Electronic Journal of Statistics
Vol. 13 (2019) 3161–3194
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1601

High-dimensional generalized linear models incorporating graphical structure among predictors*

Shengbin Zhou†

College of Information Engineering
Lingnan Normal University
Zhanjiang 524048, China
e-mail: zhoushengbin@lingnan.edu.cn

Jingke Zhou‡

School of Statistics
Southwestern University of
Finance and Economics
Chengdu, China
e-mail: jkzhou@swufe.edu.cn

Bo Zhang§

School of Statistics
Renmin University of China
Beijing, China
e-mail: mabzhang@ruc.edu.cn

Abstract: In this paper, we propose a sparse generalized linear model incorporating graphical structure among predictors (sGLMg), which is an extension of [37] where they exploit the structure information among predictors to improve the performance for the linear regression. There is an explicit expression between the coefficient and the predictor graph measured by the precision matrix in the linear regression, however, this structure does not exist in generalized linear model for the explicit expression of the coefficient in generalized linear model is usually hard to be obtained. To incorporate the graphical structure among predictors for generalized linear models, we make use of the sufficient reduction techniques to reestablish the relationship between the coefficient and the precision matrix. The oracle inequalities of the estimator for sGLMg are also presented and the finite sample performance of the proposed methods is examined via numerical simulations and a breast cancer data analysis.

*We are very grateful to the referees for their constructive comments.

†Shengbin Zhou is supproted by the grants Equipment Preresearch Foundation Project 61400010303 and the Competitive Allocation of Special Funds and Technology Innovation Strategy in Guangdong Province of China 2018A06001.

‡Jingke Zhou is supported by the grant NSFC 71801137 and the Joint Lab of Data Science and Business Intelligence at Southwest University of Finance and Economics.

§Bo Zhang is supported by the grant NSFC 71471173.

Keywords and phrases: Graphical structure, sparse regression, generalized linear models, sufficient dimension reduction, oracle inequalities.

Received February 2018.

Contents

1. Introduction

With the development of science and technology, the problem with high dimensionality has become increasingly important over the recent years. Regularization is fundamental in analysis of high-dimensional data. A well-known example for regularization is Lasso ([33]), however, in some applications, such as ANOVA or multi-task regression, the selection of important predictors corresponds to the selection of the groups of predictors. As a natural extension of Lasso, the group Lasso, which is proposed by [1] and further developed by [38], exploits a weighted sum of ℓ_2 norms of the coefficients associated with a group of features and leads to feature selection at group level. For more details about the group Lasso we refer to [13]. An obvious limitation of group Lasso is the non-overlapping structure which introduces a barrier to its applicability in practice where features may be encoded in more than one group. A solution to this problem is the overlapping group Lasso which proposed by [15] and further studied by [27] in the linear regression model setting.

In many studies, discrete data, such as categorical data or count data, is frequently encountered. Generalized linear models (GLMs, [22]) are the most commonly-used regression models for discrete data. Regularization method has been proposed to manipulate the small n and big p problems in GLMs [30, 23, 29, 3, 40]. The main shortcoming of these procedures is that the group structure of predictors must be pre-specified and this is not always possible in practice. The graphical structure of predictors, however, can be obtained from prior information, for example, in biological studies, the massive information

about gene interaction can be used to construct the predictor graph where nodes represent genes and edges indicate regulatory relationships [20, 31]. If the prior information cannot be obtained in some applications, we can construct the predictor graph by sparse estimation of the covariance (or precision) matrix of the predictors [39, 11, 6]. On the other hand, it is reasonable to assume that two neighboring genes in a network are more likely to participate together in the same biological process than two genes far away in the network [28, 16]. In particular, as demonstrated in cancer marker discovery [7], changes in expression of some causal genes governing metastatic potential (e.g., ERBB2 and MYC) may be only subtle and nonsignificant while some of their neighbors have much stronger alterations. Hence, it is reasonable and helpful to take the neighbors of a predictor in the graph as a group.

There are now a lot of methods to utilize the graphical information of predictors. Recently, Yu and Liu [37] propose a node-by-node method to incorporate the graphical information among predictors for linear regression model. By motivating by the least square estimator, they note that the true coefficients of the model can be expressed as

$$\beta^0 = \Sigma^{-1}\Sigma_{xy} = \Omega\Sigma_{xy}, \tag{1}$$

where β^0 is the true coefficients of the model, $\Omega = \Sigma^{-1}$ is the precision matrix, Σ_{xy} is the cross-covariance vector. Thus, there are a natural relationship between the predictors and the graphical structures in the linear regression model for the graphical structure of predictors can be defined by the precision matrix, Ω. However, this strategy won't work when we seek to construct such relationship between the predictors and the graphical structures in generalized linear models, since the closed form for the estimator of the generalized linear models, such as logistic regression model or poisson regression model, usually is hard to be obtained in practice. Hence, how to incorporate the graphical structure among predictors for generalized linear models becomes a very interesting and challenging problem. In this paper, we note that the true coefficients of the generalized linear models can also be expressed as the form as (1) by using the sufficient dimension reduction (SDR, [8]) techniques:

$$\beta^* \propto \Sigma^{-1}(\eta(y) - \mu) = \Omega(\eta(y) - \mu), \tag{2}$$

where β^* is the true coefficients of the generalized linear models, μ is the expectation of the predictors and $\eta(y)$ is a function of y, which is a given value of the response. Based on the equation (2) and motivated by Yu and Liu [37], we propose a sparse generalized linear models incorporating graphical structure among predictors (sGLMg) to model the graphical structure information of predictors for generalized linear models. The oracle inequality and the model selection consistency of the proposed sGLMg method is presented in this paper by assuming that the predictors graphical structure G is given. In fact, when the graphical structure G is unknown, it also can be obtained by sparse estimation of the covariance (or precision) matrix of the predictors [39, 11, 6]. In simulation studies, we compare both cases when the graphical structure of

predictors is given and it is estimated. The sGLMg method proposed in this paper can utilize the neighborhood information of the graph directly and many popular methods such as adaptive Lasso, group Lasso and ridge regression can be included as special cases.

The remainder of this paper is organized as follows. In section 2, we introduce our proposed sGLMg model. In section 3, we study the theoretical properties of sGLMg. In section 4 and 5, Monte Carlo simulation studies and a breast cancer data analysis are conducted to examine the performance of sGLMg method in finite samples. Finally, we conclude this paper with some discussion in section 6.

2. Sparse generalized linear models incorporating graphical structure

We consider the generalized linear models (GLMs) introduced by [22]. Let F be a probability distribution on $\mathbb{R}$ not concentrated on a point and $(\mathbf{X}, Y)$ be a pair of random variables, where $\mathbf{X} \in \mathbb{R}^p$ and $Y \in \mathbb{R}$. We assume that $\mathbf{X}$ follows some multivariate distribution with mean $0_{p\times 1}$ and covariance matrix Σ. The conditional distribution of Y given $\mathbf{X} = \mathbf{x}$ is $P(Y|(\beta_0^*, \beta^*), \mathbf{x}) = \exp\{y(\beta_0^* + \beta^{*\top}\mathbf{x}) - \phi(\beta_0^* + \beta^{*\top}\mathbf{x})\}$, where $\beta_0^* + \beta^{*\top}\mathbf{x} \in \Theta$ with $\Theta := \{\theta \in \mathbb{R} : \int \exp(\theta x)F(dx) < \infty\}$ and ϕ is normalized function. Note that $\mu = \mathbb{E}(Y|\mathbf{X}) \overset{\text{a.s.}}{=} \phi'(\beta_0^* + \beta^{*\top}\mathbf{X})$, that is $\beta_0^* + \beta^{*\top}\mathbf{X} \overset{\text{a.s.}}{=} g(\mu)$, where $g = \phi'^{-1}$ is the so called link function. In fact, the link function g can be any strictly monotone differentiable functions and we only consider canonical link functions, that is, $g(\mu) = \beta_0^* + \beta^{*\top}\mathbf{X}$. The standard linear regression model is obviously an example of GLMs, in addition, the common examples of GLMs including: logistic regression model, poisson regression model, gamma model and exponential (or weibull) model.

Let $(\mathbf{X}_1, Y_1), \cdots, (\mathbf{X}_n, Y_n)$ be the i.i.d. copies of the population $(\mathbf{X}, Y)$, where $\mathbf{X}_i = (X_{i,1}, \cdots, X_{i,p})^\top$, $i = 1, \cdots, n$. We consider the case of high dimensional regression and assume that

- (A1) the variable $\mathbf{X}$ is almost surely bounded by a counstant K, that is, there exists a constant $K > 0$ such that $\|\mathbf{X}\|_\infty \leq K$ a.s.;
- (A2) For all $\mathbf{x} \in [-K, K]^p$, $\beta_0^* + \beta^{*\top}\mathbf{x} \in Int(\Theta)$, where $Int(\Theta)$ denote the set of all interior point in Θ;
- (A3) the sample size n and the number of predictors p satisfy $\frac{\log(2p)}{n} \leq 1$.

Remark: Technically, Assumption (A1) is not a reasonable condition when the graphical structure among predictors is considered, since, in that case, it usually assume that $\mathbf{X}$ follows multivariate normal distribution in order to measure the graphical structure by precision matrix conveniently. However, in fact, Assumption (A1) may be extended to that $\mathbf{X}$ is bounded by $O(\sqrt{\frac{n}{\log(n)}})$ in the light of the method described in [32]. In this paper, for the sake of simplicity, we assume that $\mathbf{X}$ is bounded by a constant rather than a bound proportional to $\sqrt{\frac{n}{\log(n)}}$. Further, we note that Assumption (A1) is also a technique condition required in [3, 34].

The log-likelihood for GLMs is given by $\mathcal{L}(\beta_0,\beta)=\sum_{i=1}^n\{Y_i(\beta_0+\beta^\top\mathbf{X}_i)-\phi(\beta_0+\beta^\top\mathbf{X}_i)\}$. We denote the loss function for GLMs by $\ell(\beta_0,\beta):=\ell(\beta_0,\beta;\mathbf{x},y):=-y(\beta_0+\beta^\top\mathbf{x})+\phi(\beta_0+\beta^\top\mathbf{x})$. Notice that $\ell(\beta_0,\beta)$ is convex in β (as ϕ is convex). The associated risk is denoted by $\mathbb{P}\ell(\beta_0,\beta):=\mathbb{E}\ell(\beta_0,\beta;\mathbf{X},Y)$ and the empirical risk by $\mathbb{P}_n\ell(\beta_0,\beta):=\frac{1}{n}\sum_{i=1}^n\{-Y_i(\beta_0+\beta^\top\mathbf{X}_i)+\phi(\beta_0+\beta^\top\mathbf{X}_i)\}$. We consider

$$\Xi=\left\{(\beta_0,\beta)\in\mathbb{R}^{(p+1)}:\forall\mathbf{x}\in[-K,K]^p,\beta_0+\beta^\top\mathbf{x}\in\Theta\right\},$$

and it is obvious that $(\beta_0^*,\beta^*)=\arg\min_{(\beta_0,\beta)\in\Xi}\mathbb{P}\ell(\beta)$.

From the theorem 2.1 of [9] and the condition 3.1 of [19] we have

$$\begin{aligned}\mathbb{E}(\mathbf{X}|Y=y)&=\mathbb{E}[\mathbb{E}(\mathbf{X}|\beta^{*\top}\mathbf{X})|Y=y]=\mathbb{E}\left[\left\{\mu+\frac{\Sigma\beta^*\beta^{*\top}(\mathbf{X}-\mu)}{\beta^{*\top}\Sigma\beta^*}\right\}\Big|Y=y\right]\\&=\mu+\frac{\Sigma\beta^*\mathbb{E}[\beta^{*\top}(\mathbf{X}-\mu)|Y=y]}{\beta^{*\top}\Sigma\beta^*}\\&=\mu+\Sigma\beta^*k(y),\end{aligned}$$

where $\mu=\mathbb{E}(\mathbf{X}),\Sigma=\mathrm{Var}(\mathbf{X})$ and $k(y)=\frac{\mathbb{E}[\beta^{*\top}(\mathbf{X}-\mu)|Y=y]}{\beta^{*\top}\Sigma\beta^*}$. Let $\eta(y)=\mu+\Sigma\beta^*k(y)$ then

$$\beta^*\propto\Sigma^{-1}(\eta(y)-\mu).\tag{3}$$

Let $\Sigma^{-1}=\Omega=(\omega_{ij})_{i,j=1,2,\cdots,p}$, where Ω is the precision matrix which measures partial correlations among predictors, then by (3) we know that β^* can be reformulated as $\beta^*=\Omega\gamma$, thus we have

$$\begin{aligned}\beta_1^*&=\gamma_1\omega_{11}+\gamma_2\omega_{12}+\cdots+\gamma_j\omega_{1j}+\cdots+\gamma_p\omega_{1p}\\\beta_2^*&=\gamma_1\omega_{21}+\gamma_2\omega_{22}+\cdots+\gamma_j\omega_{2j}+\cdots+\gamma_p\omega_{2p}\\&\vdots\\\beta_p^*&=\gamma_1\omega_{p1}+\gamma_2\omega_{p2}+\cdots+\gamma_j\omega_{1j}+\cdots+\gamma_p\omega_{pp}.\end{aligned}\tag{4}$$

Notice that β^* can be formulated as the sum of p parts, $\{(\gamma_j\omega_{1j},\gamma_j\omega_{2j},\cdots,\gamma_j\omega_{pj})^\top:1\le j\le p\}$, by (4). For the jth part, $(\gamma_j\omega_{1j},\gamma_j\omega_{2j},\cdots,\gamma_j\omega_{pj})^\top$, there is a common factor γ_j. If γ_j is equal to 0, then all the components in the jth part of β^* will be 0 simultaneously. On the other hand, if γ_j is not zero and the graphical structure of predictors is defined by Ω, then the support of $(\gamma_j\omega_{1j},\gamma_j\omega_{2j},\cdots,\gamma_j\omega_{pj})^\top$ becomes the $\mathcal{N}_j$, which is the set including predictor j and its neighbors in the predictor graph. From the above analysis, we can conclude that it is reasonable to incorporate the graphical structure of the predictors into generalized linear models. Furthermore, the results in [41] and [14] indicate that it is indeed helpful to incorporate the graphical structure among the predictors in logistic regression. In this paper, we treat the neighbors of each predictor as a group, since the predictor graph can generally not be represented as some no-overlapping groups, these groups are overlapping. Therefore, we consider a latent decomposition of β^* into p parts based on the overlapping groups $\mathcal{N}_1,\mathcal{N}_2,\cdots,\mathcal{N}_p$. After choosing the non-zero candidate

component in each part according to $\mathcal{N}_1, \mathcal{N}_2, \cdots, \mathcal{N}_p$, we use the ordinary group Lasso penalty to encourage the selected components in each part to be zero or nonzero simultaneously.

