry over to the high-dimensional

multinomial regression model, which is essentially a *multi-response* model. In particular, for inference, more work needs to be done to extend previous methods and theoretical analysis to the multinomial regression model.

In this paper, we focus on the following high-dimensional multinomial regression model. Suppose we have a sample $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} (X, Y)$ generated from the distribution

$$\mathbb{P}(Y = k|X = \boldsymbol{x}) = \frac{\exp((\boldsymbol{\beta}^{(k)*})^\top \boldsymbol{x})}{\sum_{k=1}^K \exp((\boldsymbol{\beta}^{(k)*})^\top \boldsymbol{x})}, \; k = 1, \ldots, K, \quad X \sim \mathbb{P}_X, \tag{1}$$

where $\boldsymbol{x}, \boldsymbol{\beta}^{(k)*} \in \mathbb{R}^p$, $k = 1 : K$, $K \geq 2$ and $\boldsymbol{\beta}^{(K)*} = \boldsymbol{0}_p$. Here, the class K is set to be the "reference class", and each $\boldsymbol{\beta}^{(k)*}$ represents the contrast coefficient between class k and the reference class. When $K = 2$, the model reduces to the logistic regression model. We assume that $\|\boldsymbol{\beta}^{(k)*}\|_0 \leq s \ll p$ for all $k = 1 : (K-1)$. This means that although there may be more features than observations, only a few contribute to the model. Under this high-dimensional setting, we impose the Lasso penalty on the contrast coefficients and solve the following optimization problem:

$$\{\widehat{\boldsymbol{\beta}}^{(k)}\}_{k=1}^{K-1} = \underset{\{\boldsymbol{\beta}^{(k)}\}_{k=1}^{K-1}}{\arg\min} \left\{ \underbrace{-\frac{1}{n} \sum_{k=1}^{K-1} \sum_{i=1}^n y_i^{(k)} (\boldsymbol{\beta}^{(k)})^\top \boldsymbol{x}_i + \frac{1}{n} \sum_{i=1}^n \log \left[\sum_{k=1}^K \exp((\boldsymbol{\beta}^{(k)})^\top \boldsymbol{x}_i) \right]}_{\mathcal{L}_n(\{\boldsymbol{\beta}^{(k)}\}_{k=1}^{K-1})} + \lambda \sum_{k=1}^{K-1} \|\boldsymbol{\beta}^{(k)}\|_1 \right\}, \tag{2}$$

where $y_i^{(k)} = \mathbb{1}(y_i = k)$ is an indicator of whether the i-th observation belongs to class k. Our goal is to estimate the sparse contrast coefficients $\{\boldsymbol{\beta}^{(k)*}\}_{k=1}^{K-1}$ and derive a valid confidence interval for each coordinate (or equivalently, deriving a *p*-value for the hypothesis test on each coordinate) by extending the debiased Lasso to the multinomial regression setting.

In many implementations of the penalized multinomial regression (for example, R package `glmnet` and Python package `scikit-learn`), an over-parameterized version of (2) is considered, where each class is assumed to be associated with a sparse coefficient vector. This is not equivalent to the contrast-based penalization problem (2). Although the over-parameterization with the penalty does not present the identifiability issue, it can be difficult to interpret these coefficient vectors in practice. As illustrated in the empirical studies in Section 3, the contrast-based penalization method outperforms the over-parameterized version by yielding lower estimation error.

1.1 | Related works

Besides the literature mentioned above, there are a few other works specifically focused on high-dimensional multinomial regression. [24] discussed the computation of penalized multinomial regression and analyzed the misclassification error rate in the binary case. [25] proposed an intuitive extension of the debiased Lasso to the multinomial case without theoretical justification. We build on the same idea to extend the debiasing method by "vectorizing the coefficient matrix in multinomial regression models" and provide a theoretical guarantee for the technique. More recently, [26] studied the misclassification error rate on sparse high-dimensional multinomial regression and derived some minimax results. Our work differs from theirs in at least two significant ways. First, they assumed the coefficient matrix with columns $\{\boldsymbol{\beta}^{(k)*}\}_{k=1}^{K-1}$ to be "row-wise sparse", meaning that many of the non-zero components of contrast coefficients associated with different classes overlap. We assume each column of the coefficient matrix, i.e., each of $\{\boldsymbol{\beta}^{(k)*}\}_{k=1}^{K-1}$, to be sparse, and their non-zero components can be very different. Second, while their focus is on prediction performance, we address estimation, prediction, and inference, with a primary focus on inference. Other works related to penalized multinomial regression include inference on Jeffreys prior-based penalized multinomial regression, which removes first-order bias [27,28], penalized semi-varying coefficient multinomial regression [29], and practical studies of penalized multinomial regression, such as those by [30].

In this paper, we focus on the high-dimensional regime $n \gg s$, $s \ll p$, and p can grow exponentially as a function of n, commonly referred to as the *sparse high-dimensional regime*. Another popular high-dimensional regime in the literature is the *fixed-ratio high-dimensional regime*, where $n/p \to c_1$, $s/p \to c_2$ with some constants $c_1, c_2 \in (0, 1)$. We briefly compare these two regimes and discuss their differences in three perspectives below[†].

[†] Thanks to one reviewer for the suggestion to add these discussions.

(i) Studies of the fixed-ratio regime can provide a precise characterization of the mean-square error (MSE) and conditions for variable selection or signal reconstruction (e.g.,[31,32,33]). Within this regime, the bias caused by penalties such as the Lasso can be explicitly quantified and managed to describe the asymptotic behavior of the penalized estimator and its MSE (e.g.,[34,35,36]). However, this often requires stronger distribution assumptions or prior knowledge about the nuisance parameters in the model, such as the covariance matrices or noise variances. Additionally, the results are typically asymptotic, and finite-sample bounds are usually challenging to obtain. In contrast, in the sparse high-dimensional regime, after removing the non-vanishing part of the bias from the penalized estimator, the remaining terms become negligible as $n \to \infty$, which allows us to obtain the asymptotic normality, making the inference procedure cleaner and more tractable. Moreover, it is easier to provide non-asymptotic results of estimation error and misclassification error in this regime.

(ii) Different applications might motivate the analysis of different regimes. For instance, in compressed sensing, signals can be dense, making the fixed-ratio regime a natural choice for study[37]. In genomic, clinical, and biomedical research, where there can be many biomarkers and diagnostic variables but limited observations from certain populations (e.g.,[38,39]), the number of variables typically exceeds the sample size, and signals are usually very sparse, fitting the sparse high-dimensional regime well. In our dementia study application (to be introduced in the next subsection), the sample size of specific dementia subtypes is very small, and the number of important predictors is generally much smaller than the total number of predictors. Therefore, we focus on the sparse high-dimensional regime in this paper.

Overall, the two regimes complement each other, and the results obtained in each regime are not generally comparable. In addition,[40] proved that an optimal adaptive inference procedure (i.e., one adaptive to unknown nuisance parameters such as the sparsity level s and the covariance matrix of X) does not exist for high-dimensional linear regression models when $n \lesssim (s \log p)^2$. Therefore, most studies on inference in the sparse high-dimensional regime (e.g.,[23,22,21,41,42]) focus on the case where $n \gg (s \log p)^2$. We conjecture that the same conclusion holds for the high-dimensional multinomial regression. In this paper, we focus on the same regime $n \gg (s \log p)^2$ and develop an adaptive inference procedure ready for practical use. If our conjecture is true, then $n \gg (s \log p)^2$ is the necessary condition for an adaptive inference procedure to work effectively, further justifying our study of this regime.

1.2 | A real example: Identifying important variables associated with different subtypes of MCI/dementia in NACC database

Before introducing the details of the inference method, we first look at a real example that motivates the development of our methodology. This example extends the study conducted by[43].

As new approaches to dementia treatment and prevention are developed, it is increasingly important to identify which group of older adults are at a high risk of developing dementia. In the past decade, research on dementia risk prediction has rapidly intensified. A recent literature review focuses on 39 studies that model late-life dementia risk with a follow-up duration of 1-39 years[44]. Age, sex, education, MMSE, neuropsychological test score, body mass index, and alcohol intake are the most common predictors included in these recently developed models. Most models have about 10 predictor variables, 4 years prediction interval, and moderate predictive ability (Figure 2,[44]).

Using an ℓ_1-penalized multinomial regression model improves upon prior work. Due to the high dimensionality of the dataset, many previous works pre-select a few potentially important variables, then fit the model only on these variables[45,46,47]. The p-values calculated on the sub-model may be unreliable since the variable selection process can be subjective. With penalization, there is no need to manually pre-select potentially important variables before modeling.

The Uniform dataset (UDS, versions 1-3) collected by the National Alzheimer Coordinating Center (NACC) from June 2005 to November 2021 is used for modeling. We are interested in identifying significant predictors for the progression from a cognitively normal (CN) status to four common subtypes of mild cognitive impairment (MCI) and dementia, which include Alzheimer's disease (AD), frontotemporal lobar degeneration (FTLD), Lewy body/Parkinson's disease (LBD/PD), and vascular brain injury (VBI). The primary etiologic diagnosis at each patient's last visit is the response variable. We include only participants who visited the center at least twice and had a normal diagnosis on their initial visit. Without further data cleaning and pre-processing, there were 12797 participants, 1936 predictors, and 1 response variable. At their follow-up visits, among these participants, 1682 (13.14%) were diagnosed with AD, 42 (0.33%) with FTLD, 133 (1.04%) were diagnosed with LBD/PD, 261 (2.04%) were diagnosed with VBI, and 10679 (83.45%) remained normal (CN). After exclusionary pre-processing steps, including removing highly correlated predictors (over 95% correlation) and predictors with more than 10% missing values, there remained

466 predictors for model fitting. We used the same strategy to impute the missing values as in[43]. Since multiple imputations were used, we applied the Licht-Rubin procedure[48,49] to combine the inference results obtained in each replication through our inference method. Due to the imbalance of different subtypes, the sample sizes of dementia/MCI subtypes like FTLD, LBD/PD, and VBI are smaller than the number of predictors, which can make the ordinary multinomial regression model without penalization very unstable.

The contrasts, odds ratios, and p-values of significant predictors between AD and CN obtained by the debiasing method (to be introduced), are reported in Table 1. The results for other subtypes compared to CN will be presented in Section 3.2. Notably, the number of days from the initial visit to the most recent visit is a strong predictor for AD against CN. With other factors fixed, a long time from the initial visit indicates a higher likelihood of progressing into AD instead of remaining normal. APOE4 allele carriage is a significant risk factor, which is well-documented[50]. Psychometric factors at baseline include vegetable and animal naming (measures of semantic fluency), Trails B (a measure of executive function and attention), and immediate recall of the Craft story (a measure of attention). Subjective cognitive decline has long been associated with an increased risk of MCI/dementia[51,52], especially when reinforced by informant report of memory decline[53], and this is also detected as an important predictor in this study.

1.3 | Organization of this paper

The remaining parts of this paper are organized as follows. In Section 2, we introduce the ℓ_1-penalized multinomial regression model and discuss the details of estimation, prediction, statistical inference, and the analysis under model misspecification and non-identically distributed data. Detailed algorithms for computation and the associated theories are provided. In Section 3, we demonstrate the effectiveness of the inference method compared to other popular approaches, such as the bootstrap method and multiple splitting method. We also compare the estimation mean square errors of our contrast-based method and the over-parameterized method, as well as the misclassification error rate of our method against other prediction benchmark methods. Moreover, we provide additional results on the NACC real dataset. In Section 4, we conclude the paper and discuss the limitations of the method, along with future research directions. Additional details omitted in the main text, extra numerical results, and all technical proofs can be found in the appendix. The tables in the main text are placed after the references, while the tables in the appendix are positioned at the end of the appendix.

1.4 | Notations

We summarize the mathematical notations here and make some remarks on certain terminologies.

For a set S, we use $|S|$ to denote its cardinality. For a vector $\boldsymbol{x} = (x_1, \ldots, x_p) \in \mathbb{R}^p$, we define its ℓ_0-pseudo-norm, ℓ_1-norm, ℓ_2-norm and ℓ_∞-norm as $\|\boldsymbol{x}\|_0 = |\{j \in [p] : x_j \neq 0\}|$, $\|\boldsymbol{x}\|_1 = \sum_{j=1}^p |x_j|$, $\|\boldsymbol{x}\|_2 = \sqrt{\sum_{j=1}^p x_j^2}$ and $\|\boldsymbol{x}\|_\infty = \max_{j=1:p} |x_j|$, respectively. For a symmetric real matrix $\boldsymbol{A}$, we use $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ to denote the smallest and largest eigenvalues of $\boldsymbol{A}$, respectively. $\mathbb{P}$ and $\mathbb{E}$ represent the probability and expectation, respectively. For two real series $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$, we write $a_n \lesssim b_n$ when there exists a sufficiently large absolute constant $C > 0$ such that $a_n \leq Cb_n$ for all n, and we write $a_n \gtrsim b_n$ when there exists a sufficiently large absolute constant $C > 0$ such that $a_n \geq Cb_n$ for all n. $a_n \ll b_n$ or $a_n = o(b_n)$ means $a_n/b_n \to 0$, and $a_n \asymp b_n$ means $a_n/b_n, b_n/a_n \leq C < \infty$. For a series of random variables $\{x_n\}_{n=1}^\infty$ and a distribution law μ, we write $x_n \xrightarrow{d} \mu$ if the distribution of x_n weakly converges to μ as $n \to \infty$. For a series of random variables $\{x_n\}_{n=1}^\infty$ and a random variable x, we write $x_n \xrightarrow{\mathbb{P}} x$ if for any $\epsilon > 0$, $\mathbb{P}(|x_n - x| > \epsilon) \to 0$ as $n \to \infty$. For two random variable series $\{x_n\}_{n=1}^\infty$ and $\{y_n\}_{n=1}^\infty$ with $y_n \neq 0$, we write $x_n = \mathcal{O}_\mathbb{P}(y_n)$ if for any $\epsilon > 0$, $\exists C > 0$, such that $\mathbb{P}(|x_n/y_n| > C) < \epsilon$ holds for all $n \geq 1$. We use $C, C', C'', c, \ldots$, to denote absolute constants which could differ from line to line.

Throughout this paper, we may not distinguish between the terms "debiased Lasso"[21,23] and "desparsified Lasso"[22]. All the confidence intervals and p-values mentioned in this paper are in the asymptotic sense. In other words, they are approximate confidence intervals and p-values, which only work well when the sample size n is large.

2 | METHODS AND THEORIES

2.1 | Estimation and prediction

The optimization problem (2) can be solved with a cyclic coordinate descent algorithm similar to the one described in [54]. Cyclic coordinate descent algorithms for ℓ_1-penalized problems like (2) have been shown to converge to the global minimizer in literature [55,56]. The iterative procedure with a fixed penalty parameter λ is detailed in Algorithm 1. In practical implementation, we start with a large λ and apply the "warm-start" strategy by calculating subsequent solutions for a decreasing sequence of λ to speed up computation. The penalty parameter λ can be chosen through a cross-validation procedure [57].

We concatenate the coefficients $\{\beta^{(k)*}\}_{k=1}^{K-1}$ into a single long vector β^* of dimension $(K-1)p$. Similarly, we concatenate the estimates $\{\widehat{\beta}^{(k)}\}_{k=1}^{K-1}$ into a $(K-1)p$-dimensional vector $\widehat{\beta}$. Given any estimate $\beta = \{\beta^{(k)}\}_{k=1}^{K-1}$, we define the conditional probability $\mathbb{P}(Y = k|X = \boldsymbol{x}) = \frac{\exp(\boldsymbol{x}^\top \beta^{(k)})}{1+\sum_{k=1}^{K-1}\exp(\boldsymbol{x}^\top \beta^{(k)})}$ as $p_k(\boldsymbol{x}; \beta)$ for $k = 1 : (K-1)$.

Algorithm 1 Cyclic coordinate descent algorithm to solve (2)

Input: Data $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$, penalty parameter λ, initial estimates $\{\widehat{\beta}^{(k)}\}_{k=1}^{K-1}$, maximum iteration number T
for $t_1 = 1$ to T **do**
   for $k = 1$ to $K-1$ **do**
      for $t_2 = 1$ to T **do**
         $w_i^{(k)} \leftarrow p_k(\boldsymbol{x}_i; \widehat{\beta})$ for $i = 1 : n$
         for $j = 1$ to p **do**
            $v_j^{(k)} \leftarrow \frac{1}{n}\sum_{i=1}^n w_i^{(k)} x_{ij}^2$
            $r_{ji}^{(k)} \leftarrow \frac{y_i^{(k)}-p_k(\boldsymbol{x}_i;\widehat{\beta})}{w_i^{(k)}}$
            $\widehat{\beta}_j^{(k)} \leftarrow \frac{\mathcal{S}\left(n^{-1}\sum_{i=1}^n w_i^{(k)} x_{ij} r_{ji}^{(k)} + v_j^{(k)}\widehat{\beta}_j^{(k)}, \lambda\right)}{v_j^{(k)}}$, where $\mathcal{S}(a, b) = \text{sign}(a)\max\{|a| - b, 0\}$
         end for
         if converged **then**
            break
         end if
      end for
   end for
   if converged **then**
      break
   end if
end for
Output: Final estimates $\{\widehat{\beta}^{(k)}\}_{k=1}^{K-1}$

The point estimate obtained from (2) enjoys certain non-asymptotic properties under specific assumptions.

Assumption 1. $\boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim} X$ with $\mathbb{E}X = \boldsymbol{0}_p$ and $\text{Cov}(X) = \Sigma_X$ which satisfies $\lambda_{\min}(\Sigma_X) \geq c > 0$ for some constant c. And for any $\boldsymbol{u} \in \mathbb{R}^p$, $\mathbb{E}e^{\boldsymbol{u}^\top X} \leq \exp(\|\boldsymbol{u}\|_2^2 C/2)$ for some constant $C > 0$, i.e., $\boldsymbol{u}^\top X$ is sub-Gaussian with a bounded variance proxy.

Assumption 2. $\max_k \|\beta^{(k)*}\|_0 \leq s \leq p$, $\max_k \|\beta^{(k)*}\|_2 \leq C < \infty$, with some constant $C > 0$.

Theorem 1 (Upper bound of the estimation error). *Let* $\lambda = C\sqrt{\frac{\log p}{n}}$ *with a large constant* $C > 0$. *Under Assumptions 1 and 2, when* $n \geq C'K^5 s\log(Kp)$ *with a large constant* $C' > 0$, *we have*

$$\|\widehat{\beta} - \beta^*\|_2 \leq C'''K^{5/2}\sqrt{\frac{s\log(Kp)}{n}}, \quad \|\widehat{\beta} - \beta^*\|_1 \leq C'''K^3 s\sqrt{\frac{\log(Kp)}{n}},$$

with probability at least $1 - C''''(Kp)^{-1}$.

To see the complete picture, we have the following lower bound for the estimation error.

Theorem 2 (Lower bound of the estimation error). *Under Assumption 1, when $n \geq CKs(\log(ep/s))(\log K)$, $\log p \geq C \log K$, and $\log p \geq C \log s$, with a large constant $C > 0$, we have*

$$\inf_{\widehat{\beta}} \sup_{\beta \in \mathcal{B}} \mathbb{P}\left(\|\widehat{\beta} - \beta\|_2 \geq C'K\sqrt{\frac{s\log(Kp)}{n}}\right) \geq \frac{1}{4}, \quad \inf_{\widehat{\beta}} \sup_{\beta \in \mathcal{B}} \mathbb{P}\left(\|\widehat{\beta} - \beta\|_1 \geq C'K^{3/2}s\sqrt{\frac{\log(Kp)}{n}}\right) \geq \frac{1}{4},$$

where $\mathcal{B} = \{\beta = \{\beta^{(k)}\}_{k=1}^{K-1} : \|\beta^{(k)}\|_0 \leq s, \max_k \|\beta^{(k)}\|_2 \leq C''\}$, and $C', C'' > 0$ are some constants.

Remark 1. The failure probability 1/4 indicates the possibility that the estimation error exceeds the lower bound, and it can be further refined to any constant smaller than 1/2. Theorem 1 establishes upper bounds that hold with probability $1 - o(1)$, while the lower bounds aim to identify the rate at which the $1 - o(1)$ probability does not hold (i.e., there is a constant probability of failure). Thus, the smaller the gap between the upper and lower bounds, the closer the method is to the optimal one. Comparing Theorems 1 and 2, we observe that the dependence on n, s, and p matches in upper and lower bounds. However, the dependence on the number of classes K in Theorem 1 may not be optimal.

Remark 2. It might be surprising that the order of K in the ℓ_2-estimation error is greater than 1/2. Intuitively, there are n samples, and we are estimating a $(K-1)p$-dimensional parameter with at most $(K-1)s$ non-zero components. According to the classical Lasso theory, the optimal ℓ_2-estimation error of β should be of order $\sqrt{\frac{Ks\log(Kp)}{n}}$. An intuitive interpretation is that the samples of different classes do not contribute equally to the estimation of all $\beta^{(k)*}$'s. This can be seen from the expression of the conditional distribution (1). The denominator on the right-hand side of (1) is the same for different k's, but the numerator only contains one specific $\beta^{(k)*}$. In the simplest setting, suppose the sample sizes of different classes are the same. If the contribution of samples from class k to the estimation of $\beta^{(k)*}$ dominates the contribution of samples from other classes, then the ℓ_2-estimation error of $\beta^{(k)*}$ could be $\mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{s\log p}{n/K}}\right) = \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{Ks\log(Kp)}{n}}\right)$ because the effective sample size is n/K instead of n and $\log p \gtrsim \log K$. Summing up the errors of all $(K-1)$ $\beta^{(k)*}$'s, we obtain the rate $\mathcal{O}_{\mathbb{P}}\left(K\sqrt{\frac{s\log(Kp)}{n}}\right)$ for the ℓ_2-estimation error bound.

With the estimate $\widehat{\beta} = \{\widehat{\beta}^{(k)}\}_{k=1}^{K-1}$, we can construct a classifier to predict the label for a new observation x:

$$\mathcal{C}_{\widehat{\beta}}(x) = \arg\max_{k=1:K} p_k(x; \widehat{\beta}).$$

It is well-known that the Bayes classifier $\mathcal{C}_{\beta^*}(x) = \arg\max_{k=1:K} p_k(x; \beta^*)$ achieves the smallest misclassification error rate. For any classifier $\mathcal{C}$, denote the misclassification error rate $\mathbb{P}(\mathcal{C}(X) \neq Y)$ as $R(\mathcal{C})$. The difference between the error rates of a classifier $\mathcal{C}$ and the Bayes classifier $\mathcal{C}_{\beta^*}$ is called the excess misclassification error rate. We have the following upper bound for the *excess misclassification error rate* of $\mathcal{C}_{\widehat{\beta}}$.

Theorem 3 (Upper bound of the excess misclassification error). *Under the same conditions imposed in Theorem 1, we have*

$$R(\mathcal{C}_{\widehat{\beta}}) - R(\mathcal{C}_{\beta^*}) \leq CK^{5/2}\sqrt{\frac{s\log(Kp)}{n}},$$

with probability at least $1 - C'(Kp)^{-1}$, where $C, C' > 0$ are some constants.

2.2 | Statistical inference

As we mentioned in the introduction, it is well-known that the Lasso estimator is biased, with its bias being non-negligible compared to its variance and difficult to quantify. Therefore, the Lasso estimator (2) cannot be directly used to construct a confidence interval or conduct a hypothesis test. In this paper, we modify the desparsified Lasso proposed in [22] to adapt to the multinomial regression setting, allowing for valid statistical inference.

For presentation convenience, we denote $\Sigma = \mathbb{E}[\nabla^2 \mathcal{L}_n(\beta^*)]$ and $\Theta = \Sigma^{-1}$, both of which play an essential role in the statistical inference. Next, we derive the debiased Lasso estimator in a heuristic way.

Since (2) is a convex optimization problem and the objective function is coercive, the optimal solution $\widehat{\beta}$ must satisfy the first-order condition

$$\nabla \mathcal{L}_n(\widehat{\beta}) + \lambda\kappa = \mathbf{0}_{(K-1)p}, \quad \kappa = (\kappa_1, \ldots, \kappa_{(K-1)p})^\top, \tag{3}$$

where $\kappa_j = \text{sign}(\widehat{\beta}_j)$ when $\widehat{\beta}_j \neq 0$ and $\kappa_j \in \partial|\beta|\big|_{\beta=0} = [-1, 1]$. $\partial|\beta|\big|_{\beta=0}$ is the subgradient of $|\cdot|$ at 0. By plugging the expression of $\nabla\mathcal{L}_n(\widehat{\beta})$ based on (2), we have

$$
\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} [p_1(\boldsymbol{x}_i;\widehat{\beta}) - p_1(\boldsymbol{x}_i;\beta^*)]\boldsymbol{x}_i \\ \vdots \\ [p_{K-1}(\boldsymbol{x}_i;\widehat{\beta}) - p_{K-1}(\boldsymbol{x}_i;\beta^*)]\boldsymbol{x}_i \end{pmatrix} - \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} \epsilon_i^{(1)}\boldsymbol{x}_i \\ \vdots \\ \epsilon_i^{(K-1)}\boldsymbol{x}_i \end{pmatrix} + \lambda\boldsymbol{\kappa} = \mathbf{0}_{(K-1)p},
$$

where we define the "residual" of the i-th observation corresponding to class k as $\epsilon_i^{(k)} = y_i^{(k)} - p_k(\boldsymbol{x}_i;\beta^*)$, with $y_i^{(k)} = \mathbb{1}(y_i = k)$, $i = 1:n$, $k = 1:(K-1)$. By (3), we replace $\lambda\boldsymbol{\kappa}$ with $-\nabla\mathcal{L}_n(\widehat{\beta})$ in the equation above, apply the mean-value theorem, and use vectorized notations to obtain

$$
\boldsymbol{B}_n(\widehat{\beta} - \beta^*) + \frac{1}{n}\begin{pmatrix} \boldsymbol{X}^\top(\boldsymbol{y}^{(1)} - \boldsymbol{p}_1(\boldsymbol{X};\widehat{\beta})) \\ \vdots \\ \boldsymbol{X}^\top(\boldsymbol{y}^{(K-1)} - \boldsymbol{p}_{K-1}(\boldsymbol{X};\widehat{\beta})) \end{pmatrix} = \frac{1}{n}\begin{pmatrix} \boldsymbol{X}\boldsymbol{\epsilon}^{(1)} \\ \vdots \\ \boldsymbol{X}\boldsymbol{\epsilon}^{(K-1)} \end{pmatrix}, \tag{4}
$$

where $\boldsymbol{X} = (\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_n^\top) \in \mathbb{R}^{n \times p}$ is the concatenated predictor matrix, $\boldsymbol{y}^{(k)} = (y_1^{(k)}, \ldots, y_n^{(k)})^\top \in \mathbb{R}^n$ is the concatenated response vector corresponding to class k, $\boldsymbol{p}_k(\boldsymbol{X}, \widehat{\beta}) = (p_k(\boldsymbol{x}_1, \widehat{\beta}), \ldots, p_k(\boldsymbol{x}_n, \widehat{\beta}))^\top$ is the concatenated conditional probability for class k, and $\boldsymbol{\epsilon}^{(k)} := (\epsilon_1^{(k)}, \ldots, \epsilon_n^{(k)})^\top$ is the concatenated residual vector corresponding to class k, with $k = 1:(K-1)$. Here, $\boldsymbol{B}_n = \frac{1}{n}\sum_{i=1}^{n}\int_0^1 \boldsymbol{B}(\boldsymbol{x}_i;\beta^* + t(\widehat{\beta} - \beta^*))dt$, where $\forall \boldsymbol{x} \in \mathbb{R}^p$, $\beta = \{\beta^{(k)}\}_{k=1}^{K-1} \in \mathbb{R}^{(K-1)p}$,

$$
\boldsymbol{B}(\boldsymbol{x};\beta) := \begin{pmatrix} p_1(\boldsymbol{x};\beta)[1 - p_1(\boldsymbol{x};\beta)]\boldsymbol{x}\boldsymbol{x}^\top & -p_1(\boldsymbol{x};\beta)p_2(\boldsymbol{x};\beta)\boldsymbol{x}\boldsymbol{x}^\top & \ldots & -p_1(\boldsymbol{x};\beta)p_{K-1}(\boldsymbol{x};\beta)\boldsymbol{x}\boldsymbol{x}^\top \\ -p_1(\boldsymbol{x};\beta)p_2(\boldsymbol{x};\beta)\boldsymbol{x}\boldsymbol{x}^\top & p_2(\boldsymbol{x};\beta)[1 - p_2(\boldsymbol{x};\beta)]\boldsymbol{x}\boldsymbol{x}^\top & \ldots & -p_2(\boldsymbol{x};\beta)p_{K-1}(\boldsymbol{x};\beta)\boldsymbol{x}\boldsymbol{x}^\top \\ \vdots & \vdots & \vdots & \vdots \\ -p_1(\boldsymbol{x};\beta)p_{K-1}(\boldsymbol{x};\beta)\boldsymbol{x}\boldsymbol{x}^\top & -p_2(\boldsymbol{x};\beta)p_{K-1}(\boldsymbol{x};\beta)\boldsymbol{x}\boldsymbol{x}^\top & \ldots & p_{K-1}(\boldsymbol{x};\beta)[1 - p_{K-1}(\boldsymbol{x};\beta)]\boldsymbol{x}\boldsymbol{x}^\top \end{pmatrix} \in \mathbb{R}^{[(K-1)p] \times [(K-1)p]}.
$$
$$\tag{5}$$

With some algebraic calculations, it is not hard to show that $\mathbb{E}[\boldsymbol{B}(X;\beta)] = \Sigma$. Heuristically, since $\|\widehat{\beta} - \beta^*\|_2, \|\widehat{\beta} - \beta^*\|_1 \xrightarrow{\mathbb{P}} 0$ due to Theorem 1, by some concentration inequalities, we would expect $\boldsymbol{B}_n(\widehat{\beta} - \beta^*) \approx \widetilde{\boldsymbol{B}}_n(\widehat{\beta} - \beta^*) \approx \Sigma(\widehat{\beta} - \beta^*)$ in the entry-wise convergence sense with high probability, where $\widetilde{\boldsymbol{B}}_n := \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{B}(\boldsymbol{x}_i;\beta^*)$. Combing with (4), it implies

$$
\widehat{\beta} \approx \beta^* - \frac{1}{n}\Theta\begin{pmatrix} \boldsymbol{X}^\top(\boldsymbol{y}^{(1)} - \boldsymbol{p}_1(\boldsymbol{X};\widehat{\beta})) \\ \vdots \\ \boldsymbol{X}^\top(\boldsymbol{y}^{(K-1)} - \boldsymbol{p}_{K-1}(\boldsymbol{X};\widehat{\beta})) \end{pmatrix} + \frac{1}{n}\Theta\begin{pmatrix} \boldsymbol{X}\boldsymbol{\epsilon}^{(1)} \\ \vdots \\ \boldsymbol{X}\boldsymbol{\epsilon}^{(K-1)} \end{pmatrix}. \tag{6}
$$

Noticing that the last term on the RHS of (6) has mean zero, it motivates us to define the "debiased" Lasso estimator by removing the second bias term as

$$
\widehat{\boldsymbol{b}} := \widehat{\beta} + \frac{1}{n}\Theta\begin{pmatrix} \boldsymbol{X}^\top(\boldsymbol{y}^{(1)} - \boldsymbol{p}_1(\boldsymbol{X};\widehat{\beta})) \\ \vdots \\ \boldsymbol{X}^\top(\boldsymbol{y}^{(K-1)} - \boldsymbol{p}_{K-1}(\boldsymbol{X};\widehat{\beta})) \end{pmatrix}.
$$

In fact, we can make the previous heuristics rigorous and show that

$$
\sqrt{n}(\widehat{b}_j - \beta_j^*) = \text{remaining bias terms} + \frac{1}{\sqrt{n}}\Theta_j^\top\begin{pmatrix} \boldsymbol{X}\boldsymbol{\epsilon}^{(1)} \\ \vdots \\ \boldsymbol{X}\boldsymbol{\epsilon}^{(K-1)} \end{pmatrix}, \tag{7}
$$

where remaining bias terms refer to the small offset terms caused by the previous "$\approx$" argument. Note that by Lindeberg-Feller central limit theorem (e.g., Theorem 3.4.10 in[58]), we have

$$
\frac{1}{\sqrt{n}\sqrt{\Theta_j^\top \Sigma \Theta_j}}\Theta_j^\top\begin{pmatrix} \boldsymbol{X}\boldsymbol{\epsilon}^{(1)} \\ \vdots \\ \boldsymbol{X}\boldsymbol{\epsilon}^{(K-1)} \end{pmatrix} \xrightarrow{\text{d}} N(0, 1), \quad n \to \infty.
$$

As we will show later, in the regime that we are interested in, where $n \gg s^2(\log p)^2$, the remaining bias terms in (7) converge to zero in probability as $n \to \infty$. This leads to the asymptotic normality

$$\frac{\sqrt{n}(\widehat{b}_j - \beta_j^*)}{\sqrt{\Theta_j^\top \Sigma \Theta_j}} \xrightarrow{d} N(0, 1), \tag{8}$$

which can be used to construct confidence intervals for β_j^*'s and test statistics of the hypothesis test with the null $H_0 : \beta_j^* = 0$. Note that the remaining bias terms are negligible due to the sparsity assumption $n \gg s^2(\log p)^2$. As discussed in Section 1.1, this is not true in the fixed-ratio regime where n/p converges to a fixed constant [34]. The clean asymptotic normality (8) benefits from the sparse regime we study.

In practice, since Θ and Σ are unknown, we have to estimate them. For Σ, we can use the empirical estimator $\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n \boldsymbol{B}(\boldsymbol{x}_i; \widehat{\boldsymbol{\beta}})$. For Θ, we estimate each row via the nodewise regression [59]. For convenience, we denote

$$\Theta = \operatorname{diag}((\tau_1^*)^{-2}, \ldots, (\tau_{(K-1)p}^*)^{-2}) \begin{pmatrix} 1 & -\gamma_{1,1}^* & -\gamma_{1,2}^* & \cdots & -\gamma_{1,(K-1)p-1}^* \\ -\gamma_{2,1}^* & 1 & -\gamma_{2,2}^* & \cdots & -\gamma_{2,(K-1)p-1}^* \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -\gamma_{(K-1)p-1,1}^* & -\gamma_{(K-1)p-1,2}^* & -\gamma_{(K-1)p-1,3}^* & \cdots & 1 \end{pmatrix},$$

where $\gamma_j^* = (\gamma_{j,1}^*, \ldots, \gamma_{j,(K-1)p-1}^*)^\top = (\Sigma_{-j,-j})^{-1}\Sigma_{-j,j}$ by the formula of the inverse block matrix. Note that, unlike the case of GLMs [22], here it is difficult to estimate γ_j^* through regression on the j-th column of a weighted X matrix versus the other columns. Instead, we solve the following quadratic program (QP):

$$\widehat{\gamma}_j = \underset{\gamma \in \mathbb{R}^{(K-1)p-1}}{\arg\min} \left\{ \widehat{\Sigma}_{j,j} - \widehat{\Sigma}_{-j,j}\gamma + \frac{1}{2}\gamma^\top \widehat{\Sigma}_{-j,-j}\gamma + \lambda_j \|\gamma\|_1 \right\}, \tag{9}$$

and let

$$\widehat{\tau}_j^2 = \widehat{\Sigma}_{j,j} - \widehat{\Sigma}_{-j,j}\widehat{\gamma}_j,$$

for $j = 1, \ldots, (K-1)p$. There are many ways to solve this QP problem. For example, we can rewrite it as a standard QP and solve it using any QP solvers. For other methods, see the introduction part of [60]. Here we solve it by the coordinate descent algorithm outlined in Algorithm 2.

Algorithm 2 Coordinate descent algorithm to solve (9)

Input: Empirical estimate $\widehat{\Sigma}$, penalty parameter λ_j, maximum iteration number T, initial estimate $\widehat{\gamma}_j$, index j
for $t = 1$ to T **do**
   for $r = 1$ to $(K-1)p - 1$ **do**
      $\widehat{\gamma}_{j,r} \leftarrow \mathcal{S}\big((\widehat{\Sigma}_{-j,j})_r - (\widehat{\Sigma}_{-j,-j})_{r,-r}\widehat{\gamma}_{j,-r}, \lambda\big)/\widehat{\Sigma}_{j,j}$, where $(\widehat{\Sigma}_{-j,j})_r$ is the r-th component of $\widehat{\Sigma}_{-j,j}$, $(\widehat{\Sigma}_{-j,-j})_{r,-r}$
is the r-th row of $\widehat{\Sigma}_{-j,-j}$ without the r-th column, and $\mathcal{S}(a, b) = \operatorname{sign}(a)\max\{|a| - b, 0\}$
   end for
   if converged **then**
      break
   end if
end for
Output: Final estimates $\widehat{\gamma}_j$

Note that Algorithm 2 only shows the procedure to calculate the solution to (9) for one j and one specific λ_j. In practice, we run this algorithm for different λ_j choices and iterate from $j = 1$ to $j = (K-1)p$, and the iteration can be parallelized to speed up computation in practice. Each λ_j can be selected through a cross-validation procedure.

With the estimate $\widehat{\Sigma}$ and $\widehat{\Theta}$, if the estimated variance $\widehat{\Theta}_j^\top \widehat{\Sigma} \widehat{\Theta}_j$ converges to the true asymptotic variance $\Theta_j^\top \Sigma \Theta_j$ in (8) in probability, then by (8) and the Slutsky's lemma, we have

$$\frac{\sqrt{n}(\widehat{b}_j - \beta_j^*)}{\sqrt{\widehat{\Theta}_j^\top \widehat{\Sigma} \widehat{\Theta}_j}} \xrightarrow{d} N(0, 1), \tag{10}$$

which can be used to build the confidence interval for β_j^* and compute the p-value for the hypothesis testing problem $H_0 : \beta_j^* = 0$ v.s. $H_1 : \beta_j^* \neq 0$. We summarize the procedures of CI construction and p-value calculation in Algorithms 3 and 4, respectively.

Algorithm 3 Construct the $(1 - \alpha)$-confidence interval of β_j^*

Input: Debiased estimate $\widehat{b}$, empirical estimate (of Σ) $\widehat{\Sigma}$, estimate (of Θ) $\widehat{\Theta}$ calculated from (9), index j, CI level $(1 - \alpha)$

Output: $(1 - \alpha)$-confidence interval $\left[\widehat{b}_j - q_{\alpha/2}\sqrt{\frac{\widehat{\Theta}_j^\top \widehat{\Sigma} \widehat{\Theta}_j}{n}}, \widehat{b}_j + q_{\alpha/2}\sqrt{\frac{\widehat{\Theta}_j^\top \widehat{\Sigma} \widehat{\Theta}_j}{n}} \right]$, where $q_{\alpha/2}$ is the $\alpha/2$-upper tail quantile of $N(0, 1)$

Algorithm 4 Calculate p-value for the hypothesis test $H_0 : \beta_j^* = 0$ v.s. $H_1 : \beta_j^* \neq 0$

Input: Debiased estimate $\widehat{b}$, empirical estimate (of Σ) $\widehat{\Sigma}$, estimate (of Θ) $\widehat{\Theta}$ calculated from (9), index j, significance level α

Output: p-value $2\Phi\left(-\sqrt{n}|\widehat{b}_j|/\sqrt{\widehat{\Theta}_j^\top \widehat{\Sigma} \widehat{\Theta}_j} \right)$, where Φ is the CDF of $N(0, 1)$

We are now ready to rigorously state (10). Before doing so, we impose the following sparsity assumption on Θ, which is necessary for the nodewise regression (9) to work. Similar conditions can be found in [22,41,61].

Assumption 3. Suppose the following conditions hold for $\Sigma = \mathbb{E}[\nabla^2 \mathcal{L}_n(\beta^*)]$ and $\Theta = \Sigma^{-1}$:

(i) $\max_{j=1:[(K-1)p]} \|\Theta_j\|_0 \leq K s_0$, $\max_{j=1:[(K-1)p]} \|\Theta_j\|_2 \leq C < \infty$ with some constant $C > 0$;

(ii) $\max_{i=1:n} \max_{k=1:(K-1)} |x_i^\top \beta^{(k)*}| \leq C' < \infty$ a.s. with some constant $C' > 0$.

Remark 3. The sparsity assumption of Θ_j in Condition (i) can be removed for a modified inference procedure, which, however, requires sample spitting. Condition (ii) can also be removed at the cost of a more stringent sample size requirement. Interested readers can find the details in Sections A and B of the appendix.

We are now ready to present the following theorem, which characterizes the asymptotic distribution of $\widehat{b}_j$.

Theorem 4. *Let* $\lambda = C\sqrt{\frac{\log(Kp)}{n}}$ *in (2) and* $\lambda_j = C\sqrt{\frac{\log(Kp)}{n}}$ *in (9) with a large constant $C > 0$. Under Assumptions 1-3, when* $n \gg K^{18} s^2 (\log(Kp))^2 (\log(np))^2 + K^{10} s_0^2 (\log(Kp))^2$, *we have*

$$\frac{\sqrt{n}(\widehat{b}_j - \beta_j^*)}{\sqrt{\widehat{\Theta}_j^\top \widehat{\Sigma} \widehat{\Theta}_j}} \xrightarrow{d} N(0, 1),$$

for $j = 1, \ldots, (K-1)p$.

Remark 4. As discussed in Section 1.1, [40] proved that the adaptive inference procedure does not exist unless $n \gg (s \log p)^2$ for high-dimensional linear regression models. We conjecture that a similar conclusion holds for multinomial regression models when K is fixed. If this is true, when $s_0 \lesssim s$, the current sample size requirement in Theorem 4 is optimal up to a $(\log(np))^2$ factor, because our inference procedure does not rely on the knowledge of s or Σ_X and is thus adaptive. However, the sample size

requirement may not be optimal when K grows with n. Interested readers are encouraged to relax the condition by improving our proof in the appendix.

We also want to comment on the impact of estimation error of Σ and Θ on inference accuracy. As noted in our previous heuristic analysis, by (8) and the Slutsky's lemma, it suffices to find estimators $\widehat{\Sigma}$ and $\widehat{\Theta}$ that can consistently estimate the asymptotic variance (even at an arbitrarily slow rate), i.e. $|\widehat{\Theta}_j^\top \widehat{\Sigma} \widehat{\Theta}_j - \Theta_j^\top \Sigma \Theta_j| \overset{\mathbb{P}}{\to} 0$. In the proof of Theorem 4, we show that $|\widehat{\Theta}_j^\top \widehat{\Sigma} \widehat{\Theta}_j - \Theta_j^\top \Sigma \Theta_j|$ can be bounded by some term involving $\|\widehat{\Theta}_j - \Theta_j\|_1$, $\|\widehat{\Theta}_j - \Theta_j\|_2$, and $\|\widehat{\Sigma} - \Sigma\|_{\max}$. It can be derived that $\|\widehat{\Theta}_j - \Theta_j\|_1$, $\|\widehat{\Theta}_j - \Theta_j\|_2$, and $\|\widehat{\Sigma} - \Sigma\|_{\max}$ all converge to zero in probability, under the current assumptions. In this way, the estimation errors of Σ and Θ do not explicitly appear in $n^{-1/2}$-inference rate. However, the estimation errors of Σ and Θ still have some impact on the inference accuracy by contributing higher-order terms that shrink faster than $n^{-1/2}$, which are ignored in the use of Slutsky's lemma. Hence, as one reviewer pointed out, estimating the high-dimensional covariance matrix Σ and its inverse Θ is very difficult in general, but we only need consistency of the asymptotic variance estimator (where the estimation error rate can be arbitrarily slow) for the asymptotic normality to be true, which indeed holds under current assumptions.

2.3 | Model misspecification and non-identically distributed data

In the previous sections, we assumed that the data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} (X, Y)$ and y_i given $\boldsymbol{x}_i$ is generated from a multinomial regression model, which might be too restrictive in practice. In this subsection, we will investigate the performance of ℓ_1-penalized multinomial regression 1 and the debiased Lasso estimator 3 in two situations:

(i) The model is misspecified. More specifically, the data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} (X, Y)$ is generated from an unknown distribution, where Y given X is not necessarily generated from a multinomial regression model.
(ii) The data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ is independent but not identically distributed. More specifically, y_i given $\boldsymbol{x}_i$ is generated from a multinomial regression model, but $\boldsymbol{x}_i$'s are not necessarily generated from the same distribution.

Note that it is possible to consider a mix of two situations, but for simplicity, we will discuss each scenario individually.

2.3.1 | Model misspecification

First, we consider case (i), where the distribution of Y given X is misspecified. Recall the conditional probability $p_k(\boldsymbol{x}; \boldsymbol{\beta}) = \frac{\exp(\boldsymbol{x}^\top \boldsymbol{\beta}^{(k)})}{1 + \sum_{k=1}^{K-1} \exp(\boldsymbol{x}^\top \boldsymbol{\beta}^{(k)})}$ under the multinomial regression model, for $\boldsymbol{\beta} = \{\boldsymbol{\beta}^{(k)}\}_{k=1}^{K-1}$ and $k = 1 : (K-1)$. Consider the best approximater

$$\boldsymbol{\beta}^* := \underset{\boldsymbol{\beta} = \{\boldsymbol{\beta}^{(k)}\}_{k=1}^{K-1} \in \mathbb{R}^{(K-1)p}}{\arg \min} \mathbb{E}[-\log p_Y(X; \boldsymbol{\beta})], \tag{11}$$

where the expectation is with respect to the true underlying distribution of (X, Y). Minimizing $\mathbb{E}[-\log p_Y(X; \boldsymbol{\beta})]$ can be shown to be equivalent to minimizing the cross-entropy or the Kullback-Leibler divergence between the true distribution and the multinomial logistic regression model with any marginal distribution of X. We will show that the same conclusions derived in the previous sections hold for the best approximator $\boldsymbol{\beta}^*$, when the model is misspecified.

Since Y given X is not necessarily generated from the multinomial regression model, we need to replace the empirical estimator $\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n \boldsymbol{B}(\boldsymbol{x}_i; \widehat{\boldsymbol{\beta}})$ used in Section 2.2 by $\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n \widehat{\boldsymbol{B}}(y_i, \boldsymbol{x}_i; \widehat{\boldsymbol{\beta}})$ with $\widehat{\boldsymbol{\beta}} = \{\widehat{\boldsymbol{\beta}}^{(k)}\}_{k=1}^{K-1}$ from (2), where for any $\boldsymbol{x} \in \mathbb{R}^p$, $y \in [K]$, and $\boldsymbol{\beta} = \{\boldsymbol{\beta}^{(k)}\}_{k=1}^{K-1} \in \mathbb{R}^{(K-1)p}$,

$$\widehat{\boldsymbol{B}}(y, \boldsymbol{x}; \boldsymbol{\beta}) := \begin{pmatrix} \boldsymbol{x}\boldsymbol{x}^\top \widehat{\epsilon}_1^2(y, \boldsymbol{x}, \boldsymbol{\beta}) & \boldsymbol{x}\boldsymbol{x}^\top \widehat{\epsilon}_1(y, \boldsymbol{x}, \boldsymbol{\beta}) \widehat{\epsilon}_2(y, \boldsymbol{x}, \boldsymbol{\beta}) & \dots & \boldsymbol{x}\boldsymbol{x}^\top \widehat{\epsilon}_1(y, \boldsymbol{x}, \boldsymbol{\beta}) \widehat{\epsilon}_{K-1}(y, \boldsymbol{x}, \boldsymbol{\beta}) \\ \boldsymbol{x}\boldsymbol{x}^\top \widehat{\epsilon}_1(y, \boldsymbol{x}, \boldsymbol{\beta}) \widehat{\epsilon}_2(y, \boldsymbol{x}, \boldsymbol{\beta}) & \boldsymbol{x}\boldsymbol{x}^\top \widehat{\epsilon}_2^2(y, \boldsymbol{x}, \boldsymbol{\beta}) & \dots & \boldsymbol{x}\boldsymbol{x}^\top \widehat{\epsilon}_2(y, \boldsymbol{x}, \boldsymbol{\beta}) \widehat{\epsilon}_{K-1}(y, \boldsymbol{x}, \boldsymbol{\beta}) \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{x}\boldsymbol{x}^\top \widehat{\epsilon}_1(y, \boldsymbol{x}, \boldsymbol{\beta}) \widehat{\epsilon}_{K-1}(y, \boldsymbol{x}, \boldsymbol{\beta}) & \boldsymbol{x}\boldsymbol{x}^\top \widehat{\epsilon}_2(y, \boldsymbol{x}, \boldsymbol{\beta}) \widehat{\epsilon}_{K-1}(y, \boldsymbol{x}, \boldsymbol{\beta}) & \dots & \boldsymbol{x}\boldsymbol{x}^\top \widehat{\epsilon}_{K-1}^2(y, \boldsymbol{x}, \boldsymbol{\beta}) \end{pmatrix},$$

$$\widehat{\epsilon}_k(y, \boldsymbol{x}, \boldsymbol{\beta}) := y^{(k)} - p_k(\boldsymbol{x}; \boldsymbol{\beta}), \quad \text{with} \quad y^{(k)} := \mathbb{1}(y = k).$$

Then, we can show that the inference procedure leads to valid inference for the best approximator $\boldsymbol{\beta}^*$ when the model is misspecified.

Theorem 5. *Under Assumptions 1-3, Theorems 1, 3, and 4 still hold for β^* defined in (11).*

In the low-dimensional case, the convergence of MLE to the best approximator defined in (11) under model misspecification has been extensively studied in the classical inference literature (e.g.,[62,63,64,65]).[41] provides a similar result for a score test-based inference procedure for high-dimensional linear regression. Their result relies on the assumption that $Y - \mathbb{E}[Y|X]$ is sub-Gaussian with a bounded variance proxy. In the context of multinomial regression models, the sub-Gaussianity is implied automatically by the boundedness of $Y - \mathbb{E}[Y|X]$; thus, no additional assumption is required. Our result complements this line of literature by addressing the high-dimensional multinomial regression case.

2.3.2 | Non-identically distributed data

Next, we consider case (ii), where the data $\{(x_i, y_i)\}_{i=1}^n$ is independent but not identically distributed. In this case, we redefine the covariance matrix of predictors by the average as $\Sigma_X := \frac{1}{n} \sum_{i=1}^n \mathbb{E}(x_i x_i^\top)$, and the covariance matrix for the multinomial model by the average as $\Sigma := \frac{1}{n} \sum_{i=1}^n \mathbb{E}[B(x_i; \beta^*)]$, where $B(x_i; \beta^*)$ is defined in (5).

In the case of non-identically distributed data, we need to replace Assumption 1 with the following modified assumption.

Assumption 4. For all $i = 1 : n$, $\mathbb{E}x_i = 0_p$ and $\text{Cov}(x_i)$ satisfies $\lambda_{\min}(\text{Cov}(x_i)) \geq c > 0$ for some constant c. And for all $i = 1 : n$ and any $u \in \mathbb{R}^p$, $\mathbb{E}e^{u^\top \text{Cov}(x_i)} \leq \exp(\|u\|_2^2 C/2)$ for some constant $C > 0$.

The first observation is that the non-asymptotic analysis used to prove the estimation error and misclassification error bounds in Theorems 1 and 3 can still apply with the redefined Σ_X and Σ, because the main tools we used are concentration inequalities for sub-Gaussian and sub-exponential variables which can apply to the sum of non-identically distributed variables. Hence, we can obtain the same bounds as in Theorems 1 and 3. The second remark is that the asymptotic normality for the debiased Lasso estimator in Theorem 4 was proved by Lindeberg-Feller central limit theorem, which applies to the sum of non-identically distributed variables satisfying certain conditions. It is not difficult to show that the required conditions (often called Lindeberg's conditions) indeed hold; therefore, the same results in Theorem 4 can be obtained.

We summarize these intuitions into the following theorem.

Theorem 6. *Under Assumption 2, 3, and 4, Theorems 1, 3, and 4 still hold for the non-identically distributed data $\{(x_i, y_i)\}_{i=1}^n$.*

Theorem 6 takes the heterogeneity of the data into account without requiring the data to be i.i.d. generated from the same distribution. The heterogeneity in data is very common in many applications, such as healthcare and medicine[66], finance and economics[67], and social sciences[68]. A further interesting extension is to deal with dependent observations. There are a few recent studies on the high-dimensional regression for dependent data with certain structures[69,70,71], and most of them focus on estimation and prediction. It remains unknown how to generalize our framework for high-dimensional multinomial regression to the dependent case.

3 | NUMERICAL STUDIES

We have presented partial results of the application of ℓ_1-penalized multinomial regression on the NACC dataset in Section 1.2. In this section, we explore the approach under different simulation settings and discuss the remaining real-data results. Our focus will be on statistical inference, for example, confidence interval (CI) construction and hypothesis testing.

Under each model, we will first compare the estimation error of our contrast-based parameterization and the over-parameterization approach[54] on the sparse contrast coefficients. Then, we will compare the misclassification error rates of the contrast-based ℓ_1-penalized multinomial regression and other benchmark methods, including Support Vector Machines (SVMs), Linear Discriminant Analysis (LDA), Random Forests, and k-Nearest Neighbors (kNN). We fit SVMs with RBF kernel and run kNN with the number of neighbors $k = \lfloor \sqrt{n} \rfloor$, i.e., the largest integer that is smaller than or equal to $\sqrt{n}$. We will also compare the CIs produced by the debiased-Lasso method with the CIs produced by the vector bootstrap method[15] in terms of coverage probability and length. In addition, we will compare the debiased-Lasso method with the vector bootstrap method in individual testing by the type-I error rate and the power. Finally, we adjust the p-values produced by the debiased-Lasso with a Bonferroni correction and compare the method with the multiple splitting approach[17], for multiple testing. We will describe the evaluation metrics used in detail, including average coverage probability, type-I error rate, and family-wise error rate (FWER), in Section 3.1.1 for

Model 1. For Models 2-8, we only state the main findings, and the details of these metrics are omitted. Due to space constraints, we only display the results of Models 1-4 in the main text, and leave the results of Models 5-8 to Section C.1 of the appendix.

All experiments are conducted in R[72]. The contrast-based ℓ_1-penalized multinomial regression with the debiasing method is implemented in our new R package `pemultinom`, which is available on CRAN. The over-parameterized ℓ_1-penalized multinomial regression is implemented through package `glmnet`[54]. We use cross-validation to find the best penalty parameter in all penalized regressions, including the estimation of the contrast coefficients and the nodewise regression. For the vector bootstrap method, the same number of observations are resampled with replacement from the data in each of the 200 replications. The empirical quantiles are used to calculate the CI[15]. We use vector bootstrap because it works better than the residual resampling bootstrap under random design[13]. For the multiple-splitting method[17], 200 splits are used in each simulation. SVMs, Random Forests, and kNN are implemented using R packages `e1071`, `randomForests`, and `caret`, respectively. All the codes to reproduce the results in this paper can be found at https://github.com/ytstat/multinom-inf.

3.1 | Simulations

3.1.1 | Model 1

In this model, for independent observations $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, we first generate the predictor $\boldsymbol{x}_i$ from $N(\boldsymbol{0}_p, \boldsymbol{\Sigma}_{p \times p})$ with $\boldsymbol{\Sigma} = (0.75^{|i-j|})_{p \times p}$, then generate y_i from (1) with $K = 3$ classes and contrast coefficients $\boldsymbol{\beta}^{(1)*} = (1, 1, 1, \boldsymbol{0}_{p-3})$ and $\boldsymbol{\beta}^{(2)*} = (0, 0, 0, 1, 1, 1, \boldsymbol{0}_{p-6})$, where $p = 200$. We consider different sample size settings $n = 100, 200, 400$.

First, we compare the contrast-based penalization and the over-parameterization method in terms of the sum of squared ℓ_2-estimation error of $\boldsymbol{\beta}^{(1)*}$ and $\boldsymbol{\beta}^{(2)*}$ in Table 2. It can be seen that the contrast-based penalization has lower estimation errors than the over-parameterization method.

We also compare the prediction performance of the contrast-based ℓ_1-penalized multinomial regression with other benchmarks such as Support Vector Machines (SVMs), Linear Discriminant Analysis (LDA), Random Forests, and k-Nearest Neighbors (kNN). The average misclassification error rates of different methods computed on an independent test set of size 1000 over 200 replications are reported in Table 3. We can see that the ℓ_1-penalized multinomial regression performs the best among all the methods in all settings.

Next, we compare the debiased Lasso with the vector bootstrap method on confidence intervals. We compute the average coverage probability of 95% CIs on the coefficients corresponding to the signal set $S = \{\{1, 2, 3\}, \{4, 5, 6\}\}$ and noise set $S^c = \{\{4, 5, 6, \ldots, p\}, \{1, 2, 3, 7, 8, \ldots, p\}\}$, where $\{1, 2, 3\}$ and $\{4, 5, 6\}$ are signal sets of $\boldsymbol{\beta}^{(1)*}$ and $\boldsymbol{\beta}^{(2)*}$, respectively. The average CI lengths are calculated separately for the CI of signals and the CI of noises. The average is taken over 200 simulations. The results are summarized in Table 4. It can be seen that the debiasing method has a larger coverage probability on S than the bootstrap method when $n = 100$. Their coverage probabilities on S in the case of $n = 200$ and 400 are similar, although the CIs produced by the debiased Lasso are shorter. For the coverage probability on S^c, the bootstrap suffers from the super-efficiency phenomenon, where the coverage rate of CIs whose length is close to zero equals one. This is because the Lasso shrinks these coefficients to zero in most replications. In contrast, the debiased Lasso provides reasonable CIs with high coverage probability. This is consistent with the findings of[22] in the case of linear regression.

Then, we compare the debiased Lasso with the vector bootstrap method in the individual testing $H_0 : \beta_j^{(k)*} = 0$ v.s. $H_1 : \beta_j^{(k)*} \neq 0$ for $j = 1 : p$ and $k = 1 : (K-1)$, at level 5%. The Type-I error rate is calculated by taking the average of Type-I error rates on all noises, and the power is calculated by taking the average of power on all signals. The results can be found in Table 5. Note that both the debiased Lasso and the vector bootstrap method can successfully control the Type-I error rate under 5% level, but the bootstrap method has much lower power than the debiased Lasso when $n = 100$ and 200.

We also compare the debiased Lasso with the multiple splitting method for multiple testing, where we want to control the family-wise error rate (FWER), i.e., the probability of at least one false discovery, under 5% level. The p-values provided by the debiased Lasso are adjusted by a Bonferroni correction. The FWER is the proportion of 200 simulations where at least one false discovery is reported. The power is calculated by taking the average of the proportion of correctly rejected null hypotheses on the signals over 200 simulations, which is motivated by[22]. The results are summarized in Table 6. Due to the Bonferroni correction, the debiased Lasso is too conservative in terms of the FWER control, hence does not have immense power. On the other hand, the multiple splitting method successfully controls the FWER with considerably higher statistical power than the debiased Lasso.

In Section C.1.1 of the appendix, we explore a variant of Model 1 where the signal strength is stronger than the original model. We observe a substantial increase in the power of the debiased Lasso, especially in the context of multiple testing. It is

not surprising that the power may be low when the sample size n is small, which implies that some weak signals may be missed. In the literature on Lasso theory for variable selection consistency (e.g.,[8,73]), the signal strength $\min_{j \in S} |\beta_j^*|$ is typically required to be larger than a polynomial function of the penalty parameter λ, which usually decreases as n grows. As one reviewer noted, missing important variables can be costly in some applications, especially when ℓ_1-penalized multinomial regression is used for preliminary variable screening. In practice, using smaller λ values can help avoid missing weak signals.

Finally, we present the QQ plot of sample quantiles of z-values $\left\{ \frac{\sqrt{n}(\widehat{b}_j - \beta_j^*)}{\widehat{\Theta}_j^\top \widehat{\Sigma} \widehat{\Theta}_j} \right\}_{j=1}^{(K-1)p}$ and the histogram of z-values in Figure 1. It is evident that the sample quantiles of z values approximately align with the quantiles of the standard normal distribution, and the histogram fits well with the standard normal density. We also plot the empirical CDF of computed p-values (restricted to zero entries of β^*), in Figure 2. We can see that the computed p-values are approximately uniformly distributed on $[0, 1]$. These empirical diagnostic results align with our theoretical findings.
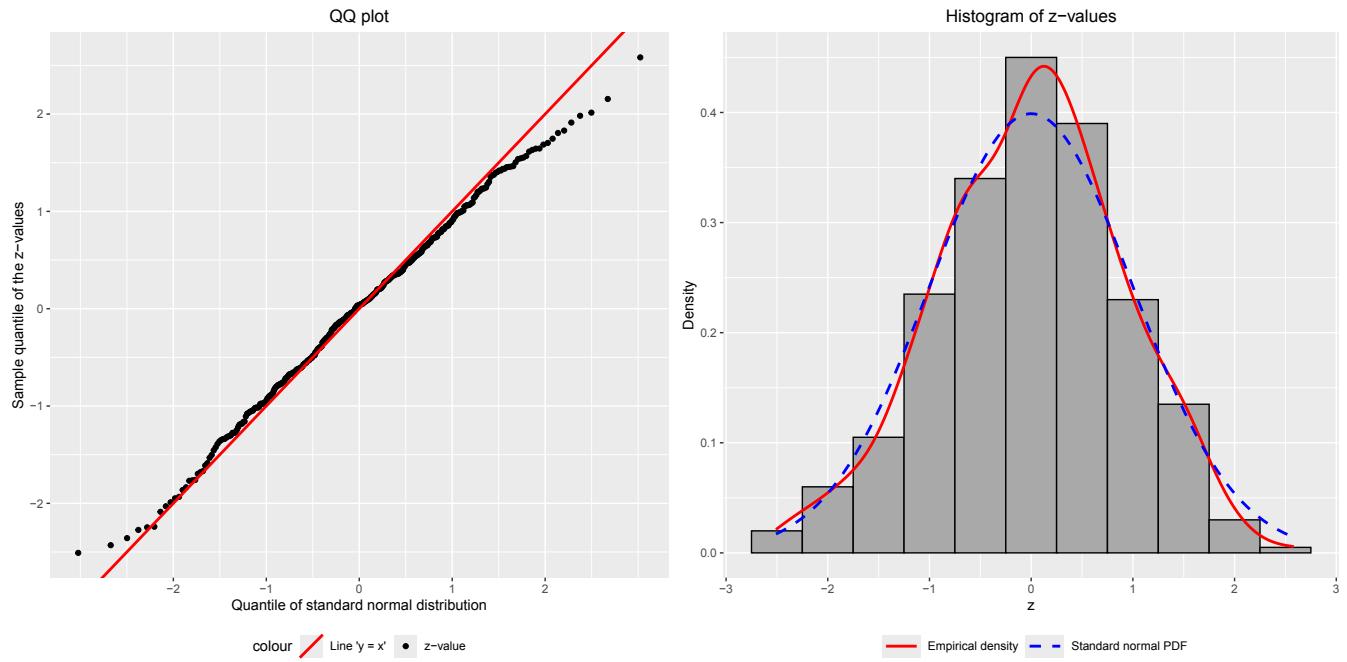

FIGURE 1 The QQ plot of the sample quantiles of z-values $\left\{ \frac{\sqrt{n}(\widehat{b}_j - \beta_j^*)}{\widehat{\Theta}_j^\top \widehat{\Sigma} \widehat{\Theta}_j} \right\}_{j=1}^{(K-1)p}$ in the left panel, and the histogram of z-values in the right panel, where $\beta^* = ((\beta^{(1)*})^\top, (\beta^{(2)*})^\top)^\top$. Both are derived from one realization of the setting $(n, K, p) = (400, 3, 200)$ in Model 1.

3.1.2 | Model 2

In this model, we generate independent observations $\{(x_i, y_i)\}_{i=1}^n$ similarly to Model 1. We first generate the predictor x_i from $N(\mathbf{0}_p, \Sigma_{p \times p})$ with $\Sigma = (0.75^{|i-j|})_{p \times p}$, then generate y_i from (1) with $K = 4$ classes and contrast coefficients $\beta^{(1)*} = (1, 1, 1, \mathbf{0}_{p-3})$, $\beta^{(2)*} = (\mathbf{0}_3, 1, 1, 1, \mathbf{0}_{p-6})$, $\beta^{(3)*} = (\mathbf{0}_6, 1, 1, 1, \mathbf{0}_{p-9})$, where $p = 200$. We consider different sample size settings $n = 100, 200, 400$.