The above idea can be extended to arbitrary graphs constructed by the priori information or estimated from the data. Now we suppose that the predictor graph G is given. We define a $p \times p$ adjacency matrix $E = (E_{ij})_{i,j=1,\cdots,p}$ where $E_{ij} = 1$ if the predictor i and j are connected and $E_{ij} = 0$ otherwise. We set $E_{jj} = 1$ for $j = 1, \cdots, p$, then the neighborhood $\mathcal{N}_j$ can be defined as $\mathcal{N}_j = \{k : E_{jk} = 1\}$. We assume that β^* can be decomposed into

$$\begin{aligned} \beta_1^* &= W_1^{(1)} E_{11} + W_1^{(2)} E_{12} + \cdots + W_1^{(j)} E_{1j} + \cdots + W_1^{(p)} E_{1p} \\ \beta_2^* &= W_2^{(1)} E_{21} + W_2^{(2)} E_{22} + \cdots + W_2^{(j)} E_{2j} + \cdots + W_2^{(p)} E_{2p} \\ &\vdots \\ \beta_p^* &= W_p^{(1)} E_{p1} + W_p^{(2)} E_{p2} + \cdots + W_p^{(j)} E_{pj} + \cdots + W_p^{(p)} E_{pp}. \end{aligned} \tag{5}$$

According to the definition of E_{ij}, the candidate nonzero components of the jth part, $(W_1^{(j)} E_{1j}, W_2^{(j)} E_{2j}, \cdots, W_p^{(j)} E_{pj})^\top$, are $\{W_k^{(j)} E_{kj} : k \in \mathcal{N}_j\}$. Note that the factor $\{W_k^{(j)} : k \in \mathcal{N}_j\}$ in each part can be viewed as the effect arising form the marginal correlation between the jth predictor and the response variable. If they are uncorrelated, $W_k^{(j)}$ will be zero for each $k \in \mathcal{N}_j$ and the components in the set $\{W_k^{(j)} E_{kj} : k \in \mathcal{N}_j\}$ will be zero together. Thus, it is reasonable to use the group Lasso penalty to encourage the selected components in each part to be zero or nonzero simultaneously if the the candidate nonzero components in each part have been selected based on $\mathcal{N}_1, \mathcal{N}_2, \cdots, \mathcal{N}_p$. Based on this idea which is motivated by [37], given the graph of the predictors and the training data $(\mathbf{X}_1, Y_1), \cdots, (\mathbf{X}_n, Y_n)$, we propose the following sparse generalized linear models incorporating graphical structure among predictors (sGLMg).

- Find the neighborhoods $\mathcal{N}_1, \mathcal{N}_2, \cdots, \mathcal{N}_p$ based on the predictor graph G (note that $j \in \mathcal{N}_j$, $j = 1, \cdots, p$)
- Solve the following optimization problem:

$$\min_{(\beta_0,\beta), W^{(1)}, \cdots, W^{(p)}} \mathbb{P}_n \ell(\beta_0, \beta) + \lambda \sum_{j=1}^{p} d_j \|W^{(j)}\|_2, \tag{6}$$

  subject to $\sum_{j=1}^{p} W^{(j)} = \beta$ and $\text{supp}(W^{(j)}) \subset \mathcal{N}_j, \forall j = 1, \cdots, p$, where $\text{supp}(W^{(j)})$ is the support of vector $W^{(j)}$ and $\|\cdot\|_2$ is the ℓ_2 norm.

Here, λ is a tuning parameter which can be determined by cross validation and d_j is the positive weight for the jth group and the choice of d_j will be discussed in section 4.

The optimization problem of (6) can be solved by the predictor duplication method proposed in [26]. More precisely, let $W_{\mathcal{N}_j}^{(j)}$ and $\mathbf{X}_{i\mathcal{N}_j}$ denote the $|\mathcal{N}_j| \times 1$ sub-vector of $W^{(j)}$ and the $|\mathcal{N}_j| \times 1$ sub-vector of $\mathbf{X}_i$ with indices in $\mathcal{N}_j$, respectively, where $i = 1, \cdots, n, j = 1, \cdots, p$. Let $\tilde{\mathbf{X}}_i = (\mathbf{X}_{i\mathcal{N}_1}^\top, \mathbf{X}_{i\mathcal{N}_2}^\top, \cdots, \mathbf{X}_{i\mathcal{N}_p}^\top)^\top$

and $\tilde{W} = ({W_{\mathcal{N}_1}^{(1)}}^\top, {W_{\mathcal{N}_2}^{(2)}}^\top, \cdots, {W_{\mathcal{N}_p}^{(p)}}^\top)^\top$, then, it is easy to verify that $\beta^\top \mathbf{X}_i = \tilde{W}^\top \tilde{\mathbf{X}}_i$. Therefore, the optimization problem (6) is equivalent to the following ordinary group Lasso problem:

$$\min_{(\beta_0, \tilde{W})} \frac{1}{n} \sum_{i=1}^{n} \left[-Y_i(\beta_0 + {\tilde{W}}^\top \tilde{\mathbf{X}}_i) + \phi(\beta_0 + {\tilde{W}}^\top \tilde{\mathbf{X}}_i) \right] + \lambda \sum_{j=1}^{p} d_j \|W_{\mathcal{N}_j}^{(j)}\|_2. \tag{7}$$

There are now a lot of efficient R packages, such as grpLasso [23], grpreg [4] and gglasso [36], can be used to solve the optimal problem (7). Recently, Zeng and Breheny [40] develop an R package called grpregOverlap based on grpreg, which can be used to solve the overlapping group Lasso directly.

By setting $\hat{W}_{\mathcal{N}_j^c}^{(j)} = 0$ for $j = 1, \cdots, p$, we can get $\hat{\beta} = \sum_{j=1}^{p} \hat{W}^{(j)}$. Notice that in some special graphical structures, there may exists some exactly same neighborhood. Then, the vector $\{W_{\mathcal{N}_j}^{(j)} : j \in F\}$ is indistinguishable and therefore the decomposition of β is not unique (i.e. $\{W^{(1)}, W^{(2)}, \cdots, W^{(p)}\}$ is not unique). In this case, the vector in $\{W_{\mathcal{N}_j}^{(j)} : j \in F\}$ can not be estimated stably, however, we can estimate $\sum_{j \in F} W_{\mathcal{N}_j}^{(j)}$ directly and stably using the penalty term $(\min_{j \in F} d_j) \| \sum_{j \in F} W_{\mathcal{N}_j}^{(j)} \|_2$. Because $\hat{\beta} = \sum_{j=1}^{p} \hat{W}^{(j)}$, different decompositions of β lead to the same estimation of β.

3. Theoretical properties of sGLMg

In this section we study the theoretical properties of the proposed sGLMg and the Oracle inequalities for the estimator of sGLMg will be presented in the finite sample setting. Given the predictor graph G and positive weights d_j, for $\beta \in \mathbb{R}^p$, we denote

$$\|\beta\|_{G,d} = \min_{\sum_{j=1}^{p} W^{(j)} = \beta, \operatorname{supp}(W^{(j)}) \subset \mathcal{N}_j} \sum_{j=1}^{p} d_j \|W^{(j)}\|_2. \tag{8}$$

Note that $\|\beta\|_{G,d}$ defined in (8) is similar to the latent group Lasso penalty defined in [26], however, it is very different in motivation between them since our proposed method is a graph based penalization problem. To give a geometric illustration of this norm we consider the case of $p = 4$. We consider that $\mathcal{N}_1 = \{1, 2, 3\}, \mathcal{N}_2 = \{1, 2\}, \mathcal{N}_3 = \{1, 3, 4\}$ and $\mathcal{N}_4 = \{3, 4\}$, then the norm $\|\beta\|_{G,d}$ we defined is a unit ball in $\mathbb{R}^4$ that has two circular sets of singularities corresponding to cases where (β_1, β_2) only or (β_3, β_4) only is nonzero and two spherical sets of singularities corresponding to cases where $(\beta_1, \beta_2, \beta_3)$ only or $(\beta_1, \beta_3, \beta_4)$ only is nonzero. The graph of this norm in $\mathbb{R}^3$ can refer to figure 2 of [26]. Thus, the minimum in (6) is equivalent to

$$\min_{(\beta_0, \beta) \in \mathbb{R}^{(p+1)}} \mathbb{P}_n \ell(\beta_0, \beta) + \lambda \|\beta\|_{G,d}. \tag{9}$$

Note that the optimal decomposition of β minimizing $\|\beta\|_{G,d}$ always exists, but may not be unique [26]. We introduce the following notations: denote $J^* = \{j :$

$\beta_j^* \neq 0\}$ as the true nonzero coefficient set and $J^{*c} = \{j : \beta_j^* = 0\}$ as the true zero coefficient set. Let $s^* = |J^*|$ denote the number of true nonzero coefficients. For each $\beta \in \mathbb{R}^p$, we denote $\mathfrak{W}(\beta)$ as the set of all optimal decompositions of β. Define $\mathcal{K}_G(\beta) = \min_{(W^{(1)},W^{(2)},\cdots,W^{(p)})\in\mathfrak{W}(\beta)} \left|\{j : \|W^{(j)}\|_2 \neq 0\}\right|$, which denotes the number of nonzero $W^{(j)}$ in the optimal decomposition of β that has the minimal number of nonzero $W^{(j)}$. Denote $\mathcal{K}_G = \sup_{\text{supp}(\beta)\subset J^*} \mathcal{K}_G(\beta)$. It is easy to check that $\mathcal{K}_G = s^*$ if the graph G has no edge, $\mathcal{K}_G = K_0$ if G consists of some disconnected complete subgraph and J^* is the union of K_0 nodes sets of those disconnected subgraph. Denote $N_{\max} = \max\{|\mathcal{N}_j| : j = 1, \cdots, p\}$ as the number of variables in the neighborhood which contains the maximum number of predictors. We make the following assumption for the neighborhood $\mathcal{N}_j$:

- (A4) For each $j \in J^*$, $\mathcal{N}_j \subset J^*$.

This condition assumes that predictors connected to the useful predictor are also useful.

3.1. The sub-gradient conditions for sGLMg

We introduce the following sub-gradient conditions for the problem (9).

Proposition 1. *A vector $\beta \in \mathbb{R}^p$ is a solution of (9) if and only if β can be decomposed as $\beta = \sum_{j=1}^p W^{(j)}$ where $W^{(j)}$ satisfy:*

- *(a) $W^{(j)}_{\mathcal{N}_j^c} = 0$;*
- *(b) either $W^{(j)}_{\mathcal{N}_j} \neq 0$ and $\nabla_{\mathcal{N}_j}\mathcal{L}(\beta) = n\lambda d_j \frac{W^{(j)}_{\mathcal{N}_j}}{\left\|W^{(j)}_{\mathcal{N}_j}\right\|_2}$, or $W^{(j)}_{\mathcal{N}_j} = 0$ and $d_j^{-1}\|\nabla_{\mathcal{N}_j}\mathcal{L}(\beta)\|_2 \leq n\lambda$, where $\nabla_{\mathcal{N}_j}\mathcal{L}(\beta) \in \mathbb{R}^{|\mathcal{N}_j|}$ denote as the gradient of $\mathcal{L}(\beta)$ with respect to the predictors in $\mathcal{N}_j, j = 1, \cdots, p$.*

The proof of Proposition 1 is similar to the Lemma 11 of [26]. The Proposition 1 shows that if $(\hat{W}^{(1)}, \hat{W}^{(2)}, \cdots, \hat{W}^{(p)})$ is a solution of the problem (6), then for each j, either $\hat{W}^{(j)} = 0_{p\times 1}$ or $\text{supp}(\hat{W}^{(j)}) = \mathcal{N}_j$. Thus, the estimate $\hat{\beta} = \sum_{j=1}^p \hat{W}^{(j)}$ acquired by our proposed sGLMg has the same decomposition as (5).

3.2. The connections between sGLMg and some existing methods

Some existing methods, such as the adaptive Lasso, group Lasso and ridge regression, can be included as special cases of our proposed sGLMg method when the given predictor graph has some special structures. The following proposition shows this connections.

Proposition 2.

- *(a) If the predictor graph has no edge, the proposed sGLMg method is identical to the adaptive Lasso mehtod for each tuning parameter λ;*

- *(b) If the predictor graph is composed of T disconnected complete subgraphs, our proposed sGLMg method is the same as the ordinary group Lasso method for each λ;*
- *(c) If the predictor graph is a complete graph, our proposed sGLMg method has the same nonzero solution set as the ridge regression, i.e. for each nonzero solution acquired by ridge regression (or sGLMg), sGLMg (or ridge regression) could acquired the same solution using a different tuning parameter.*

The proof of this proposition is parallel to [37]. The proposition 2 indicates that our proposed sGLMg method is much more general than Adaptive Lasso, Group Lasso and ridge regression and can deal with any arbitrary predictor graph structures.

3.3. The oracle inequalities for sGLMg

In this section we study the finite properties of the estimator of our proposed sGLMg method and present the oracle inequalities for estimation and prediction of sGLMg. For each $\beta \in \Xi$, we need to prove the concentration inequalities for the empirical process $\mathbb{P}_n\ell(\beta)$, i.e. we need to give an appropriate lower bonds on $(\mathbb{P}_n - \mathbb{P})(\ell(\hat{\beta}_n) - \ell(\beta^*))$. To do this, we decompose the empirical process into a linear part and a part which depends on the normalized parameter ϕ, i.e.

$$(\mathbb{P}_n - \mathbb{P})(\ell(\beta)) = (\mathbb{P}_n - \mathbb{P})(\ell_l(\beta)) + (\mathbb{P}_n - \mathbb{P})(\ell_\phi(\beta)),$$

where $\ell_l(\beta) := \ell_l(\beta, \mathbf{x}, y) = -y\beta^\top \mathbf{x}$ and $\ell_\phi(\beta) := \ell_\phi(\beta, \mathbf{x}) = \phi(\beta^\top \mathbf{x})$. In addition, we need to make the follow assumption for β^*:

- (A5) There exists a constant $b > 0$ such that $\|\beta^*\|_{G,d} \leq b$.

We define the following events

$$\mathfrak{A} = \left\{ \left\| \frac{1}{n} \sum_{i=1}^{n} \left[Y_i \mathbf{X}_{i\mathcal{N}_j} - \mathbb{E}(Y \mathbf{X}_{\mathcal{N}_j}) \right] \right\|_2 \leq \frac{\lambda d_j}{2},\ j = 1, \cdots, p \right\},$$

$$\mathfrak{B} = \left\{ \sup_{\beta : \|\beta - \beta^*\|_G \leq a} \left| \frac{(\mathbb{P}_n - \mathbb{P})(\ell_\phi(\beta^*) - \ell_\phi(\beta))}{\|\beta - \beta^*\|_G + \epsilon_n} \right| \leq \frac{\lambda}{2} \right\},$$

where $a = 8b + \epsilon_n, \epsilon_n = \frac{1}{n}$.

First, we want to show that the events $\mathfrak{A}$ and $\mathfrak{B}$ occur with high probability, i.e., we will give a lower bound for the probabilities of the events $\mathfrak{A}$ and $\mathfrak{B}$, which is equivalent to prove the concentration inequalities for the linear and nonlinear part of the empirical process.

Lemma 1. *Let $(\mathbf{X}, Y)$ be a pair of random variables whose conditional distribution is $P(Y; \beta^*|\mathbf{X} = \mathbf{x}) = \exp(y{\beta^*}^\top \mathbf{x} - \phi({\beta^*}^\top \mathbf{x}))$ and assume assumptions (A.1)-(A.3) are fulfilled. For all $l \in \mathbb{N}^*$ there exists a constant $C_{K,b}$ (which depends only on K and b) such that $\mathbb{E}(|Y|^l) \leq l!(C_{K,b})^l$.*

Note that to prove this Lemma we need to use the assumption (A.1), the details of the proof of this Lemma can refer to [3].

Theorem 1. *Let* $\lambda d_j \geq A^2 16 K C_{K,b} \frac{\log(2p)}{n} \vee A 8\sqrt{2} K C_{K,b} \sqrt{\frac{\log(2p)}{n}}$, *then*

$$\mathbb{P}(\mathfrak{A}) \geq 1 - 2N_{\max}(2p)^{1-A^2},$$

where $A > 1$ *and* $C_{K,b}$ *is the same as Lemma 1.*

To prove the concentration inequalities for the nonlinear part of the empirical process, we need to use the boundedness assumption for $\mathbf{X}$ and to show that we can restrict the study of ϕ to a suitable compact set. Since ϕ is Lipchitzian on this compact set, we can use the concentration results for Lipchitzian loss funcitons [18] to bound the probability of event $\mathfrak{B}$. Thus, a lower bound for the probability of event $\mathfrak{B}$ can be obtained.

Theorem 2. *Let* $d_{\min} = \min_{1 \leq j \leq p} d_j$ *and* $\delta_n = 17b + \frac{2}{n} = 2a + b$. *If*

$$\lambda d_{\min} \geq A 20 \sqrt{N_{\max}} K \alpha \max_{\left\{|x| \leq \frac{K\sqrt{N_{\max}}}{d_{\min}} \delta_n\right\} \cap \Theta} \{|\phi'(x)|\} \sqrt{\frac{2\log(2p)}{n}},$$

where $A \geq 1$, *then there exists a constant* C *such that*

$$\mathbb{P}(\mathfrak{B}) \geq 1 - C(2p)^{-\frac{A^2}{2}}.$$

After obtaining the lower bonds for the probability of events $\mathfrak{A}$ and $\mathfrak{B}$, the following corollary can be easily inferred.

Corollary 1. *If* $\lambda d_j \geq AK\mu \left\{C_{K,b} \vee \max_{\{|x| \leq K\delta_n\} \cap \Theta} |\phi'(x)|\right\} \sqrt{\frac{2\log(2p)}{n}}$ *then*

$$\mathbb{P}(\mathfrak{A} \cap \mathfrak{B}) \geq 1 - (2N_{\max} + C)(2p)^{-\frac{A^2}{2}},$$

where μ *and* C *are universal constants and* $A \geq \sqrt{2}$. *The definition of* $C_{K,b}$ *is the same as Lemma 1.*

Thus, according to the Theorem 1 and Corollary 1 we can deduce the upper bounds for the linear part and nonlinear part of the empirical process, i.e., $\|(\mathbb{P}_n - \mathbb{P})(\ell_l(\beta^*) - \ell_l(\hat{\beta}_n))\|_2$ and $(\mathbb{P}_n - \mathbb{P})(\ell_\phi(\beta^*) - \ell_\phi(\hat{\beta}_n))$, respectively.

Theorem 3. *On the event* $\mathfrak{A}$,

$$(\mathbb{P}_n - \mathbb{P})(\ell_l(\beta^*) - \ell_l(\hat{\beta}_n)) \leq \frac{\lambda}{2} \|\hat{\beta}_n - \beta^*\|_{G,d}.$$

Theorem 3 shows that the difference between the linear part of the empirical process and its expectation is bounded above by the tuning parameter multiplied by the norm (defined by (8)) of the difference between the estimator of sGLMg and the true parameter. Note that the norm defined by (8) is associated to the predictor graph. A similar result for the nonlinear part of the empirical process can be also stated, the key of the proof is based on the following lemma which show that the estimator of sGLMg, $\hat{\beta}_n$, is in the neighborhood of the target parameter β^* on the event $\mathfrak{A} \cap \mathfrak{B}$.

Lemma 2. *On the event* $\mathfrak{A} \cap \mathfrak{B}$*, we have* $\|\hat{\beta}_n - \beta^*\|_{G,d} \leq a$ *where* $a = 8b + \epsilon_n$ *and* $\epsilon_n = \frac{1}{n}$.

An upper bound for $(\mathbb{P}_n - \mathbb{P})(\ell_\phi(\beta^*) - \ell_\phi(\hat{\beta}_n))$ can be directly obtained based on the definition of the event $\mathfrak{B}$ and the Lemma 2.

Theorem 4. *On the event* $\mathfrak{A} \cap \mathfrak{B}$ *we have*

$$(\mathbb{P}_n - \mathbb{P})(\ell_\phi(\beta^*) - \ell_\phi(\hat{\beta}_n)) \leq \frac{\lambda}{2}\left(\|\hat{\beta}_n - \beta^*\|_{G,d} + \epsilon_n\right).$$

According to the restricted strong convexity condition for M-estimators in [25], we need to ensure that the the loss function is not too flat after stating the concentration of the loss function around its mean, i.e., there exists $\varepsilon > 0$, when $|\ell(\hat{\beta}_n) - \ell(\beta^*)| \leq \varepsilon$, the corresponding estimation error satisfies $|\hat{\beta}_n - \beta^*| \leq \varepsilon$. Notice that the boundedness assumption on the components of $\mathbf{X}$ is not required to obtain such kind of strong convexity, however, we need it to establish Theorem 1. As stated by [25], if the tail of the covariates is sub-gaussian and the covariance matrix is positive definite then the loss function satisfies a kind of restricted strong convexity property with high probability. Therefore, the primary condition to prove the oracle inequalities for the estimator of sGLMg rests on the correlation between the covariates, i.e., on the behaviour of the Gram matrix $\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top$ which is necessarily singular when $p > n$. Meire et al. [23] show that the group lasso is consistent under the logistic regression model and give an upper bound for the prediction error under the assumption that $\mathbb{E}(\mathbf{X}\mathbf{X}^\top)$ is nonsingular. Blazere et al. [3] present the oracle inequalities for the estimation and prediction error of the generalized linear models under the group satbil condition which is similar to the restricted eigenvalue conditions in [24] and [21]. However, the group structure in [3] must be non-overlapping and specified in advance. The same stabil conditions are used by [27] and [37] who proved the theoretical properties of overlapping group Lasso and of linear regression model incorporating the graphical structure among predictors, respectively. In this paper, we will present the oracle inequalities for the estimation and prediction error of our proposed sGLMg under the similar conditions as we discussed above.

For a given graph G, positive weights d_j's and subset $J \subset \{1, 2, \cdots, p\}$, denote $\Omega(\beta, J)$ as the set of all optimal decomposition of β such that

$$\sum_{j \in J^c} d_j \|W^{(j)}\|_2 \leq 3 \sum_{j \in J} d_j \|W^{(j)}\|_2 + \epsilon,$$

for all $\epsilon > 0$. Denote $\Sigma := \mathbb{E}(\mathbf{X}\mathbf{X}^\top)$ as $p \times p$ covariance matrix and consider the following assumption:

- (A6) Let $\beta \in \mathbb{R}^p \setminus \{0\}, |J| \leq s^*$, then for all $(W^{(1)}, W^{(2)}, \cdots, W^{(p)}) \in \Omega(\beta, J)$ there exists $\epsilon > 0$ and $0 < \kappa < 1$ such that

$$\beta^\top \Sigma \beta \geq \kappa \sum_{j \in J} d_j^2 \|W^{(j)}\|_2^2 - \epsilon.$$

Note that assumption (A6) plays a key role in the proof of the oracle inequalities for the estimator of sGLMg. This assumption is similar to the restricted strong convexity in [25], which is considered as a key condition for ensuring the fast convergence rates and well theoretical properties of the regularized M-estimators in high dimension scaling.

The next theorem is the most important result in this paper, which present the finite sample bounds for estimation and prediction of the estimator of our proposed sGLMg.

Theorem 5. *Suppose that assumptions (A1)-(A6) are satisfied. Let* $N_{\max} = \max_{1\le j\le p}\mathcal{N}_j$ *and* $d_{\min} = \min_{1\le j\le p} d_j$. *If we choose*

$$\lambda d_{\min} \ge AK\mu\left\{C_{K,b} \vee \max_{\{|x|\le K\delta_n\}\cap\Theta}|\phi'(x)|\right\}\sqrt{\frac{2\log(2p)}{n}},$$

where μ *is the universal constant,* $A \ge \sqrt{2}$ *and the definition of* $C_{K,b}$ *is similar as Lemma 1, then, for any optimal solution* $\hat{\beta}_n$ *of problem (9), we have*

$$\|\hat{\beta}_n - \beta^*\|_{G,d} \le \frac{4\lambda\mathcal{K}_G}{c_n\kappa} + (1+\frac{1}{\lambda})\frac{\epsilon_n}{2},$$
$$\|\hat{\beta}_n - \beta^*\|_2 \le \frac{4\lambda\mathcal{K}_G}{c_n\kappa d_{\min}} + (1+\frac{1}{\lambda})\frac{\epsilon_n}{2d_{\min}},$$
$$\mathbb{E}(\hat{\beta}_n^\top\mathbf{X} - \beta^{*\top}\mathbf{X})^2 \le \frac{12}{c_n^2\kappa}\lambda^2\mathcal{K}_G + \frac{(3+4\lambda)}{c_n}\frac{\epsilon_n}{2},$$

with probability at least $1-(2N_{\max}+C)(2p)^{-\frac{A^2}{2}}$.

Where $c_n = \max_{\left\{|x|\le\frac{K\sqrt{N_{\max}}}{d_{\min}}(\alpha+b)\right\}\cap\Theta}|\phi''(x)|$ *and* C *is the universal constant.*

The results presented in Theorem 5 are very general for some existing results have close connections with it if the predictor graph has some special structure. For example, the oracle inequalities for the prediction and estimation error of GLMs obtained by group Lasso and Lasso methods, respectively, in [3] are special cases of Theorem 5. In fact, when the given graph G has no edge, we have $\mathcal{K}_G = s^*$ and $\|\hat{\beta}_n - \beta^*\|_{G,d} = \|\hat{\beta}_n - \beta^*\|_1$ if $d_j = 1$ for $j = 1, \cdots, p$. Theorem 5 indicates that the same results of the estimation and prediction error in [3] for the Lasso method can be re-derived (Theorem III.8). When the predictor graph G consists of some disconnected complete subgraphs and J^* is the union of K_0 node sets of those disconnected subgraphs, we have $\mathcal{K}_G = K_0$. In this setting, the results presented in [3] for the group Lasso can be also recovered (Theorem III.6). In addition, the results about the linear regression model in [2], [24], [21] and [37] are also connected with the result shown in Theorem 5.

Notice that if $\mathcal{K}_G = O(1)$ then the bond of the estimation error in Theorem 5 is of the order $O\left(\sqrt{\frac{\log p}{n}}\right)$ and the sGLMg estimator $\hat{\beta}_n$ still remains consistent for the estimation error $\|\hat{\beta}_n - \beta^*\|_{G,d}$ and prediction error $\mathbb{E}(\hat{\beta}_n^\top\mathbf{X} - \beta^{*\top}\mathbf{X})^2$

under the key assumption (A6) if the number of predictors p increases almost as fast as $O(\exp(n))$. Under the similar conditions, the estimation error for the linear model in [37] is of the order $O(\exp(n))$. Compare to this, the term $\sqrt{\log(p)}$ in the GLMs is the price to pay for having a large number factors and not knowing where are the nonzero ones.

4. Model selection consistency

In this section we discuss model selection consistency for the case with a fixed dimension p. For every $\beta \in \mathbb{R}^p$, denote β_{J^*} and $\beta_{J^{*c}}$ as sub-vectors of β with indices in J^* and J^{*c} respectively.

Theorem 6. *Assume assumption (A2) and (A4) hold. Suppose the tuning parameter λ and d_i are chosen such that $\sqrt{n}\lambda \to 0$ and $n^{(\gamma+1)/2}\lambda \to \infty$ for some $\gamma > 0$. Furthermore, $d_j = O(1)$ for each $j \in J^*$ and $\liminf_{n\to\infty} n^{-\gamma/2} d_j > 0$ for each $j \in J^{*c}$. Then, with dimension p fixed, as $n \to \infty$, we have*

$$\sqrt{n}(\hat{\beta}_{J^*} - \beta^*_{J^*}) \xrightarrow{d} N(0, I^{-1}_{J^*,J^*}(\beta))$$

and

$$\hat{\beta}_{J^{*c}} \xrightarrow{p} 0,$$

where $I_{J^*,J^*}(\beta)$ is the sub-matrix of $I(\beta)$ consisting of the entries with row and column indices in J^* and $I(\beta)$ is the Fisher information matrix of the model.

Note that Theorem 6 shows that our proposed sGLMg method is model selection consistent for the fixed p case. It also provide a guideline on how to choose the positive weight d_j. In fact, the choice of weights of overlapping groups is much more important and complicated than in the case of disjoint groups. Obozinski et al. [26] have made a thorough discussion on the choice of weights for the overlapping groups and proposed some guidelines on the choice of weights. They suggest to consider weights of the form $d_j = m_j^\gamma$, where $m_j = |\mathcal{N}_j|$ is the number of predictors in the neighborhood $\mathcal{N}_j$ and $\gamma \in (0, \frac{1}{2})$. And, $\gamma = 0$ and $\gamma = \frac{1}{2}$ correspond to two extreme cases that only the largest and only the smallest groups are active, respectively. Furthermore, they give a critical value with $\gamma = \frac{\log(2)}{2\log(3)}$, which is the smallest value that it is possible to select two singleton only. In our simulation studies, we suggest to choose $d_j = m_j^\gamma, \gamma = \frac{\log(2)}{2\log(3)}$.

5. Simulation study

We consider the Logistic model. In order to examine the performance of our proposed sGLMg, we compare it with some popular penalized methods such as Lasso, ridge regression, adaptive Lasso and elastic net. In the simulation, the predictor graph is defined by the precision matrix of the predictors. The performance of sGLMg using both the estimated predictor graph and the oracle

true predictor graph are evaluated on all examples. We denote sGLMg-O as the sGLMg method using the true predictor graph.

The response Y of Logistic regression is generated by $Y \in \{0,1\}$ and $P(Y = 1|\mathbf{X}) = \frac{\exp(\mathbf{X}\beta^*)}{1+\exp(\mathbf{X}\beta^*)}$. We divide the data set $\mathbf{X}$ into three separate subsets: a training data set, a validation data set and a testing data set. All the models are fitted on the training data set only. The validation data set are used to choose the tuning parameter and the test data set is used to evaluate different methods. We use the notation ./././ to show the sample size in the training, validation and test sets, respectively. For each example, we consider three cases: (A) 40/40/400, (B) 80/80/400 and (C) 120/120/400. For each case, we repeat the simulation 100 times. The predictor graph is estimated by the graphical Lasso method [11] only using the training data in all cases.