The sum of squared ℓ_2-estimation error of $\beta^{(1)*}$, $\beta^{(2)*}$, and $\beta^{(3)*}$ is summarized in Table 7, where it is evident that the contrast-based penalization achieves lower estimation error than the over-parameterized method.

We also compare the prediction performance of the contrast-based ℓ_1-penalized multinomial regression with other benchmarks, and the average misclassification error rates of different methods are reported in Table 8. The results indicate that the ℓ_1-penalized multinomial regression outperforms all the other methods.

Similar to Model 1, we compute the average coverage probability of 95% CIs on the coefficients corresponding to the signal set and noise set for Model 2 in Table 9. The debiased Lasso exhibits a higher coverage probability on the signal set S than the bootstrap method.
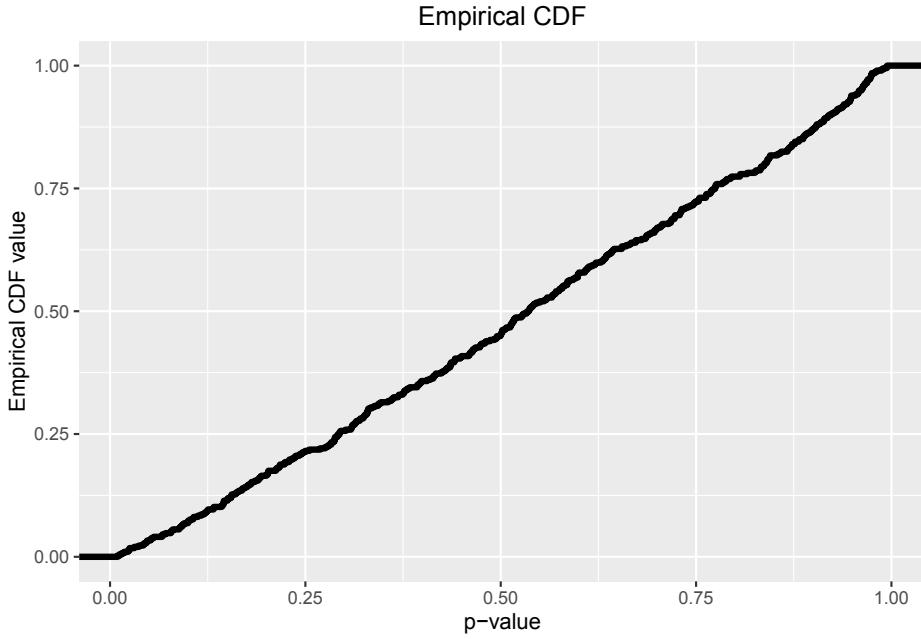
FIGURE 2   The empirical CDF of computed p-values (restricted to zero entries of $\boldsymbol{\beta}^*$), where $\boldsymbol{\beta}^* = ((\boldsymbol{\beta}^{(1)*})^\top, (\boldsymbol{\beta}^{(2)*})^\top)^\top$. It is derived from one realization of the setting $(n, K, p) = (400, 3, 200)$ in Model 1.

In Table 10, we summarize the average Type-I error rate and the power for the debiased Lasso with the vector bootstrap method in the individual testing $H_0 : \beta_j^{(k)*} = 0$ v.s. $H_1 : \beta_j^{(k)*} \neq 0$ for $j = 1 : p$ and $k = 1 : (K-1)$ at level 5%. The debiased Lasso demonstrates significantly larger power than the bootstrap method while successfully controlling the Type-I error rate under 5%.

We summarize the average FWER and power of multiple testing for the debiased Lasso with Bonferroni correction and the multiple splitting method in Table 11. The multiple splitting method fails to control the FWER under 5%, whereas the debiased Lasso successfully controls the FWER, although it might be over-conservative.

3.1.3 | Model 3

In this model, we generate each of the observations $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ independently using a linear discriminant analysis (LDA) setting. First, the class label y_i is generated from the discrete distribution on $\{1, 2, 3\}$ with probability $0.3, 0.3, 0.4$, respectively. Then given $y_i = k$, the predictor $\boldsymbol{x}_i$ is sampled from $N(\boldsymbol{\mu}^{(k)*}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu}^{(k)*} = \boldsymbol{\mu}^{(3)*} + \boldsymbol{\Sigma}\boldsymbol{\beta}^{(k)*}$ for $k = 1, 2$, where each coordinate of $\boldsymbol{\mu}^{(3)*}$ is randomly sampled from $\text{Unif}(\{-1, 1\})$, $\boldsymbol{\Sigma} = (0.5^{|i-j|})_{p \times p}$, and $p = 200$. The vectors $\boldsymbol{\beta}_{S^{(1)}}^{(k)*} = (1, -1, 1)^\top$, $\boldsymbol{\beta}_{S^{(2)}}^{(k)*} = (1, 1, -1)^\top$, where $S^{(1)}$ and $S^{(2)}$ are two sets of cardinality 3 randomly sampled from $1 : p$ without replacement. Following a simulation setting in [22], we set a random seed for the randomness in the model parameters for all 200 simulations, which gives $S^{(1)} = \{197, 92, 152\}$ and $S^{(2)} = \{173, 170, 191\}$. Thus, the data generation model remains the same across 200 simulations. In this model, the Gaussian means $\boldsymbol{\mu}^{(1)*}, \boldsymbol{\mu}^{(2)*}, \boldsymbol{\mu}^{(3)*}$ are dense, but the contrast coefficients (or discriminant coefficients, as referred to in LDA models) $\boldsymbol{\beta}^{(1)*}, \boldsymbol{\beta}^{(2)*}$ are sparse. We consider different sample size settings $n = 100, 200, 400$.

The results in Table 12 for the sum of squared ℓ_2-estimation error of $\boldsymbol{\beta}^{(1)*}$ and $\boldsymbol{\beta}^{(2)*}$ are similar to those of Models 1 and 2, where the contrast-based penalization performs better than the over-parameterization method.

We also compare the prediction performance of the contrast-based ℓ_1-penalized multinomial regression with other benchmarks, and the average misclassification error rates of different methods are reported in Table 13. We can see that the ℓ_1-penalized multinomial regression outperforms all the other methods.

Table 14 summarizes the average coverage probability of 95% CIs for the debiased Lasso and the vector bootstrap method. It shows that the debiased Lasso has a higher coverage probability on S than the multiple splitting method.

Table 15 displays the average Type-I error rate and power of individual testing for the debiased Lasso and the vector bootstrap method. Both the debiased Lasso and the vector bootstrap can control the Type-I error rate under the 5% level, but the debiased Lasso exhibits greater power than the bootstrap method when $n = 100$ and 200.

In Table 16, we show the average FWER and the power of the multiple testing for the debiased Lasso and the multiple splitting method. The multiple splitting method fails to control the FWER when $n = 100$ and 200 but achieves overwhelming power and successful control for FWER when $n = 400$. In contrast, the debiased Lasso with Bonferroni correction consistently maintains good control of FWER but might be too conservative for large n, leading to lower statistical power.

3.1.4 | Model 4

Similar to Model 3, we generate each of the observations $\{(x_i, y_i)\}_{i=1}^n$ independently using a linear discriminant analysis (LDA) framework. First, we sample the class label y_i from the discrete distribution on $\{1, 2, 3, 4\}$ with probability $0.3, 0.2, 0.3, 0.2$, respectively. Given $y_i = k$, the predictor x_i is sampled from $N(\mu^{(k)*}, \Sigma)$ with $\mu^{(k)*} = \mu^{(4)*} + \Sigma\beta^{(k)*}$ for $k = 1, 2, 3$, where each coordinate of $\mu^{(4)*}$ is randomly sampled from Unif($\{-1, 1\}$), $\Sigma = (0.5^{|i-j|})_{p \times p}$, and $p = 200$. The vectors $\beta_{S^{(1)}}^{(k)*} = (1, -1, 1)^\top$, $\beta_{S^{(2)}}^{(k)*} = (1, 1, -1)^\top$, $\beta_{S^{(3)}}^{(k)*} = (-1, 1, 1)^\top$, where $S^{(1)}$, $S^{(2)}$, and $S^{(3)}$ are three sets of cardinality 3 randomly sampled from $1 : p$ without replacement. We fix a random seed to ensure consistency in model parameters across all 200 simulations, resulting in $S^{(1)} = \{197, 92, 152\}$, $S^{(2)} = \{173, 170, 191\}$, and $S^{(3)} = \{23, 73, 148\}$. We consider different sample size settings with $n = 100, 200, 400$.

Table 17 summarizes the sum of squared ℓ_2-estimation error of the contrast coefficients, demonstrating that contrast-based penalization performs better than over-parameterization.

We also compare the prediction performance of the contrast-based ℓ_1-penalized multinomial regression with other benchmarks, and the average misclassification error rates of different methods are reported in Table 18. The results show that the ℓ_1-penalized multinomial regression outperforms all other methods.

Table 19 presents the coverage probability of 95% CIs generated by the debiased Lasso and the vector bootstrap method. The debiased Lasso provides CIs for signals in S with higher coverage probabilities compared to those of the bootstrap method.

We summarize the average Type-I error rate and power of individual testing in Table 20. It is evident that both the debiased Lasso and the vector bootstrap method control the Type-I error rate under 5%. However, the debiased Lasso exhibits significantly greater power than the bootstrap method when $n = 100$ and 200.

Finally, Table 21 shows the results for the multiple testing. The multiple splitting method fails to control the FWER, whereas the debiased Lasso successfully does so while maintaining relatively high power.

3.2 | A real example: Identifying important variables associated with different subtypes of MCI/dementia in NACC database

We have described the dataset background, pre-processing procedure, and the contrast coefficient between AD and CN in Section 1.2. In this section, we continue by presenting and discussing the results of contrast coefficients between FTLD and CN, LBD/PD and CN, and VBI and CN, which can be found in Tables 22, 23, and 24, respectively.

Similar to the effect on AD and CN, the number of days from the initial visit to the most recent visit is also a strong predictor for LBD/PD and CN, and VBI and CN. With other factors fixed, a long time from the initial visit means more likely progressing into AD, LBD/PD, and VBI. For FTLD, the presence of primary supranuclear palsy and family FTLD immutation are significant predictors. This aligns with the findings in[74] that around 30% of FTLD patients have a strong family history, and many FTLD cases are related to genetics. Additionally, the presumptive etiologic diagnosis of primary supranuclear palsy (PSP) is a strong predictor. Such a relationship has been found in the literature. For example, it has been known that PSP shares pathology with FTLD[75]. Moreover, difficulties with traveling, playing games, and engaging in hobbies are also predictors of FTLD, suggestive of the subtle, non-memory cognitive symptoms that may arise in preclinical stages. For LBD/PD, hallucinations, REM sleep behavior disorder, and depression are significant predictors against CN, which aligns with findings in the literature[76,77]. For VBI, a history of vascular events (stroke, transient ischemic attack) and risk factors (sleep apnea, diabetes) appear across several variables as predictors, as does diabetes, a strong indicator of vascular risk. Similar conclusions are noted in[78] and[79].

Finally, similar to[43], we group the predictors with individual p-values less than 5% into 10 categories, including demographics, subject medications, psychometric tests, etc., and plot them with their p-values in Figure 3. We also present radar charts of significant predictors (under 5% level) in different categories between each dementia subtype and CN in Figure 4, where each percentage represents the number of significant predictors in a category relative to the total number of predictors in that category. This reflects the significance of predictors in each category for each subtype. Another series of bar charts illustrates significant
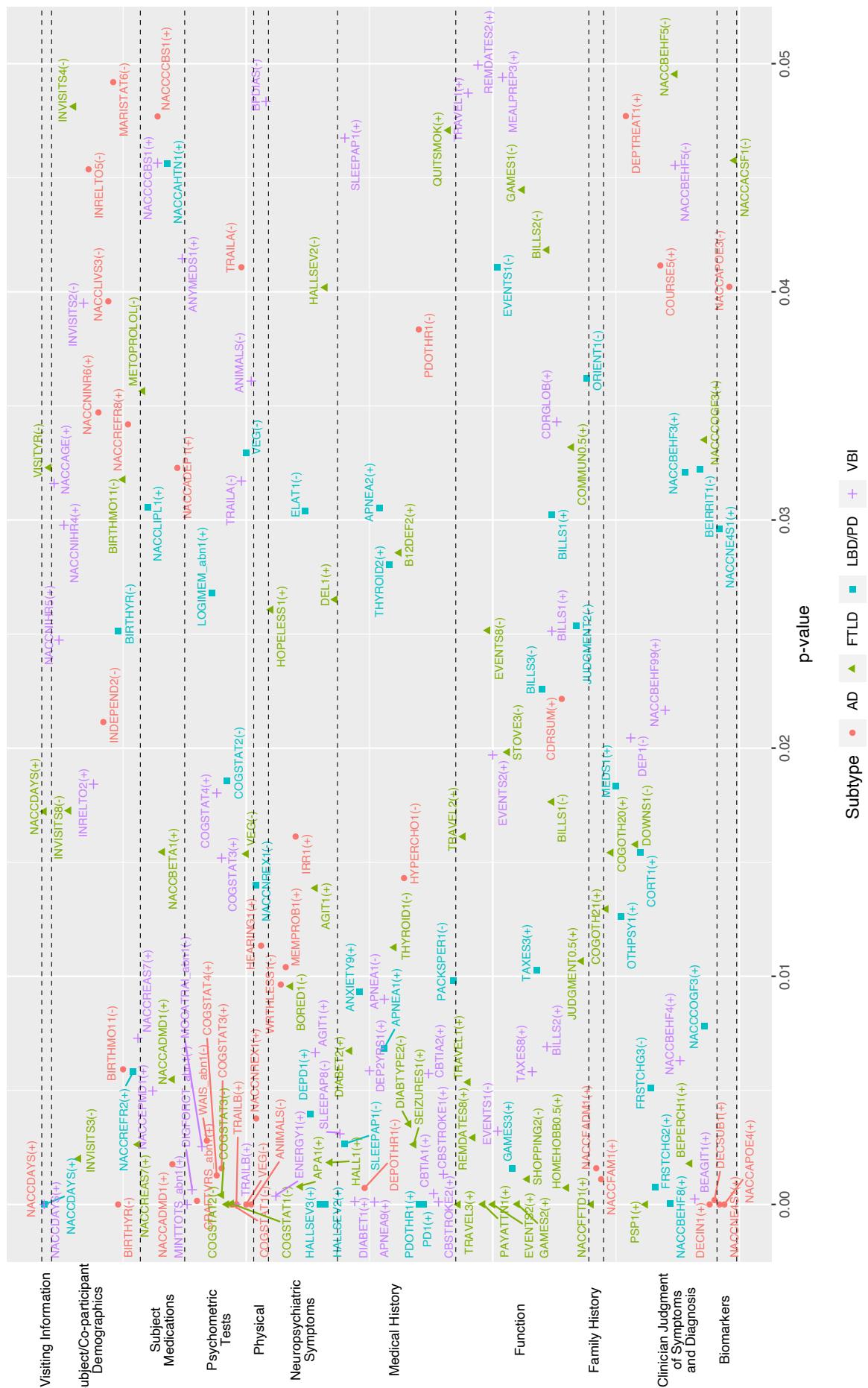
FIGURE 3   *p*-values of significant predictors (under 5% significance level) associated with the contrast between 4 dementia subtypes and CN, grouped into 11 categories

predictors (under the 5% significance level) across 11 categories for four dementia subtypes, where each percentage represents the number of significant predictors in a category relative to the total number of significant predictors for the subtype, reflecting the composition of significant predictors for each subtype.

Combining these three figures, we can obtain some interesting findings. From Figure 4, we can see that over 20% of family history variables are important for AD and FTLD, while no family history variables are significant for LBD/PD and VBI. On the other hand, over 10% of function predictors are significant for FTLD, LBD/PD, and VBI, while none are important for AD. From Figure 5, we can conclude that psychometric tests and demographics contain the most useful information to identify AD, as over 40% significant predictors belong to these two categories. Similarly, function, neuropsychiatric symptoms, clinician diagnosis, and medical history are most useful for identifying FTLD; Clinician diagnosis, medical history, and function predictors are most important for identifying LBD/PD; and Medical history, function, and psychometric tests are most useful for identifying VBI.
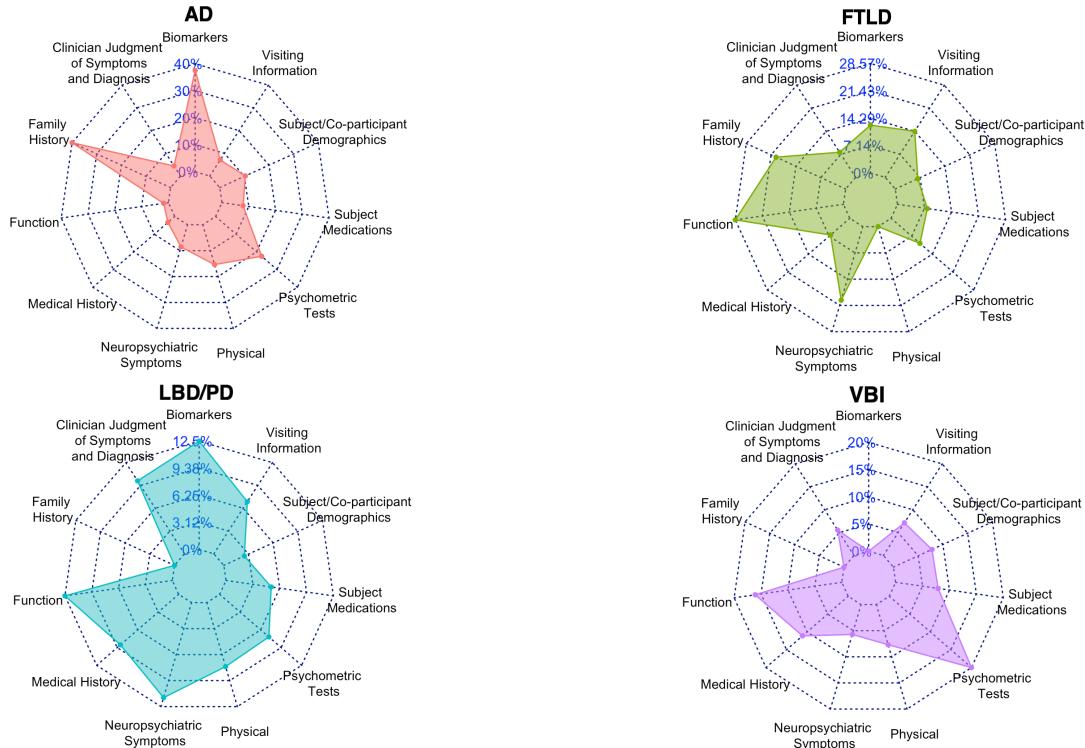
FIGURE 4 Radar charts of significant predictors (under 5% significance level) in 11 categories for 4 dementia subtypes. Each percentage represents the number of significant predictors in a category relative to the total number of predictors in that category, reflecting the significance of predictors in each category for each subtype.

4 | DISCUSSION

This paper analyzes the estimation and prediction error of the ℓ_1-penalized multinomial regression model. We also extend the debiased Lasso to the multinomial case to provide valid statistical inference, including confidence intervals and p-values for hypothesis testing. Additionally, we address cases of model misspecification and non-identically distributed data. The debiased Lasso is applied to a real dementia dataset, where we identify important variables associated with the progression into different subtypes of dementia, yielding highly interpretable results. The estimation, inference, and prediction methods for the contrast-based ℓ_1-penalized multinomial regression model have been implemented in a new R package `pemultinom`, which is available on CRAN. We believe this package will be beneficial to practitioners working with multi-label data.
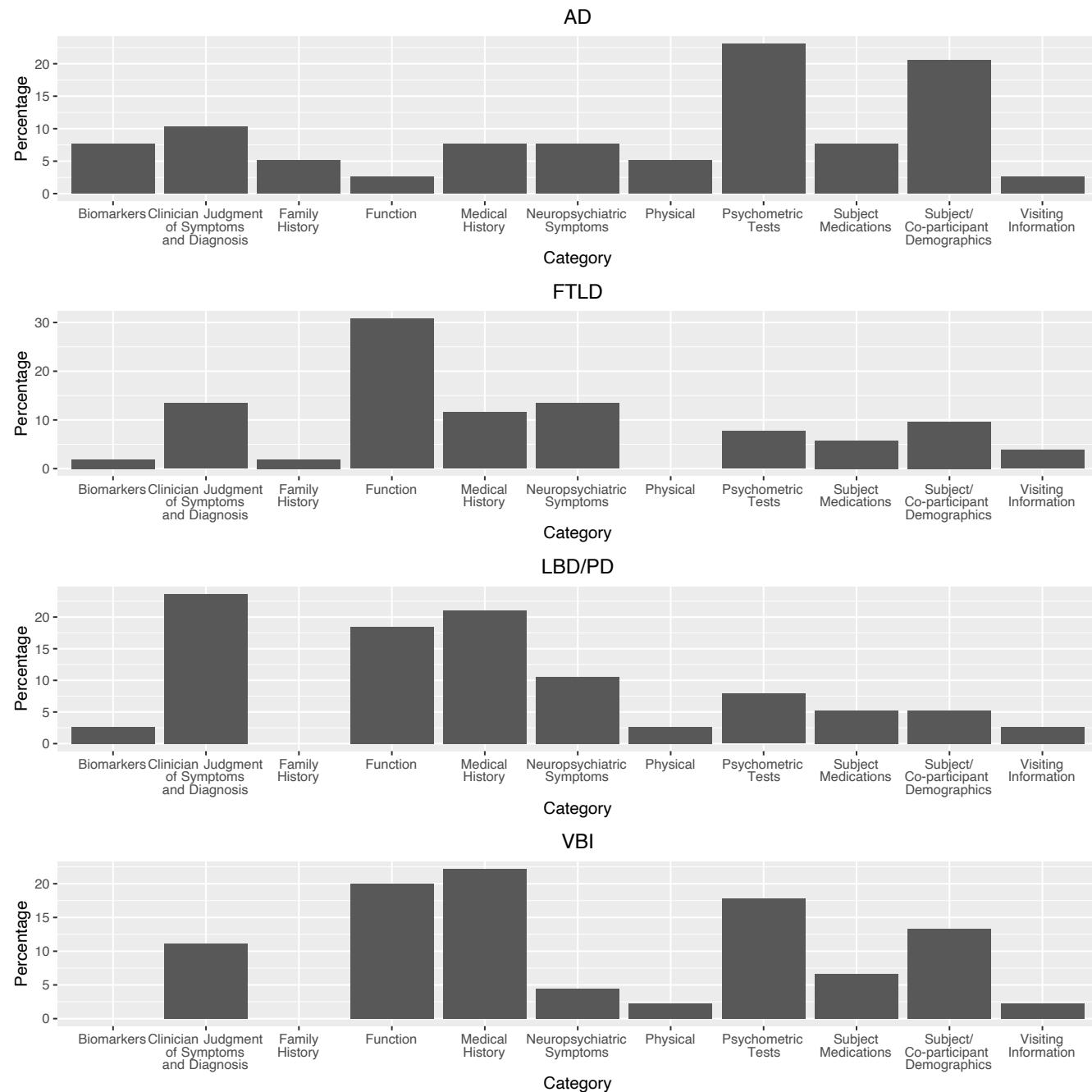
FIGURE 5 Bar charts of significant predictors (under 5% significance level) in 11 categories for 4 dementia subtypes. Each percentage represents the number of significant predictors in a category relative to the total number of significant predictors for the subtype, reflecting the composition of significant predictors for each subtype.

There are several interesting future research avenues. First, extending the debiasing method and theory to a more general high-dimensional multi-response setting would be valuable. Second, in this work, we do not consider the sparsity structure between the contrasts associated with different classes. As mentioned in Section 1.1,[26] discussed a row-wise sparsity structure, and we believe it is possible to extend the current analysis to that case. Finally,[80] suggested combining the debiasing method with a residual bootstrap procedure, which theoretically requires weaker conditions and empirically performs better than the debiasing method, particularly in simultaneous hypothesis testing problems. Exploring the application of this idea to the multinomial regression version of the debiased Lasso would be an intriguing direction for future research.

ACKNOWLEDGMENTS

REFERENCES

1. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological).* 1996;58(1):267–288.
2. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association.* 2001;96(456):1348–1360.
3. Zhang CH. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics.* 2010;38(2):894–942.
4. Raskutti G, Wainwright MJ, Yu B. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE transactions on information theory.* 2011;57(10):6976–6994.
5. Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology).* 2005;67(2):301–320.
6. Zou H. The adaptive lasso and its oracle properties. *Journal of the American statistical association.* 2006;101(476):1418–1429.
7. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* 2006;68(1):49–67.
8. Zhao P, Yu B. On model selection consistency of Lasso. *The Journal of Machine Learning Research.* 2006;7:2541–2563.
9. Loh PL, Wainwright MJ. Support recovery without incoherence: A case for nonconvex regularization. *The Annals of Statistics.* 2017;45(6):2455–2482.
10. Negahban SN, Ravikumar P, Wainwright MJ, Yu B. A unified framework for high-dimensional analysis of *M*-estimators with decomposable regularizers. *Statistical science.* 2012;27(4):538–557.
11. Loh PL, Wainwright MJ. Regularized M-estimators with Nonconvexity: Statistical and Algorithmic Theory for Local Optima. *Journal of Machine Learning Research.* 2015;16:559–616.
12. Fu W, Knight K. Asymptotics for lasso-type estimators. *The Annals of statistics.* 2000;28(5):1356–1378.
13. Chatterjee A, Lahiri SN. Bootstrapping lasso estimators. *Journal of the American Statistical Association.* 2011;106(494):608–625.
14. Chatterjee A, Lahiri SN. Rates of convergence of the adaptive LASSO estimators to the oracle distribution and higher order refinements by the bootstrap. *The Annals of Statistics.* 2013;41(3):1232–1259.
15. Sartori S. Penalized regression: Bootstrap confidence intervals and variable selection for high-dimensional data sets. 2011.
16. Wasserman L, Roeder K. High dimensional variable selection. *Annals of statistics.* 2009;37(5A):2178.
17. Meinshausen N, Meier L, Bühlmann P. P-values for high-dimensional regression. *Journal of the American Statistical Association.* 2009;104(488):1671–1681.
18. Zhang D, Khalili A, Asgharian M. Post-model-selection inference in linear regression models: An integrated review. *Statistics Surveys.* 2022;16:86–136.
19. Taylor J, Tibshirani RJ. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences.* 2015;112(25):7629–7634.
20. Lee JD, Sun DL, Sun Y, Taylor JE. Exact post-selection inference, with application to the lasso. *The Annals of Statistics.* 2016;44(3):907–927.
21. Javanmard A, Montanari A. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research.* 2014;15(1):2869–2909.
22. Geer V. dS, Bühlmann P, Ritov Y, Dezeure R. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics.* 2014;42(3):1166–1202.
23. Zhang CH, Zhang SS. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* 2014;76(1):217–242.
24. Krishnapuram B, Carin L, Figueiredo MA, Hartemink AJ. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE transactions on pattern analysis and machine intelligence.* 2005;27(6):957–968.
25. Pencer M. *Feature Selection and Post-Selection Statistical Inference in Multinomial Models*. PhD thesis. McGill University, Montreal, ; 2016.
26. Abramovich F, Grinshtein V, Levy T. Multiclass classification by sparse multinomial logistic regression. *IEEE Transactions on Information Theory.* 2021;67(7):4637–4646.
27. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika.* 1993;80(1):27–38.
28. Bull SB, Lewinger JP, Lee SS. Confidence intervals for multinomial logistic regression in sparse data. *Statistics in Medicine.* 2007;26(4):903–918.
29. Ke Y, Fu B, Zhang W. Semi-varying coefficient multinomial logistic regression for disease progression risk prediction. *Statistics in medicine.* 2016;35(26):4764–4778.
30. Jong dVM, Eijkemans MJ, Calster vB, et al. Sample size considerations and predictive performance of multinomial logistic prediction models. *Statistics in medicine.* 2019;38(9):1601–1619.
31. Donoho DL, Maleki A, Montanari A. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences.* 2009;106(45):18914–18919.
32. Donoho DL, Maleki A, Montanari A. Message passing algorithms for compressed sensing: I. motivation and construction. In: IEEE. 2010:1–5.
33. Donoho DL, Maleki A, Montanari A. Message passing algorithms for compressed sensing: II. Analysis and validation. In: IEEE. 2010:1–5.
34. Javanmard A, Montanari A. Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory.* 2014;60(10):6522–6554.
35. Huang H. Controlling the false discoveries in LASSO. *Biometrics.* 2017;73(4):1102–1110.
36. Huang H. Asymptotic risk and phase transition of ℓ_1-penalized robust estimator. *The Annals of Statistics.* 2020;48(5):3090–3111.

37. Kabashima Y, Wadayama T, Tanaka T. A typical reconstruction limit for compressed sensing based on lp-norm minimization. *Journal of Statistical Mechanics: Theory and Experiment.* 2009;2009(09):L09003.

38. Ghosh D, Chinnaiyan AM. Classification and selection of biomarkers in genomic data using LASSO. *BioMed Research International.* 2005;2005(2):147–154.

39. Ryali S, Supekar K, Abrams DA, Menon V. Sparse logistic regression for whole-brain classification of fMRI data. *NeuroImage.* 2010;51(2):752–764.

40. Cai TT, Guo Z. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of statistics.* 2017;45(2):615–646.

41. Ning Y, Liu H. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Annals of Statistics.* 2017;45(1):158–195.

42. Ma R, Guo Z, Cai TT, Li H. Statistical inference for genetic relatedness based on high-dimensional logistic regression. *Statistica Sinica.* 2024;34:1023–1043.

43. Tian Y, Rusinek H, V. Masurkar A, Feng Y. Risk factor to predict dementia subtypes: an analysis of the NACC database. *submitted.* 2023.

44. Hou XH, Feng L, Zhang C, Cao XP, Tan L, Yu JT. Models for predicting risk of dementia: a systematic review. *Journal of Neurology, Neurosurgery & Psychiatry.* 2019;90(4):373–379.

45. Battista P, Salvatore C, Castiglioni I. Optimizing neuropsychological assessments for cognitive, behavioral, and functional impairment classification: a machine learning study. *Behavioural neurology.* 2017;2017.

46. James C, Ranson JM, Everson R, Llewellyn DJ. Performance of machine learning algorithms for predicting progression to dementia in memory clinic patients. *JAMA network open.* 2021;4(12):e2136553–e2136553.

47. Thabtah F, Ong S, Peebles D. Detection of Dementia Progression from Functional Activities Data using Machine Learning Techniques. *Intelligent Decision Technologies.* 2022;16(3):615–630.

48. Rubin DB. *Multiple imputation for nonresponse in surveys.* 81. John Wiley & Sons, 2004.

49. Licht C. *New methods for generating significance levels from multiply-imputed data.* PhD thesis. Otto-Friedrich-Universität Bamberg, Fakultät Sozial-und Wirtschaftswissenschaften, ; 2010.

50. Corder EH, Saunders AM, Strittmatter WJ, et al. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science.* 1993;261(5123):921–923.

51. Reisberg B, Ferris SH, Leon dMJ, Crook T. The Global Deterioration Scale for assessment of primary degenerative dementia.. *The American journal of psychiatry.* 1982.

52. Reisberg B, Shulman MB, Torossian C, Leng L, Zhu W. Outcome over seven years of healthy adults with and without subjective cognitive impairment. *Alzheimer's & Dementia.* 2010;6(1):11–24.

53. Jessen F, Amariglio RE, Van Boxtel M, et al. A conceptual framework for research on subjective cognitive decline in preclinical Alzheimer's disease. *Alzheimer's & dementia.* 2014;10(6):844–852.

54. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software.* 2010;33(1):1.

55. Tseng P, Yun S. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming.* 2009;117:387–423.

56. Saha A, Tewari A. On the nonasymptotic convergence of cyclic coordinate descent methods. *SIAM Journal on Optimization.* 2013;23(1):576–601.

57. Hastie T, Tibshirani R, Friedman JH, Friedman JH. *The elements of statistical learning: data mining, inference, and prediction.* 2. Springer, 2009.

58. Durrett R. *Probability: theory and examples.* 49. Cambridge university press, 2019.

59. Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the lasso. *The annals of statistics.* 2006;34(3):1436–1462.