Example 1 (Ω is block diagonal)**.** $p = 100$, $s^* = 15$, and the true coefficient vector $\beta^* = (0.3, 0.3, \cdots, 0.3, 0, 0, \cdots, 0)^\top$. The predictors are generated as:

$$\begin{aligned} X_j &= Z_1 + 0.4\varepsilon_j, \quad Z_1 \sim N(0,1), \quad 1 \le j \le 5; \\ X_j &= Z_2 + 0.4\varepsilon_j, \quad Z_2 \sim N(0,1), \quad 6 \le j \le 10; \\ X_j &= Z_3 + 0.4\varepsilon_j, \quad Z_3 \sim N(0,1), \quad 11 \le j \le 15; \\ X_j &\overset{i.i.d}{\sim} N(0,1), \quad 16 \le j \le 100, \end{aligned}$$

where $\varepsilon_j \overset{i.i.d}{\sim} N(0,1), j = 1, 2, \cdots, 15$.

Example 2 (Ω is banded)**.** $p = 100$ and β^* is the same as β^* used in Example 1. The predictors $(X_1, X_2, \cdots, X_p)^\top \sim N(0, \Sigma)$, with $\Sigma_{ij} = 0.5^{|i-j|}$. For this example, we have $\omega_{ii} = 1.333$, $\omega_{ij} = -0.677$, if$|i-j| = 1$ and $\omega_{ij} = 0$, if$|i-j| > 1$.

Example 3 (Ω is sparse)**.** $p = 100$ and the predictors $(X_1, X_2, \cdots, X_p)^\top \sim N(0, \Omega^{-1})$ where $\Omega^{-1} = B + \delta I$. Each off-diagonal entry in B is generated independently and equals to 0.5 with probability 0.05, or 0 with probability 0.95. The diaonal entry of B is 0. Here, δ is chosen such that the conditional number of Ω is equal to p. Finally, Ω is standardized to have unit diagonals. We set $\beta^* = \Omega\gamma$, where $\gamma = (\gamma_1, \gamma_2, \cdots, \gamma_p)^\top$ with $\gamma_i = 0.1, i = 1, 2, 3, 4$ and $\gamma_i = 0$ otherwise.

To evaluate the different methods, we use the following measures:

- ℓ_2 distance: $\|\hat{\beta} - \beta^*\|_2$;
- Prediction error: $\frac{1}{N_{test}}(\hat{\beta} - \beta^*)^\top X_{test}^\top X_{test}(\hat{\beta} - \beta^*)$, where X_{test} is the test samples and N_{test} is the number of test samples;
- False Positive Rate (FPR: the rate of irrelevant variables incorrectly identified as relevant) and False Negaive Rate (FNR: the rate of relevant variables incorrectly identified as irrelevant);
- Nonzero match ratio:

$$\text{NMR} = \frac{|\{(i,j) : \Omega_{ij} \neq 0, \hat{\beta}_i \neq 0, \hat{\beta}_j \neq 0\}|}{|\{(i,j) : \Omega_{ij} \neq 0, \beta_i^* \neq 0, \beta_j^* \neq 0\}|};$$

Zero match ratio:

$$\text{ZMR} = \frac{|\{(i,j) : \Omega_{ij} \neq 0, \hat{\beta}_i = 0, \hat{\beta}_j = 0\}|}{|\{(i,j) : \Omega_{ij} \neq 0, \beta_i^* = 0, \beta_j^* = 0\}|},$$

where NMR (or ZMR) is used to check whether the estimated coefficients of two connected useful (or useless) predictors are both nonzero (or zero). Note that we use NMR and ZMR when there is at least one edge connecting two useful predictors and one edge connecting two useless predictors. Thus, these two ratios are well defined and always between 0 and 1.

TABLE 1
Performance comparison of estimation and prediction for Example 1

Methods	ℓ_2 distance			Prediction error		
	(A)	(B)	(C)	(A)	(B)	(C)
Lasso	2.211(0.109)	1.989(0.112)	1.977(0.171)	6.127(0.464)	4.489(0.536)	3.810(0.950)
Ridge	1.029(0.002)	0.999(0.001)	0.972(0.039)	5.291(0.045)	5.081(0.042)	2.535(0.102)
Alasso	2.557(0.142)	2.129(0.118)	2.013(0.245)	8.336(0.790)	4.866(0.631)	3.449(0.519)
Enet	1.756(0.067)	1.614(0.081)	1.543(0.103)	4.715(0.232)	3.876(0.313)	3.203(0.147)
sGLMg-O	1.091(0.041)	0.933(0.036)	0.836(0.026)	2.252(0.111)	1.297(0.060)	0.924(0.039)
sGLMg	1.106(0.044)	0.935(0.036)	0.836(0.026)	2.322(0.123)	1.298(0.060)	0.924(0.039)

TABLE 2
Performance comparison of model selection for Example 1

Methods	FPR			FNR		
	(A)	(B)	(C)	(A)	(B)	(C)
Lasso	0.151(0.007)	0.228(0.009)	0.281(0.013)	0.679(0.010)	0.542(0.010)	0.468(0.010)
Ridge	1.000(0.000)	1.000(0.000)	1.000(0.000)	0.000(0.000)	0.000(0.000)	0.000(0.000)
ALasso	0.091(0.005)	0.113(0.006)	0.194(0.009)	0.723(0.012)	0.554(0.010)	0.532(0.010)
ENet	0.301(0.011)	0.365(0.014)	0.420(0.014)	0.317(0.017)	0.234(0.011)	0.191(0.010)
SGLMG-O	0.097(0.006)	0.129(0.007)	0.155(0.007)	0.020(0.008)	0.000(0.000)	0.000(0.000)
SGLMG	0.102(0.006)	0.130(0.007)	0.155(0.007)	0.048(0.011)	0.001(0.001)	0.000(0.000)

Figure 2, 3 and 4 in Appendix A show the true predictor graphs (defined by Ω) of these three examples. The numbers of edges for these three graphs are 30, 99 and 256. Such graphs are also studied in Yang et al. (2012), [6] and [37]. Table 1 and Table 2 show the performance comparison of estimation, prediction and model selection of Example 1. The comparison results indicate that the ridge regression method obtains the better estimation than Lasso, adpative Lasso and Elastic net methods, however, the ridge regression method can not select the predictors automatically. Comparing with Lasso, adaptive Lasso and ridge regression methods, the Elastic net method acquires the overall optimal estimation, prediction and model selection by using the combination of ℓ_1 and ridge penalty. Specifically, the estimation acquired by elastic net method is almost the same with ridge regression and the prediction obtained by elastic net method is better than Lasso and adaptive Lasso. Although the elastic net method has relatively high FPR than Lasso and adaptive Lasso, the FNR of elastic net is the smallest of the three. Compared with the other methods our proposed sGLMg method delivers the best performance of estimation and prediction. For the cases with smaller sample sizes (condition A and B), our proposed sGLMg (sGLMg-O) method has slightly higher FPR than adaptive Lasso method, however, with the increase of the sample size (C), our proposed sGLMg (sGLMg-O) method acquire the lowest FPR than the other methods compared

with it. Especially, the FNR obtained by our proposed sGLMg (sGLMg-O) method is much lower than the other methods. The reason is that our proposed sGLMg (sGLMg-O) method using the information of the predictor graph and the predictors connected in the graph has much more chances to be selected or removed simultaneously.

When the signal strength is weak, the sGLMg (sGLMg-O) method tend to select more predictors as the significant predictors. If the signal strength becomes stronger, the FPR of our proposed sGLMg (sGLMg-O) method will be decreased gradually even smaller than the adaptive Lasso method. The results of Table 10 in Appendix A indicates that the the performance of the model selection of our proposed sGLMg (sGLMg-O) method may be achieved the optimal results compared with the other methods when the signal is increased properly. For this example, since the estimated predictor graph is almost the same as the true predictor graph, the performance of sGLMg method is similar to sGLMg-O method.

TABLE 3
Performance comparison of estimation and prediction for Example 2

Methods	ℓ_2 distance			Prediction error		
	(A)	(B)	(C)	(A)	(B)	(C)
Lasso	2.175(0.125)	2.075(0.140)	2.003(0.229)	6.751(0.684)	6.219(0.894)	6.010(0.446)
Ridge	1.048(0.002)	1.017(0.002)	1.421(0.042)	2.911(0.025)	2.751(0.024)	2.363(0.122)
ALasso	2.310(0.127)	2.158(0.124)	2.439(0.296)	7.372(0.749)	6.437(0.877)	6.105(0.673)
ENet	1.829(0.077)	1.932(0.101)	1.763(0.162)	4.717(0.324)	4.969(0.486)	4.090(0.446)
SGLMG-O	1.468(0.050)	1.363(0.053)	1.268(0.055)	3.325(0.173)	2.477(0.150)	2.075(0.161)
SGLMG	1.463(0.051)	1.474(0.056)	1.301(0.050)	3.292(0.160)	2.725(0.161)	2.004(0.127)

TABLE 4
Performance of model selection for Example 2

Methods	FPR			FNR		
	(A)	(B)	(C)	(A)	(B)	(C)
Lasso	0.138(0.006)	0.228(0.010)	0.294(0.013)	0.636(0.013)	0.426(0.011)	0.310(0.010)
Ridge	1.000(0.000)	1.000(0.000)	1.000(0.000)	0.000(0.000)	0.000(0.000)	0.000(0.000)
ALasso	0.083(0.004)	0.136(0.006)	0.195(0.010)	0.677(0.013)	0.453(0.011)	0.395(0.011)
ENet	0.276(0.012)	0.381(0.014)	0.408(0.016)	0.437(0.016)	0.284(0.011)	0.193(0.010)
SGLMG-O	0.292(0.013)	0.416(0.015)	0.484(0.016)	0.354(0.018)	0.168(0.011)	0.109(0.007)
SGLMG	0.272(0.011)	0.368(0.013)	0.437(0.014)	0.436(0.019)	0.282(0.012)	0.187(0.009)

The performance comparison for Example 2 is displayed in Table 3 and Table 4. As Example 1, the ridge regression method has better performance of estimation and prediction than the other methods when the sample size is relative small. However, our proposed sGLMg (sGLMg-O) method may acquire better performance of estimation or prediction than the ridge regression for the relative large sample size. For example, for the case C, the sGLMg (sGLMg-O) method obtains better performance of estimation than the ridge regression method; for the cases B and C, the sGLMg (sGLMg-O) method acquires better performance of prediction than the ridge regression method.

Compared with Lasso, adaptive Lasso and Elastic net methods, the adaptive Lasso and Elastic net method acquire the lowest FPR and FNR respectively. Our proposed sGLMgsGLMg (sGLMg-O) method has the lowest FNR in the performance of model selection than the other methods, although the FPR obtained by sGLMg (sGLMg-O) is little higher than the other methods, which is mainly because of the weak signal strength. The results of Table 11

in Appendix A shows that the performance of model selection of our proposed sGLMg (sGLMg-O) method may be significantly improved if the signal intensity is increased, especially the FNR is almost 0 when the sample size is relative large. However, the other methods can benefit a little from increasing the signal strength. For example, the performance of FPR obtained by the Elastic net method will be worse when the signal strength is increased. Overall, the performance of sGLMg-O method is better than the sGLMg method in estimation, prediction and model selection.

TABLE 5
Performance comparison of estimation and prediction for Example 3

Methods	ℓ_2 distance			Prediction error		
	(A)	(B)	(C)	(A)	(B)	(C)
Lasso	0.771(0.089)	0.517(0.088)	0.382(0.032)	2.246(0.544)	1.449(0.698)	0.297(0.081)
Ridge	0.271(0.003)	0.257(0.002)	0.237(0.020)	0.108(0.007)	0.075(0.005)	0.018(0.043)
ALasso	0.971(0.127)	0.475(0.067)	0.435(0.062)	4.309(1.027)	1.019(0.443)	0.653(0.368)
ENet	0.644(0.067)	0.496(0.080)	0.438(0.125)	1.603(0.379)	1.282(0.516)	1.103(0.371)
SGLMG-O	0.478(0.044)	0.368(0.028)	0.339(0.032)	0.861(0.203)	0.337(0.084)	0.316(0.077)
SGLMG	0.606(0.065)	0.428(0.046)	0.369(0.041)	1.539(0.341)	0.552(0.206)	0.389(0.160)

TABLE 6
Performance comparison of model selection for Example 3

Methods	FPR			FNR		
	(A)	(B)	(C)	(A)	(B)	(C)
Lasso	0.086(0.009)	0.074(0.013)	0.073(0.013)	0.915(0.012)	0.921(0.013)	0.910(0.014)
Ridge	1.000(0.000)	1.000(0.000)	1.000(0.000)	0.000(0.000)	0.000(0.000)	0.000(0.000)
ALasso	0.059(0.007)	0.039(0.008)	0.035(0.010)	0.953(0.007)	0.949(0.009)	0.937(0.012)
ENet	0.137(0.015)	0.093(0.017)	0.088(0.019)	0.864(0.018)	0.897(0.019)	0.870(0.020)
SGLMG-O	0.250(0.023)	0.212(0.027)	0.208(0.030)	0.757(0.026)	0.805(0.026)	0.743(0.030)
SGLMG	0.398(0.033)	0.315(0.033)	0.307(0.031)	0.639(0.030)	0.733(0.029)	0.720(0.027)

Table 5 and Table 6 display the results for Example 3. Our proposed sGLMg (sGLMg-O) method delivers the best performance of estimation and prediction compared with the other methods (not including the ridge regression method). Note that all the methods here are not good at the performance of model selection, especially the FNR acquired by these methods is too high, however, our proposed sGLMg (sGLMg-O) method still has the lowest FNR than the other methods. This results indicates that it is more difficult to do model selection for generalized linear models than for linear regression model when the predictor graph is complicated. The reason may be that the generalized linear models have larger number of unknowing factors than the linear regression factors. As the previous two examples, our proposed sGLMg methods has the overall optimal results for both estimation, prediction and model selection compared with the other methods.