60. Solntsev S, Nocedal J, Byrd RH. An algorithm for quadratic ℓ_1-regularized optimization with a flexible active-set strategy. *Optimization Methods and Software.* 2015;30(6):1213–1237.

61. Guo Z, Rakshit P, Herman DS, Chen J. Inference for the case probability in high-dimensional logistic regression. *Journal of Machine Learning Research.* 2021;22(254):1–54.

62. Cox DR. Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society Series B: Statistical Methodology.* 1962;24(2):406–424.

63. Huber PJ, others . The behavior of maximum likelihood estimates under nonstandard conditions. In: . 1. Berkeley, CA: University of California Press. 1967:221–233.

64. Kent JT. Robust properties of likelihood ratio tests. *Biometrika.* 1982;69(1):19–27.

65. White H. Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the econometric society.* 1982:1–25.

66. Celi LA, Mark RG, Stone DJ, Montgomery RA. "Big data" in the intensive care unit. Closing the data loop. *American journal of respiratory and critical care medicine.* 2013;187(11):1157.

67. Stock JH, Watson MW. *Introduction to econometrics.* Pearson, 2020.

68. Smith TW. Developing and evaluating cross-national survey instruments. *Methods for testing and evaluating survey questionnaires.* 2004:431–452.

69. Han Y, Tsay RS. High-dimensional linear regression for dependent data with applications to nowcasting. *Statistica Sinica.* 2020;30(4):1797–1827.

70. Ing CK. Model selection for high-dimensional linear regression with dependent observations. *The Annals of Statistics.* 2020;48(4):1959–1980.

71. Fan J, Guo Y, Jiang B. Adaptive Huber regression on Markov-dependent data. *Stochastic processes and their applications.* 2022;150:802–818.

72. R Core Team . *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing; Vienna, Austria: 2020.

73. Wainwright MJ. Sharp thresholds for High-Dimensional and noisy sparsity recovery using ℓ_1-Constrained Quadratic Programming (Lasso). *IEEE transactions on information theory.* 2009;55(5):2183–2202.

74. Greaves CV, Rohrer JD. An update on genetic frontotemporal dementia. *Journal of neurology.* 2019;266(8):2075–2086.

75. Giagkou N, Höglinger GU, Stamelou M. Progressive supranuclear palsy. *International review of neurobiology.* 2019;149:49–86.

76. Boeve BF, Silber M, Ferman T, et al. REM sleep behavior disorder and degenerative dementia: an association likely reflecting Lewy body disease. *Neurology.* 1998;51(2):363–370.

77. Onofrj M, Taylor J, Monaco D, et al. Visual hallucinations in PD and Lewy body dementias: old and new hypotheses. *Behavioural neurology.* 2013;27(4):479–493.

78. Ahtiluoto S, Polvikoski T, Peltonen M, et al. Diabetes, Alzheimer disease, and vascular dementia: a population-based neuropathologic study. *Neurology.* 2010;75(13):1195–1202.

79. Kalaria RN, Akinyemi R, Ihara M. Stroke injury, cognitive impairment and vascular dementia. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease.* 2016;1862(5):915–925.

80. Dezeure R, Bühlmann P, Zhang CH. High-dimensional simultaneous inference with the bootstrap. *Test.* 2017;26(4):685–719.

81. Candes E, Tao T. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics.* 2007;35(6):2313–2351.

82. Ledoux M, Talagrand M. *Probability in Banach Spaces: isoperimetry and processes.* 23. Springer Science & Business Media, 1991.

83. Wainwright MJ. *High-dimensional statistics: A non-asymptotic viewpoint.* 48. Cambridge University Press, 2019.

84. Plan Y, Vershynin R. One-bit compressed sensing by linear programming. *Communications on Pure and Applied Mathematics.* 2013;66(8):1275–1297.

85. Mendelson S, Pajor A, Tomczak-Jaegermann N. Uniform uncertainty principle for Bernoulli and subgaussian ensembles. *Constructive Approximation.* 2008;28(3):277–289.

86. Zhou S. Restricted eigenvalue conditions on subgaussian random matrices. *arXiv preprint arXiv:0912.4045.* 2009.

87. Tsybakov AB. *Introduction to nonparametric estimation.* Springer, New York, 2009.

T A B L E 1 Contrasts, odds ratios, and p-values of top 10 predictors between AD and CN obtained by the debiased ℓ_1-penalized multinomial regression. Brackets indicate dummy variables for categorical variables.

Predictor	Coefficient	Odds ratio	p-value	Description
NACCDAYS	0.6245	1.8674	$< 10^{-4}$	Days from initial visit to most recent visit
NACCAPOE4	1.4692	4.3456	$< 10^{-4}$	APOE genotype [4 = (e4, e4)]
BIRTHYR	-0.9941	0.37	$< 10^{-4}$	Subject's year of birth
VEG	-0.2555	0.7745	$< 10^{-4}$	Total number of vegetables named in 60 seconds
DECIN1	0.6145	1.8487	$< 10^{-4}$	Does the co-participant report a decline in subject's memory? [1= Yes]
ANIMALS	-0.2141	0.8073	$< 10^{-4}$	Total number of animals named in 60 seconds
COGSTAT1	-0.6955	0.4988	$< 10^{-4}$	Per clinician, based on the neuropsychological examination, the subject's cognitive status is deemed [1 = better than normal for age]
NACCNE4S1	0.5057	1.6581	$< 10^{-4}$	Number of APOE e4 alleles
TRAILB	0.2074	1.2305	$< 10^{-4}$	Trail Making Test Part b - Total number of seconds to complete
CRAFTVRS_abn1	4.0484	57.3031	0.0002	Craft Story 21 Recall (Immediate) - Total story units recalled, verbatim scoring [abn1 = Abnormal]

T A B L E 2 The sum of squared ℓ_2-estimation error of $\beta^{(1)*}$ and $\beta^{(2)*}$ in Model 1, for the contrast-based penalization and the over-parameterization method.

Method	$n = 100$	$n = 200$	$n = 400$
Contrast	2.77 (0.58)	1.787 (0.566)	0.973 (0.377)
Over-parameterization	3.314 (0.755)	2.285 (0.507)	1.352 (0.363)

T A B L E 3 The average misclassification error rates on test data in Model 1, for ℓ_1-penalized multinomial regression, SVMs, LDA, Random Forests, and kNN.

Method	$n = 100$	$n = 200$	$n = 400$
Penalized multinomial regression	32.37% (2.23%)	30.08% (1.80%)	28.52% (1.51%)
SVMs	52.86% (3.65%)	45.74% (2.23%)	40.03% (1.82%)
LDA	50.36% (2.49%)	63.26% (2.32%)	46.87% (2.08%)
Random Forests	39.67% (3.08%)	35.52% (2.41%)	32.42% (1.62%)
kNN	57.32% (2.52%)	54.70% (2.20%)	51.66% (2.03%)

T A B L E 4 The average coverage probability of 95% CIs on the coefficients corresponding to the signal set $S = \{\{1,2,3\}, \{4,5,6\}\}$ and noise set $S^c = \{\{4,5,6,\ldots,p\}, \{1,2,3,7,8,\ldots,p\}\}$ in Model 1, for the debiased Lasso and the vector bootstrap method.

Method	Measure	$n = 100$	$n = 200$	$n = 400$
Debiased-Lasso	Average coverage probability on S	0.954 (0.082)	0.928 (0.108)	0.939 (0.107)
	Average coverage probability on S^c	0.991 (0.007)	0.983 (0.008)	0.976 (0.009)
	Average length of CI (S)	2.415 (1.311)	1.52 (0.136)	1.157 (0.066)
	Average length of CI (S^c)	3.116 (1.529)	1.599 (0.102)	1.227 (0.044)
Bootstrap	Average coverage probability on S	0.877 (0.132)	0.914 (0.125)	0.928 (0.116)
	Average coverage probability on S^c	1 (0)	1 (0)	1 (0.001)
	Average length of CI (S)	1.831 (0.259)	1.658 (0.16)	1.314 (0.098)
	Average length of CI (S^c)	0.192 (0.013)	0.23 (0.013)	0.269 (0.019)

T A B L E 5 The average Type-I error rate and power of individual testing $H_0 : \beta_j^{(k)*} = 0$ v.s. $H_1 : \beta_j^{(k)*} \neq 0$ at level 5% in Model 1, for the debiased Lasso and the vector bootstrap method.

Method	Measure	$n = 100$	$n = 200$	$n = 400$
Debiased-Lasso	Type-I error	0.009 (0.007)	0.016 (0.008)	0.024 (0.009)
	Power	0.463 (0.161)	0.712 (0.158)	0.895 (0.116)
Bootstrap	Type-I error	0 (0)	0 (0)	0 (0.001)
	Power	0.215 (0.148)	0.504 (0.165)	0.863 (0.131)

T A B L E 6 The average family-wise error rate (FWER) and power of multiple testing $H_0 : \beta_j^{(k)*} = 0$ v.s. $H_1 : \beta_j^{(k)*} \neq 0$ for $j = 1 : p$ and $k = 1 : (K - 1)$ in Model 1, for the debiased Lasso and the multiple splitting method.

Method	Measure	$n = 100$	$n = 200$	$n = 400$
Debiased-Lasso	FWER	0.005 (0.071)	0.005 (0.071)	0 (0)
	Power	0.027 (0.061)	0.12 (0.121)	0.318 (0.144)
Multiple-Spliting	FWER	0.015 (0.122)	0.01 (0.1)	0.04 (0.196)
	Power	0.233 (0.144)	0.329 (0.111)	0.496 (0.146)

TABLE 7 The sum of squared ℓ_2-estimation error of $\beta^{(1)*}$, $\beta^{(2)*}$, and $\beta^{(3)*}$ in Model 2, for the contrast-based penalization and the over-parameterization method.

Method	$n = 100$	$n = 200$	$n = 400$
Contrast	4.607 (0.926)	2.93 (0.789)	1.696 (0.496)
Over-parameterization	5.559 (1.292)	3.563 (0.795)	2.134 (0.482)

TABLE 8 The average misclassification error rates on test data in Model 2, for ℓ_1-penalized multinomial regression, SVMs, LDA, Random Forests, and kNN.

Method	$n = 100$	$n = 200$	$n = 400$
Penalized multinomial regression	40.02% (2.61%)	36.33% (2.09%)	34.46% (1.59%)
SVMs	63.59% (4.01%)	55.30% (2.61%)	48.74% (1.92%)
LDA	59.65% (2.24%)	71.59% (2.02%)	56.46% (1.94%)
Random Forests	50.01% (3.65%)	43.79% (2.58%)	40.28% (1.90%)
kNN	66.11% (2.24%)	63.53% (1.91%)	60.80% (1.91%)

TABLE 9 The average coverage probability of 95% CIs on the coefficients corresponding to the signal set $S = \{\{1,2,3\},\{4,5,6\},\{7,8,9\}\}$ and noise set $S^c = \{\{4,5,6,\ldots,p\},\{1,2,3,7,8,\ldots,p\},\{1,2,3,4,5,6,10,11,\ldots,p\}\}$ in Model 2, for the debiased Lasso and the vector bootstrap method.

Method	Measure	$n = 100$	$n = 200$	$n = 400$
Debiased-Lasso	Average coverage probability on S	0.978 (0.049)	0.934 (0.083)	0.946 (0.078)
	Average coverage probability on S^c	0.993 (0.006)	0.986 (0.006)	0.979 (0.006)
	Average length of CI (S)	3.156 (2.158)	1.697 (0.175)	1.348 (0.075)
	Average length of CI (S^c)	2.761 (0.711)	1.723 (0.104)	1.398 (0.047)
Bootstrap	Average coverage probability on S	0.848 (0.117)	0.895 (0.112)	0.918 (0.112)
	Average coverage probability on S^c	1 (0)	1 (0)	1 (0.001)
	Average length of CI (S)	1.689 (0.199)	1.598 (0.145)	1.312 (0.078)
	Average length of CI (S^c)	0.174 (0.01)	0.21 (0.011)	0.25 (0.014)

TABLE 10 The average Type-I error rate and power of individual testing $H_0 : \beta_j^{(k)*} = 0$ v.s. $H_1 : \beta_j^{(k)*} \neq 0$ in Model 2 at level 5%, for the debiased Lasso and the vector bootstrap method.

Method	Measure	$n = 100$	$n = 200$	$n = 400$
Debiased-Lasso	Type-I error	0.007 (0.006)	0.014 (0.006)	0.021 (0.006)
	Power	0.308 (0.122)	0.629 (0.141)	0.778 (0.12)
Bootstrap	Type-I error	0 (0)	0 (0)	0 (0.001)
	Power	0.159 (0.1)	0.43 (0.126)	0.791 (0.124)

TABLE 11 The average family-wise error rate (FWER) and power of multiple testing $H_0 : \beta_j^{(k)*} = 0$ v.s. $H_1 : \beta_j^{(k)*} \neq 0$ for $j = 1 : p$ and $k = 1 : (K - 1)$ in Model 2, for the debiased Lasso and the multiple splitting method.

Method	Measure	$n = 100$	$n = 200$	$n = 400$
Debiased-Lasso	FWER	0.005 (0.071)	0 (0)	0 (0)
	Power	0.006 (0.024)	0.058 (0.069)	0.146 (0.096)
Multiple-Splitting	FWER	0.44 (0.498)	0.61 (0.489)	0.895 (0.307)
	Power	0.089 (0.093)	0.132 (0.084)	0.224 (0.09)

TABLE 12 The sum of squared ℓ_2-estimation error of $\beta^{(1)*}$ and $\beta^{(2)*}$, in Model 3, for the contrast-based penalization and the over-parameterization method.

Method	$n = 100$	$n = 200$	$n = 400$
Contrast	9.291 (4.103)	5.46 (2.577)	3.379 (1.508)
Over-parameterization	11.691 (4.578)	6.986 (2.967)	4.163 (1.783)

TABLE 13 The average misclassification error rates on test data in Model 3, for ℓ_1-penalized multinomial regression, SVMs, LDA, Random Forests, and kNN.

Method	$n = 100$	$n = 200$	$n = 400$
Penalized multinomial regression	54.58% (4.12%)	52.95% (2.54%)	51.74% (1.95%)
SVMs	62.81% (3.78%)	60.16% (4.82%)	57.96% (4.85%)
LDA	60.66% (3.73%)	64.86% (3.09%)	59.49% (4.75%)
Random Forests	58.76% (3.37%)	56.89% (3.18%)	55.69% (2.39%)
kNN	60.51% (2.85%)	59.08% (2.78%)	58.00% (2.67%)

TABLE 14 The average coverage probability of 95% CIs on the coefficients corresponding to the signal set $S = \{\{197, 92, 152\}, \{173, 170, 191\}\}$ and noise set $S^c = \{(1:p)\backslash\{197, 92, 152\}, (1:p)\backslash\{173, 170, 191\}\}$ in Model 3, for the debiased Lasso and the vector bootstrap method.

Method	Measure	$n = 100$	$n = 200$	$n = 400$
Debiased-Lasso	Average coverage probability on S	0.963 (0.085)	0.952 (0.089)	0.945 (0.085)
	Average coverage probability on S^c	0.995 (0.005)	0.987 (0.006)	0.977 (0.009)
	Average length of CI (S)	1.838 (0.454)	1.172 (0.124)	0.841 (0.045)
	Average length of CI (S^c)	1.958 (0.363)	1.174 (0.077)	0.831 (0.027)
Bootstrap	Average coverage probability on S	0.738 (0.176)	0.887 (0.121)	0.925 (0.112)
	Average coverage probability on S^c	1 (0)	1 (0.001)	0.999 (0.001)
	Average length of CI (S)	1.183 (0.135)	1.024 (0.075)	0.795 (0.059)
	Average length of CI (S^c)	0.197 (0.01)	0.234 (0.01)	0.259 (0.014)

TABLE 15 The average Type-I error rate and power of individual testing $H_0 : \beta_j^{(k)*} = 0$ v.s. $H_1 : \beta_j^{(k)*} \neq 0$ at level 5% in Model 3, for the debiased Lasso and the vector bootstrap method.

Method	Measure	$n = 100$	$n = 200$	$n = 400$
Debiased-Lasso	Type-I error	0.005 (0.005)	0.013 (0.006)	0.023 (0.009)
	Power	0.546 (0.186)	0.893 (0.127)	0.992 (0.036)
Bootstrap	Type-I error	0 (0)	0 (0.001)	0.001 (0.001)
	Power	0.368 (0.188)	0.855 (0.134)	0.998 (0.017)

TABLE 16 The average family-wise error rate (FWER) and power of multiple testing $H_0 : \beta_j^{(k)*} = 0$ v.s. $H_1 : \beta_j^{(k)*} \neq 0$ for $j = 1:p$ and $k = 1:(K-1)$ in Model 3, for the debiased Lasso and the multiple splitting method.

Method	Measure	$n = 100$	$n = 200$	$n = 400$
Debiased-Lasso	FWER	0 (0)	0.005 (0.071)	0.005 (0.071)
	Power	0.039 (0.077)	0.267 (0.164)	0.75 (0.165)
Multiple-Spliting	FWER	0.105 (0.307)	0.06 (0.238)	0.03 (0.171)
	Power	0.157 (0.141)	0.658 (0.152)	0.995 (0.029)

TABLE 17 The sum of squared ℓ_2-estimation error of $\boldsymbol{\beta}^{(1)*}$, $\boldsymbol{\beta}^{(2)*}$, and $\boldsymbol{\beta}^{(3)*}$, in Model 4, for the contrast-based penalization and the over-parameterization method.

Method	$n = 100$	$n = 200$	$n = 400$
Contrast	7.331 (2.772)	4.433 (1.706)	2.543 (0.864)
Over-parameterization	9.858 (5.624)	5.902 (2.717)	3.465 (1.435)

TABLE 18 The average misclassification error rates on test data in Model 4, for ℓ_1-penalized multinomial regression, SVMs, LDA, Random Forests, and kNN.

Method	$n = 100$	$n = 200$	$n = 400$
Penalized multinomial regression	56.01% (5.05%)	50.70% (4.44%)	47.50% (3.42%)
SVMs	70.02% (3.80%)	66.46% (4.21%)	62.92% (5.48%)
LDA	68.63% (3.42%)	73.33% (2.89%)	65.76% (4.71%)
Random forests	64.50% (3.84%)	61.04% (3.72%)	58.43% (3.90%)
k-nearest neighbors	68.13% (2.58%)	66.10% (2.62%)	64.61% (2.86%)

TABLE 19 The average coverage probability of 95% CIs on the coefficients corresponding to the signal set $S = \{\{197, 92, 152\}, \{173, 170, 191\}, \{23, 73, 148\}\}$ and noise set $S^c = \{(1:p)\backslash\{197, 92, 152\}, (1:p)\backslash\{173, 170, 191\}, (1:p)\backslash\{23, 73, 148\}\}$ in Model 4, for the debiased Lasso and the vector bootstrap method.

Method	Measure	$n = 100$	$n = 200$	$n = 400$
Debiased-Lasso	Average coverage probability on S	0.968 (0.064)	0.958 (0.063)	0.962 (0.065)
	Average coverage probability on S^c	0.996 (0.003)	0.99 (0.005)	0.981 (0.006)
	Average length of CI (S)	2.146 (0.673)	1.242 (0.12)	0.936 (0.046)
	Average length of CI (S^c)	2.096 (0.399)	1.277 (0.085)	0.961 (0.034)
Bootstrap	Average coverage probability on S	0.66 (0.148)	0.814 (0.129)	0.921 (0.086)
	Average coverage probability on S^c	1 (0)	1 (0.001)	1 (0.001)
	Average length of CI (S)	1.115 (0.119)	1.018 (0.067)	0.803 (0.043)
	Average length of CI (S^c)	0.174 (0.009)	0.213 (0.008)	0.242 (0.01)

TABLE 20 The average Type-I error rate and power of individual testing $H_0 : \beta_j^{(k)*} = 0$ v.s. $H_1 : \beta_j^{(k)*} \neq 0$ at level 5% in Model 4, for the debiased Lasso and the vector bootstrap method.

Method	Measure	$n = 100$	$n = 200$	$n = 400$
Debiased-Lasso	Type-I error	0.004 (0.003)	0.01 (0.005)	0.019 (0.006)
	Power	0.433 (0.15)	0.854 (0.104)	0.981 (0.052)
Bootstrap	Type-I error	0 (0)	0 (0.001)	0 (0.001)
	Power	0.246 (0.126)	0.75 (0.129)	0.991 (0.031)

TABLE 21 The average family-wise error rate (FWER) and power of multiple testing $H_0 : \beta_j^{(k)*} = 0$ v.s. $H_1 : \beta_j^{(k)*} \neq 0$ for $j = 1 : p$ and $k = 1 : (K-1)$ in Model 4, for the debiased Lasso and the multiple splitting method.

Method	Measure	$n = 100$	$n = 200$	$n = 400$
Debiased-Lasso	FWER	0 (0)	0.005 (0.071)	0 (0)
	Power	0.012 (0.034)	0.163 (0.104)	0.516 (0.166)
Multiple-Splitting	FWER	0.37 (0.484)	0.78 (0.415)	1 (0)
	Power	0.087 (0.119)	0.227 (0.14)	0.616 (0.068)

TABLE 22 Contrasts, odds ratios, and p-values of top 10 predictors between FTLD and CN obtained by the debiased ℓ_1-penalized multinomial regression. Brackets indicate dummy variables for categorical variables.

Feature	Coefficient	Odds ratio	p-value	Description
TRAVEL3	33.9214	5.39×10^{14}	0	In the past four weeks, did the subject have any difficulty or need help with: Traveling out of the neighborhood, driving, or arranging to take public transportation [3 = Dependent]
PSP1	48.3045	9.51×10^{20}	0	Presumptive etiologic diagnosis - primary supranuclear palsy (PSP) [1 = Yes]
NACCFFTD1	4.5809	97.6059	0	In this family, is there evidence for an FTLD mutation? [1 = Yes]
EVENTS2	28.9003	3.56×10^{12}	0	In the past four weeks, did the subject have any difficulty or need help with: Keeping track of current events [2 = Requires assistance]
COGSTAT2	-1.953	0.1418	0	Per clinician, based on the neuropsychological examination, the subject's cognitive status is deemed [2=Normal for age]
PAYATTN1	6.3784	589.0052	0	In the past four weeks, did the subject have any difficulty or need help with: Paying attention to and understanding a TV program, book, or magazine [1 = Has difficulty, but does by self]
GAMES2	20.4844	7.88×10^8	0	In the past four weeks, did the subject have any difficulty or need help with: Playing a game of skill such as bridge or chess, working on a hobby [2 = Requires assistance]
COGSTAT1	-2.6038	0.074	0	Per clinician, based on the neuropsychological examination, the subject's cognitive status is deemed [1=better than normal for age]
COGSTAT3	-1.9584	0.1411	0.0004	Per clinician, based on the neuropsychological examination, the subject's cognitive status is deemed [3 = One or two test scores abnormal]
HOMEHOBB0.5	5.877	356.7285	0.0007	Home and hobbies [0.5 = Questionable impairment]

TABLE 23 Contrasts, odds ratios, and p-values of top 10 predictors between LBD/PD and CN obtained by the debiased ℓ_1-penalized multinomial regression. Brackets indicate dummy variables for categorical variables.

Feature	Coefficient	Odds ratio	p-value	Description
HALLSEV3	55.7257	1.59×10^{24}	$< 10^{-4}$	Hallucinations severity [3 = Severe (very marked or prominent; a dramatic change)]
PDOTHR1	6.486	655.9019	$< 10^{-4}$	Other parkinsonian disorder [1 = Recent/Active]
PD1	4.4647	86.8966	$< 10^{-4}$	Parkinson's disease (PD) [1 = Recent/Active]
NACCDAYS	0.5998	1.8218	$< 10^{-4}$	Days from initial visit to most recent visit
HALLSEV2	10.8981	54071.5757	$< 10^{-4}$	Hallucinations severity [2 = Moderate (significant, but not a dramatic change)]
NACCBEHF8	4.4825	88.4524	$< 10^{-4}$	Indicate the predominant symptom that was first recognized as a decline in the subject's behavior [8 = REM sleep behavior disorder]
FRSTCHG2	2.3303	10.2807	0.0008	Indicate the predominant domain that was first recognized as changed in the subject [2 = behavior]
GAMES3	11.6056	1.10×10^5	0.0016	In the past four weeks, did the subject have any difficulty or need help with: Playing a game of skill such as bridge or chess, working on a hobby [3 = Dependent]
SLEEPAP1	-3.8154	0.022	0.0027	Sleep apnea present [1 = Yes]
DEPD1	1.1913	3.2915	0.004	Depression or dysphoria in the last month [1 = Yes]

TABLE 24   Contrasts, odds ratios, and *p*-values of top 10 predictors between VBI and CN obtained by the debiased ℓ_1-penalized multinomial regression. Brackets indicate dummy variables for categorical variables.

Feature	Coefficient	Odds ratio	*p*-value	Description
NACCDAYS	0.6244	1.8672	0	Days from initial visit to most recent visit
MINTTOTS_abn1	8.471	4774.4576	0	Multilingual Naming Test (MINT) - Total score [abn1 = Abnormal]
TRAILB	0.3822	1.4655	0	Trail Making Test Part b - Total number of seconds to complete
APNEA9	4.6957	109.4747	0.0001	Sleep apnea history reported at Initial Visit [9 = Unknown]
CBSTROKE2	1.3345	3.7981	0.0001	Stroke [2 = Remote/Inactive]
DIABET1	8.546	5146.2553	0.0001	Diabetes present at visit [1 = Yes, Type I]
BEAGIT1	4.5007	90.0779	0.0002	Subject currently manifests meaningful change in behavior - Agitation [1 = Yes]
ENERGY1	0.6031	1.8277	0.0004	Do you feel full of energy? [1 = No]
CBTIA1	1.8537	6.3837	0.0005	Transient ischemic attack (TIA) [1 = Recent/Active]
DIGFORCT_abn1	-7.5341	0.005	0.0006	Number Span Test: Forward - Number of correct trials [abn1 = Abnormal]

□

APPENDIX

We stack the coefficients $\{\beta^{(k)*}\}_{k=1}^{K-1}$ into a single long vector β^* of dimension $(K-1)p$. Similarly, we stack the estimates $\{\widehat{\beta}^{(k)}\}_{k=1}^{K-1}$ into a $(K-1)p$-dimensional vector $\widehat{\beta}$. Given any estimate $\beta = \{\beta^{(k)}\}_{k=1}^{K-1}$, we define the conditional probability $\mathbb{P}(Y = k|X = x) = \frac{\exp(x^\top \beta^{(k)})}{1+\sum_{k=1}^{K-1} \exp(x^\top \beta^{(k)})}$ as $p_k(x; \beta)$, for $k = 1 : (K-1)$. Similarly, we define $\widehat{\Delta}^{(k)} = \widehat{\beta}^{(k)} - \beta^{(k)*}$ and $\widehat{\Delta} = \{\widehat{\Delta}^{(k)}\}_{k=1}^{K-1}$.

A ANOTHER DEBIASED PROCEDURE

In Remark 3, we mentioned that the ℓ_0-sparsity condition of Θ_j in Assumption 3.(i) can be relaxed. In this section, we develop a variant of the inference procedure introduced in the main text and demonstrate its validity without the ℓ_0-sparsity requirement of Θ_j.

We begin by splitting the data into two equal halves, $(X^{[1]}, Y^{[1]}) \cup (X^{[2]}, Y^{[2]})$. For simplicity, we assume n is even in this section.

We obtain two point estimators of $\beta^{(t)*}$ on two folds of data through (2), respectively, and denote them as $\widehat{\beta}^{[1]}$ and $\widehat{\beta}^{[2]}$. For Σ, consider the empirical estimator $\widehat{\Sigma}^{[r]} := \frac{1}{n/2} \sum_{i=1}^{n/2} B(x_i^{[r]}; \widehat{\beta}^{[r]})$, with $r = 1, 2$.

Consider another debiased Lasso estimator given by

$$\widehat{b} := \widehat{\beta}^{[1]} + \frac{1}{n} M \begin{pmatrix} (X^{[1]})^\top (Y^{[1](1)} - p_1(X^{[1]}; \widehat{\beta}^{[1]})) \\ \vdots \\ (X^{[1]})^\top (Y^{[1](K-1)} - p_{K-1}(X^{[1]}; \widehat{\beta}^{[1]})) \end{pmatrix},$$

where $M \in \mathbb{R}^{[(K-1)p] \times [(K-1)p]}$, and the j-th row of M is the solution to the following linear program:

$$\min_{m \in \mathbb{R}^{(K-1)p}} \|m\|_1$$
$$\text{s.t. } \|\widehat{\Sigma}^{[2]} m - e_j\|_\infty \le \eta. \tag{A1}$$

The choice of η will be specified later. The form of (A1) is similar to the Dantzig selector [81], although no ℓ_0-sparsity of the estimnond corresponding to m is required.

We impose the following assumption to replace Assumption 3 in the main text. Note that the new assumption is weaker because the ℓ_0-sparsity of $\gamma_j^* = (\gamma_{j,1}^*, \dots, \gamma_{j,(K-1)p-1}^*)^\top = (\Sigma_{-j,-j})^{-1} \Sigma_{-j,j}$ is not necessary, and the requirement on $\max_{j=1:[(K-1)p]} \|\Theta_j\|_1$ is less stringent.

Assumption 5. Denote $\Sigma = \mathbb{E}[\nabla^2 \mathcal{L}_n(\beta^*)]$ and $\Theta = \Sigma^{-1}$. Then $\max_{j=1:[(K-1)p]} \|\Theta_j\|_2 \le C < \infty$.

We are now ready to present the following theorem.

Theorem 7. *Let $\lambda = C\sqrt{\frac{\log p}{n}}$ in (2) and $\eta = CK^{5/2}\sqrt{\frac{s \log p}{n}}$ in (9) with a large constant $C > 0$. Under Assumptions 1, 2, and 5, when $n \gg \|\Theta^*\|_1^2 K^{10} s^2 (\log p)^2 + K^{11} s^3 (\log p)^2$,*

$$\frac{\sqrt{n/2}(\widehat{b}_j - \beta_j^*)}{\sqrt{M_j^\top \widehat{\Sigma}^{[1]} M_j}} \xrightarrow{d} N(0, 1),$$

for $j = 1, \dots, (K-1)p$.

B BEYOND THE ALMOST SURELY BOUNDEDNESS ASSUMPTION

In Remark 3, we mentioned that Assumption 3.(ii) can be removed with a more stringent sample size requirement. The following theorem justifies this claim.

Theorem 8. *Let $\lambda = C\sqrt{\frac{\log(Kp)}{n}}$ in (2) and $\lambda_j = CK^{5/2}\sqrt{\frac{s\log(Kp)}{n}}$ in (9) with a large constant $C > 0$. Under Assumptions 1-3 but without Assumption 3.(iii), when $n \gg K^{20}(ss_0)^2(\log(Kp))^2 + K^{11}ss_0^2\log(Kp)$,*

$$\frac{\sqrt{n}(\widehat{b}_j - \beta_j^*)}{\sqrt{\widehat{\Theta}_j^\top \widehat{\Sigma} \widehat{\Theta}_j}} \xrightarrow{\text{d}} N(0,1),$$

for $j = 1, \ldots, (K-1)p$.

C   ADDITIONAL NUMERICAL STUDIES

C.1   Additional Simulations

C.1.1   Model 5

In this subsection, we explore a simulation example that is identical to Model 1, except that $\beta^{(1)*} = (2, 2, 2, \mathbf{0}_{p-3})$ and $\beta^{(2)*} = (0, 0, 0, 2, 2, 2, \mathbf{0}_{p-6})$.

Table C1 summarizes the sum of squared ℓ_2-estimation error of the contrast-based penalization and the over-parameterization method, where we can see that the contrast-based penalization leads to a smaller estimation error rate.

The average misclassification error rates for different methods are summarized in Table C2, where ℓ_1-penalized multinomial regression performs the best among all approaches.

Table C3 demonstrates that the debiased Lasso achieves a higher average coverage probability on S than the bootstrap method when $n = 100$ and 200, but has a lower average coverage probability on S when $n = 400$.

Table C4 shows that the debiased Lasso and the bootstrap method achieve comparable power while successfully controlling the Type-I error under 5%, with the debiased Lasso slightly outperforming the bootstrap method when $n = 100$. Compared to the results in Table 5 for Model 1, the power for Model 5 is significantly higher due to the increased signal strength.