The comparison results on NMR and ZMR for the cases with sample sizes 40/40/400, 80/80/400 and 120/120/400 are shown in Table 7, 8 and 9. Compared with the other methods, our proposed sGLMg-O method acquires the best performance in NMR and competitive performance in ZMR. From the above analysis, the reason for that the sGLMg-O method acquires the lower NMR than the other methods may be that our proposed sGLMg method tend to select more predictors as significant predictors, which lesson the zero match ratio of the predictors. Our proposed sGLMg method acquires very competitive performance in NMR and ZMR when the predictor graph is not very complicated (Example 1 and 2). However, even in these cases the NMR acquired by

TABLE 7
Performance comparison of NMR and ZMR (Sample size: 40/40/400)

Methods	NMR			ZMR		
	Example 1	Example 2	Example 3	Example 1	Example 2	Example 3
Lasso	0.083(0.007)	0.130(0.011)	0.016(0.004)	-	0.817(0.014)	0.847(0.016)
Ridge	1.000(0.000)	1.000(0.000)	1.000(0.000)	-	0.000(0.000)	0.000(0.000)
ALasso	0.068(0.011)	0.120(0.014)	0.003(0.001)	-	0.925(0.011)	0.902(0.012)
ENet	0.501(0.022)	0.356(0.019)	0.042(0.009)	-	0.476(0.020)	0.777(0.023)
SGLMG-O	0.980(0.008)	0.545(0.017)	0.140(0.020)	-	0.671(0.017)	0.614(0.034)
SGLMG	0.942(0.012)	0.427(0.019)	0.237(0.024)	-	0.676(0.018)	0.467(0.041)

TABLE 8
Performance comparison of NMR and ZMR (Sample size: 80/80/400)

Methods	NMR			ZMR		
	Example 1	Example 2	Example 3	Example 1	Example 2	Example 3
Lasso	0.184(0.009)	0.315(0.016)	0.017(0.004)	-	0.629(0.016)	0.880(0.020)
Ridge	1.000(0.000)	1.000(0.000)	1.000(0.000)	-	0.000(0.000)	0.000(0.000)
ALasso	0.176(0.010)	0.295(0.016)	0.009(0.003)	-	0.784(0.013)	0.939(0.013)
ENet	0.582(0.017)	0.524(0.017)	0.040(0.012)	-	0.416(0.020)	0.858(0.023)
SGLMG-O	1.000(0.000)	0.749(0.015)	0.114(0.020)	-	0.499(0.018)	0.694(0.035)
SGLMG	0.997(0.002)	0.565(0.017)	0.202(0.025)	-	0.536(0.018)	0.577(0.043)

TABLE 9
Performance comparison of NMR and ZMR (Sample size: 120/120/400)

Methods	NMR			ZMR		
	Example 1	Example 2	Example 3	Example 1	Example 2	Example 3
Lasso	0.258(0.011)	0.460(0.017)	0.028(0.008)	-	0.521(0.018)	0.864(0.020)
Ridge	1.000(0.000)	1.000(0.000)	1.000(0.000)	-	0.000(0.000)	0.000(0.000)
ALasso	0.192(0.010)	0.348(0.015)	0.019(0.006)	-	0.680(0.017)	0.912(0.016)
ENet	0.648(0.016)	0.660(0.016)	0.055(0.016)	-	0.381(0.018)	0.822(0.026)
SGLMG-O	1.000(0.000)	0.825(0.010)	0.168(0.027)	-	0.417(0.018)	0.610(0.036)
SGLMG	1.000(0.000)	0.681(0.015)	0.215(0.023)	-	0.444(0.018)	0.533(0.040)

Lasso and adaptive Lasso is almost 0, which because that the Lasso method tend to select only some predictors from the highly correlated predictors. When the predictor graph is complicated (Example 3), the NMR acquired by all these methods is not very well, however, our proposed sGLMg method is still the best. The NMR's and ZMR's of sGLMg-O indicate that our proposed sGLMg method incorporate the predictor graph information to group the predictors and make use of most edges between useful and useless predictors efficiently. Therefore, our proposed sGLMg method can choose those connected useful predictors simultaneously and exclude those connected useless predictors jointly.

In conclusion, the simulation results indicate that when the group structure is unknown our proposed sGLMg method can make use of the structure information among predictors to group the predictors efficiently and performs well for both estimation, prediction and model selection.

6. Application

In this section, we consider a real example to compare the performance of our proposed sGLMg method with Lasso, adaptive Lasso, ridge regression and Elastic net. The breast cancer data consists of 22,283 gene expression levels of 133 subjects, including 34 subjects with pathological complete response (pCR) and 99 subjects with residual disease (RD). The dataset were analysed by [12] and are available at http://bioinformatics.mdanderson.org/pubdata.html. The pCR is defined as no evidence of viable, invasive tumor cells left in surgical specimen, which has been considered to have a high chance of can-

cer free survival in the long term, justifying its use as a surrogate marker of chemosensitivity [17]. Thus, it is of considerable interest to study the response states of the patients (pCR or RD) to neoadjuvant (preoperative) chemotherapy. [10] and [6] apply linear discriminant analysis to predict wether or not a subject can achieve the pCR state by estimating the inverse covariance matrix (or precision matix) of the gene expression levels. In this paper, we follow the same analysis scheme used by [10] and [6] to estimate the precision matrix and then compare the performance of our proposed sGLMg method with Lasso, adaptive Lasso, ridge regression and Elastic net based on the estimated precision matrix.

To estimate the precision matrix, we randomly divide the data into the training and testing sets of sizes 112 and 21, respectively, and repeat the whole process 100 times. A stratified sampling method is used in order to maintain a similar class proportion for the training and testing datasets. We randomly select 5 pCR subjects and 16 RD subjects each time from the corresponding groups to form the the testing data (both are roughly 1/6 of the subjects in each group) and the remaining subjects will be used to constitute the the training set. For each training set, a two-sample t test is performed between the two groups and the most significant 113 genes that have the smallest p-values are selected as the predictors for prediction. We note that the training sample size $n = 112$ is slightly smaller than the variable dimensionality $p = 113$, which allows us to examine the performance when $p > n$. Then, a gene-wise standardization is performed by dividing the data with the corresponding standard deviation, estimated from the training dataset. Finally, we estimate the precision matrix Ω using the training data and the predictor graph G is estimated by the graphical Lasso [11] based on the estimated precision matrix. Note that all the models are fitted using training data and evaluated by the mean squared error (MSE) calculated from the testing data. We perform a 10-fold CV to choose the tuning parameters of different methods. Figure 1 shows the box plot of the averaged mean squared errors of different methods. The results indicates that our proposed sGLMg method acquires better performance on MSE score than Lasso, adaptive Lasso and Elastic net methods and slightly worse than the ridge regression method, which consistent with the simulation results.

7. Conclusion

In this paper, we propose the sparse generalized linear models incorporating graphical structure among predictors (sGLMg) which can be used to analysis the sparse graphical or overlapping structure data with generalized linear models. For sGLMg model, the overlapping structure does not need to be specified in advance and it can be obtained by the graphical structure among predictors. Even the graphical structure is unknown, we can also construct it by the sparse estimation of the covariance matrix of predictors. Since the closed form of estimator for generalized linear models usually can not be obtained, the graphical structure among predictors can not be incorporated into the estimation process of GLMs as the linear regression model where the estimator can be formulated as the multiplication of the inverse of covariance matrix, which can be defined

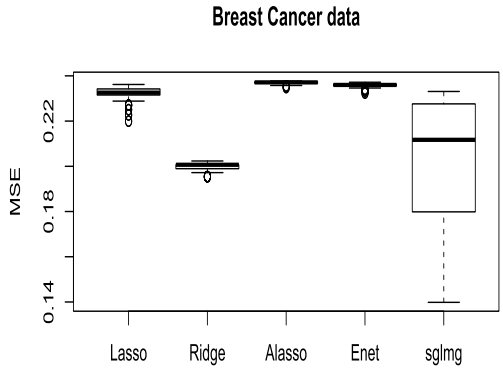

FIG 1. *Comparison of MSE for various methods on the Breast Cancer data*

as the predictor graph, and a vector between predictors and the response from least square estimation. In this paper, we use the sufficient dimension reduction techniques to show that we can also formulate the estimators of GLMs as the multiplication of the inverse of covariance matrix and a vector. Thus, the graphical structure of predictors can be also incorporated into the GLMs. In order to utilize the neighborhood information of the graph we apply a node-by-node strategy to convert the graphical structure to the overlapping group structure. Furthermore, our proposed method is very general and some popular methods such as Lasso, group Lasso and ridge regression can be included as special cases. The theoretical results we obtained are still true when the overlapping group structure is pre-specified.

Appendix A: Additional figures and tables

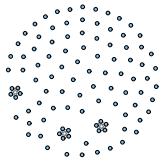

FIG 2. *True predictor graph of Example 1*

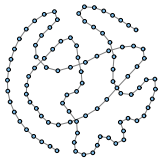

FIG 3. *True predictor graph of Example 2*

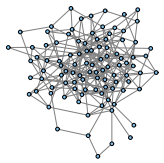

FIG 4. *True predictor graph of Example 3*

TABLE 10
Performance comparison of model selection for $\beta^* = (3, \cdots, 3, 0, \cdots, 0)$ *in Example 1*

Methods	FPR			FNR		
	(A)	(B)	(C)	(A)	(B)	(C)
Lasso	0.138(0.004)	0.220(0.004)	0.288(0.006)	0.548(0.009)	0.368(0.009)	0.263(0.009)
Ridge	1.000(0.000)	1.000(0.000)	1.000(0.000)	0.000(0.000)	0.000(0.000)	0.000(0.000)
ALasso	0.067(0.002)	0.074(0.003)	0.138(0.004)	0.560(0.010)	0.340(0.009)	0.277(0.009)
ENet	0.317(0.007)	0.399(0.006)	0.509(0.009)	0.100(0.007)	0.034(0.004)	0.012(0.002)
SGLMG-O	0.051(0.002)	0.046(0.002)	0.035(0.002)	0.000(0.000)	0.000(0.000)	0.000(0.000)
SGLMG	0.062(0.003)	0.046(0.002)	0.035(0.002)	0.018(0.005)	0.001(0.001)	0.000(0.000)

TABLE 11
Performance comparison of model selection for $\beta^* = (3, \cdots, 3, 0, \cdots, 0)$ *in Example 2*

Methods	FPR			FNR		
	(A)	(B)	(C)	(A)	(B)	(C)
Lasso	0.137(0.004)	0.221(0.004)	0.300(0.006)	0.419(0.011)	0.162(0.008)	0.068(0.006)
Ridge	1.000(0.000)	1.000(0.000)	1.000(0.000)	0.000(0.000)	0.000(0.000)	0.000(0.000)
ALasso	0.064(0.003)	0.082(0.003)	0.135(0.004)	0.000(0.000)	1.000(0.000)	1.000(0.000)
ENet	0.320(0.007)	0.394(0.006)	0.505(0.010)	0.182(0.010)	0.056(0.005)	0.013(0.002)
SGLMG-O	0.250(0.007)	0.342(0.008)	0.379(0.009)	0.158(0.011)	0.037(0.004)	0.006(0.002)
SGLMG	0.274(0.008)	0.314(0.008)	0.339(0.008)	0.286(0.011)	0.125(0.007)	0.052(0.006)

Appendix B: Proofs

Proof of Theorem 1. From the definition of $\mathfrak{A}$, we have

$$\begin{aligned}\mathfrak{A}^c &= \bigcup_{j=1}^{p}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}(Y_i\mathbf{X}_{i\mathcal{N}_j} - \mathbb{E}Y\mathbf{X}_{\mathcal{N}_j})\right\|_2^2 > \frac{\lambda^2 d_j^2}{4}\right\} \\ &= \bigcup_{j=1}^{p}\left\{\sum_{k\in\mathcal{N}_j}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i\mathbf{X}_{ik} - \mathbb{E}Y\mathbf{X}_k)\right]^2 > \frac{\lambda^2 d_j^2}{4}\right\},\end{aligned}$$

thus,

$$\begin{aligned}\mathbb{P}(\mathfrak{A}^c) &\le \sum_{j=1}^{p}\mathbb{P}\left\{\sum_{k\in\mathcal{N}_j}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i\mathbf{X}_{ik} - \mathbb{E}Y\mathbf{X}_k)\right]^2 > \frac{\lambda^2 d_j^2}{4}\right\} \\ &\le \sum_{j=1}^{p}\sum_{k\in\mathcal{N}_j}\mathbb{P}\left\{\frac{1}{n}\left|\sum_{i=1}^{n}(Y_i\mathbf{X}_{ik} - \mathbb{E}Y\mathbf{X}_k)\right| > \frac{\lambda d_j}{2}\right\}. \qquad (10)\end{aligned}$$

For $j = 1, \cdots, p,\ i = 1, \cdots, n$, let

$$\chi_{ik}^{\mathcal{N}_j} = Y_i\mathbf{X}_{ik}^{\mathbb{N}_j} - \mathbb{E}Y\mathbf{X}_k^{\mathcal{N}_j}, \quad k \in \mathcal{N}_j.$$

The random variables $\left\{\chi_{ik}^{\mathcal{N}_j}\right\}_{i=1,\cdots,n}$ are independent, identically distributed and centered and for all $m \geq 2$, we have

$$\mathbb{E}\left|\chi_{ik}^{\mathcal{N}_j}\right|^m \leq \sum_{l=1}^{m} \binom{m}{l} \mathbb{E}|Y_i \mathbf{X}_{ik}|^l \left(\mathbb{E}|Y_i \mathbf{X}_{ik}|\right)^{m-l}, \quad k \in \mathcal{N}_j.$$

By using Jensen inequality, for each $l \in \mathbb{N}$, we obtain

$$\mathbb{E}\left|\chi_{ik}^{\mathcal{N}_j}\right|^m \leq 2^m \max_{l=1,\cdots,m} \left\{\mathbb{E}|Y_i \mathbf{X}_{ik}|^l \mathbb{E}|Y_i \mathbf{X}_{ik}|^{m-l}\right\}.$$

By assumption (A1) and Lemma 1, we have

$$\mathbb{E}|Y_i \mathbf{X}_{ik}|^l \leq K^l l! (C_{K,b})^l.$$

Thus,

$$\mathbb{E}\left|\chi_{ik}^{\mathcal{N}_j}\right|^m \leq m! (2KC_{K,b})^m.$$

Therefore, by the Bernstein concentration inequality (Bennett 1962), we obtain

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n} \chi_{ik}^{\mathcal{N}_j}\right| > \frac{\lambda d_j}{2}\right) \leq 2\left\{\exp\left(-\frac{n\lambda d_j}{16KC_{K,b}}\right) + \exp\left(-\frac{n\lambda^2 d_j^2}{32(2KC_{K,b})^2}\right)\right\}. \tag{11}$$

Finally, from (10) and (11), we have

$$\mathbb{P}(\mathfrak{A}^c) \leq 2pN_{\max}\left\{\exp\left(-\frac{n\lambda d_j}{16KC_{K,b}}\right) + \exp\left(-\frac{n\lambda^2 d_j^2}{32(2KC_{K,b})^2}\right)\right\}.$$

Hence, if

$$\lambda d_j \geq A^2 16KC_{Kb}\frac{\log(2p)}{n} \vee A8\sqrt{2}KC_{K,b}\sqrt{\frac{\log(2p)}{n}},$$

with $A > 1$, then

$$\mathbb{P}(\mathfrak{A}^c) \leq 2N_{\max}(2p)^{1-A^2}. \qquad \blacksquare$$

Proof of Theorem 2. First we prove the following Lemma:

Lemma S.1. *Let $\alpha > 0$ be given. Define*

$$Z_\alpha := \sup_{\|\beta-\beta^*\|_G \leq \alpha} \{\|(\mathbb{P}_n - \mathbb{P})(\ell_\phi(\beta^*) - \ell_\phi(\beta))\|_2\}.$$

If $A \geq 1$ then

$$\mathbb{P}\left(Z_\alpha \geq \frac{A5\sqrt{N_{\max}}K\eta\alpha}{d_{\min}}\sqrt{\frac{2\log(2p)}{n}}\right) \leq (2p)^{-A^2}, \tag{12}$$

where $\eta = \max_{\{|x| \leq K\sqrt{N_{\max}}/d_{\min}(\alpha+b)\}\cap\Theta}\{|\phi'(x)|\}$, $d_{\min} = \min_{1\leq j\leq p} d_j$.

Proof of Lemma S.1. Let $\beta \in \Xi$ satisfy $\|\beta - \beta^*\|_G \leq \alpha$. We note that if we change $\mathbf{X}_i$ while keepinf the others fixed then Z_α is modified of at most $\frac{2K\alpha\eta\sqrt{N_{\max}}}{nd_{\min}}$. In fact, let $\mathbb{P}_n = \frac{1}{n}\sum_{j=1}^n 1_{\mathbf{X}_j, Y_j}$ and $\mathbb{P}'_n = \frac{1}{n}\sum_{j\neq i} 1_{\mathbf{X}_j, Y_j} + 1_{\mathbf{X}'_i, Y'_i}$, then we have

$$
\begin{aligned}
&\|(\mathbb{P}_n - \mathbb{P})(\ell_\phi(\beta^*) - \ell_\phi(\beta))\|_2 - \|(\mathbb{P}'_n - \mathbb{P})(\ell_\phi(\beta^*) - \ell_\phi(\beta))\|_2 \\
&\leq \|(\mathbb{P}_n - \mathbb{P})(\ell_\phi(\beta^*) - \ell_\phi(\beta)) - (\mathbb{P}'_n - \mathbb{P})(\ell_\phi(\beta^*) - \ell_\phi(\beta))\|_2 \\
&= \|\frac{1}{n}\{\ell_\phi(\beta^*, \mathbf{X}_i) - \ell_\phi(\beta, \mathbf{X}_i) - \ell_\phi(\beta^*, \mathbf{X}'_i) + \ell_\phi(\beta, \mathbf{X}'_i)\}\|_2 \\
&\leq \frac{1}{n}\|\phi'(\tilde{\beta}^\top \mathbf{X}_i)\|_2 \|\beta^{*\top}\mathbf{X}_i - \beta^\top \mathbf{X}_i\|_2 + \frac{1}{n}\|\phi'(\tilde{\beta}^\top \mathbf{X}'_i)\|_2 \|\beta^{*\top}\mathbf{X}'_i - \beta^\top \mathbf{X}'_i\|_2
\end{aligned}
$$

where $\tilde{\beta}^\top \mathbf{X}_i$ is an intermediate point between $\beta^\top \mathbf{X}_i$ and $\beta^{*\top}\mathbf{X}_i$. Let $U^{(1)}, U^{(2)}, \cdots, U^{(p)}$ and $V^{(1)}, V^{(2)}, \cdots, V^{(p)}$ are arbitrary optimal decompositions of $\beta - \beta^*$ and β^*, respectively, then,

$$
\begin{aligned}
\|\tilde{\beta}^\top \mathbf{X}_i\|_2 &\leq \|(\beta - \beta^*)^\top \mathbf{X}_i\|_2 + \|\beta^{*\top}\mathbf{X}_i\|_2 \\
&= \left\|\sum_{j=1}^p {U_{\mathcal{N}_j}^{(j)}}^\top \mathbf{X}_{i\mathcal{N}_j}\right\|_2 + \left\|\sum_{j=1}^p {V_{\mathcal{N}_j}^{(j)}}^\top \mathbf{X}_{i\mathcal{N}_j}\right\|_2 \\
&= \left\|\sum_{j=1}^p \frac{1}{d_j} d_j {U_{\mathcal{N}_j}^{(j)}}^\top \mathbf{X}_{i\mathcal{N}_j}\right\|_2 + \left\|\sum_{j=1}^p \frac{1}{d_j} d_j {V_{\mathcal{N}_j}^{(j)}}^\top \mathbf{X}_{i\mathcal{N}_j}\right\|_2 \\
&\leq \sum_{j=1}^p \|d_j U_{\mathcal{N}_j}^{(j)}\|_2 \|\mathbf{X}_{i\mathcal{N}_j}/d_j\|_2 + \sum_{j=1}^p \|d_j V_{\mathcal{N}_j}^{(j)}\|_2 \|\mathbf{X}_{i\mathcal{N}_j}/d_j\|_2 \\
&\leq \frac{K\sqrt{N_{\max}}}{d_{\min}}\|\beta - \beta^*\|_{G,d} + \frac{K\sqrt{N_{\max}}}{d_{\min}}\|\beta^*\|_{G,d} \\
&\leq \frac{K\sqrt{N_{\max}}}{d_{\min}}(\alpha + b).
\end{aligned}
$$

For $\beta, \beta^* \in \Xi$, where Ξ is convex set, we have $\tilde{\beta} \in \Xi$ and $\tilde{\beta}^\top \mathbf{X}_i \in \Theta$, a.s. Hence,

$$
\begin{aligned}
&\|(\mathbb{P}_n - \mathbb{P})(\ell_\phi(\beta^*) - \ell_\phi(\beta))\|_2 - \|(\mathbb{P}'_n - \mathbb{P})(\ell_\phi(\beta^*) - \ell_\phi(\beta))\|_2 \\
&\leq \frac{2K\alpha\sqrt{N_{\max}}}{nd_{\min}} \max_{\left\{|x| \leq \frac{K\sqrt{N_{\max}}}{d_{\min}}(\alpha+b)\right\} \cap \Theta} |\phi'(x)| \\
&= \frac{2K\alpha\eta\sqrt{N_{\max}}}{nd_{\min}}.
\end{aligned}
$$

By McDiarmid inequality (Devroye and Lugosi 2001), we have

$$
\mathbb{P}(Z_\alpha - \mathbb{E}Z_\alpha \geq t) \leq \exp\left(-\frac{nt^2 d_{\min}^2}{2K^2\alpha^2\eta^2 N_{\max}}\right).
$$

Therefore, for $A > 0$, if we choose $\lambda \geq \frac{A\sqrt{N_{\max}}K\alpha\eta}{d_{\min}}\sqrt{\frac{\log(2p)}{n}}$, then we have

$$\mathbb{P}(Z_\alpha - \mathbb{E}Z_\alpha \geq \lambda) \leq (2p)^{-A^2}. \tag{13}$$

Now we have to bound the mean $\mathbb{E}Z_\alpha$. Let $\varepsilon_1, \cdots, \varepsilon_n$ be Rademacher sequence independent of $\mathbf{X}_1, \cdots, \mathbf{X}_n$ and let $S_\alpha := \{\beta \in \mathbb{R}^p : \|\beta - \beta^*\|_{G,d} \leq \alpha\}$. Then, by the symmetrization theorem [35] and contraction theorem [18] (note that $|\phi(x) - \phi(x')| \leq \eta|x - x'|$, i.e., ϕ is η-Lipschitz on the compact set S_α), we have

$$\begin{aligned}
\mathbb{E}Z_\alpha &\leq 4\eta\mathbb{E}\left(\sup_{\beta\in S_\alpha}\frac{1}{n}\sum_{i=1}^{n}\|\varepsilon_i({\beta^*}^\top\mathbf{X}_i - \beta\top\mathbf{X}_i)\|_2\right)\\
&= 4\eta\mathbb{E}\left(\sup_{\beta\in S_\alpha}\frac{1}{n}\sum_{i=1}^{n}\left\|\sum_{j=1}^{p}\varepsilon_i {U_{\mathcal{N}_j}^{(j)}}^\top\mathbf{X}_{i\mathcal{N}_j}\right\|_2\right)\\
&= 4\eta\mathbb{E}\left(\sup_{\beta\in S_\alpha}\frac{1}{n}\sum_{i=1}^{n}\left\|\sum_{j=1}^{p}\varepsilon_i\frac{1}{d_j}d_j{U_{\mathcal{N}_j}^{(j)}}^\top\mathbf{X}_{i\mathcal{N}_j}\right\|_2\right)\\
&\leq 4\eta\mathbb{E}\left(\sup_{\beta\in S_\alpha}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}\left\|d_jU_{\mathcal{N}_j}^{(j)}\right\|_2\left\|\varepsilon_i\mathbf{X}_{i\mathcal{N}_j}/d_j\right\|_2\right)\\
&= 4\eta\mathbb{E}\left(\sup_{\beta\in S_\alpha}\sum_{j=1}^{p}\left\|d_jU_{\mathcal{N}_j}^{(j)}\right\|_2\frac{1}{n}\sum_{i=1}^{n}\left\|\varepsilon_i\mathbf{X}_{i\mathcal{N}_j}/d_j\right\|_2\right)\\
&\leq 4\eta\alpha\mathbb{E}\left(\max_{1\leq j\leq p}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\|\mathbf{X}_{i\mathcal{N}_j}/d_j\|_2\right|\right).
\end{aligned}$$

To bound the mean $\mathbb{E}\left(\max_{1\leq j\leq p}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\|\mathbf{X}_{i\mathcal{N}_j}/d_j\|_2\right|\right)$ we need the following Lemma.

Lemma S.2. *Let $\mathbf{X}_1, \cdots, \mathbf{X}_n$ be independent random variables on $\mathcal{X}$ and $f_1, \cdots, f_n$ be real-valued functions on $\mathcal{X}$ which satisfies for all $i = 1, \cdots, n$ and all $j = 1, \cdots, p$*

$$\mathbb{E}f_j(\mathbf{X}_i) = 0, \quad |f_j(\mathbf{X}_i)| \leq a_{ij}.$$

Then,

$$\mathbb{E}\left(\max_{1\leq j\leq p}\left|\sum_{i=1}^{n}f_j(\mathbf{X}_i)\right|\right) \leq \sqrt{2\log(2p)}\max_{1\leq j\leq p}\sqrt{\sum_{i=1}^{n}a_{ij}^2}.$$

Proof of Lemma S.2. The proof of this Lemma can be directly deduced by Hoeffding inequality. ■

By $\left|\varepsilon_i \left\| \frac{\mathbf{X}_{\mathcal{N}_j}}{d_j} \right\|_2 \right| \leq \frac{K\sqrt{|N_j|}}{d_{\min}}$ and Lemma S.2 we have

$$\mathbb{E}\left(\max_{1\leq j\leq p} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \|\mathbf{X}_{i\mathcal{N}_j}/d_j\|_2 \right| \right) \leq \frac{K\sqrt{N_{\max}}}{d_{\min}} \sqrt{\frac{2\log(2p)}{n}}.$$

Hence,

$$\mathbb{E}Z_\alpha \leq \frac{4\sqrt{N_{\max}} K\eta\alpha}{d_{\min}} \sqrt{\frac{2\log(2p)}{n}}. \tag{14}$$

Therefore, we can conclude from (13) and (14) that if $A \geq 1$ then for all $\alpha > 0$, we have

$$\mathbb{P}\left(Z_\alpha \geq \frac{A5\sqrt{N_{\max}} K\eta\alpha}{d_{\min}} \sqrt{\frac{2\log(2p)}{n}} \right) \leq (2p)^{-A^2}. \tag{15}$$

■

Second, we split up $\{\beta \in \mathbb{R}^p : \|\beta - \beta^*\|_G \leq a\}$ into the following two sets:

$$\begin{aligned} E_1 &= \{\beta \in \mathbb{R}^p : \|\beta - \beta^*\|_G \leq \varepsilon_n\}, \\ E_2 &= \{\beta \in \mathbb{R}^p : \varepsilon_n \leq \|\beta - \beta^*\|_G \leq a\} \\ &\subset \bigcup_{j=1}^{j_n} \{\beta \in \mathbb{R}^p : 2^{j-1}\varepsilon_n < \|\beta - \beta^*\|_G \leq 2^j \varepsilon_n\}, \end{aligned}$$

where $j_n := \lfloor \log_2(na) \rfloor + 1$ is the smaller integer such that $2^{j_n}\varepsilon_n \geq a$. Let

$$\nu_n(\beta, \beta^*) := \frac{(\mathbb{P}_n - \mathbb{P})(\ell_\phi(\beta^*) - \ell_\phi(\beta))}{\|\beta - \beta^*\|_{G,d} + \varepsilon_n},$$

and to simplify notation let

$$\kappa_n(\beta, \beta^*) := (\mathbb{P}_n - \mathbb{P})(\ell_\phi(\beta^*) - \ell_\phi(\beta))$$

and

$$\Phi(t) := \max_{\{|x|\leq t\}\cap\Theta} |\phi'(x)|.$$

Let $A \geq 1$. We recall that $\delta_n := 17b + \frac{2}{n} = 2a + b$. On the event E_1, we have

$$\begin{aligned} &\mathbb{P}\left(\sup_{\beta\in E_1} |\nu_n(\beta, \beta^*)| \geq \frac{A10\sqrt{N_{\max}} K\Phi\left(\frac{K\sqrt{N_{\max}}}{d_{\min}}\delta_n\right)}{d_{\min}} \sqrt{\frac{2\log(2p)}{n}} \right) \\ &\leq \mathbb{P}\left(\sup_{\beta\in E_1} |\kappa_n(\beta, \beta^*)| \geq \frac{A10\sqrt{N_{\max}} K\Phi\left(\frac{K\sqrt{N_{\max}}}{d_{\min}}\delta_n\right)\varepsilon_n}{d_{\min}} \sqrt{\frac{2\log(2p)}{n}} \right) \\ &\leq \mathbb{P}\left(\sup_{\beta\in E_1} |\kappa_n(\beta, \beta^*)| \geq \frac{A5\sqrt{N_{\max}} K\Phi\left(\frac{K\sqrt{N_{\max}}}{d_{\min}}(\varepsilon_n + b)\right)\varepsilon_n}{d_{\min}} \sqrt{\frac{2\log(2p)}{n}} \right), \end{aligned}$$

where we suppose that $2a \geq \varepsilon_n$. From Lemma S.1 with $\alpha = \varepsilon_n$ we obtain

$$\mathbb{P}\left(\sup_{\beta\in E_1}|\nu_n(\beta,\beta^*)| \geq \frac{A10\sqrt{N_{\max}}K\Phi\left(\frac{K\sqrt{N_{\max}}}{d_{\min}}\delta_n\right)}{d_{\min}}\sqrt{\frac{2\log(2p)}{n}}\right) \leq (2p)^{-A^2}. \tag{16}$$

Similarly, on the event E_2, with $\alpha = 2^j\varepsilon_n$ (given that $2a \geq 2^j\varepsilon_n$) for all $j = 1, \cdots, j_n$, we have

$$\mathbb{P}\left(\sup_{\beta\in E_2}|\nu_n(\beta,\beta^*)| \geq \frac{A10\sqrt{N_{\max}}K\Phi\left(\frac{K\sqrt{N_{\max}}}{d_{\min}}\delta_n\right)}{d_{\min}}\sqrt{\frac{2\log(2p)}{n}}\right) \leq j_n(2p)^{-A^2}.$$