The results of multiple testing are summarized in Table C5.

C.1.2   Model 6

In this subsection, we explore a simulation example that is identical to Model 1, except that $\beta^{(1)*} = (\mathbf{1}_{10}, \mathbf{0}_{p-10})$ and $\beta^{(2)*} = (\mathbf{0}_{10}, \mathbf{1}_{10}, \mathbf{0}_{p-20})$.

The sum of squared ℓ_2-estimation errors of the contrast-based penalization and the over-parameterization method are summarized in Table C6, which demonstrates the superiority of the contrast-based penalization.

Table C7 reports the average misclassification error rates on test data, showing that ℓ_1-penalized multinomial regression outperforms the other methods across all sample size settings.

Table C8 presents the average coverage probability of the debiased Lasso and the bootstrap method. The results indicate that the debiased Lasso provides better coverage than the bootstrap method on S.

Tables C9 and C10 summarize the results of individual testing and multiple testing, respectively.

C.1.3   Model 7

In this subsection, we explore a setting that is the same as Model 2, except that $\beta^{(1)*} = (2, 2, 2, \mathbf{0}_{p-3})$, $\beta^{(2)*} = (\mathbf{0}_3, 2, 2, 2, \mathbf{0}_{p-6})$, and $\beta^{(3)*} = (\mathbf{0}_6, 2, 2, 2, \mathbf{0}_{p-9})$.

Table C11 summarizes the sum of squared ℓ_2-estimation errors of the contrast-based penalization and the over-parameterization method, showing that the contrast-based penalization performs better.

Table C12 reports the average misclassification error rates on test data, demonstrating that ℓ_1-penalized multinomial regression dominates the other methods across all sample size settings.

The average coverage probability of 95% CIs is summarized in Table C13, where the debiased Lasso shows a larger coverage probability than the bootstrap method.

Tables C14 and C15 report the results for individual testing and multiple testing, respectively. The debiased Lasso outperforms the bootstrap and multiple-splitting methods by successfully controlling the Type I error and FWER while achieving higher power.

C.1.4   Model 8

In this subsection, we explore a setting that is identical to Model 2, except that $\beta^{(1)*} = (\mathbf{1}_{10}, \mathbf{0}_{p-10})$, $\beta^{(2)*} = (\mathbf{0}_{10}, \mathbf{1}_{10}, \mathbf{0}_{p-6})$, and $\beta^{(3)*} = (\mathbf{0}_{20}, \mathbf{1}_{10}, \mathbf{0}_{p-30})$.

Table C16 summarizes the sum of squared ℓ_2-estimation errors of the contrast-based penalization and the over-parameterization method, showing that contrast-based penalization performs better.

Table C17 reports the average misclassification error rates on test data, indicating that ℓ_1-penalized multinomial regression dominates the other methods across all sample size settings.

Table C18 summarizes the average coverage probability of 95% CIs for the debiased Lasso and the vector bootstrap method, where the debiased Lasso provides better coverage than the bootstrap method.

Tables C19 and C20 summarize the results of individual testing and multiple testing. For multiple testing, both debiased Lasso and multiple-splitting are overly conservative, successfully controlling the FWER but resulting in near-zero power.

D PROOFS

D.1 Lemmas and their proofs

Lemma 1. *Denote*

$$\mathcal{E}_n(\mathbf{\Delta}) = \mathcal{L}_n(\boldsymbol{\beta}^* + \mathbf{\Delta}) - \mathcal{L}_n(\boldsymbol{\beta}^*) - \nabla \mathcal{L}_n(\boldsymbol{\beta}^*)^\top \mathbf{\Delta}.$$

Under Assumptions 1 and 2, there exist constants $c_1, c_2 > 0$, such that

$$\mathcal{E}_n(\mathbf{\Delta}) \geq \frac{1}{K^2}\left(c_1\|\mathbf{\Delta}\|_2^2 - c_2\|\mathbf{\Delta}\|_2\|\mathbf{\Delta}\|_1\sqrt{\frac{\log p}{n}}\right), \quad \forall \|\mathbf{\Delta}\|_2 \leq 1,$$

with probability at least $1 - C(Kp)^{-1}$.

Lemma 2. *Suppose $K \geq 2$. For any $\{p_k\}_{k=1}^K$ with all $p_k \geq 0$, $\sum_{k=1}^K p_k = 1$ and any $\{x_k\}_{k=1}^{K-1} \in \mathbb{R}^{K-1}$,*

$$(x_1, \ldots, x_{K-1})^\top \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_{K-1} \\ -p_1p_2 & p_2(1-p_2) & \cdots & -p_2p_{K-1} \\ \vdots & \vdots & \vdots & \vdots \\ -p_1p_{K-1} & -p_2p_{K-1} & \cdots & p_{K-1}(1-p_{K-1}) \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_{K-1} \end{pmatrix} \geq p_K \sum_{k=1}^{K-1} p_k x_k^2.$$

Lemma 3. *Given $\boldsymbol{\beta} = \left((\boldsymbol{\beta}^{(1)})^\top, \ldots, (\boldsymbol{\beta}^{(K-1)})^\top\right)^\top \in \mathbb{R}^{(K-1)p}$, denote the joint distribution of $(\mathbf{x}, y)$ as $\mathbb{P}_{\mathbf{x},y;\boldsymbol{\beta}}$. Under Assumption 1, for any $\boldsymbol{\beta} = \left((\boldsymbol{\beta}^{(1)})^\top, \ldots, (\boldsymbol{\beta}^{(K-1)})^\top\right)^\top \in \mathbb{R}^{(K-1)p}$ and $\widetilde{\boldsymbol{\beta}} = \left((\widetilde{\boldsymbol{\beta}}^{(1)})^\top, \ldots, (\widetilde{\boldsymbol{\beta}}^{(K-1)})^\top\right)^\top \in \mathbb{R}^{(K-1)p}$, it holds that*

$$\text{KL}(\mathbb{P}_{\mathbf{x},y;\boldsymbol{\beta}} \| \mathbb{P}_{\mathbf{x},y;\widetilde{\boldsymbol{\beta}}}) \leq C\sigma^2 \|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}\|_2^2 \cdot \left(\frac{1}{K}e^{2T} + \exp\left\{-\frac{C'T^2}{\min_k \|\boldsymbol{\beta}^{(k)}\|_2^2}\right\}\right),$$

with any $T > 0$.

Lemma 4. *Under Assumptions 1-3, when $n \gg K^{18}s^2(\log(Kp))^2(\log(np))^2 + K^{10}s_0^2(\log(Kp))^2$, for any $j = 1, \ldots, (K-1)p$,*

$$\|\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j^*\|_2 \leq CK^{9/2}\sqrt{\frac{s(\log(Kp))(\log(np))}{n}} + CK^{5/2}\sqrt{\frac{s_0\log(Kp)}{n}}, \quad \|\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j^*\|_1 \leq CK^7 s\sqrt{\frac{\log(Kp)}{n}}\log(np) + CK^3 s_0 \sqrt{\frac{\log(Kp)}{n}},$$

$$|\widehat{\tau}_j^2 - (\tau_j^*)^2|, |\widehat{\tau}_j^{-2} - (\tau_j^*)^{-2}| \leq CK^{9/2}\sqrt{\frac{s(\log(Kp))(\log(np))}{n}} + CK^{5/2}\sqrt{\frac{s_0\log(Kp)}{n}},$$

with probability $1 - C'\exp\{-C''n\} - C'(np)^{-1} - C'(Kp)^{-1}$.

Proof of Lemma 1. By Taylor expansion, we know that

$$\mathcal{E}_n(\mathbf{\Delta}) = \mathbf{\Delta}^\top \left(\frac{1}{n}\sum_{i=1}^n \int_0^1 B(\mathbf{x}_i; \boldsymbol{\beta}^* + t\mathbf{\Delta})dt\right) \mathbf{\Delta},$$

where

$$B(\mathbf{x}; \boldsymbol{\beta}) := \begin{pmatrix} p_1(\mathbf{x};\boldsymbol{\beta})[1-p_1(\mathbf{x};\boldsymbol{\beta})]\mathbf{x}\mathbf{x}^\top & -p_1(\mathbf{x};\boldsymbol{\beta})p_2(\mathbf{x};\boldsymbol{\beta})\mathbf{x}\mathbf{x}^\top & \cdots & -p_1(\mathbf{x};\boldsymbol{\beta})p_{K-1}(\mathbf{x};\boldsymbol{\beta})\mathbf{x}\mathbf{x}^\top \\ -p_1(\mathbf{x};\boldsymbol{\beta})p_2(\mathbf{x};\boldsymbol{\beta})\mathbf{x}\mathbf{x}^\top & p_2(\mathbf{x};\boldsymbol{\beta})[1-p_2(\mathbf{x};\boldsymbol{\beta})]\mathbf{x}\mathbf{x}^\top & \cdots & -p_2(\mathbf{x};\boldsymbol{\beta})p_{K-1}(\mathbf{x};\boldsymbol{\beta})\mathbf{x}\mathbf{x}^\top \\ \vdots & \vdots & \vdots & \vdots \\ -p_1(\mathbf{x};\boldsymbol{\beta})p_{K-1}(\mathbf{x};\boldsymbol{\beta})\mathbf{x}\mathbf{x}^\top & -p_2(\mathbf{x};\boldsymbol{\beta})p_{K-1}(\mathbf{x};\boldsymbol{\beta})\mathbf{x}\mathbf{x}^\top & \cdots & p_{K-1}(\mathbf{x};\boldsymbol{\beta})[1-p_{K-1}(\mathbf{x};\boldsymbol{\beta})]\mathbf{x}\mathbf{x}^\top \end{pmatrix}.$$

By Lemma 2,

$$\mathcal{E}_n(\mathbf{\Delta}) \geq \frac{1}{n}\sum_{i=1}^n \int_0^1 p_K(\mathbf{x}_i; \boldsymbol{\beta}^* + t\mathbf{\Delta}) \sum_{k=1}^{K-1} p_k(\mathbf{x}_i; \boldsymbol{\beta}^* + t\mathbf{\Delta})((\mathbf{\Delta}^{(k)})^\top \mathbf{x}_i)^2 dt.$$

Fix $\boldsymbol{\Delta}$ with $\|\boldsymbol{\Delta}\|_2 = \delta \in (0, 1]$, and set $\tau = \tilde{k}\delta$, $\phi_\tau(u) = u^2 \mathbb{1}(|u| \le 2\tau)$, which leads to

$$
\begin{aligned}
\mathcal{E}_n(\boldsymbol{\Delta}) &\ge \frac{1}{n} \sum_{i=1}^n \int_0^1 p_K(\boldsymbol{x}_i; \boldsymbol{\beta}^* + t\boldsymbol{\Delta}) \sum_{k=1}^{K-1} p_k(\boldsymbol{x}_i; \boldsymbol{\beta}^* + t\boldsymbol{\Delta}) \varphi_\tau((\boldsymbol{\Delta}^{(k)})^\top \boldsymbol{x}_i) \mathbb{1}\left(\sum_{k=1}^{K-1} e^{(\boldsymbol{\beta}^{(k)*})^\top \boldsymbol{x}_i} \le T_1 \right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \cdot \mathbb{1}(|(\boldsymbol{\beta}^{(k)*})^\top \boldsymbol{x}_i| \le T_1)dt \\
&\ge \gamma \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^{K-1} \varphi_\tau((\boldsymbol{\Delta}^{(k)})^\top \boldsymbol{x}_i) \mathbb{1}\left(\sum_{k=1}^{K-1} e^{(\boldsymbol{\beta}^{(k)*})^\top \boldsymbol{x}_i} \le T_1 \right) \mathbb{1}(|(\boldsymbol{\beta}^{(k)*})^\top \boldsymbol{x}_i| \le T_2),
\end{aligned}
$$

where

$$
\begin{aligned}
\gamma &:= \min_{\sum_{k=1}^{K-1} e^{(\boldsymbol{\beta}^{(k)*})^\top \boldsymbol{x}_i} \le T_1, |(\boldsymbol{\beta}^{(k)*})^\top \boldsymbol{x}_i| \le T_2, |(\boldsymbol{\Delta}^{(k)})^\top \boldsymbol{x}_i| \le \tilde{k}} \min_k p_K(\boldsymbol{x}_i; \boldsymbol{\beta}^* + t\boldsymbol{\Delta}) p_k(\boldsymbol{x}_i; \boldsymbol{\beta}^* + t\boldsymbol{\Delta}) \\
&\ge \frac{e^{-(T_2 + \tilde{k})}}{(T_1 e^{\tilde{k}})^2}.
\end{aligned}
$$

Define

$$
\begin{aligned}
Z_n(r, \delta) = \sup_{\substack{\|\boldsymbol{\Delta}\|_2 \le \delta \\ \|\boldsymbol{\Delta}\|_1 \le r}} \Bigg| &\frac{1}{n} \sum_{i=1}^n \sum_{k=1}^{K-1} \varphi_\tau((\boldsymbol{\Delta}^{(k)})^\top \boldsymbol{x}_i) \mathbb{1}\left(\sum_{k=1}^{K-1} e^{(\boldsymbol{\beta}^{(k)*})^\top \boldsymbol{x}_i} \le T_1 \right) \mathbb{1}(|(\boldsymbol{\beta}^{(k)*})^\top \boldsymbol{x}_i| \le T_2) \\
&- \mathbb{E}\left[\sum_{k=1}^{K-1} \varphi_\tau((\boldsymbol{\Delta}^{(k)})^\top \boldsymbol{x}) \mathbb{1}\left(\sum_{k=1}^{K-1} e^{(\boldsymbol{\beta}^{(k)*})^\top \boldsymbol{x}} \le T_1 \right) \mathbb{1}(|(\boldsymbol{\beta}^{(k)*})^\top \boldsymbol{x}| \le T_2) \right] \Bigg|.
\end{aligned}
$$

Note that

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{K-1} \varphi_\tau((\boldsymbol{\Delta}^{(k)})^\top \boldsymbol{x}) \mathbb{1}\left(\sum_{k=1}^{K-1} e^{(\boldsymbol{\beta}^{(k)*})^\top \boldsymbol{x}} \le T_1 \right) \mathbb{1}(|(\boldsymbol{\beta}^{(k)*})^\top \boldsymbol{x}| \le T_2) \right] \\
&= \underbrace{\mathbb{E}\left[\sum_{k=1}^{K-1} \varphi_\tau((\boldsymbol{\Delta}^{(k)})^\top \boldsymbol{x}) \right]}_{[1]} - \underbrace{\mathbb{E}\left[\sum_{k=1}^{K-1} \varphi_\tau((\boldsymbol{\Delta}^{(k)})^\top \boldsymbol{x}) \mathbb{1}\left(\sum_{k=1}^{K-1} e^{(\boldsymbol{\beta}^{(k)*})^\top \boldsymbol{x}} > T_1 \right) \right]}_{[2]} \\
&\quad - \underbrace{\mathbb{E}\left[\sum_{k=1}^{K-1} \varphi_\tau((\boldsymbol{\Delta}^{(k)})^\top \boldsymbol{x}) \mathbb{1}(|(\boldsymbol{\beta}^{(k)*})^\top \boldsymbol{x}| > T_2) \right]}_{[3]}.
\end{aligned}
\tag{D2}
$$

And

$$
\begin{aligned}
[1] &\ge \mathbb{E}\left[\sum_{k=1}^{K-1} ((\boldsymbol{\Delta}^{(k)})^\top \boldsymbol{x})^2 \cdot \mathbb{1}(|(\boldsymbol{\Delta}^{(k)})^\top \boldsymbol{x}| \le \tau) \right] \\
&\ge \mathbb{E}\left[\sum_{k=1}^{K-1} ((\boldsymbol{\Delta}^{(k)})^\top \boldsymbol{x})^2 \right] - \mathbb{E}\left[\sum_{k=1}^{K-1} ((\boldsymbol{\Delta}^{(k)})^\top \boldsymbol{x})^2 \cdot \mathbb{1}(|(\boldsymbol{\Delta}^{(k)})^\top \boldsymbol{x}| > \tau) \right] \\
&\ge \|\boldsymbol{\Delta}\|_2^2 \lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{x}}) - \sum_{k=1}^{K-1} \sqrt{\mathbb{E}((\boldsymbol{\Delta}^{(k)})^\top \boldsymbol{x})^4} \sqrt{\mathbb{P}(|(\boldsymbol{\Delta}^{(k)})^\top \boldsymbol{x}| > \tau)} \\
&\ge \|\boldsymbol{\Delta}\|_2^2 \lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{x}}) - \sum_{k=1}^{K-1} C\sigma^2 \|\boldsymbol{\Delta}^{(k)}\|_2^2 \cdot \exp\left\{ -\frac{\tau^2}{2\|\boldsymbol{\Delta}^{(k)}\|_2^2 \sigma^2} \right\} \\
&\ge \|\boldsymbol{\Delta}\|_2^2 \cdot \left[\lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{x}}) - C\sigma^2 \exp\left\{ -\frac{\tau^2}{2\max_k \|\boldsymbol{\Delta}^{(k)}\|_2^2 \sigma^2} \right\} \right] \\
&\ge \|\boldsymbol{\Delta}\|_2^2 \cdot c\lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{x}}),
\end{aligned}
\tag{D3}
$$

if $\tau = \tilde{k}\delta \geq C'\delta\sigma^2[\log(\lambda_{\min}^{-1}(\boldsymbol{\Sigma_x})) + \log\sigma^2]$ with a large constant $C' > 0$.

$$
\begin{aligned}
[2] &\leq \sum_{k=1}^{K-1} \sqrt{\mathbb{E}((\boldsymbol{\Delta}^{(k)})^\top \boldsymbol{x})^4} \sqrt{\mathbb{P}\left(\sum_{k=1}^{K-1} e^{(\boldsymbol{\beta}^{(k)*})^\top \boldsymbol{x}} > T_1\right)} \\
&\leq \sum_{k=1}^{K-1} C\sigma^2 \|\boldsymbol{\Delta}^{(k)}\|_2^2 \cdot \sqrt{\frac{\sum_{k=1}^{K-1} \mathbb{E}e^{(\boldsymbol{\beta}^{(k)*})^\top \boldsymbol{x}}}{T_1}} \\
&\leq C\sigma^2 \|\boldsymbol{\Delta}\|_2^2 \cdot \sqrt{\frac{K}{T_1}} \cdot \exp\left\{C \max_k \|\boldsymbol{\beta}^{(k)*}\|_2^2\right\} \\
&\leq \|\boldsymbol{\Delta}\|_2^2 \cdot \frac{1}{4}c\lambda_{\min}(\boldsymbol{\Sigma_x}),
\end{aligned}
\tag{D4}
$$

if $T_1 = C'K\sigma^4 \exp\left\{C' \max_k \|\boldsymbol{\beta}^{(k)*}\|_2^2\right\} \lambda_{\min}^{-2}(\boldsymbol{\Sigma_x})$ with a large constant $C' > 0$.

$$
\begin{aligned}
[3] &\leq \sum_{k=1}^{K-1} C\sigma^2 \|\boldsymbol{\Delta}^{(k)}\|_2^2 \cdot \sqrt{\mathbb{P}(|(\boldsymbol{\beta}^{(k)*})^\top \boldsymbol{x}| > T_2)} \\
&\leq C\|\boldsymbol{\Delta}\|_2^2 \cdot \exp\left\{-\frac{T_2^2}{2 \max_k \|\boldsymbol{\beta}^{(k)*}\|_2^2 \sigma^2}\right\} \\
&\leq \|\boldsymbol{\Delta}\|_2^2 \cdot \frac{1}{4}c\lambda_{\min}(\boldsymbol{\Sigma_x}),
\end{aligned}
\tag{D5}
$$

if $T_2 = C'\sigma^2 \max_k \|\boldsymbol{\beta}^{(k)*}\|_2^2[1 + \log(\lambda_{\min}^{-1}(\boldsymbol{\Sigma_x}))]$.

Putting (D2), (D3), (D4) and (D5) together,

$$
C_e(\boldsymbol{\Delta}) := \mathbb{E}\left[\sum_{k=1}^{K-1} \varphi_\tau((\boldsymbol{\Delta}^{(k)})^\top \boldsymbol{x})\mathbb{1}\left(\sum_{k=1}^{K-1} e^{(\boldsymbol{\beta}^{(k)*})^\top \boldsymbol{x}} \leq T_1\right)\mathbb{1}(|(\boldsymbol{\beta}^{(k)*})^\top \boldsymbol{x}| \leq T_2)\right] \geq \|\boldsymbol{\Delta}\|_2^2 \cdot \frac{1}{2}c\lambda_{\min}(\boldsymbol{\Sigma_x}).
$$

By symmetrization,

$$
\begin{aligned}
&\mathbb{E}Z_n(r, \delta) \\
&\leq 2\mathbb{E}_{\boldsymbol{x},\boldsymbol{\epsilon}}\left[\sup_{\substack{\|\boldsymbol{\Delta}\|_2 \leq \delta \\ \|\boldsymbol{\Delta}\|_1 \leq r}} \frac{1}{n}\sum_{i=1}^n \sum_{k=1}^{K-1} \epsilon_{i,k}\varphi_\tau((\boldsymbol{\Delta}^{(k)})^\top \boldsymbol{x}_i)\mathbb{1}\left(\sum_{k=1}^{K-1} e^{(\boldsymbol{\beta}^{(k)*})^\top \boldsymbol{x}_i} \leq T_1\right)\mathbb{1}(|(\boldsymbol{\beta}^{(k)*})^\top \boldsymbol{x}_i| \leq T_2)\right] \\
&\leq 8\tilde{k}\delta\mathbb{E}_{\boldsymbol{x},\boldsymbol{\epsilon}}\left[\sup_{\substack{\|\boldsymbol{\Delta}\|_2 \leq \delta \\ \|\boldsymbol{\Delta}\|_1 \leq r}} \frac{1}{n}\sum_{i=1}^n \sum_{k=1}^{K-1} \epsilon_{i,k} \cdot ((\boldsymbol{\Delta}^{(k)})^\top \boldsymbol{x}_i)\right] \\
&= 8\tilde{k}\delta\mathbb{E}_{\boldsymbol{x},\boldsymbol{\epsilon}}\left[\sup_{\substack{\|\boldsymbol{\Delta}\|_2 \leq \delta \\ \|\boldsymbol{\Delta}\|_1 \leq r}} \sum_{k=1}^{K-1} \left\langle \boldsymbol{\Delta}^{(k)}, \frac{1}{n}\sum_{i=1}^n \epsilon_{i,k}\boldsymbol{x}_i\right\rangle\right] \\
&\leq 8\tilde{k}\delta\mathbb{E}_{\boldsymbol{x},\boldsymbol{\epsilon}}\left[\sup_{\substack{\|\boldsymbol{\Delta}\|_2 \leq \delta \\ \|\boldsymbol{\Delta}\|_1 \leq r}} \sum_{k=1}^{K-1} \|\boldsymbol{\Delta}^{(k)}\|_1 \left\|\frac{1}{n}\sum_{i=1}^n \epsilon_{i,k}\boldsymbol{x}_i\right\|_\infty\right] \\
&\leq 8\tilde{k}\delta r \cdot \max_k \left\|\frac{1}{n}\sum_{i=1}^n \epsilon_{i,k}\boldsymbol{x}_i\right\|_\infty,
\end{aligned}
\tag{D6}
$$

where (D6) is by the contraction inequality for Rademacher variables [82,83]. And by Theorem 3.26 in [83], for any $\delta' > 0$,

$$
\mathbb{P}(Z_n(r, \delta) \geq \mathbb{E}Z_n(r, \delta) + \delta') \leq \exp\left\{-\frac{n\delta'^2}{4 \cdot (2K\tilde{k}\delta^2)^2}\right\},
\tag{D7}
$$

implying that

$$
Z_n(r, \delta) \leq 8\tilde{k}\delta r \cdot \max_k \left\| \frac{1}{n} \sum_{i=1}^n \epsilon_{i,k} \boldsymbol{x}_i \right\|_\infty + \delta'
$$
$$
\leq C\tilde{k}\delta r \cdot \sqrt{\frac{\log(Kp)}{n}} + CK\tilde{k}\delta^2 \sqrt{\frac{\log(Kp)}{n}},
$$

with probability at least $1 - \exp\left\{ -\frac{n(\delta')^2}{16K^2\tilde{k}^2\delta^2} \right\} = 1 - C'(Kp)^{-1}$, by letting $\delta' = C''K\tilde{k}\delta\sqrt{\frac{\log(Kp)}{n}}$ with a large constant $C'' > 0$. This implies that

$$
\mathcal{E}_n(\boldsymbol{\Delta}) \geq \gamma \left(\frac{1}{2}c\delta^2\lambda_{\min}(\boldsymbol{\Sigma_x}) - CK\tilde{k}\delta^2\sqrt{\frac{\log(Kp)}{n}} - C\tilde{k}\delta r \cdot \sqrt{\frac{\log(Kp)}{n}} \right)
$$
$$
\geq \frac{e^{-(T_2+\tilde{k}\delta)}}{(T_1 e^{\tilde{k}\delta})^2} \cdot \left(\frac{1}{2}c\lambda_{\min}(\boldsymbol{\Sigma_x}) - CK\tilde{k}\sqrt{\frac{\log(Kp)}{n}} - C\tilde{k}r \cdot \sqrt{\frac{\log(Kp)}{n}} \right)
$$
$$
\geq \underbrace{\frac{c'\lambda_{\min}^4(\boldsymbol{\Sigma_x})}{\exp\left\{ C'\max_k \|\boldsymbol{\beta}^{(k)*}\|_2^2 \right\} K^2\sigma^8} \exp\{-C\sigma^2[\log(\lambda_{\min}^{-1}(\boldsymbol{\Sigma_x})) + \log\sigma^2]\}}_{C_\star}
$$
$$
\cdot \left(\frac{1}{2}c\lambda_{\min}(\boldsymbol{\Sigma_x})\delta^2 - CK\tilde{k}\delta^2\sqrt{\frac{\log(Kp)}{n}} - C\tilde{k}\delta r \cdot \sqrt{\frac{\log(Kp)}{n}} \right).
$$

Denote a random event

$$
\mathcal{A} = \left\{ \mathcal{E}_n(\boldsymbol{\Delta}) \geq C_\star \left[\left(\frac{1}{2}c\lambda_{\min}(\boldsymbol{\Sigma_x}) - CK\tilde{k}\sqrt{\frac{\log(Kp)}{n}} \right) \|\boldsymbol{\Delta}\|_2^2 - C\tilde{k}r \cdot \sqrt{\frac{\log(Kp)}{n}} \|\boldsymbol{\Delta}\|_2\|\boldsymbol{\Delta}\|_1 \right] \right\}.
$$

Consider the following peeling argument. Let the event $\mathcal{S}_{k,l} = \{ \boldsymbol{\Delta} \in \mathbb{R}^{(K-1)p} : 2^{k-1} \leq \frac{\|\boldsymbol{\Delta}\|_1}{\|\boldsymbol{\Delta}\|_2} \leq 2^k, 2^{l-1} \leq \|\boldsymbol{\Delta}\|_2 \leq 2^l \}$, then it is easy to see that $\mathcal{A}^c \subseteq \bigcup_{k=1}^N \bigcup_{l=1}^\infty \mathcal{S}_{k,l}$ with $N = \lceil \frac{\log(Kp)}{2\log 2} \rceil$. Suppose $\widehat{\boldsymbol{\Delta}} \in \mathcal{S}_{k,l}$ for some k and l is not in $\mathcal{A}$. Then since

$$
C_\star \left[\left(\frac{1}{2}c\lambda_{\min}(\boldsymbol{\Sigma_x}) - CK\tilde{k}\sqrt{\frac{\log(Kp)}{n}} \right) \|\widehat{\boldsymbol{\Delta}}\|_2^2 - C\tilde{k} \cdot \sqrt{\frac{\log(Kp)}{n}} \|\widehat{\boldsymbol{\Delta}}\|_2\|\widehat{\boldsymbol{\Delta}}\|_1 \right]
$$
$$
\geq \mathcal{E}_n(\widehat{\boldsymbol{\Delta}})
$$
$$
\geq C_\star C_e(\widehat{\boldsymbol{\Delta}}) - Z_n(2^{k+l}, 2^l)
$$
$$
\geq C_\star \|\widehat{\boldsymbol{\Delta}}\|_2^2 \cdot \frac{1}{2}c\lambda_{\min}(\boldsymbol{\Sigma_x}) - Z_n(2^{k+l}, 2^l),
$$

we must have

$$
Z_n(2^{k+l}, 2^l) \geq C_\star CK\tilde{k}\sqrt{\frac{\log(Kp)}{n}} \cdot 2^{2l-2} + C_\star C\tilde{k}\sqrt{\frac{\log(Kp)}{n}} \cdot 2^{k-1} \cdot 2^{2l-2}.
$$

Therefore by (D7), for any $k \leq N$ and $l \geq 1$,

$$
\mathbb{P}(\mathcal{A}^c \cap \mathcal{S}_{k,l}) \leq \mathbb{P} \left(Z_n(2^{k+l}, 2^l) \geq C_\star CK\tilde{k}\sqrt{\frac{\log(Kp)}{n}} \cdot 2^{2l-2} + C_\star C\tilde{k}\sqrt{\frac{\log(Kp)}{n}} \cdot 2^{k-1} \cdot 2^{2l-2} \right)
$$
$$
\leq \exp\left\{ -C'K^2 \log(Kp) \cdot 2^{4l-4} \right\}
$$
$$
\leq C'' \frac{1}{Kp\log(Kp)} \cdot 16^{-l}.
$$

Hence

$$
\mathbb{P}(\mathcal{A}^c) \leq \sum_{k=1}^N \sum_{l=1}^\infty \mathbb{P}(\mathcal{A}^c \cap \mathcal{S}_{k,l})
$$

$$\leq N \cdot C'' \frac{1}{Kp \log(Kp)} \cdot \sum_{l=1}^{\infty} 16^{-l}$$

$$\leq C \cdot \frac{1}{Kp},$$

which completes our proof. $\square$

Proof of Lemma 2. By Cauchy-Schwarz inequality,

$$\text{LHS} \geq \sum_{k=1}^{K-1} p_k x_k^2 - \left(\sum_{k=1}^{K-1} p_k x_k \right)^2 = \left(\sum_{k=1}^{K-1} p_k x_k^2 \right) \left(\sum_{k=1}^{K-1} p_k \right) - \left(\sum_{k=1}^{K-1} p_k x_k \right)^2 + p_K \sum_{k=1}^{K-1} p_k x_k^2 \geq p_K \sum_{k=1}^{K-1} p_k x_k^2.$$

$\square$

Proof of Lemma 3. First, notice that

$$\text{KL}(\mathbb{P}_{X,Y;\boldsymbol{\beta}} \| \mathbb{P}_{X,Y;\widetilde{\boldsymbol{\beta}}}) = -\sum_{k=1}^{K-1} \mathbb{E}_X \left[\mathbb{P}(Y = k|X = \boldsymbol{x}; \boldsymbol{\beta}) \log \left[\frac{\mathbb{P}(Y = k|X = \boldsymbol{x}; \widetilde{\boldsymbol{\beta}})}{\mathbb{P}(Y = k|X = \boldsymbol{x}; \boldsymbol{\beta})} \right] \right],$$

where

$$\log \left[\frac{\mathbb{P}(Y = k|X = \boldsymbol{x}; \widetilde{\boldsymbol{\beta}})}{\mathbb{P}(Y = k|X = \boldsymbol{x}; \boldsymbol{\beta})} \right]$$

$$= \frac{(\nabla_{\boldsymbol{\beta}} \mathbb{P}(Y = k|X = \boldsymbol{x}; \boldsymbol{\beta})|_{\boldsymbol{\beta}})^{\top}}{\mathbb{P}(Y = k|X = \boldsymbol{x}; \boldsymbol{\beta})} (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \frac{1}{2} (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\top} \nabla_{\boldsymbol{\beta}}^2 \mathbb{P}(Y = k|X = \boldsymbol{x}; \boldsymbol{\beta})|_{\boldsymbol{\beta} = \widetilde{\boldsymbol{\beta}}_t} (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}),$$

and $\widetilde{\boldsymbol{\beta}}_t = (1 - t)\widetilde{\boldsymbol{\beta}} + \boldsymbol{\beta}$ with some $t \in [0, 1]$. It implies that

$$\text{KL}(\mathbb{P}_{X,Y;\boldsymbol{\beta}} \| \mathbb{P}_{X,Y;\widetilde{\boldsymbol{\beta}}})$$

$$= -\sum_{k=1}^{K-1} \mathbb{E}_X \left[\mathbb{P}(Y = k|X = \boldsymbol{x}; \boldsymbol{\beta})(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\top} \nabla_{\boldsymbol{\beta}}^2 \mathbb{P}(Y = k|X = \boldsymbol{x}; \boldsymbol{\beta})|_{\boldsymbol{\beta} = \widetilde{\boldsymbol{\beta}}_t} (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \right].$$

And by basic algebra,

$$-\nabla_{\boldsymbol{\beta}}^2 \mathbb{P}(Y = k|X = \boldsymbol{x}; \boldsymbol{\beta})|_{\boldsymbol{\beta} = \widetilde{\boldsymbol{\beta}}_t}$$

$$= \begin{pmatrix} p_1(\boldsymbol{x}; \widetilde{\boldsymbol{\beta}}_t)(1 - p_1(\boldsymbol{x}; \widetilde{\boldsymbol{\beta}}_t)) & \dots & -p_1(\boldsymbol{x}; \widetilde{\boldsymbol{\beta}}_t)p_{K-1}(\boldsymbol{x}; \widetilde{\boldsymbol{\beta}}_t) \\ \vdots & \vdots & \vdots \\ -p_1(\boldsymbol{x}; \widetilde{\boldsymbol{\beta}}_t)p_{K-1}(\boldsymbol{x}; \widetilde{\boldsymbol{\beta}}_t) & \dots & p_{K-1}(\boldsymbol{x}; \widetilde{\boldsymbol{\beta}}_t)(1 - p_{K-1}(\boldsymbol{x}; \widetilde{\boldsymbol{\beta}}_t)) \end{pmatrix} \otimes \boldsymbol{x}\boldsymbol{x}^{\top}$$

$$\leq \text{diag}(p_1(\boldsymbol{x}; \widetilde{\boldsymbol{\beta}}_t)(1 - p_1(\boldsymbol{x}; \widetilde{\boldsymbol{\beta}}_t)), \dots, p_{K-1}(\boldsymbol{x}; \widetilde{\boldsymbol{\beta}}_t)(1 - p_{K-1}(\boldsymbol{x}; \widetilde{\boldsymbol{\beta}}_t))) \otimes \boldsymbol{x}\boldsymbol{x}^{\top},$$

which leads to

$$\text{KL}(\mathbb{P}_{X,Y;\boldsymbol{\beta}} \| \mathbb{P}_{X,Y;\widetilde{\boldsymbol{\beta}}})$$

$$\leq \frac{1}{2} \sum_{k=1}^{K-1} (\widetilde{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^{(k)})^{\top} \mathbb{E}_{\boldsymbol{x}} \left[\mathbb{P}(Y = k|X = \boldsymbol{x}; \boldsymbol{\beta})\boldsymbol{x}\boldsymbol{x}^{\top} \right] (\widetilde{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^{(k)})$$

$$\leq \frac{1}{2} \sum_{k=1}^{K-1} \mathbb{E}_{\boldsymbol{x}} \left[\mathbb{P}(Y = k|X = \boldsymbol{x}; \boldsymbol{\beta})(\boldsymbol{x}^{\top}(\widetilde{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^{(k)}))^2 \mathbb{1} \left(\max_k |\boldsymbol{x}^{\top}\boldsymbol{\beta}^{(k)}| \leq T \right) \right]$$

$$+ \frac{1}{2} \sum_{k=1}^{K-1} \mathbb{E}_{\boldsymbol{x}} \left[\mathbb{P}(Y = k|X = \boldsymbol{x}; \boldsymbol{\beta})(\boldsymbol{x}^{\top}(\widetilde{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^{(k)}))^2 \mathbb{1} \left(\max_k |\boldsymbol{x}^{\top}\boldsymbol{\beta}^{(k)}| > T \right) \right]$$

$$\leq \frac{e^T}{2Ke^T} \sum_{k=1}^{K-1} \mathbb{E}_x \left[(x^\top (\widetilde{\beta}^{(k)} - \beta^{(k)}))^2 \right] + \frac{1}{2} \sum_{k=1}^{K-1} \sqrt{\mathbb{E}_x \left[(x^\top (\widetilde{\beta}^{(k)} - \beta^{(k)}))^4 \right] \cdot \mathbb{P} \left(\max_k |x^\top \beta^{(k)}| > T \right)}$$

$$\leq C\sigma^2 \left(\frac{1}{K} e^{2T} \|\widetilde{\beta} - \beta\|_2^2 + \sum_{k=1}^{K-1} \|\widetilde{\beta}^{(k)} - \beta^{(k)}\|_2^2 \cdot \exp \left\{ -\frac{C'T^2}{\min_k \|\beta^{(k)}\|_2^2} \right\} \right)$$

$$\leq C\sigma^2 \|\widetilde{\beta} - \beta\|_2^2 \cdot \left(\frac{1}{K} e^{2T} + \exp \left\{ -\frac{C'T^2}{\min_k \|\beta^{(k)}\|_2^2} \right\} \right).$$

$\square$

Proof of Lemma 4. Denote $\widehat{\Sigma}_\beta = \frac{1}{n} \sum_{i=1}^n B(x_i; \beta)$. By the optimality of $\widehat{\gamma}_j$, we have

$$\frac{1}{2} \widehat{\gamma}_j^\top \widehat{\Sigma}_{-j,-j} \widehat{\gamma}_j - \widehat{\Sigma}_{-j,j} + \lambda_j \|\widehat{\gamma}_j\|_1 \leq \frac{1}{2} (\gamma_j^*)^\top \widehat{\Sigma}_{-j,-j} \gamma_j^* - \widehat{\Sigma}_{-j,j} + \lambda_j \|\gamma_j^*\|_1,$$

which implies that

$$\frac{1}{2} (\widehat{\gamma}_j - \gamma_j^*)^\top \widehat{\Sigma}_{-j,-j} (\widehat{\gamma}_j - \gamma_j^*)$$

$$\leq (\widehat{\Sigma}_{-j,j} - \widehat{\Sigma}_{-j,-j} \gamma_j^*)^\top (\widehat{\gamma}_j - \gamma_j^*) + \lambda_j (\|\gamma_j^*\|_1 - \|\widehat{\gamma}_j\|_1)$$

$$\leq (\widehat{\Sigma}_{-j,j} - \widehat{\Sigma}_{\beta^*,-j,j} - \widehat{\Sigma}_{-j,-j} \gamma_j^* + \widehat{\Sigma}_{\beta^*,-j,-j} \gamma_j^*)^\top (\widehat{\gamma}_j - \gamma_j^*) + (\widehat{\Sigma}_{\beta^*,-j,j} - \widehat{\Sigma}_{\beta^*,-j,-j} \gamma_j^*)^\top (\widehat{\gamma}_j - \gamma_j^*) + \lambda_j (\|\gamma_j^*\|_1 - \|\widehat{\gamma}_j\|_1). \quad \text{(D8)}$$

Let us bound $(\widehat{\Sigma}_{-j,j} - \widehat{\Sigma}_{\beta^*,-j,j})^\top (\widehat{\gamma}_j - \gamma_j^*)$ on the RHS. WLOG, consider the case $j = 1$. Denote $\widehat{\gamma}_j = ((\widehat{\gamma}_j^{(1)})^\top, \ldots, (\widehat{\gamma}_j^{(K-1)})^\top)^\top$, $\gamma_j^* = ((\gamma_j^{(1)*})^\top, \ldots, (\gamma_j^{(K-1)*})^\top)^\top$, $\beta^* = ((\beta^{(1)*})^\top, \ldots, (\beta^{(K-1)*})^\top)^\top$, and $\widehat{\beta} = ((\widehat{\beta}^{(1)})^\top, \ldots, (\widehat{\beta}^{(K-1)})^\top)^\top$. Then

$$\left| (\widehat{\Sigma}_{-j,j} - \widehat{\Sigma}_{\beta^*,-j,j})^\top (\widehat{\gamma}_j - \gamma_j^*) \right| = \left| \frac{1}{n} \sum_{k=1}^{K-1} \sum_{i=1}^n (\widehat{\gamma}_j^{(k)} - \gamma_j^{(k)*})^\top x_i \cdot \left[g_k(\{x_i^\top \widehat{\beta}^{(k)}\}_{k=1}^{K-1}) - g_k(\{x_i^\top \beta^{(k)*}\}_{k=1}^{K-1}) \cdot x_{i,1} \right] \right|$$

$$\leq \frac{c}{2nK^2} \sum_{k=1}^{K-1} \sum_{i=1}^n \left[(\widehat{\gamma}_j^{(k)} - \gamma_j^{(k)*})^\top x_i \right]^2 + \frac{K^2}{2nc} \sum_{k=1}^{K-1} \sum_{i=1}^n [x_i^\top (\widehat{\beta}^{(k)} - \beta^{(k)*})]^2 \cdot [\nabla g_k(x_i^\top \widetilde{\beta}_{k,i})]^2 \cdot |x_{i,1}|^2,$$

$$\text{(D9)}$$

where $g_1(\{x_i^\top \widehat{\beta}^{(k)}\}_{k=1}^{K-1}) = p_1(x_i; \widehat{\beta})[1 - p_1(x_i; \widehat{\beta})]$, $g_k(\{x_i^\top \widehat{\beta}^{(k)}\}_{k=1}^{K-1}) = -p_1(x_i; \widehat{\beta}) p_k(x_i; \widehat{\beta})$ for $k \geq 2$. Define five events

$$\mathcal{A}_1 = \left\{ \frac{1}{n} \sum_{i=1}^n (x_i^\top u)^2 \leq C \|u\|_2^2, \forall u \in \Pi_1(Ks) \right\}, \quad \mathcal{A}_2 = \left\{ \max_{i=1:n} \|x_i\|_\infty \leq C' \sqrt{\log(np)} \right\},$$

$$\mathcal{A}_3 = \left\{ \|\widehat{\beta} - \beta^*\|_2 \leq CK^{5/2} \sqrt{\frac{s \log(Kp)}{n}}, \|\widehat{\beta} - \beta^*\|_1 \leq CK^3 \sqrt{\frac{s \log(Kp)}{n}} \right\}$$

$$\mathcal{A}_4 = \left\{ \frac{1}{n} \sum_{i=1}^n (x_i^\top u)^2 \leq C \|u\|_2^2, \forall u \in \Pi_1(K^7 s) \right\}, \quad \mathcal{A}_5 = \left\{ \lambda_j = C \sqrt{\frac{\log(Kp)}{n}} \geq 2 \|\widehat{\Sigma}_{\beta^*,-j,j} - \widehat{\Sigma}_{\beta^*,-j,-j} \gamma_j^*\|_\infty \right\}.$$

where $\Pi_1(Ks) = \{u \in \mathbb{R}^p : \|u\|_2 \leq 1, \|u\|_1 \leq \sqrt{Ks}\}$, C and C' are some large positive constants. We want to argue that $\mathbb{P}(\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3 \cap \mathcal{A}_4 \cap \mathcal{A}_5) \geq 1 - 2C'' \exp\{-C'''n\} - C''(np)^{-1} - C''(Kp)^{-1}$. Note that by the sub-Gaussianity of x_{i1}, we have $\mathbb{P}(\mathcal{A}_2) \geq 1 - C''(np)^{-1}$. By Lemma 3.1 of [84], $\Pi_1(Ks) \subseteq 2\overline{\text{conv}}(\Pi_0(Ks))$ where $\Pi_0(Ks) = \{u \in \mathbb{R}^p : \|u\|_2 \leq 1, \|u\|_0 \leq Ks\}$. Then by Lemma 2.3 in [85], $\Pi_0(Ks) \subseteq 2\overline{\text{conv}}(\Lambda)$, where Λ is a 1/2-cover of $\Pi_0(Ks)$ with $|\Lambda| \leq C\binom{p}{Ks} 5^{Ks} \leq \exp\{C'n\}$ when $n \geq C'Ks \log p$. By Theorem 2.4 in [85], $\mathbb{P}(\mathcal{A}_1) \geq 1 - C'' \exp\{-C'''n\}$ when $n \geq C'Ks \log p$. By a similar argument, $\mathbb{P}(\mathcal{A}_4) \geq 1 - C'' \exp\{-C'''n\}$ when $n \geq C''s \log(pK)$. Additionally, by Theorem 1, we have $\mathbb{P}(\mathcal{A}_3) \geq 1 - C''(Kp)^{-1}$. Regarding $\mathbb{P}(\mathcal{A}_5)$, note that

$$\widehat{\Sigma}_{\beta^*,-j,j} - \widehat{\Sigma}_{\beta^*,-j,-j} \gamma_j^* = \widehat{\Sigma}_{\beta^*,-j,j} - \Sigma_{-j,j} + (\Sigma_{-j,-j} - \widehat{\Sigma}_{\beta^*,-j,-j}) \Sigma_{-j,-j}^{-1} \Sigma_{-j,j},$$

leading to

$$\|\widehat{\Sigma}_{\beta^*,-j,j} - \widehat{\Sigma}_{\beta^*,-j,-j} \gamma_j^*\|_\infty \lesssim \|\widehat{\Sigma}_{\beta^*,-j,j} - \Sigma_{-j,j}\|_\infty + \|\Sigma_{-j,-j} - \widehat{\Sigma}_{\beta^*,-j,-j}\|_{\max}.$$

Denote $B_{ij} = (\widehat{\Sigma}_{\beta^*,-j,-j} - \Sigma_{-j,-j})_{ij}$. It holds that

$$\left| (\widehat{\Sigma}_{\beta^*,-j,-j} - \Sigma_{-j,-j})_{11} \right| \leq \left| \frac{1}{n} \sum_{i=1}^{n} x_{i1}^2 p_1(\boldsymbol{x}_i; \beta^*)(1 - p_1(\boldsymbol{x}_i; \beta^*)) - \mathbb{E}[x_{i1}^2 p_1(\boldsymbol{x}_i; \beta^*)(1 - p_1(\boldsymbol{x}_i; \beta^*))] \right|.$$

Since $p_1(\boldsymbol{x}_i; \beta^*)(1 - p_1(\boldsymbol{x}_i; \beta^*))$ is always bounded, by the tail bound of sub-exponential variables, we have

$$\mathbb{P}(B_{ij} \geq \delta) \leq C \exp\{-C' n \delta^2\}, \ \forall i,j.$$

Therefore

$$\mathbb{P}\left(\max_{i,j} B_{ij} \geq C\sqrt{\frac{\log(Kp)}{n}} \right) \leq C'[p(K-1)]^2 \exp\{-C'' \log(Kp)\} \leq C'''(pK)^{-1}.$$

Hence $\mathbb{P}\left(\|\Sigma_{-j,-j} - \widehat{\Sigma}_{\beta^*,-j,-j}\|_{\max} \leq C\sqrt{\frac{\log(Kp)}{n}} \right) \geq 1 - C'''(pK)^{-1}$. Similarly, we can show $\mathbb{P}\left(\|\widehat{\Sigma}_{\beta^*,-j,j} - \Sigma_{-j,j}\|_{\infty} \leq C\sqrt{\frac{\log(Kp)}{n}} \right) \geq 1 - C'''(pK)^{-1}$. Therefore $\mathbb{P}(\mathcal{A}_5) \geq 1 - C'''(pK)^{-1}$.

Then $\mathbb{P}(\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3 \cap \mathcal{A}_4 \cap \mathcal{A}_5) \geq 1 - C'' \exp\{-C''' n\} - C''(np)^{-1} - C''(Kp)^{-1}$ follows by the basic inequality $\mathbb{P}(\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3 \cap \mathcal{A}_4 \cap \mathcal{A}_5) \geq 1 - \mathbb{P}(\mathcal{A}_1^c) - \mathbb{P}(\mathcal{A}_2^c) - \mathbb{P}(\mathcal{A}_3^c) - \mathbb{P}(\mathcal{A}_4^c) - \mathbb{P}(\mathcal{A}_5^c)$.

Under $\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3 \cap \mathcal{A}_4 \cap \mathcal{A}_5$, we have

$$\begin{aligned}
|\boldsymbol{x}_i^\top \widehat{\beta}^{(k)}| &\leq |\boldsymbol{x}_i^\top (\widehat{\beta}^{(k)} - \beta^{(k)*})| + |\boldsymbol{x}_i^\top \beta^{(k)*}| \\
&\leq \|\boldsymbol{x}_i\|_\infty \|\widehat{\beta}^{(k)} - \beta^{(k)*}\|_1 + C'' \\
&\lesssim C'\sqrt{\log(np)} \cdot K^{5/2} \sqrt{\frac{s \log(Kp)}{n}} + 1 \\
&\lesssim 1,
\end{aligned}$$

for all $i \in [n]$ and $k \in [K-1]$, when $n \gtrsim K^5 s(\log(Kp))(\log(np))^2$. Therefore, conditioned on $\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3$, we have $p_k(\boldsymbol{x}_i; \widehat{\beta}) \gtrsim 1/K$. Further by Lemma 2, conditioned on $\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3 \cap \mathcal{A}_4 \cap \mathcal{A}_5$, we have

$$\begin{aligned}
\text{LHS of (D8)} &= \frac{1}{2}(\widehat{\gamma}_j - \gamma_j^*)^\top \widehat{\Sigma}_{-j,-j}(\widehat{\gamma}_j - \gamma_j^*) \\
&\geq \frac{1}{n}\sum_{i=1}^{n} p_K(\boldsymbol{x}_i; \widehat{\beta}) \sum_{k=1}^{K-1} p_k(\boldsymbol{x}_i; \widehat{\beta})[(\widehat{\gamma}_j^{(k)} - \gamma_j^{(k)*})^\top \boldsymbol{x}_i]^2 \\
&\geq \frac{c}{nK^2} \sum_{i=1}^{n} \sum_{k=1}^{K-1} [(\widehat{\gamma}_j^{(k)} - \gamma_j^{(k)*})^\top \boldsymbol{x}_i]^2.
\end{aligned}$$

Denote $S_j = \text{supp}(\gamma_j^*)$. Similar to the arguments in the proof of Theorem 1, when $\lambda_j \geq 2\|\widehat{\Sigma}_{\beta^*,-j,j} - \widehat{\Sigma}_{\beta^*,-j,-j}\gamma_j^*\|_\infty$, it follows from (D8) and (D9) that

$$\frac{c}{nK^2} \sum_{i=1}^{n} \sum_{k=1}^{K-1} [(\widehat{\gamma}_j^{(k)} - \gamma_j^{(k)*})^\top \boldsymbol{x}_i]^2 \leq CK^7 \frac{s\log(Kp)}{n}\log(np) + \lambda_j \left(\frac{3}{2}\|(\widehat{\gamma}_j - \gamma_j^*)_{S_j}\|_1 - \frac{1}{2}\|(\widehat{\gamma}_j - \gamma_j^*)_{S_j^c}\|_1 \right).$$

Plugging $\lambda_j \asymp \sqrt{\frac{\log p}{n}}$ into above, we have

$$\|(\widehat{\gamma}_j - \gamma_j^*)_{S_j^c}\|_1 \leq 3\|(\widehat{\gamma}_j - \gamma_j^*)_{S_j}\|_1 + CK^7 s\sqrt{\frac{\log(Kp)}{n}}\log(np).$$

Case 1: $\|(\widehat{\gamma}_j - \gamma_j^*)_{S_j}\|_1 > CK^7 s\sqrt{\frac{\log(Kp)}{n}}\log(np)$.

We have $\|(\widehat{\gamma}_j - \gamma_j^*)_{S_j^c}\|_1 \leq 4\|(\widehat{\gamma}_j - \gamma_j^*)_{S_j}\|_1$ and

$$\frac{c'}{K^2}\|\widehat{\gamma}_j - \gamma_j^*\|_2^2 \leq \frac{c}{nK^2} \sum_{i=1}^{n} \sum_{k=1}^{K-1} [(\widehat{\gamma}_j^{(k)} - \gamma_j^{(k)*})^\top \boldsymbol{x}_i]^2 \leq C\sqrt{\frac{\log(Kp)}{n}}\|(\widehat{\gamma}_j - \gamma_j^*)_{S_j}\|_1 \leq C\sqrt{\frac{Ks_0 \log(Kp)}{n}}\|\widehat{\gamma}_j - \gamma_j^*\|_2,$$

where the first inequality comes from Theorem 1.6 of[86]. It implies that

$$\|\widehat{\gamma}_j - \gamma_j^*\|_2 \le C' K^{5/2} \sqrt{\frac{s_0 \log(Kp)}{n}},$$

which entails that

$$\|\widehat{\gamma}_j - \gamma_j^*\|_1 \le C' K^3 s_0 \sqrt{\frac{\log(Kp)}{n}}.$$

Case 2: $\|(\widehat{\gamma}_j - \gamma_j^*)_{S_j^c}\|_1 \le CK^7 s \sqrt{\frac{\log(Kp)}{n}} \log(np)$.

We must have $\|(\widehat{\gamma}_j - \gamma_j^*)_{S_j}\|_1 \le 4CK^7 s \sqrt{\frac{\log(Kp)}{n}} \log(np) \le K^{7/2} \sqrt{s}$, when $n \gtrsim K^7 s(\log(Kp))(\log(np))^2$. By Lemma 3.1 of[84], $\Pi_1(K^7 s) \subseteq 2\overline{\mathrm{conv}}(\Pi_0(K^7 s))$. Then by Lemma 2.3 in[85], $\Pi_0(K^7 s) \subseteq 2\overline{\mathrm{conv}}(\Lambda)$, where Λ is a 1/2-cover of $\Pi_0(K^7 s)$ with $|\Lambda| \le C\binom{p}{K^7 s} 5^{K^7 s} \le \exp\{C'n\}$ when $n \ge C'K^7 s \log p$. By Theorem 2.4 in[85], with probability at least $1 - C'' \exp\{-C'''n\}$, we have $\frac{c'}{K^2} \|\widehat{\gamma}_j - \gamma_j^*\|_2^2 \le \frac{c}{nK^2} \sum_{i=1}^n \sum_{k=1}^{K-1} [(\widehat{\gamma}_j^{(k)} - \gamma_j^{(k)*})^\top x_i]^2$. Therefore,

$$\frac{c'}{K^2} \|\widehat{\gamma}_j - \gamma_j^*\|_2^2 \le CK^7 \frac{s \log(Kp)}{n} \log(np) + C\sqrt{\frac{Ks_0 \log(Kp)}{n}} \|\widehat{\gamma}_j - \gamma_j^*\|_2,$$

which implies that

$$\|\widehat{\gamma}_j - \gamma_j^*\|_2 \lesssim K^{9/2} \sqrt{\frac{s(\log(Kp))(\log(np))}{n}} + K^{5/2} \sqrt{\frac{s_0 \log(Kp)}{n}}, \quad \|\widehat{\gamma}_j - \gamma_j^*\|_1 \lesssim K^7 s \sqrt{\frac{\log(Kp)}{n}} \log(np).$$

Combining two cases, we have

$$\|\widehat{\gamma}_j - \gamma_j^*\|_2 \lesssim K^{9/2} \sqrt{\frac{s(\log(Kp))(\log(np))}{n}} + K^{5/2} \sqrt{\frac{s_0 \log(Kp)}{n}}, \quad \|\widehat{\gamma}_j - \gamma_j^*\|_1 \lesssim K^7 s \sqrt{\frac{\log(Kp)}{n}} \log(np) + K^3 s_0 \sqrt{\frac{\log(Kp)}{n}},$$

with probability $1 - C' \exp\{-C''n\} - C'(np)^{-1} - C'(Kp)^{-1}$.

Next we will bound $|\widehat{\tau}_j^2 - (\tau_j^*)^2|$ and $|\widehat{\tau}_j^{-2} - (\tau_j^*)^{-2}|$. Note that $\widehat{\tau}_j^2 = \widehat{\Sigma}_{j,j} - \widehat{\Sigma}_{-j,j}^\top \widehat{\gamma}_j$, and $(\tau_j^*)^2 = \Sigma_{j,j} - \Sigma_{-j,j}^\top \gamma_j^*$. Then

$$\begin{aligned}
|\widehat{\tau}_j^2 - (\tau_j^*)^2| &\le |\widehat{\Sigma}_{j,j} - \Sigma_{j,j}| + |\widehat{\Sigma}_{-j,j}^\top \widehat{\gamma}_j - \Sigma_{-j,j}^\top \gamma_j^*| \\
&\le |\widehat{\Sigma}_{j,j} - \Sigma_{j,j}| + |(\widehat{\Sigma}_{-j,j} - \Sigma_{-j,j})^\top \widehat{\gamma}_j| + |\Sigma_{-j,j}^\top (\widehat{\gamma}_j - \gamma_j^*)| \\
&\le |\widehat{\Sigma}_{j,j} - \Sigma_{j,j}| + |(\widehat{\Sigma}_{-j,j} - \Sigma_{-j,j})^\top \gamma_j^*| + \|\widehat{\Sigma}_{-j,j} - \Sigma_{-j,j}\|_\infty \|\widehat{\gamma}_j - \gamma_j^*\|_1 + \|\Sigma_{-j,j}\|_2 \cdot \|\widehat{\gamma}_j - \gamma_j^*\|_2,
\end{aligned}$$

where

$$|\widehat{\Sigma}_{j,j} - \Sigma_{j,j}| \lesssim \|\widehat{\beta} - \beta^*\|_2 \lesssim K^{5/2} \sqrt{\frac{s \log(Kp)}{n}},$$

$$|(\widehat{\Sigma}_{-j,j} - \Sigma_{-j,j})^\top \gamma_j^*| \lesssim \|\widehat{\beta} - \beta^*\|_2 \|\gamma_j^*\|_2 \lesssim K^{5/2} \sqrt{\frac{s \log(Kp)}{n}},$$

$$\|\widehat{\Sigma}_{-j,j} - \Sigma_{-j,j}\|_\infty \lesssim \|\widehat{\beta} - \beta^*\|_2 \lesssim K^{5/2} \sqrt{\frac{s \log(Kp)}{n}},$$

$$\|\widehat{\gamma}_j - \gamma_j^*\|_1 \lesssim K^7 s \sqrt{\frac{\log(Kp)}{n}} \log(np) + K^3 s_0 \sqrt{\frac{\log(Kp)}{n}},$$

$$\|\Sigma_{-j,j}\|_2 \lesssim 1,$$

with probability $1 - C'(Kp)^{-1}$, which can be similarly derived as before. This implies

$$|\widehat{\tau}_j^2 - (\tau_j^*)^2| \lesssim K^{9/2} \sqrt{\frac{s(\log(Kp))(\log(np))}{n}} + K^{5/2} \sqrt{\frac{s \log(Kp)}{n}} \cdot \left(K^7 s \sqrt{\frac{\log(Kp)}{n}} \log(np) + K^3 s_0 \sqrt{\frac{\log(Kp)}{n}} \right) + K^{5/2} \sqrt{\frac{s_0 \log(Kp)}{n}},$$

$$\lesssim K^{9/2} \sqrt{\frac{s(\log(Kp))(\log(np))}{n}} + K^{5/2} \sqrt{\frac{s_0 \log(Kp)}{n}},$$

$$|\widehat{\tau}_j^{-2} - (\tau_j^*)^{-2}| = \frac{|\widehat{\tau}_j^2 - (\tau_j^*)^2|}{\widehat{\tau}_j^2 (\tau_j^*)^2} \lesssim |\widehat{\tau}_j^2 - (\tau_j^*)^2| \lesssim K^{9/2} \sqrt{\frac{s(\log(Kp))(\log(np))}{n}} + K^{5/2} \sqrt{\frac{s_0 \log(Kp)}{n}},$$

with probability $1 - C' \exp\{-C'' n\} - C'(np)^{-1} - C'(Kp)^{-1}$, where the second inequality of $|\widehat{\tau}_j^2 - (\tau_j^*)^2|$ is due to the condition $n \gtrsim K^{10} s^2 \log(Kp) + K^6 s_0^2 \log(Kp)$. Since $\widehat{\Theta}_j$ consists of $\widehat{\tau}_j^{-2}$ and $\widehat{\gamma}_j$, we have

$$\|\widehat{\Theta}_j - \Theta_j^*\|_2 \lesssim \|\widehat{\gamma}_j - \gamma_j^*\|_2 + |\widehat{\tau}_j^{-2} - (\tau_j^*)^{-2}| \lesssim K^{9/2} \sqrt{\frac{s(\log(Kp))(\log(np))}{n}} + K^{5/2} \sqrt{\frac{s_0 \log(Kp)}{n}},$$

$$\|\widehat{\Theta}_j - \Theta_j^*\|_1 \lesssim \|\widehat{\gamma}_j - \gamma_j^*\|_1 + |\widehat{\tau}_j^{-2} - (\tau_j^*)^{-2}| \lesssim K^7 s \sqrt{\frac{\log(Kp)}{n}} \log(np) + K^3 s_0 \sqrt{\frac{\log(Kp)}{n}},$$

with probability $1 - C' \exp\{-C'' n\} - C'(np)^{-1} - C'(Kp)^{-1}$. $\qquad\square$

D.2 Proofs of theorems

Proof of Theorem 1. Denote $\widehat{\Delta} = \widehat{\beta} - \beta^*$. The negative log-likelihood is

$$\mathcal{L}_n(\beta) = -\frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K y_i^{(k)} \cdot x_i^\top \beta^{(k)} + \frac{1}{n} \sum_{i=1}^n \log \left(1 + \sum_{k=1}^{K-1} e^{x_i^\top \beta^{(k)}} \right).$$

The gradient is

$$\nabla \mathcal{L}_n(\beta) = - \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n \left(y_i^{(1)} - \frac{e^{x_i^\top \beta^{(1)}}}{1 + \sum_{k=1}^{K-1} e^{x_i^\top \beta^{(k)}}} \right) x_i \\ \vdots \\ \frac{1}{n} \sum_{i=1}^n \left(y_i^{(k)} - \frac{1}{1 + \sum_{k=1}^{K-1} e^{x_i^\top \beta^{(k)}}} \right) x_i \end{pmatrix}_{[(K-1)p] \times 1}.$$