Because $j_n := \lfloor \log_2(na)\rfloor + 1$ and $n \ll p$, there exists a constant C' such that

$$\mathbb{P}\left(\sup_{\beta\in E_2}|\nu_n(\beta,\beta^*)| \geq \frac{A10\sqrt{N_{\max}}K\Phi\left(\frac{K\sqrt{N_{\max}}}{d_{\min}}\delta_n\right)}{d_{\min}}\sqrt{\frac{2\log(2p)}{n}}\right) \leq C'(2p)^{-A^2}. \tag{17}$$

Let $C = 1 + C'$, by (16) and (17) we have

$$\mathbb{P}(\mathfrak{B}^c) \leq C(2p)^{-A^2}. \qquad \blacksquare$$

Proof of Theorem 3. Let $\{W^{(1)}, W^{(2)}, \cdots, W^{(p)}\}$ be an optimal decomposition of $\hat{\beta}_n - \beta^*$. Then,

$$\begin{aligned}
(\mathbb{P}_n - \mathbb{P})(\ell_l(\hat{\beta}_n) - \ell_l(\beta^*)) &= \frac{1}{n}\sum_{i=1}^{n}(\hat{\beta}_n - \beta^*)^\top(Y_i\mathbf{X}_i - \mathbb{E}Y\mathbf{X}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}{W_{\mathcal{N}_j}^{(j)}}^\top(Y_i\mathbf{X}_{i\mathcal{N}_j} - \mathbb{E}Y\mathbf{X}_{\mathcal{N}_j}) \\
&= \sum_{j=1}^{p}{W_{\mathcal{N}_j}^{(j)}}^\top\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i\mathbf{X}_{i\mathcal{N}_j} - \mathbb{E}Y\mathbf{X}_{\mathcal{N}_j})\right] \\
&= \sum_{j=1}^{p}d_j{W_{\mathcal{N}_j}^{(j)}}^\top\left[\frac{1}{nd_j}\sum_{i=1}^{n}(Y_i\mathbf{X}_{i\mathcal{N}_j} - \mathbb{E}Y\mathbf{X}_{\mathcal{N}_j})\right] \\
&\leq \sum_{j=1}^{p}\left\|d_jW_{\mathcal{N}_j}^{(j)}\right\|_2\left\|\frac{1}{nd_j}\sum_{i=1}^{n}(Y_i\mathbf{X}_{i\mathcal{N}_j} - \mathbb{E}Y\mathbf{X}_{\mathcal{N}_j})\right\|_2
\end{aligned}$$

On the event $\mathfrak{A}$, we have $\left\|\frac{1}{nd_j}\sum_{i=1}^{n}(Y_i\mathbf{X}_{i\mathcal{N}_j} - \mathbb{E}Y\mathbf{X}_{\mathcal{N}_j})\right\|_2 \leq \frac{\lambda}{2}$. $\blacksquare$

Proof of Lemma 2. Define $t := \frac{a}{a+\|\hat{\beta}_n - \beta^*\|_{G,d}}$ and $\tilde{\beta} = t\hat{\beta}_n + (1-t)\beta^*$. Then

$$\|\tilde{\beta} - \beta^*\|_{G,d} = t\|\hat{\beta}_n - \beta^*\|_{G,d} = \frac{a\|\hat{\beta}_n - \beta^*\|_{G,d}}{a + \|\hat{\beta}_n - \beta^*\|_{G,d}} \leq a.$$

On the event $\mathfrak{A} \cap \mathfrak{B}$, we have $(\mathbb{P}_n - \mathbb{P})(\ell(\beta^*) - \ell(\tilde{\beta})) \leq \lambda \|\tilde{\beta} - \beta^*\|_{G,d} + \lambda \frac{\epsilon_n}{2}$. In addition, by the definition of $\hat{\beta}_n$, we have

$$\mathbb{P}_n \ell(\hat{\beta}_n) + 2\lambda \|\hat{\beta}_n\|_{G,d} \leq \mathbb{P}_n \ell(\beta^*) + 2\lambda \|\beta^*\|_{G,d}.$$

Note that $\ell_\phi(\beta)$ and $\|\beta\|_{G,d}$ are both convex function, we have

$$\mathbb{P}_n \ell(\tilde{\beta}) + 2\lambda \|\tilde{\beta}\|_{G,d} \leq \mathbb{P}_n \ell(\beta^*) + 2\lambda \|\beta^*\|_{G,d}.$$

Thus,

$$\begin{aligned} \mathbb{P}(\ell(\tilde{\beta}) - \ell(\beta^*)) + 2\lambda \|\tilde{\beta}\|_{G,d} &\leq (\mathbb{P}_n - \mathbb{P})(\ell(\beta^*) - \ell(\tilde{\beta})) + 2\lambda \|\beta^*\|_{G,d} \\ &\leq \lambda \|\tilde{\beta} - \beta^*\|_{G,d} + \lambda \frac{\epsilon_n}{2} + 2\lambda \|\beta^*\|_{G,d}. \end{aligned}$$

Because $\mathbb{P}(\ell(\tilde{\beta}) - \ell(\beta^*)) \geq 0$, by adding to both sides of the above inequality $2\lambda \|\beta^*\|_{G,d}$ we have

$$\begin{aligned} 2\lambda \|\tilde{\beta} - \beta^*\|_{G,d} &\leq 2\lambda \|\tilde{\beta}\|_{G,d} + 2\lambda \|\beta^*\|_{G,d} \\ &\leq \lambda \|\tilde{\beta} - \beta^*\|_{G,d} + 2\lambda \|\beta^*\|_{G,d} + 2\lambda \|\beta^*\|_{G,d} + \lambda \frac{\epsilon_n}{2}, \end{aligned}$$

that is,

$$\|\tilde{\beta} - \beta^*\|_{G,d} \leq 4\|\beta^*\|_{G,d} + \frac{\epsilon_n}{2}.$$

Therefore, using assumption (A5), we have

$$\|\tilde{\beta} - \beta^*\|_{G,d} \leq 4b + \frac{\epsilon_n}{2} = \frac{a}{2}.$$

Note that $\tilde{\beta} - \beta^* = t(\hat{\beta}_n - \beta^*)$, thus

$$t\|\hat{\beta}_n - \beta^*\|_{G,d} \leq \frac{a}{2}.$$

Therefore, by the definition of t, we have

$$\|\hat{\beta}_n - \beta^*\|_{G,d} \leq a. \qquad \blacksquare$$

Proof of Theorem 5. First we introduce the following Lemma.

Lemma S.3. *For any predictor graph G and positive weights $d_1, d_2, \cdots, d_p$, suppose $W^{(1)}, W^{(2)}, \cdots, W^{(p)}$ is an optimal decomposition of $\beta \in \mathbb{R}^p$, then for any $S \subset \{1, 2, \cdots, p\}$, $\{W^{(j)} : j \in S\}$ is also an optimal decomposition of $\sum_{j \in S} W^{(j)}$.*

Proof of Lemma S.3. See Lemma 2 from [37]. ■

Then we prove the oracle inequalities. By the definition of $\hat{\beta}_n$ we have

$$\mathbb{P}_n \ell(\hat{\beta}_n) + 2\lambda \|\hat{\beta}_n\|_{G,d} \leq \mathbb{P}_n \ell(\beta^*) + 2\lambda \|\beta^*\|_{G,d}.$$

By adding $\mathbb{P}(\ell(\hat{\beta}_n) - \ell(\beta^*))$ to the both sides of the inequality above, we have

$$\mathbb{P}(\ell(\hat{\beta}_n) - \ell(\beta^*)) + 2\lambda\|\hat{\beta}_n\|_{G,d} \leq (\mathbb{P}_n - \mathbb{P})(\ell(\beta^*) - \ell(\hat{\beta}_n)) + 2\lambda\|\beta^*\|_{G,d}. \quad (18)$$

We decompose the empirical process $(\mathbb{P}_n - \mathbb{P})(\ell(\beta^*) - \ell(\hat{\beta}_n))$ into a linear part and a part which depends on the normalized parameter ϕ, then by Theorem 3 and Theorem 4 we have

$$\begin{aligned}
&(\mathbb{P}_n - \mathbb{P})(\ell(\beta^*) - \ell(\hat{\beta}_n)) \\
&\quad = (\mathbb{P}_n - \mathbb{P})(\ell_l(\beta^*) - \ell_l(\hat{\beta}_n)) + (\mathbb{P}_n - \mathbb{P})(\ell_\phi(\beta^*) - \ell_\phi(\hat{\beta}_n)) \\
&\quad \leq \frac{\lambda}{2}\|\hat{\beta}_n - \beta^*\|_{G,d} + \frac{\lambda}{2}(\|\hat{\beta}_n - \beta^*\|_{G,d} + \epsilon_n) \\
&\quad = \lambda\|\hat{\beta}_n - \beta^*\|_{G,d} + \frac{\lambda}{2}\epsilon_n. \qquad (19)
\end{aligned}$$

Substituting $(\mathbb{P}_n - \mathbb{P})(\ell(\beta^*) - \ell(\hat{\beta}_n))$ in (18) for (19) and adding $\lambda\|\hat{\beta}_n - \beta^*\|_{G,d}$ to both sides of the resulting inequality we find that

$$\mathbb{P}(\ell(\hat{\beta}_n) - \ell(\beta^*)) + \lambda\|\hat{\beta}_n - \beta^*\|_{G,d} \leq 2\lambda(\|\hat{\beta}_n - \beta^*\|_{G,d} + \|\beta^*\|_{G,d} - \|\hat{\beta}_n\|_{G,d}) + \frac{\lambda}{2}\epsilon_n. \quad (20)$$

Denote $U^{(1)}, U^{(2)}, \cdots, U^{(p)}$ and $V^{(1)}, V^{(2)}, \cdots, V^{(p)}$ as the arbitrary optimal decompositions of $\hat{\beta}_n - \beta^*$ and β^*, respectively. Applying assumption (A4), we find that for each $j \in J^*$, $\mathcal{N}_j \subset J^*$ we have

$$\begin{aligned}
&\|\hat{\beta}_n - \beta^*\|_{G,d} + \|\beta^*\|_{G,d} - \|\hat{\beta}_n\|_{G,d} \\
&\quad = \left\|\sum_{j\in J^*} U^{(j)}\right\|_{G,d} + \left\|\sum_{j\in J^{*c}} U^{(j)}\right\|_{G,d} + \left\|\sum_{j\in J^*} V^{(j)}\right\|_{G,d} - \|\hat{\beta}_n\|_{G,d}. \qquad (21)
\end{aligned}$$

Further, we note that

$$\begin{aligned}
\|\hat{\beta}_n\|_{G,d} = \|\hat{\beta}_n - \beta^* + \beta^*\|_{G,d} &= \left\|\sum_{j\in J^*} U^{(j)} + \sum_{j\in J^{*c}} U^{(j)} + \sum_{j\in J^*} V^{(j)}\right\|_{G,d} \\
&\geq \left\|\sum_{j\in J^{*c}} U^{(j)} + \sum_{j\in J^*} V^{(j)}\right\|_{G,d} - \left\|\sum_{j\in J^*} U^{(j)}\right\|_{G,d} \\
&= \left\|\sum_{j\in J^{*c}} U^{(j)}\right\|_{G,d} + \left\|\sum_{j\in J^*} V^{(j)}\right\|_{G,d} - \left\|\sum_{j\in J^*} U^{(j)}\right\|_{G,d}, \qquad (22)
\end{aligned}$$

thus, by (21) and (22), we have

$$\|\hat{\beta}_n - \beta^*\|_{G,d} + \|\beta^*\|_{G,d} - \|\hat{\beta}_n\|_{G,d} \leq 2\left\|\sum_{j\in J^*} U^{(j)}\right\|_{G,d}.$$

By the definition of β^* we have $\mathbb{P}(\ell(\hat{\beta}_n)-\ell(\beta^*)) \geq 0$, hence from (20) we deduce that

$$\|\hat{\beta}_n-\beta^*\|_{G,d} \leq 2(\|\hat{\beta}_n-\beta^*\|_{G,d}+\|\beta^*\|_{G,d}-\|\hat{\beta}_n\|_{G,d})+\frac{\epsilon_n}{2} \leq 4\left\|\sum_{j\in J^*} U^{(j)}\right\|_{G,d}+\frac{\epsilon_n}{2}.$$

By Lemma S.3 we have $\|\sum_{j\in J^*} U^{(j)}\|_{G,d} = \sum_{j\in J^*} d_j\|U^{(j)}\|_2$, further, we note that

$$\|\hat{\beta}_n - \beta^*\|_{G,d} = \sum_{j\in J^*} d_j\|U^{(j)}\|_2 + \sum_{j\in J^{*c}} d_j\|U^{(j)}\|_2.$$

This yields

$$\sum_{j\in J^{*c}} d_j\|U^{(j)}\|_2 \leq 3\sum_{j\in J^*} d_j\|U^{(j)}\|_2 + \frac{\epsilon_n}{2}.$$

Now we have to bound the empirical process $\mathbb{P}(\ell(\hat{\beta}_n) - \ell(\beta^*))$.

Lemma S.4. *On the event $\mathfrak{A}\cap\mathfrak{B}$, we have*

$$\mathbb{P}(\ell(\hat{\beta}_n) - \ell(\beta^*)) \geq c_n\mathbb{E}\left[\left\{f_{\hat{\beta}_n}(\mathbf{X}) - f_{\beta^*}(\mathbf{X})\right\}^2\right],$$

where $c_n = \max_{\left\{|x|\leq \frac{K\sqrt{N_{\max}}}{d_{\min}}(\alpha+b)\right\}\cap\Theta} |\phi''(x)|$.