By the convexity of $\mathcal{L}_n$,

$$0 \le \mathcal{L}_n(\beta^* + \widehat{\Delta}) - \mathcal{L}_n(\beta^*) - \nabla \mathcal{L}_n(\beta^*)^\top \widehat{\Delta}$$
$$\le -\nabla \mathcal{L}_n(\beta^*)^\top \widehat{\Delta} + \lambda(\|\beta^*\|_1 - \|\widehat{\beta}\|_1)$$
$$\le \frac{3}{2} \lambda \|\widehat{\Delta}_S\|_1 - \frac{1}{2} \lambda \|\widehat{\Delta}_{S^c}\|_1,$$

implying that

$$\|\widehat{\Delta}_{S^c}\|_1 \le 3 \|\widehat{\Delta}_S\|_1. \tag{D10}$$

Define $F_n(\Delta) = \mathcal{L}_n(\beta^* + \Delta) - \mathcal{L}_n(\beta^*) + \lambda \|\beta^* + \Delta\|_1 - \lambda \|\beta^*\|_1$, which is a convex function of Δ. By the optimality of $\widehat{\beta}$, $F_n(\widehat{\Delta}) \le 0$. If $\|\widehat{\Delta}\|_2 > 1$, then $\exists t \in (0, 1)$ such that $\widetilde{\Delta} = t \widehat{\Delta}$. By convexity $F_n(\widetilde{\Delta}) \le t F_n(\widehat{\Delta}) + (1 - t) F_n(\mathbf{0}) \le 0$. And by (D10), $\|\widetilde{\Delta}_S\|_1 \le 3 \|\widetilde{\Delta}_{S^c}\|_1$. Then

$$F_n(\widetilde{\Delta}) \ge \mathcal{E}_n(\widetilde{\Delta}) + \nabla \mathcal{L}_n(\beta^*)^\top \widetilde{\Delta} + \lambda \|\beta^* + \widetilde{\Delta}\|_1 - \lambda \|\beta^*\|_1$$
$$\ge \mathcal{E}_n(\widetilde{\Delta}) - \|\nabla \mathcal{L}_n(\beta^*)\|_\infty \|\widetilde{\Delta}\|_1 + \lambda(\|\beta_S^* + \widetilde{\Delta}_S\|_1 - \|\beta_S^*\|_1 + \|\widetilde{\Delta}_{S^c}\|_1)$$
$$\ge \mathcal{E}_n(\widetilde{\Delta}) - \frac{3}{2} \lambda \|\widetilde{\Delta}_S\|_1 + \frac{1}{2} \lambda \|\widetilde{\Delta}_{S^c}\|_1,$$

because $\|\nabla \mathcal{L}_n(\beta^*)\|_\infty \le C \sqrt{\frac{\log(Kp)}{n}}$ with probability at least $1 - C'(Kp)^{-1}$. Since $\|\widetilde{\Delta}\|_2 = 1$, by Lemma 1,

$$F_n(\widetilde{\Delta}) \ge \frac{1}{K^2} \left(c_1 - c_2 \sqrt{\frac{\log(Kp)}{n}} \|\widetilde{\Delta}\|_1 \right) - \frac{3}{2} \lambda \|\widetilde{\Delta}_S\|_1 + \frac{1}{2} \lambda \|\widetilde{\Delta}_{S^c}\|_1$$
$$\ge \frac{c_1}{K^2} - \frac{4 c_2}{K^2} \cdot \sqrt{\frac{\log(Kp)}{n}} \cdot \sqrt{Ks} - \frac{3}{2} \lambda \sqrt{Ks}$$

$$\geq \frac{c_1}{K^2} - \frac{3}{2K^2}C\sqrt{\frac{Ks\log(Kp)}{n}} - c_2\sqrt{\frac{Ks\log(Kp)}{n}}$$
$$> 0,$$

when $n \geq C'K^5 s\log p$ with a large constant $C' > 0$. But this is contradicted by the fact that $F_n(\widetilde{\Delta}) \leq 0$. Therefore, we must have $\|\widehat{\Delta}\|_2 \leq 1$. Then we can apply a similar argument above with Lemma 1 to show that

$$\frac{1}{K^2}\left(c_1\|\widehat{\Delta}\|_2^2 - c_2\sqrt{\frac{\log(Kp)}{n}}\|\widehat{\Delta}\|_1\|\widehat{\Delta}\|_2\right) \leq \mathcal{E}_n(\widehat{\Delta})$$
$$\leq \frac{3}{2}\lambda\|\widehat{\Delta}_S\|_1 - \frac{1}{2}\lambda\|\widehat{\Delta}_{S^c}\|_1$$
$$\leq \frac{3}{2}C\sqrt{\frac{Ks\log(Kp)}{n}}\|\widehat{\Delta}\|_2,$$

with probability at least $1 - C'(Kp)^{-1}$. Recalling (D10), we have

$$\frac{1}{2K^2}c_1\|\widehat{\Delta}\|_2^2 \leq \frac{1}{K^2}\left(c_1 - c_2\sqrt{\frac{\log(Kp)}{n}}4\sqrt{Ks}\right)\|\Delta\|_2^2 \leq \frac{3}{2}C\sqrt{\frac{Ks\log(Kp)}{n}}\|\widehat{\Delta}\|_2,$$

when $n \geq 64\frac{c_1^2}{c_2^2}Ks\log(Kp)$. Therefore,

$$\|\widehat{\Delta}\|_2 \leq \frac{3C}{c_1}K^{5/2}\sqrt{\frac{s\log(Kp)}{n}}, \quad \|\widehat{\Delta}\|_1 \leq 4\sqrt{Ks}\|\widehat{\Delta}\|_2 \leq \frac{12C}{c_1}K^3 s\sqrt{\frac{\log(Kp)}{n}},$$

with probability at least $1 - C'(Kp)^{-1}$. □

Proof of Theorem 2. Consider $\mathcal{H} = \delta \cdot \{z \in \{-1, 0, 1\}^p : \|z\|_0 = s\}$, then by Lemma 5 in [4], there exists $\widetilde{\mathcal{H}} \subseteq \mathcal{H}$ such that for any $z', z \in \widetilde{\mathcal{H}}$, we have $\|z - z'\|_0 \geq s/2$ and $\log |\widetilde{\mathcal{H}}| \geq C's\log(ep/s)$. Take $\beta, \widetilde{\beta} \in \widetilde{\mathcal{H}}^{\otimes(K-1)}$, by Lemma 3,

$$\text{KL}(\mathbb{P}_{x,y;\beta}^{\otimes n}\|\mathbb{P}_{x,y;\widetilde{\beta}}^{\otimes n}) \leq Cn\sigma^2\|\widetilde{\beta} - \beta\|_2^2 \cdot \left(\frac{1}{K}e^{2T} + \exp\left\{-\frac{C'T^2}{\min_k\|\beta^{(k)}\|_2^2}\right\}\right)$$
$$\leq CnK\delta^2 s\left(\frac{1}{K}e^{2T} + \exp\left\{-\frac{C'T^2}{\delta^2 s}\right\}\right).$$

Then by Fano's lemma [87],

$$\inf_{\widehat{\beta}}\sup_{\beta}\mathbb{P}\left(\|\widehat{\beta} - \beta\|_2^2 \geq \frac{Ks}{4}\delta^2\right) \geq 1 - \frac{\log 2 + CnK\delta^2 s\left(\frac{1}{K}e^{2T} + \exp\left\{-\frac{C'T^2}{\delta^2 s}\right\}\right)}{K\log|\widetilde{\mathcal{H}}|}$$
$$\geq 1 - \frac{\log 2 + CnK\frac{K\log(ep/s)}{n}s\left(\frac{1}{K}C'' + \exp\left\{-\frac{C'''n}{Ks\log(ep/s)}\right\}\right)}{C'Ks\log(ep/s)}$$
$$\geq \frac{1}{4},$$

if T is a small constant and $\delta = c\sqrt{\frac{K\log(ep/s)}{n}}$ with a small constant $c > 0$, when $n \geq CKs(\log(ep/s))(\log K)$, $p \gg s$. The lower bound for $\|\widehat{\beta} - \beta\|_1$ can be similarly shown. □

Proof of Theorem 3. It is easy to see that for any classifier $\mathcal{C}$,

$$R(\mathcal{C}) = 1 - \mathbb{E}_X[\mathbb{P}_{Y|X}(\mathcal{C}(X) = Y)].$$

This implies that

$$R(\mathcal{C}_{\beta^*}) = 1 - \mathbb{E}_x\left[\max_k \mathbb{P}(Y = k|X = x)\right].$$

Since

$$\mathbb{P}_{Y|X}(\mathcal{C}_{\widehat{\boldsymbol{\beta}}}(X) = Y) \geq \widehat{\mathbb{P}}_{Y|X}(\mathcal{C}_{\widehat{\boldsymbol{\beta}}}(X) = Y) - \max_k \left| \widehat{\mathbb{P}}(Y = k|X = \boldsymbol{x}) - \mathbb{P}(Y = k|X = \boldsymbol{x}) \right|$$

$$= \max_k \widehat{\mathbb{P}}(Y = k|X = \boldsymbol{x}) - \max_k \left| \widehat{\mathbb{P}}(Y = k|X = \boldsymbol{x}) - \mathbb{P}(Y = k|X = \boldsymbol{x}) \right|,$$

we have

$$R(\mathcal{C}_{\widehat{\boldsymbol{\beta}}}) \leq 1 - \mathbb{E}_{\boldsymbol{x}}\left[\max_k \widehat{\mathbb{P}}(Y = k|X = \boldsymbol{x}) \right] + \mathbb{E}_{\boldsymbol{x}}\left[\max_k \left| \widehat{\mathbb{P}}(Y = k|X = \boldsymbol{x}) - \mathbb{P}(Y = k|X = \boldsymbol{x}) \right| \right].$$

Hence

$$R(\mathcal{C}_{\widehat{\boldsymbol{\beta}}}) - R(\mathcal{C}_{\boldsymbol{\beta}^*})$$

$$\leq 2\mathbb{E}_{\boldsymbol{x}}\left[\max_k \left| \widehat{\mathbb{P}}(Y = k|X = \boldsymbol{x}) - \mathbb{P}(Y = k|X = \boldsymbol{x}) \right| \right]$$

$$= 2\mathbb{E}_{\boldsymbol{x}}\left[\max_k \left| p_k(\boldsymbol{x}; \widehat{\boldsymbol{\beta}}) - p_k(\boldsymbol{x}; \boldsymbol{\beta}^*) \right| \right]$$

$$= 2\mathbb{E}_{\boldsymbol{x}}\left[\max_k \left| \nabla g_k(t\boldsymbol{x}^\top\widehat{\boldsymbol{\beta}}^{(1)} + (1-t)\boldsymbol{x}^\top\boldsymbol{\beta}^{(1)*}, \ldots, t\boldsymbol{x}^\top\boldsymbol{\beta}^{(K-1)*} + (1-t)\boldsymbol{x}^\top\boldsymbol{\beta}^{(K-1)*}) \begin{pmatrix} \boldsymbol{x}^\top(\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}_1^*) \\ \vdots \\ \boldsymbol{x}^\top(\widehat{\boldsymbol{\beta}}_{K-1} - \boldsymbol{\beta}_{K-1}^*) \end{pmatrix} \right| \right]$$

$$\leq 2\mathbb{E}_{\boldsymbol{x}}\left[\max_k \| \nabla g_k(t\boldsymbol{x}^\top\widehat{\boldsymbol{\beta}}^{(1)} + (1-t)\boldsymbol{x}^\top\boldsymbol{\beta}^{(1)*}, \ldots, t\boldsymbol{x}^\top\widehat{\boldsymbol{\beta}}^{(K-1)} + (1-t)\boldsymbol{x}^\top\boldsymbol{\beta}^{(K-1)*}) \|_2 \right.$$

$$\left. \cdot \sqrt{\sum_{k=1}^{K-1} (\boldsymbol{x}^\top(\widehat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^{(k)*}))^2} \right]$$

$$\leq C\mathbb{E}_{\boldsymbol{x}}\sqrt{\sum_{k=1}^{K-1} (\boldsymbol{x}^\top(\widehat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^{(k)*}))^2}$$

$$\leq C\sqrt{\sum_{k=1}^{K-1} \mathbb{E}_{\boldsymbol{x}}(\boldsymbol{x}^\top(\widehat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^{(k)*}))^2}$$

$$\leq C\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$$

$$\leq CK^{5/2}\sqrt{\frac{s\log(Kp)}{n}},$$

with probability at least $1 - C'(Kp)^{-1}$, which completes the proof. $\qquad\square$

Proof of Theorem 4. Note that

$$\widehat{\boldsymbol{b}} - \boldsymbol{\beta}^* = (\boldsymbol{I} - \widehat{\boldsymbol{\Theta}}\boldsymbol{B}_n)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \frac{1}{n}(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})\begin{pmatrix} \boldsymbol{X}^\top\boldsymbol{\epsilon}_1 \\ \boldsymbol{X}^\top\boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{X}^\top\boldsymbol{\epsilon}_{K-1} \end{pmatrix} + \frac{1}{n}\boldsymbol{\Theta}\begin{pmatrix} \boldsymbol{X}^\top\boldsymbol{\epsilon}_1 \\ \boldsymbol{X}^\top\boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{X}^\top\boldsymbol{\epsilon}_{K-1} \end{pmatrix},$$

where

$$\boldsymbol{B}_n = \frac{1}{n}\sum_{i=1}^n \int_0^1 B(\boldsymbol{x}_i; \boldsymbol{\beta}^* + t(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*))dt,$$

$$B(\boldsymbol{x}; \boldsymbol{\beta}) := \begin{pmatrix} p_1(\boldsymbol{x}; \boldsymbol{\beta})[1 - p_1(\boldsymbol{x}; \boldsymbol{\beta})]\boldsymbol{x}\boldsymbol{x}^\top & -p_1(\boldsymbol{x}; \boldsymbol{\beta})p_2(\boldsymbol{x}; \boldsymbol{\beta})\boldsymbol{x}\boldsymbol{x}^\top & \ldots & -p_1(\boldsymbol{x}; \boldsymbol{\beta})p_{K-1}(\boldsymbol{x}; \boldsymbol{\beta})\boldsymbol{x}\boldsymbol{x}^\top \\ -p_1(\boldsymbol{x}; \boldsymbol{\beta})p_2(\boldsymbol{x}; \boldsymbol{\beta})\boldsymbol{x}\boldsymbol{x}^\top & p_2(\boldsymbol{x}; \boldsymbol{\beta})[1 - p_2(\boldsymbol{x}; \boldsymbol{\beta})]\boldsymbol{x}\boldsymbol{x}^\top & \ldots & -p_2(\boldsymbol{x}; \boldsymbol{\beta})p_{K-1}(\boldsymbol{x}; \boldsymbol{\beta})\boldsymbol{x}\boldsymbol{x}^\top \\ \vdots & \vdots & \vdots & \vdots \\ -p_1(\boldsymbol{x}; \boldsymbol{\beta})p_{K-1}(\boldsymbol{x}; \boldsymbol{\beta})\boldsymbol{x}\boldsymbol{x}^\top & -p_2(\boldsymbol{x}; \boldsymbol{\beta})p_{K-1}(\boldsymbol{x}; \boldsymbol{\beta})\boldsymbol{x}\boldsymbol{x}^\top & \ldots & p_{K-1}(\boldsymbol{x}; \boldsymbol{\beta})[1 - p_{K-1}(\boldsymbol{x}; \boldsymbol{\beta})]\boldsymbol{x}\boldsymbol{x}^\top \end{pmatrix}.$$

Then

$$\widehat{b}_j - \beta_j^* = \underbrace{(e_j - \widehat{\Theta}_j^\top B_n)(\widehat{\beta} - \beta^*)}_{[1]} + \underbrace{\frac{1}{n}(\widehat{\Theta}_j - \Theta_j)^\top \begin{pmatrix} X^\top \epsilon_1 \\ X^\top \epsilon_2 \\ \vdots \\ X^\top \epsilon_{K-1} \end{pmatrix}}_{[2]} + \underbrace{\frac{1}{n}\Theta_j^\top \begin{pmatrix} X^\top \epsilon_1 \\ X^\top \epsilon_2 \\ \vdots \\ X^\top \epsilon_{K-1} \end{pmatrix}}_{[3]}.$$

And

$$[1] \le \underbrace{\left| (\widehat{\Theta}_j - \Theta_j)^\top B_n(\widehat{\beta} - \beta^*) \right|}_{[1].1} + \underbrace{\left| \Theta_j^\top (\Sigma - B_n)(\widehat{\beta} - \beta^*) \right|}_{[1].2},$$

where

$$\begin{aligned}
[1].1 &\le \frac{1}{2}(\widehat{\Theta}_j - \Theta_j)^\top B_n(\widehat{\Theta}_j - \Theta_j) + \frac{1}{2}(\widehat{\beta} - \beta^*)^\top B_n(\widehat{\beta} - \beta^*) \\
&\le \frac{1}{2}(\widehat{\Theta}_j - \Theta_j)^\top \operatorname{diag}(XX^\top, \ldots, XX^\top)(\widehat{\Theta}_j - \Theta_j) \\
&\quad + \frac{1}{2}(\widehat{\beta} - \beta^*)^\top \operatorname{diag}(XX^\top, \ldots, XX^\top)(\widehat{\beta} - \beta^*) \\
&\lesssim_p C\|\widehat{\Theta}_j - \Theta_j\|_2^2 + C\|\widehat{\beta} - \beta^*\|_2^2 \\
&\lesssim_p K^9 \frac{s(\log(Kp))(\log(np))}{n} + K^5 \frac{s_0 \log(Kp)}{n},
\end{aligned}$$

where the second last inequality comes from a similar argument in the proof of Lemma 4 to analyze event $\mathcal{A}_1$, and the last inequality comes of Theorem 1 and Lemma 4. In addition, by a similar argument to prove (D9), we have

$$\begin{aligned}
[1].2 &\lesssim \frac{1}{n}\sum_{k=1}^{K-1}\sum_{i=1}^{n} \left[(\widehat{\gamma}_j^{(k)} - \gamma_j^{(k)*})^\top x_i \right]^2 + \frac{1}{n}\sum_{k=1}^{K-1}\sum_{i=1}^{n} [x_i^\top(\widehat{\beta}^{(k)} - \beta^{(k)*})]^2 \cdot |x_i^\top \Theta_j^{(k)*}|^2 \\
&\lesssim \|\widehat{\gamma}_j - \gamma_j^*\|_2^2 + \|\widehat{\beta} - \beta^*\|_2^2 \log(nK) \\
&\lesssim_\mathbb{P} K^9 \frac{s(\log(Kp))(\log(np))}{n} + K^5 \frac{s_0 \log(Kp)}{n}.
\end{aligned}$$

Therefore,

$$[1] \lesssim_\mathbb{P} K^9 \frac{s(\log(Kp))(\log(np))}{n} + K^5 \frac{s_0 \log(Kp)}{n} \ll n^{-1/2},$$

when $n \gg K^{18} s^2 (\log(Kp))^2 (\log(np))^2 + K^{10} s_0^2 (\log(Kp))^2$.

Similarly,

$$[2] \le \frac{1}{n}\|\widehat{\Theta}_j - \Theta_j\|_1 \cdot \left\| \begin{pmatrix} X^\top \epsilon_1 \\ X^\top \epsilon_2 \\ \vdots \\ X^\top \epsilon_{K-1} \end{pmatrix} \right\|_\infty \lesssim_p \left[K^7 s \sqrt{\frac{\log(Kp)}{n}} \log(np) + K^3 s_0 \sqrt{\frac{\log(Kp)}{n}} \right] \cdot \sqrt{\frac{\log(Kp)}{n}} \ll n^{-1/2}.$$

And

$$[3] = \frac{1}{n}\sum_{i=1}^{n} x_i^\top \left(\sum_{k=1}^{K-1} \epsilon_{k,i}\theta_k \right) = \Theta_j^\top \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} \epsilon_{1,i}x_i \\ \vdots \\ \epsilon_{K-1,i}x_i \end{pmatrix}.$$

By the Lindeberg-Feller central limit theorem, we have the following asymptotic normality:

$$\frac{\sqrt{n}[3]}{\sqrt{\Theta_j^\top D \Theta_j}} \xrightarrow{d} N(0, 1),$$

where $\boldsymbol{D} = \mathrm{Cov}((\epsilon_1 X^\top, \ldots, \epsilon_{K-1} X^\top)^\top)$. Note that

$$\boldsymbol{D} = \mathbb{E} \begin{pmatrix} \epsilon_1^2 XX^\top & \ldots & \epsilon_1\epsilon_{K-1}XX^\top \\ \vdots & \vdots & \vdots \\ \epsilon_1\epsilon_{K-1}XX^\top & \ldots & \epsilon_{K-1}^2 XX^\top \end{pmatrix},$$

where

$$\mathbb{E}(\epsilon_k^2 XX^\top) = \mathbb{E}_X \left\{ \mathbb{E}[(y_k - p_k(X;\boldsymbol{\beta}^*))^2 | X] XX^\top \right\} = \mathbb{E}_X \left[XX^\top p_k(X;\boldsymbol{\beta}^*)(1 - p_k(X;\boldsymbol{\beta}^*)) \right],$$

$$\mathbb{E}(\epsilon_{k_1}\epsilon_{k_2} XX^\top) = \mathbb{E}_X \left\{ \mathbb{E}[(y_{k_1} - p_{k_1}(X;\boldsymbol{\beta}^*))(y_{k_2} - p_{k_2}(X;\boldsymbol{\beta}^*)) | X] XX^\top \right\}$$
$$= -\mathbb{E}_{\boldsymbol{x}} \left[XX^\top p_{k_1}(X;\boldsymbol{\beta}^*)p_{k_2}(X;\boldsymbol{\beta}^*) \right],$$

for any $k, k_1 \neq k_2 \in \{1, \ldots, K-1\}$. Therefore $\boldsymbol{D} = \mathbb{E}[\nabla^2 \mathcal{L}_n(\boldsymbol{\beta}^*)] = \boldsymbol{\Sigma}$. It implies that

$$\left| \widehat{\boldsymbol{\Theta}}_j^\top \widehat{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\Theta}}_j - \boldsymbol{\Theta}_j^\top \boldsymbol{\Sigma} \boldsymbol{\Theta}_j \right|$$

$$\leq \left| (\widehat{\boldsymbol{\Theta}}_j - \boldsymbol{\Theta}_j)^\top \widehat{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\Theta}}_j \right| + \left| \boldsymbol{\Theta}_j^\top (\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})^\top \widehat{\boldsymbol{\Theta}}_j \right| + \left| \boldsymbol{\Theta}_j^\top \boldsymbol{\Sigma} (\widehat{\boldsymbol{\Theta}}_j - \boldsymbol{\Theta}_j) \right|$$

$$\leq \left| (\widehat{\boldsymbol{\Theta}}_j - \boldsymbol{\Theta}_j)^\top (\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})(\widehat{\boldsymbol{\Theta}}_j - \boldsymbol{\Theta}_j) \right| + \left| (\widehat{\boldsymbol{\Theta}}_j - \boldsymbol{\Theta}_j)^\top (\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\boldsymbol{\Theta}_j \right| + \left| (\widehat{\boldsymbol{\Theta}}_j - \boldsymbol{\Theta}_j)^\top \boldsymbol{\Sigma}(\widehat{\boldsymbol{\Theta}}_j - \boldsymbol{\Theta}_j) \right| + \left| (\widehat{\boldsymbol{\Theta}}_j - \boldsymbol{\Theta}_j)^\top \boldsymbol{\Sigma} \boldsymbol{\Theta}_j \right|$$

$$+ \left| \boldsymbol{\Theta}_j^\top (\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})^\top (\widehat{\boldsymbol{\Theta}}_j - \boldsymbol{\Theta}_j) \right| + \left| \boldsymbol{\Theta}_j^\top (\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})^\top \boldsymbol{\Theta}_j \right| + \left| \boldsymbol{\Theta}_j^\top \boldsymbol{\Sigma}(\widehat{\boldsymbol{\Theta}}_j - \boldsymbol{\Theta}_j) \right|$$

$$\lesssim \|\widehat{\boldsymbol{\Theta}}_j - \boldsymbol{\Theta}_j\|_1^2 \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max} + \|\widehat{\boldsymbol{\Theta}}_j - \boldsymbol{\Theta}_j\|_1 \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max}\|\boldsymbol{\Theta}_j\|_1 + \|\widehat{\boldsymbol{\Theta}}_j - \boldsymbol{\Theta}_j\|_2^2 + \|\widehat{\boldsymbol{\Theta}}_j - \boldsymbol{\Theta}_j\|_2 \|\boldsymbol{\Sigma}\boldsymbol{\Theta}_j\|_2$$

$$\lesssim_{\mathbb{P}} \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max} \sqrt{Ks_0} + \|\widehat{\boldsymbol{\Theta}}_j - \boldsymbol{\Theta}_j\|_2$$

$$\lesssim_{\mathbb{P}} K^3 \sqrt{\frac{ss_0 \log(Kp)}{n}} + K^{9/2} \sqrt{\frac{s(\log(Kp))(\log(np))}{n}} + K^{5/2} \sqrt{\frac{s_0 \log(Kp)}{n}}$$

$$= \mathcal{O}(1),$$

because $\|\widehat{\boldsymbol{\Theta}}_j - \boldsymbol{\Theta}_j\|_1 = \mathcal{O}_{\mathbb{P}}(1)$ by Lemma 4, when $n \gg K^{18}s^2(\log(Kp))^2(\log(np))^2 + K^{10}s_0^2(\log(Kp))^2$. Hence, the desired conclusion in Theorem 4 holds due to Slutsky's Theorem. $\square$

Proof of Theorem 5. The results in Theorems 1, 3, and 4 can be derived in a similar way in the model misspecification scenario, so we only point out the difference here.

When deriving the estimation error bounds in Theorem 1, we need to show $\|\nabla \mathcal{L}_n(\boldsymbol{\beta}^*)\|_\infty \lesssim \sqrt{\frac{\log(Kp)}{n}}$ with probability at least $1 - C(Kp)^{-1}$. Note that

$$\nabla \mathcal{L}_n(\boldsymbol{\beta}^*) = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n [y_i^{(1)} - p_1(\boldsymbol{x}_i;\boldsymbol{\beta}^*)]\boldsymbol{x}_i \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n [y_i^{(K-1)} - p_{K-1}(\boldsymbol{x}_i;\boldsymbol{\beta}^*)]\boldsymbol{x}_i \end{pmatrix} = \frac{1}{n} \begin{pmatrix} X^\top \epsilon_1 \\ X^\top \epsilon_2 \\ \vdots \\ X^\top \epsilon_{K-1} \end{pmatrix}, \tag{D11}$$

where $\epsilon_k := \boldsymbol{y}^{(k)} - p_k(X;\boldsymbol{\beta}^*)$ for $k = 1 : (K-1)$. By the definition of the best approximator (11) and the first-order condition of a minimizer of a convex function, we have $\mathbb{E}[\nabla \mathcal{L}_n(\boldsymbol{\beta}^*)] = \boldsymbol{0}$, where the expectation is w.r.t. the joint distribution of (X, Y). Moreover, since $Y - p_k(X;\boldsymbol{\beta}^*)$ is bounded, it is a sub-Gaussian variable with a bounded variance proxy. Therefore, each coordinate of $\nabla \mathcal{L}_n(\boldsymbol{\beta}^*)$ is a zero-mean sub-exponential variable, which entails that $\|\nabla \mathcal{L}_n(\boldsymbol{\beta}^*)\|_\infty \lesssim \sqrt{\frac{\log(Kp)}{n}}$ with probability at least $1 - C(Kp)^{-1}$. Then, we can follow the same argument in the proof of Theorem 1 to obtain the same high-probability bounds.

The conclusion in Theorem 3 can be proved in the same way as Theorem 3, hence we omit the proof.

Finally, we comment on the proof of asymptotic normality in Theorem 4. Note that we can obtain the same results in Lemma 4 based on the same derivation. With the estimation error bounds for $\boldsymbol{\Theta}$ in Lemma 4 at hand, we can walk through the arguments

in the proof of Theorem 4. First, note that we can decompose the $\hat{b}_j - \beta_j^*$ as

$$
\hat{b}_j - \beta_j^* = \underbrace{(e_j - \widehat{\Theta}_j^\top \boldsymbol{B}_n)(\widehat{\beta} - \beta^*)}_{[1]} + \underbrace{\frac{1}{n}(\widehat{\Theta}_j - \Theta_j)^\top \begin{pmatrix} \boldsymbol{X}^\top \epsilon_1 \\ \boldsymbol{X}^\top \epsilon_2 \\ \vdots \\ \boldsymbol{X}^\top \epsilon_{K-1} \end{pmatrix}}_{[2]} + \underbrace{\frac{1}{n} \Theta_j^\top \begin{pmatrix} \boldsymbol{X}^\top \epsilon_1 \\ \boldsymbol{X}^\top \epsilon_2 \\ \vdots \\ \boldsymbol{X}^\top \epsilon_{K-1} \end{pmatrix}}_{[3]},
$$

where $\epsilon_k := \boldsymbol{y}^{(k)} - p_k(\boldsymbol{X}; \beta^*)$ for $k = 1 : (K-1)$. Our previous analysis of (D11) and the estimator error of $\widehat{\beta}$ guarantees that [1] and [2] can be bounded similarly as in the proof of Theorem 4, which finally leads to the conclusion that $[1], [2] = \mathcal{O}_{\mathbb{P}}(n^{-1/2})$.

Denote $\Sigma = \text{Cov}((X^\top \epsilon_1, X^\top \epsilon_2, \ldots, X^\top \epsilon_{K-1})^\top)$. Then by CLT, we have

$$
\frac{1}{\sqrt{\Theta_j^\top \Sigma \Theta_j}} \left[\sqrt{n}[3] - \frac{1}{\sqrt{n}} \Theta_j^\top \mathbb{E} \begin{pmatrix} \boldsymbol{X}^\top \epsilon_1 \\ \boldsymbol{X}^\top \epsilon_2 \\ \vdots \\ \boldsymbol{X}^\top \epsilon_{K-1} \end{pmatrix} \right] \xrightarrow{\text{d}} N(0, 1),
$$

where

$$
\frac{1}{n} \mathbb{E} \begin{pmatrix} \boldsymbol{X}^\top \epsilon_1 \\ \boldsymbol{X}^\top \epsilon_2 \\ \vdots \\ \boldsymbol{X}^\top \epsilon_{K-1} \end{pmatrix} = \mathbb{E}[\nabla \mathcal{L}_n(\beta^*)] = \boldsymbol{0}.
$$

Then by Slutsky's theorem, we have

$$
\frac{\sqrt{n}[3]}{\sqrt{\Theta_j^\top \Sigma \Theta_j}} \xrightarrow{\text{d}} N(0, 1).
$$

Combining the analysis on all three terms $[1], [2], [3]$, it holds that

$$
\frac{\sqrt{n}(\hat{b}_j - \beta_j^*)}{\sqrt{\Theta_j^\top \Sigma \Theta_j}} \xrightarrow{\text{d}} N(0, 1). \tag{D12}
$$

Finally, it suffices to derive $|\widehat{\Theta}_j^\top \widehat{\Sigma} \widehat{\Theta} - \Theta_j^\top \Sigma \Theta_j| = \mathcal{O}_{\mathbb{P}}(1)$, and the rest would follow from (D12) and Slutsky's theorem. Note that similar to the arguments in the proof of Theorem 4, it holds that

$$
\begin{aligned}
& \left| \widehat{\Theta}_j^\top \widehat{\Sigma} \widehat{\Theta}_j - \Theta_j^\top \Sigma \Theta_j \right| \\
& \leq \left| (\widehat{\Theta}_j - \Theta_j)^\top \widehat{\Sigma} \widehat{\Theta}_j \right| + \left| \Theta_j^\top (\widehat{\Sigma} - \Sigma)^\top \widehat{\Theta}_j \right| + \left| \Theta_j^\top \Sigma (\widehat{\Theta}_j - \Theta_j) \right| \\
& \leq \left| (\widehat{\Theta}_j - \Theta_j)^\top (\widehat{\Sigma} - \Sigma)(\widehat{\Theta}_j - \Theta_j) \right| + \left| (\widehat{\Theta}_j - \Theta_j)^\top (\widehat{\Sigma} - \Sigma)\Theta_j \right| + \left| (\widehat{\Theta}_j - \Theta_j)^\top \Sigma (\widehat{\Theta}_j - \Theta_j) \right| + \left| (\widehat{\Theta}_j - \Theta_j)^\top \Sigma \Theta_j \right| \\
& \quad + \left| \Theta_j^\top (\widehat{\Sigma} - \Sigma)^\top (\widehat{\Theta}_j - \Theta_j) \right| + \left| \Theta_j^\top (\widehat{\Sigma} - \Sigma)^\top \Theta_j \right| + \left| \Theta_j^\top \Sigma (\widehat{\Theta}_j - \Theta_j) \right| \\
& \lesssim \|\widehat{\Theta}_j - \Theta_j\|_1^2 \|\widehat{\Sigma} - \Sigma\|_{\max} + \|\widehat{\Theta}_j - \Theta_j\|_1 \|\widehat{\Sigma} - \Sigma\|_{\max} \|\Theta_j\|_1 + \|\widehat{\Theta}_j - \Theta_j\|_2^2 + \|\widehat{\Theta}_j - \Theta_j\|_2 \|\Sigma \Theta_j\|_2 \\
& \lesssim_{\mathbb{P}} \|\widehat{\Sigma} - \Sigma\|_{\max} \sqrt{K s_0} + \|\widehat{\Theta}_j - \Theta_j\|_2 \\
& \lesssim_{\mathbb{P}} K^3 \sqrt{\frac{s s_0 \log(Kp)}{n}} + K^{9/2} \sqrt{\frac{s(\log(Kp))(\log(np))}{n}} + K^{5/2} \sqrt{\frac{s_0 \log(Kp)}{n}} \\
& = \mathcal{O}(1),
\end{aligned}
$$

which completes the proof. $\qquad \square$

Proof of Theorem 6. The results in Theorems 1, 3, and 4 can be derived in a similar way for the non-identically distributed data, as we argued in Section 2.3.2, thus we omit the proof. $\qquad \square$

Proof of Theorem 7. By the definition of $\widehat{b}$:

$$\widehat{b} - \beta^* = (I_{(K-1)p} - MB_n^{[1]})(\widehat{\beta}^{[1]} - \beta^*) + \frac{1}{n}M\begin{pmatrix} (X^{[1]})^\top \epsilon^{[1](1)} \\ \vdots \\ (X^{[1]})^\top \epsilon^{[1](K-1)} \end{pmatrix}$$

$$= (I_{(K-1)p} - M\widehat{\Sigma}^{[2]}_{\widehat{\beta}^{[2]}})(\widehat{\beta}^{[1]} - \beta^*) + M(\widehat{\Sigma}^{[2]}_{\widehat{\beta}^{[2]}} - B_n^{[1]})(\widehat{\beta}^{[1]} - \beta^*) + \frac{1}{n/2}M\begin{pmatrix} (X^{[1]})^\top \epsilon^{[1](1)} \\ \vdots \\ (X^{[1]})^\top \epsilon^{[1](K-1)} \end{pmatrix},$$

where

$$B_n^{[1]} = \frac{1}{n/2}\sum_{i=1}^{n/2}\int_0^1 B(x_i^{[1]};\beta^* + t(\widehat{\beta}^{[1]} - \beta^*))dt,$$

$$B(x;\beta) := \begin{pmatrix} p_1(x;\beta)[1-p_1(x;\beta)]xx^\top & -p_1(x;\beta)p_2(x;\beta)xx^\top & \ldots & -p_1(x;\beta)p_{K-1}(x;\beta)xx^\top \\ -p_1(x;\beta)p_2(x;\beta)xx^\top & p_2(x;\beta)[1-p_2(x;\beta)]xx^\top & \ldots & -p_2(x;\beta)p_{K-1}(x;\beta)xx^\top \\ \vdots & \vdots & \vdots & \vdots \\ -p_1(x;\beta)p_{K-1}(x;\beta)xx^\top & -p_2(x;\beta)p_{K-1}(x;\beta)xx^\top & \ldots & p_{K-1}(x;\beta)[1-p_{K-1}(x;\beta)]xx^\top \end{pmatrix}.$$

Step 1: Show that when $\eta \asymp K^{5/2}\sqrt{\frac{s\log p}{n}}$, the LP (A1) is feasible with probability $1 - \mathcal{O}(1)$.

Consider $M = \Theta$ with $m_j = \Theta_j$:

$$\Theta\widehat{\Sigma}^{[2]}_{\widehat{\beta}^{[2](k)}} - I = \Theta(\widehat{\Sigma}^{[2]}_{\widehat{\beta}^{[2](k)}} - \widehat{\Sigma}^{[2]}_{\beta^*}) + \Theta\widehat{\Sigma}^{[2]}_{\beta^*} - I.$$

There exist functions $\{g_{j_1 j_2}\}$ such that

$$(\widehat{\Sigma}^{[2]}_{\widehat{\beta}^{[2](k)}})_{j_1,j_2} = \frac{1}{n/2}\sum_{i=1}^{n/2}g_{j_1 j_2}(\{x_i^\top \widehat{\beta}^{[2](k)}\})(x_i^{[2]})_{j_1 \bmod K}(x_i^{[2]})_{j_2 \bmod K}, \quad (\widehat{\Sigma}^{[2]}_{\beta^{(k)*}})_{j_1,j_2} = \frac{1}{n/2}\sum_{i=1}^{n/2}g_{j_1 j_2}(\{(x_i^{[2]})^\top \beta^{(k)*}\})(x_i^{[2]})_{j_1 \bmod K}(x_i^{[2]})_{j_2 \bmod K},$$

where $\|\nabla g_{j_1 j_2}\|_2 \lesssim 1$. Hence, by Cauchy-Schwarz inequality, we have

$$\left| e_1^\top \Theta(\widehat{\Sigma}^{[2]}_{\widehat{\beta}^{[2]}} - \widehat{\Sigma}^{[2]}_{\beta^*})e_1 \right| = \left| \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K-1}[\nabla g_{1,1}(\{(1-\nu_i)x_i^\top \widehat{\beta}^{[2]} + \nu_i x_i^\top \beta^{(k)*}\}_{k=1}^{K-1})]^\top \{x_i^\top(\widehat{\beta}^{[2](k)} - \beta^{(k)*})\}_{k=1}^{K-1} \cdot (\Theta_1^{(k)})^\top x_i \cdot x_{i,1} \right|$$

$$\lesssim \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K-1}[(\Theta_1^{(k)})^\top x_i]^2 x_{i,1}^2} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K-1}[x_i^\top(\widehat{\beta}^{[2](k)} - \beta^{(k)*})]^2}$$

$$\lesssim \|\widehat{\beta}^{[2]} - \beta^*\|_2$$

$$\lesssim K^{5/2}\sqrt{\frac{s\log(Kp)}{n}},$$

with probability at least $1 - C'(Kp)^{-1}$. Similarly, we can show that

$$\|\Theta(\widehat{\Sigma}^{[2]}_{\widehat{\beta}^{[2]}} - \widehat{\Sigma}^{[2]}_{\beta^*})\|_{\max} \lesssim K^{5/2}\sqrt{\frac{s\log(Kp)}{n}}$$

with probability at least $1 - C'(Kp)^{-1}$. Additionally, similar to the argument in the proof of Lemma 4, we can show that

$$\|\widehat{\Sigma}^{[2]}_{\beta^*} - I\|_{\max} \lesssim \sqrt{\frac{\log(Kp)}{n}},$$

with probability at least $1 - C(Kp)^{-1}$. Therefore, with probability at least $1 - C(Kp)^{-1}$, we have

$$\|\Theta^*\widehat{\Sigma}^{[2]}_{\widehat{\beta}^{[2]}} - I\|_{\max} \lesssim K^{5/2}\sqrt{\frac{s\log(Kp)}{n}},$$

which implies that the LP (A1) is feasible when $\eta \gtrsim K^{5/2}\sqrt{\frac{s\log(Kp)}{n}}$.

Step 2: We show that $\left|\boldsymbol{m}_j^\top(\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\beta}}}^{[2]} - \boldsymbol{B}_n^{[1]})(\widehat{\boldsymbol{\beta}}^{[2]} - \boldsymbol{\beta}^*)\right| \lesssim \|\boldsymbol{\Theta}_j\|_1 \cdot \frac{K^5 s \log(Kp)}{n}$ with probability $1 - o(1)$.

Since $\boldsymbol{\Theta}_j$ is feasible for the LP (A1), we must have $\|\boldsymbol{m}_j\|_1 \leq \|\boldsymbol{\Theta}_j\|_1$. Therefore,

$$\left|\boldsymbol{m}_j^\top(\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\beta}}^{[2]}}^{[2]} - \boldsymbol{B}_n^{[1]})(\widehat{\boldsymbol{\beta}}^{[2]} - \boldsymbol{\beta}^*)\right| \leq \left|\boldsymbol{m}_j^\top(\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\beta}}^{[2]}}^{[2]} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}^*}^{[2]})(\widehat{\boldsymbol{\beta}}^{[2]} - \boldsymbol{\beta}^*)\right| + \left|\boldsymbol{m}_j^\top(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}^*}^{[2]} - \boldsymbol{\Sigma})(\widehat{\boldsymbol{\beta}}^{[2]} - \boldsymbol{\beta}^*)\right| + \left|\boldsymbol{m}_j^\top(\boldsymbol{\Sigma} - \boldsymbol{B}_n^{[1]})(\widehat{\boldsymbol{\beta}}^{[2]} - \boldsymbol{\beta}^*)\right|,$$

where

$$\left|\boldsymbol{m}_j^\top(\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\beta}}^{[2]}}^{[2]} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}^*}^{[2]})(\widehat{\boldsymbol{\beta}}^{[2]} - \boldsymbol{\beta}^*)\right| \leq \|\boldsymbol{m}_j\|_1 \|\widehat{\boldsymbol{\beta}}^{[2]} - \boldsymbol{\beta}^*\|_2^2 \lesssim \|\boldsymbol{\Theta}_j\|_1 \frac{K^5 s \log(Kp)}{n},$$

$$\left|\boldsymbol{m}_j^\top(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}^*}^{[2]} - \boldsymbol{\Sigma})(\widehat{\boldsymbol{\beta}}^{[2]} - \boldsymbol{\beta}^*)\right| \leq \|\boldsymbol{m}_j\|_1 \|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}^*}^{[2]} - \boldsymbol{\Sigma}\|_{\max} \|\widehat{\boldsymbol{\beta}}^{[2]} - \boldsymbol{\beta}^*\|_1 \lesssim \|\boldsymbol{\Theta}_j\|_1 \frac{K^3 s \log(Kp)}{n},$$

$$\left|\boldsymbol{m}_j^\top(\boldsymbol{\Sigma} - \boldsymbol{B}_n^{[1]})(\widehat{\boldsymbol{\beta}}^{[2]} - \boldsymbol{\beta}^*)\right| \leq \|\boldsymbol{m}_j\|_1 \|\boldsymbol{\Sigma} - \boldsymbol{B}_n^{[1]}\|_{\max} \|\widehat{\boldsymbol{\beta}}^{[2]} - \boldsymbol{\beta}^*\|_1 \lesssim \|\boldsymbol{\Theta}_j\|_1 \frac{K^5 s \log(Kp)}{n},$$

with probability at least $1 - C'(Kp)^{-1}$. Therefore,

$$\left|\boldsymbol{m}_j^\top(\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\beta}}}^{[2]} - \boldsymbol{B}_n^{[1]})(\widehat{\boldsymbol{\beta}}^{[2]} - \boldsymbol{\beta}^*)\right| \lesssim \|\boldsymbol{\Theta}_j\|_1 \cdot \frac{K^5 s \log(Kp)}{n},$$

with probability at least $1 - C'(Kp)^{-1}$.

Step 3: We complete the proof by applying CLT.

Note that

$$\|(\boldsymbol{e}_j - \boldsymbol{m}_j^\top \widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\beta}}^{[2]}}^{[2]})(\widehat{\boldsymbol{\beta}}^{[1]} - \boldsymbol{\beta}^*)\| \leq \|\widehat{\boldsymbol{\Sigma}}_{\widehat{\boldsymbol{\beta}}^{[2]}}^{[2]}\boldsymbol{m}_j - \boldsymbol{e}_j\|_\infty \|\widehat{\boldsymbol{\beta}}^{[1]} - \boldsymbol{\beta}^*\|_1 \lesssim \eta K^3 \sqrt{\frac{s \log(Kp)}{n}} \lesssim K^{11/2} s^{3/2} \frac{\log(Kp)}{n},$$

with probability at least $1 - C'(Kp)^{-1}$. Because $\boldsymbol{m}_j$ is independent of $\boldsymbol{X}^{[1]}$ and $\boldsymbol{\epsilon}^{[1]}$, conditioned on $(\boldsymbol{X}^{[2]}, \boldsymbol{Y}^{[2]})$, by the CLT,

$$\frac{1}{\sqrt{n/2}\sqrt{\boldsymbol{m}_j^\top \boldsymbol{\Sigma} \boldsymbol{m}_j}} \boldsymbol{m}_j^\top \begin{pmatrix} (\boldsymbol{X}^{[1]})^\top \boldsymbol{\epsilon}^{[1](1)} \\ \vdots \\ (\boldsymbol{X}^{[1]})^\top \boldsymbol{\epsilon}^{[1](K-1)} \end{pmatrix} \xrightarrow{\text{d}} N(0,1).$$

By the definition of weak convergence through the expectation of continuous bounded functions, the weak convergence also holds with the σ-field generated by all the data $(\boldsymbol{X}^{[1]}, \boldsymbol{Y}^{[1]}) \cup (\boldsymbol{X}^{[2]}, \boldsymbol{Y}^{[2]})$.

Finally, we have

$$\frac{\sqrt{n/2}(\hat{b}_j - \beta_j^*)}{\sqrt{\boldsymbol{m}_j^\top \boldsymbol{\Sigma} \boldsymbol{m}_j}} = \mathcal{O}_\mathbb{P}\left(\|\boldsymbol{\Theta}_j\|_1 \cdot \frac{K^5 s \log(Kp)}{\sqrt{n}} + K^{11/2} s^{3/2} \frac{\log(Kp)}{\sqrt{n}}\right) + \frac{1}{\sqrt{n/2}} \boldsymbol{m}_j^\top \begin{pmatrix} (\boldsymbol{X}^{[1]})^\top \boldsymbol{\epsilon}^{[1](1)} \\ \vdots \\ (\boldsymbol{X}^{[1]})^\top \boldsymbol{\epsilon}^{[1](K-1)} \end{pmatrix}$$

$$\xrightarrow{\text{d}} N(0,1).$$

It suffices to prove $\left|\boldsymbol{m}_j^\top \widehat{\boldsymbol{\Sigma}}^{[1]} \boldsymbol{m}_j - \boldsymbol{m}_j^\top \boldsymbol{\Sigma} \boldsymbol{m}_j\right| = \mathcal{O}_\mathbb{P}(1)$, which follows from a similar argument used in the proof of Theorem 4, hence we omit it here. $\square$

Proof of Theorem 8. Most of the proof follows the same argument in the proof of Theorem 4, therefore we only point out the different places.

First, we need to show the high-probability bound for $\|\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j^*\|_2$ and $\|\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j^*\|_1$, respectively. Instead of the way we took to bound $\|\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j^*\|_2$ in the proof of Theorem 4, we use the fact that

$$\|\widehat{\boldsymbol{\Sigma}}_{-j,j} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}^*,-j,j} - \widehat{\boldsymbol{\Sigma}}_{-j,-j}\boldsymbol{\gamma}_j^* + \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}^*,-j,-j}\boldsymbol{\gamma}_j^*\|_\infty \lesssim \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \lesssim K^{5/2}\sqrt{\frac{s \log(Kp)}{n}},$$

$$\left|(\widehat{\boldsymbol{\Sigma}}_{-j,j} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}^*,-j,j} - \widehat{\boldsymbol{\Sigma}}_{-j,-j}\boldsymbol{\gamma}_j^* + \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}^*,-j,-j}\boldsymbol{\gamma}_j^*)^\top(\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j^*)\right| \lesssim \|\widehat{\boldsymbol{\Sigma}}_{-j,j} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}^*,-j,j} - \widehat{\boldsymbol{\Sigma}}_{-j,-j}\boldsymbol{\gamma}_j^* + \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}^*,-j,-j}\boldsymbol{\gamma}_j^*\|_\infty \|\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j^*\|_1,$$

with probability at least $1 - (Kp)^{-1}$, which implies that

$$\|(\widehat{\gamma}_j - \gamma_j^*)_{S_j^c}\|_1 \lesssim \|(\widehat{\gamma}_j - \gamma_j^*)_{S_j}\|_1. \tag{D13}$$

with probability at least $1 - (Kp)^{-1}$.

Second, using (D13), with a similar analysis in the proof of Lemma 1, we can prove that

$$\frac{1}{K^2}\|\widehat{\gamma}_j - \gamma_j^*\|_2 \lesssim \frac{1}{2n}(\widehat{\gamma}_j - \gamma_j^*)^\top \widehat{\Sigma}_{-j,-j}(\widehat{\gamma}_j - \gamma_j^*) \lesssim \lambda_j \|(\widehat{\gamma}_j - \gamma_j^*)_{S_j}\|_1 \lesssim \lambda_j \sqrt{Ks_0}\|(\widehat{\gamma}_j - \gamma_j^*)_{S_j}\|_2,$$

with probability at least $1 - (Kp)^{-1}$, which implies that

$$\|\widehat{\gamma}_j - \gamma_j^*\|_2 \lesssim K^5 \sqrt{ss_0}\sqrt{\frac{\log(Kp)}{n}}, \quad \|\widehat{\gamma}_j - \gamma_j^*\|_1 \lesssim K^{11/2}\sqrt{ss_0}\sqrt{\frac{\log(Kp)}{n}},$$

with probability at least $1 - (Kp)^{-1}$. Then, by following the same argument in the proof of Lemma 4, we can show that

$$|\widehat{\tau}_j^2 - (\tau_j^*)^2|, |\widehat{\tau}_j^{-2} - (\tau_j^*)^{-2}| \lesssim K^5 \sqrt{ss_0}\sqrt{\frac{\log(Kp)}{n}},$$

with probability at least $1 - (Kp)^{-1}$, when $n \gtrsim K^{11}ss_0^2 \log(pK)$.

Finally, we follow the steps in the proof of Theorem 4 and argue that $\|\widehat{\gamma}_j - \gamma_j^*\|_2^2 \vee |\widehat{\tau}_j^{-2} - (\tau_j^*)^{-2}|^2 \vee |\widehat{\tau}_j^2 - (\tau_j^*)^2|^2 \vee \|\widehat{\beta} - \beta^*\|_2^2 \ll n^{-1/2}$ with probability at least $1 - (Kp)^{-1}$, when $n \gg K^{20}(ss_0)^2(\log(Kp))^2$, which completes the proof. $\square$

T A B L E C1 The sum of squared ℓ_2-estimation error of $\beta^{(1)*}$, $\beta^{(2)*}$, and $\beta^{(3)*}$, in Model 5, for the contrast-based penalization and the over-parameterization method.

Method	$n = 100$	$n = 200$	$n = 400$
Contrast	9.009 (2.465)	5.682 (1.505)	3.384 (0.98)
Over-parameterization	10.859 (2.144)	7.212 (1.471)	4.469 (0.955)

T A B L E C2 The average misclassification error rates on test data in Model 5, for ℓ_1-penalized multinomial regression, SVMs, LDA, Random Forests, and kNN.

Method/n	$n = 100$	$n = 200$	$n = 400$
Penalized multinomial regression	20.64% (2.54%)	18.01% (1.67%)	16.44% (1.36%)
SVMs	47.88% (3.96%)	38.99% (2.59%)	31.59% (2.09%)
LDA	43.04% (2.71%)	61.45% (2.68%)	38.88% (1.96%)
Random forests	29.13% (3.51%)	24.66% (2.17%)	21.84% (1.65%)
k-nearest neighbors	53.94% (2.30%)	50.77% (2.22%)	47.02% (2.08%)

T A B L E C3 The average coverage probability of 95% CIs on the coefficients corresponding to the signal set $S = \{\{1,2,3\}, \{4,5,6\}\}$ and noise set $S^c = \{\{4,5,6,\ldots,p\}, \{1,2,3,7,8,\ldots,p\}\}$ in Model 5, for the debiased Lasso and the vector bootstrap method.

Method	Measure	$n = 100$	$n = 200$	$n = 400$
Debiased-Lasso	Average coverage probability on S	0.917 (0.13)	0.884 (0.132)	0.897 (0.131)
	Average coverage probability on S^c	0.999 (0.002)	0.997 (0.003)	0.992 (0.005)
	Average length of CI (S)	5.491 (3.001)	2.118 (0.586)	1.53 (0.159)
	Average length of CI (S^c)	4.841 (2.861)	2.355 (0.555)	1.645 (0.133)
Bootstrap	Average coverage probability on S	0.776 (0.166)	0.875 (0.141)	0.942 (0.107)
	Average coverage probability on S^c	1 (0)	1 (0.001)	1 (0.001)
	Average length of CI (S)	2.495 (0.294)	2.185 (0.149)	1.763 (0.111)
	Average length of CI (S^c)	0.181 (0.012)	0.217 (0.011)	0.26 (0.013)

T A B L E C4 The average Type-I error rate and power of individual testing $H_0 : \beta_j^{(k)*} = 0$ v.s. $H_1 : \beta_j^{(k)*} \neq 0$ at level 5% in Model 5, for the debiased Lasso and the vector bootstrap method.

Method	Measure	$n = 100$	$n = 200$	$n = 400$
Debiased-Lasso	Type-I error	0.001 (0.002)	0.002 (0.003)	0.009 (0.005)
	Power	0.5 (0.201)	0.892 (0.131)	0.997 (0.029)
Bootstrap	Type-I error	0 (0)	0 (0.001)	0 (0.001)
	Power	0.441 (0.149)	0.862 (0.122)	1 (0)

T A B L E C5 The average family-wise error rate (FWER) and power of multiple testing $H_0 : \beta_j^{(k)*} = 0$ v.s. $H_1 : \beta_j^{(k)*} \neq 0$ for $j = 1 : p$ and $k = 1 : (K-1)$ in Model 5, for the debiased Lasso and the multiple splitting method.

Method	Measure	$n = 100$	$n = 200$	$n = 400$
Debiased-Lasso	FWER	0 (0)	0 (0)	0 (0)
	Power	0.038 (0.076)	0.298 (0.168)	0.778 (0.152)
Multiple-Spliting	FWER	0.045 (0.208)	0.015 (0.122)	0.01 (0.1)
	Power	0.057 (0.114)	0.2 (0.186)	0.86 (0.122)

T A B L E C6 The sum of squared ℓ_2-estimation error of $\beta^{(1)*}$, $\beta^{(2)*}$, and $\beta^{(3)*}$, in Model 6, for the contrast-based penalization and the over-parameterization method.

Method	$n = 100$	$n = 200$	$n = 400$
Contrast	12.041 (1.54)	8.84 (1.195)	6.01 (1.018)
Over-parameterization	13.02 (1.523)	9.813 (1.435)	6.711 (0.87)

T A B L E C7 The average misclassification error rates on test data in Model 6, for ℓ_1-penalized multinomial regression, SVMs, LDA, Random Forests, and kNN.

Method	$n = 100$	$n = 200$	$n = 400$
Penalized multinomial regression	20.05% (2.37%)	16.03% (1.60%)	13.79% (1.19%)
SVMs	35.07% (4.43%)	25.94% (2.29%)	20.36% (1.49%)
LDA	37.96% (2.74%)	60.69% (2.70%)	36.39% (1.89%)
Random forests	23.34% (3.24%)	19.17% (1.94%)	16.86% (1.35%)
k-nearest neighbors	45.14% (2.83%)	39.99% (2.58%)	34.74% (2.27%)

T A B L E C8 The average coverage probability of 95% CIs on the coefficients corresponding to the signal set $S = \{\{1, 2, \ldots, 10\}, \{11, 12, \ldots, 20\}\}$ and noise set $S^c = \{\{11, 12, \ldots, p\}, \{1, 2, \ldots, 10, 21, 22, \ldots, p\}\}$ in Model 6, for the debiased Lasso and the vector bootstrap method.

Method	Measure	$n = 100$	$n = 200$	$n = 400$
Debiased-Lasso	Average coverage probability on S	0.993 (0.022)	0.964 (0.043)	0.923 (0.064)
	Average coverage probability on S^c	1 (0.001)	0.999 (0.001)	0.998 (0.003)
	Average length of CI (S)	4.542 (1.856)	2.801 (0.811)	1.748 (0.197)
	Average length of CI (S^c)	4.083 (1.377)	2.997 (0.882)	1.859 (0.203)
Bootstrap	Average coverage probability on S	0.74 (0.082)	0.849 (0.077)	0.892 (0.071)
	Average coverage probability on S^c	1 (0)	1 (0)	1 (0)
	Average length of CI (S)	1.512 (0.133)	1.568 (0.105)	1.477 (0.078)
	Average length of CI (S^c)	0.162 (0.011)	0.196 (0.012)	0.238 (0.011)

T A B L E C9 The average Type-I error rate and power of individual testing $H_0 : \beta_j^{(k)*} = 0$ v.s. $H_1 : \beta_j^{(k)*} \neq 0$ at level 5% in Model 6, for the debiased Lasso and the vector bootstrap method.

Method	Measure	$n = 100$	$n = 200$	$n = 400$
Debiased-Lasso	Type-I error	0 (0.001)	0.001 (0.001)	0.003 (0.003)
	Power	0.041 (0.045)	0.17 (0.078)	0.446 (0.085)
Bootstrap	Type-I error	0 (0)	0 (0)	0 (0)
	Power	0.059 (0.046)	0.198 (0.073)	0.499 (0.087)

T A B L E C10 The average family-wise error rate (FWER) and power of multiple testing $H_0 : \beta_j^{(k)*} = 0$ v.s. $H_1 : \beta_j^{(k)*} \neq 0$ for $j = 1 : p$ and $k = 1 : (K - 1)$ in Model 6, for the debiased Lasso and the multiple splitting method.

Method	Measure	$n = 100$	$n = 200$	$n = 400$
Debiased-Lasso	FWER	0 (0)	0 (0)	0 (0)
	Power	0 (0)	0.001 (0.009)	0.017 (0.03)
Multiple-Spliting	FWER	0 (0)	0.015 (0.122)	0 (0)
	Power	0 (0.005)	0.003 (0.025)	0.053 (0.07)

T A B L E C11 The sum of squared ℓ_2-estimation error of $\beta^{(1)*}$, $\beta^{(2)*}$, and $\beta^{(3)*}$, in Model 7, for the contrast-based penalization and the over-parameterization method.

Method	$n = 100$	$n = 200$	$n = 400$
Contrast	14.669 (2.627)	8.93 (2.007)	5.239 (1.41)
Over-parameterization	17.366 (2.426)	11.319 (1.916)	7.232 (1.328)

T A B L E C12 The average misclassification error rates on test data in Model 7, for ℓ_1-penalized multinomial regression, SVMs, LDA, Random Forests, and kNN.

Method/n	$n = 100$	$n = 200$	$n = 400$
Penalized multinomial regression	25.97% (2.63%)	22.06% (1.68%)	19.98% (1.41%)
SVMs	59.18% (4.28%)	48.07% (2.64%)	39.79% (2.13%)
LDA	52.67% (2.77%)	69.83% (2.65%)	47.78% (2.05%)
Random forests	38.52% (3.64%)	32.22% (2.59%)	28.17% (2.07%)
k-nearest neighbors	62.74% (2.20%)	59.46% (2.16%)	55.56% (2.15%)

T A B L E C13 The average coverage probability of 95% CIs on the coefficients corresponding to the signal set $S = \{\{1, 2, 3\}, \{4, 5, 6\}, \{7, 8, 9\}\}$ and noise set $S^c = \{\{4, 5, 6, \ldots, p\}, \{1, 2, 3, 7, 8, \ldots, p\}, \{1, 2, 3, 4, 5, 6, 10, 11, \ldots, p\}\}$ in Model 7, for the debiased Lasso and the vector bootstrap method.

Method	Measure	$n = 100$	$n = 200$	$n = 400$
Debiased-Lasso	Average coverage probability on S	0.922 (0.093)	0.901 (0.112)	0.915 (0.1)
	Average coverage probability on S^c	1 (0.001)	0.998 (0.002)	0.994 (0.004)
	Average length of CI (S)	6.145 (4.395)	2.521 (0.991)	1.758 (0.158)
	Average length of CI (S^c)	5.634 (3.871)	2.65 (0.72)	1.889 (0.132)
Bootstrap	Average coverage probability on S	0.684 (0.141)	0.824 (0.128)	0.906 (0.113)
	Average coverage probability on S^c	1 (0)	1 (0)	1 (0.001)
	Average length of CI (S)	2.298 (0.255)	2.107 (0.149)	1.696 (0.092)
	Average length of CI (S^c)	0.163 (0.009)	0.197 (0.01)	0.237 (0.011)

T A B L E C14   The average Type-I error rate and power of individual testing $H_0 : \beta_j^{(k)*} = 0$ v.s. $H_1 : \beta_j^{(k)*} \neq 0$ at level 5% in Model 7, for the debiased Lasso and the vector bootstrap method.

Method	Measure	$n = 100$	$n = 200$	$n = 400$
Debiased-Lasso	Type-1 error	0 (0.001)	0.002 (0.002)	0.006 (0.004)
	Power	0.303 (0.149)	0.77 (0.136)	0.973 (0.064)
Bootstrap	Type-1 error	0 (0)	0 (0)	0 (0.001)
	Power	0.371 (0.118)	0.795 (0.12)	0.994 (0.025)

T A B L E C15   The average family-wise error rate (FWER) and power of multiple testing $H_0 : \beta_j^{(k)*} = 0$ v.s. $H_1 : \beta_j^{(k)*} \neq 0$ for $j = 1 : p$ and $k = 1 : (K-1)$ in Model 7, for the debiased Lasso and the multiple splitting method.

Method	Measure	$n = 100$	$n = 200$	$n = 400$
Debiased-Lasso	FWER	0 (0)	0 (0)	0 (0)
	Power	0.011 (0.036)	0.134 (0.111)	0.516 (0.155)
Multiple-Spliting	FWER	0.05 (0.218)	0.085 (0.28)	0.985 (0.122)
	Power	0.013 (0.057)	0.019 (0.051)	0.398 (0.114)

T A B L E C16   The sum of squared ℓ_2-estimation error of $\boldsymbol{\beta}^{(1)*}$, $\boldsymbol{\beta}^{(2)*}$, and $\boldsymbol{\beta}^{(3)*}$, in Model 8, for the contrast-based penalization and the over-parameterization method.

Method	$n = 100$	$n = 200$	$n = 400$
Contrast	71.432 (1.891)	66.436 (1.547)	63.046 (1.132)
Over-parameterization	57.311 (26.924)	39.248 (17.762)	28.162 (13.787)

T A B L E C17   The average misclassification error rates on test data in Model 8, for ℓ_1-penalized multinomial regression, SVMs, LDA, Random Forests, and kNN.

Method	$n = 100$	$n = 200$	$n = 400$
Penalized multinomial regression	44.57% (2.65%)	38.44% (2.30%)	33.82% (2.12%)
SVMs	63.45% (6.71%)	56.71% (5.56%)	53.27% (5.96%)
LDA	64.78% (5.25%)	72.14% (3.46%)	62.49% (6.07%)
Random forests	58.48% (4.83%)	54.51% (4.67%)	51.74% (4.11%)
k-nearest neighbors	61.63% (3.64%)	59.06% (3.70%)	55.60% (3.69%)

T A B L E   C18   The average coverage probability of 95% CIs on the coefficients corresponding to the signal set $S = \{\{1, 2, \ldots, 10\}, \{11, 12, \ldots, 20\}, \{21, 22, \ldots, 30\}\}$ and noise set $S^c = \{\{11, 12, \ldots, p\}, \{1, 2, \ldots, 10, 21, 22, \ldots, p\}, \{1, 2, \ldots, 20, 41, 42, \ldots, p\}\}$ in Model 8, for the debiased Lasso and the vector bootstrap method.

Method	Measure	$n = 100$	$n = 200$	$n = 400$
Debiased-Lasso	Average coverage probability on S	0.997 (0.01)	0.994 (0.018)	0.98 (0.03)
	Average coverage probability on S^c	1 (0)	1 (0.001)	0.999 (0.001)
	Average length of CI (S)	6.203 (3.579)	3.331 (1.596)	1.79 (0.331)
	Average length of CI (S^c)	7.467 (9.405)	3.643 (1.853)	1.845 (0.342)
Bootstrap	Average coverage probability on S	0.485 (0.078)	0.603 (0.08)	0.685 (0.072)
	Average coverage probability on S^c	1 (0)	1 (0.001)	1 (0.001)
	Average length of CI (S)	0.979 (0.067)	1 (0.042)	0.88 (0.025)
	Average length of CI (S^c)	0.143 (0.008)	0.165 (0.01)	0.193 (0.012)

T A B L E C19   The average Type-I error rate and power of individual testing $H_0 : \beta_j^{(k)*} = 0$ v.s. $H_1 : \beta_j^{(k)*} \neq 0$ at level 5% in Model 8, for the debiased Lasso and the vector bootstrap method.

Method	Measure	$n = 100$	$n = 200$	$n = 400$
Debiased-Lasso	Type-I error	0 (0)	0 (0.001)	0.001 (0.001)
	Power	0.019 (0.033)	0.124 (0.08)	0.506 (0.099)
Bootstrap	Type-I error	0 (0)	0 (0.001)	0 (0.001)
	Power	0.101 (0.052)	0.408 (0.078)	0.793 (0.069)

T A B L E C20 The average family-wise error rate (FWER) and power of multiple testing $H_0 : \beta_j^{(k)*} = 0$ v.s. $H_1 : \beta_j^{(k)*} \neq 0$ for $j = 1 : p$ and $k = 1 : (K-1)$ in Model 8, for the debiased Lasso and the multiple splitting method.

Method	Measure	$n = 100$	$n = 200$	$n = 400$
Debiased-Lasso	FWER	0 (0)	0 (0)	0 (0)
	Power	0 (0)	0 (0)	0.012 (0.02)
Multiple-Spliting	FWER	0 (0)	0 (0)	0 (0)
	Power	0 (0.002)	0 (0)	0 (0)