Proof of Lemma S.4. We note that

$$\begin{aligned}\mathbb{P}(\ell(\hat{\beta}_n)) - \ell(\beta^*)) =& -\mathbb{E}\left[\mathbb{E}(Y|\mathbf{X})\{f_{\hat{\beta}_n}(\mathbf{X}) - f_{\beta^*}(\mathbf{X})\}\right]\\ &+\mathbb{E}\left[\phi'(f_{\beta^*}(\mathbf{X}))\{f_{\hat{\beta}_n}(x) - f_{\beta^*}(\mathbf{X})\}\right]\\ &+\mathbb{E}\left[\frac{\phi''(f_{\tilde{\beta}}(\mathbf{X}))}{2}\{f_{\hat{\beta}_n}(\mathbf{X}) - f_{\beta^*}(\mathbf{X})\}^2\right]\end{aligned}$$

where $\tilde{\beta}^\top\mathbf{X}$ is an intermediate point between $\hat{\beta}_n^\top\mathbf{X}$ and ${\beta^*}^\top\mathbf{X}$ given by a second order Taylor expansion of ϕ. Because $\phi'(f_{\beta^*}(\mathbf{X})) = \mathbb{E}(Y|\mathbf{X})$ we find

$$\mathbb{P}(\ell(\hat{\beta}_n)) - \ell(\beta^*)) = \mathbb{E}\left[\frac{\phi''(f_{\tilde{\beta}}(\mathbf{X}))}{2}\{f_{\hat{\beta}_n}(\mathbf{X}) - f_{\beta^*}(\mathbf{X})\}^2\right].$$

In addition, by using the same strategy in the proof of Lemma S1 we have

$$\|\tilde{\beta}^\top\mathbf{X}\|_2 \leq \frac{K\sqrt{N_{\max}}}{d_{\min}}(\alpha+b) \qquad \text{a.s.}$$

Furthermore, $\hat{\beta}_n$ and β^* belong to Ξ which is a convex set. Thus, $\tilde{\beta}\in\Xi$ and $\tilde{\beta}^\top\mathbf{X}\in\Theta$. Therefore, we have

$$\mathbb{P}(\ell(\hat{\beta}_n) - \ell(\beta^*)) \geq c_n\mathbb{E}\left[\left\{f_{\hat{\beta}_n}(\mathbf{X}) - f_{\beta^*}(\mathbf{X})\right\}^2\right],$$

where $c_n = \max_{\left\{|x|\leq \frac{K\sqrt{N_{\max}}}{d_{\min}}(\alpha+b)\right\}\cap\Theta} |\phi''(x)|$. ∎

From Lemma S.4 and (20) we deduce that

$$\lambda\|\hat{\beta}_n - \beta^*\|_{G,d} + c_n\mathbb{E}(\hat{\beta}_n^\top \mathbf{X} - \beta^{*\top}\mathbf{X})^2 \le 4\lambda \sum_{j\in J^*} d_j\|U^{(j)}\|_2 + \frac{\lambda}{2}\epsilon_n. \tag{23}$$

Let Σ be $p \times p$ covariance matrix, we have

$$\mathbb{E}(\hat{\beta}_n^\top \mathbf{X} - \beta^{*\top}\mathbf{X})^2 = (\hat{\beta}_n - \beta^*)^\top \Sigma(\hat{\beta}_n - \beta^*).$$

By (A6), we have

$$c_n(\hat{\beta}_n - \beta^*)^\top \Sigma(\hat{\beta}_n - \beta^*) \ge c_n\kappa \sum_{j\in J^*} d_j^2\|U^{(j)}\|_2^2 - \frac{\epsilon_n}{2}. \tag{24}$$

Hence, by applying (23) and (24), we have

$$\lambda\|\hat{\beta}_n - \beta^*\|_G + c_n\kappa \sum_{j\in J^*} d_j^2\|U^{(j)}\|_2^2 \le 4\lambda \sum_{j\in J^*} d_j\|U^{(j)}\|_2 + (\lambda+1)\frac{\epsilon_n}{2}. \tag{25}$$

Further, for each $j \in J^*$, there is at most $\mathcal{K}_G$ nonzero $U^{(j)}$, thus, we have

$$\sum_{j\in J^*} d_j\|U^{(j)}\|_2 \le \mathcal{K}_G^{1/2}\sqrt{\sum_{j\in J^*} d_j^2\|U^{(j)}\|_2^2}. \tag{26}$$

Therefore, by (25) and (26), we have

$$\lambda\|\hat{\beta}_n - \beta^*\|_{G,d} + c_n\kappa \sum_{j\in J^*} d_j^2\|U^{(j)}\|_2^2 \le 4\lambda\mathcal{K}_G^{1/2}\sqrt{\sum_{j\in J^*} d_j^2\|U^{(j)}\|_2^2} + (\lambda+1)\frac{\epsilon_n}{2}. \tag{27}$$

Now the fact $2xy \le tx^2 + y^2/t$ for all $t > 0$ leads to the following inequality

$$\lambda\|\hat{\beta}_n - \beta^*\|_{G,d} + c_n\kappa \sum_{j\in J^*} d_j^2\|U^{(j)}\|_2^2 \le 4t\lambda^2\mathcal{K}_G + \frac{1}{t}\sum_{j\in J^*} d_j^2\|U^{(j)}\|_2^2 + (\lambda+1)\frac{\epsilon_n}{2}. \tag{28}$$

Substituting t for $\frac{1}{c_n\kappa}$ in (28) we obtain

$$\|\hat{\beta}_n - \beta^*\|_{G,d} \le \frac{4\lambda\mathcal{K}_G}{c_n\kappa} + (1+\frac{1}{\lambda})\frac{\epsilon_n}{2}.$$

Besides, we note that

$$\|\hat{\beta}_n - \beta^*\|_2 = \left\|\sum_{j=1}^{p} \frac{1}{d_j} d_j U^{(j)}\right\|_2 \le \frac{\|\hat{\beta}_n - \beta^*\|_{G,d}}{d_{\min}} \le \frac{4\lambda\mathcal{K}_G}{c_n\kappa d_{\min}} + (1+\frac{1}{\lambda})\frac{\epsilon_n}{2d_{\min}}.$$

Then, by (23), we have

$$\lambda\|\hat{\beta}_n - \beta^*\|_{G,d} + c_n\mathbb{E}(\hat{\beta}_n^\top \mathbf{X} - \beta^{*\top}\mathbf{X})^2 \le 4\lambda \sum_{j\in J^*} d_j\|U^{(j)}\|_2 + \frac{\lambda}{2}\epsilon_n$$

$$\leq 4\lambda \sum_{j=1}^{p} d_j \|U^{(j)}\|_2 + \frac{\lambda}{2}\epsilon_n.$$

Thus, we obtain

$$c_n \mathbb{E}(\hat{\beta}_n^\top \mathbf{X} - \beta^{*\top}\mathbf{X})^2 \leq 3\lambda \|\hat{\beta}_n - \beta^*\|_{G,d} + \frac{\lambda}{2}\epsilon_n.$$

Therefore, we conclude that

$$\mathbb{E}(\hat{\beta}_n^\top \mathbf{X} - \beta^{*\top}\mathbf{X})^2 \leq \frac{12}{c_n^2 \kappa}\lambda^2 \mathcal{K}_G + \frac{(3+4\lambda)}{c_n}\frac{\epsilon_n}{2}. \qquad \blacksquare$$

Proof of Theorem 6. For each $v \in \mathbb{R}^p$, define $M_n = \underbrace{\sum_{i=1}^{n} \log \frac{\ell(\beta + v/\sqrt{n}; X_i, Y_i)}{\ell(\beta; X_i, Y_i)}}_{(I)} + \underbrace{n\lambda(\|\beta^* + v/\sqrt{n}\|_{G,d} - \|\beta^*\|_{G,d})}_{(II)}$. By the argmax theorem ([35], Example 3.2.4), we get

$$(I) = v' \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S(\beta; X_i, Y_i) - \frac{1}{2} v' I(\beta) v + o_P(1) \xrightarrow{d} v'W - \frac{1}{2} v' I(\beta) v,$$

where $S(\beta; X, Y) = \frac{1}{L(\beta;X,Y)} \frac{\partial L(\beta;X,Y)}{\partial \beta}$, $L(\beta; X, Y) = \frac{\ell(\beta + v/\sqrt{n};X,Y)}{\ell(\beta;X,Y)}$, is the score function, $I(\beta) = -\sum_{i=1}^{n} \frac{\partial^2}{\partial \beta^2} \log L(\beta; X_i, Y_i)$ is the Fisher information matrix and W is a random variable with $N_p(0, I(\beta))$-distribution.

Without loss of generality, assume that $\beta^* = ((\beta^*_{J^*})^\top, 0)$, that is, the first $|J^*|$ elements of β^* are nonzero and the other $p - |J^*|$ elements are zero. Then we have

$$(II) = (III) + (IV),$$

where $(III) = n\lambda(\left\| \begin{pmatrix} \beta^*_{J^*} + \frac{1}{\sqrt{n}} v_{J^*} \\ 0 \end{pmatrix} \right\|_{G,d} - \left\| \begin{pmatrix} \beta^*_{J^*} \\ 0 \end{pmatrix} \right\|_{G,d})$ and $(IV) = \sqrt{n}\lambda \left\| \begin{pmatrix} 0 \\ v_{J^{*c}} \end{pmatrix} \right\|_{G,d}$.

Denote $W^{(1)}, W^{(2)}, \cdots, W^{(p)}$ as an arbitrary optimal decomposition of v. Then, by the triangle inequality, we have

$$|(III)| \leq n\lambda \left\| \begin{pmatrix} \frac{1}{\sqrt{n}} v_{J^*} \\ 0 \end{pmatrix} \right\|_{G,d} = \sqrt{n}\lambda \sum_{j \in J^*} d_j \|W^{(j)}\|_2.$$

If $\sqrt{n}\lambda \to 0$ and $d_j = O(1)$ for each $j \in J^*$, then for each fixed v, we have $|(III)| \to 0$ as $n \to \infty$.

Furthermore, we observe that

$$|(IV)| = \sqrt{n}\lambda \sum_{j \in J^{*c}} d_j \|W^{(j)}\|_2 = (n^{(\gamma+1)/2}\lambda)(n^{-\gamma/2} \sum_{j \in J^{*c}} d_j \|W^{(j)}\|_2.$$

If $n^{(\gamma+1)/2}\lambda \to \infty$, $v_{J^{*c}} \neq 0$ and $\liminf_{n\to\infty} n^{-\gamma/2} d_j > 0$ for each $j \in J^{*c}$, then $|(IV)| \to \infty$, as $n \to \infty$.

Therefore, we get $M_n(v) \xrightarrow{d} M(v)$, where

$$M(v) = \begin{cases} v'W - \frac{1}{2}v'I(\beta)v, & if \quad supp(v) \subset J^* \\ \infty, & else. \end{cases}$$

Since $v^* = (I^{-1}_{J^*,J^*}(\beta)W_{J^*}, 0)^\top = \arg\max_{v\in\mathbb{R}^p} M(v)$, by the argmax theorem ([35], collary 3.2.3) and assumption (A4), we have

$$\sqrt{n}(\hat{\beta}_{J^*} - \beta^*_{J^*}) \xrightarrow{d} N(0, I^{-1}_{J^*,J^*}(\beta))$$

and

$$\sqrt{n}\hat{\beta}_{J^{*c}} \xrightarrow{d} 0 \text{ and therefore } \hat{\beta}_{J^{*c}} \xrightarrow{P} 0. \qquad \blacksquare$$

References

[1] Bakin, S. (1999). *Adaptive regression and model selection in data mining problems*, Ph.D. thesis. Australian National Univ., Canberra.

[2] Bickel, P., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics* **37** 1705–1732. MR2533469

[3] Blazère, M., Loubes, J. and Gamboa, F. (2014). Oracle inequalities for a group lasso procedure applied to generalized linear models in high dimension. *IEEE Transactions on Information Theory* **60** 2303–2318. MR3181526

[4] Breheny, P. and Huang, J. (2009). Penalized methods for bi-level variable selection. *Statistics and Its Interface* **2** 369–380. MR2540094

[5] Bülhlmann, P. and Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. New York, NY, USA: Springer-Verlag. MR2807761

[6] Cai, T., Liu, W. and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* **106** 594–607. MR2847973

[7] Chuang, H. Y., Lee, E. et al.(2007). Network-based classification of breast cancer metastasis. *Molecular Systems Biology* **3** 140.

[8] Cook, R. D.(1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics*. New York: John Wiley & Sons. MR1645673

[9] Duan, N. and Li, K.(1991). Slicing regression: a link-free regression method. *The Annals of Statistics* **19** 505–530. MR1105834

[10] Fan, J., Feng, Y. and Wu, Y. (2009). Network exploration via the adaptive lasso and SCAD penalties. *The Annals of Applied Statistics* **2** 521–541. MR2750671

[11] Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

[12] Hess, K., Anderson, K., Symmans, W. et al. (2006). Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *Journal of Clinical Oncology* **24** 4236–4244.

[13] Huang, J., Breheny, P. and Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics* **27**. 10.1214/12-STS392. MR3025130

[14] Huang, H., Liang, Y. and Liu, X. (2015). Network-based logistic classification with an enhanced $\ell_{1/2}$ solver reveals biomarker and subnetwork signatures for diagnosing lung cancer. *Biostatistics* **2015** 1–7.

[15] Jacob, L., Obozinski, G. and Vert, J. (2009). Group lasso with overlap and graph lasso. *In Proceedings of the 26th Annual International Conference on Machine Learning* **38** 1978–2004.

[16] Kim, S., Pan, W. and Shen, X. (2013). Network-based penalized regression with application to genomic data. *Biometrics* **69** 582–593. MR3106586

[17] Kuerer, H., Newman, L., Smith, T. et al (2009). Clinical course of breast cancer patients with complete pathologic primary tumor and axillary lymph node response to doxorubicin-based neoadjuvant chemotherapy. *Journal of Clinical Oncology* **17** 460–469.

[18] Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, New York. MR1102015

[19] Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86** 316–327. MR1137117

[20] Li, C. and Li, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* **24** 1175–1182.

[21] Lounici, K., Pontil, M., Geer, S. et al. (2011). Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics* **39** 2164–2204. MR2893865

[22] McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman & Hall, London U.K. MR3223057

[23] Meier, L., Geer, S. and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society* **70** 53–71. MR2412631

[24] Nardi, Y. and Rinaldo, A. (2008). On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics* **2** 605–633. MR2426104

[25] Negahban, S., Ravikumar, P., Wainwright, M. et al. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistitical Science* **27** 447–608. MR3025133

[26] Obozinski, G., Jacob, L. and Vert, J. (2011). *Group lasso with overlaps: the latent group lasso approach*. arXiv preprint arXiv:1110.0413. MR3211304

[27] Percival, D. (2012). Theoretical properties of the overlapping groups lasso. *Electronic Journal of Statistics* **6** 269–288. MR2988408

[28] Percival, D., Roeder, K., Rosenfeld, R. et al. (2011). Structured

sparse regression with application to HIV drug resistance. *Annals of Applied Statistics* **2** 628–644. MR2840168

[29] Roth, V. and Fischer, B. (2008). The group-lasso for generalized linear models: uniqueness of solutions and efficient algorithms. *In ICML '08: Proceedings of the 25th international conference on Machine learning* 848–855.

[30] Shevade, S. and Keerthi, S. (2003). A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics* **19** 2246–2253.

[31] Subramanian, A., Tamayo, P., Mootha, V. et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102** 15545–15550.

[32] Tarigan, B. and Van De Geer, S. A. (2006). Classifiers of support vector machine type with ℓ_1 complexity regularization. *Bernoulli* **12(6)** 1045–1076. MR2274857

[33] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society* **58** 267–288. MR1379242

[34] Van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics* **36(2)** 614–645. MR2396809

[35] Van Der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer. MR1385671

[36] Yang, Y. and Zou, H. (2013). *gglasso: Group Lasso Penalized Learning Using A Unified BMD Algorithm*. R package version 1.1.

[37] Yu, G. and Liu, Y. (2016). Sparse regression incorporating graphical structure among predictors. *Journal of the Americal Statistical Association* **111** 707–720. MR3538699

[38] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society* **68** 49–67. MR2212574

[39] Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Journal of the Royal Statistical Society* **94** 19–35. MR2367824

[40] Zeng, Y. and Breheny, P. (2016). Overlapping group logistic regression with application to genetic pathway selection. *Cancer Informatics* **15** 179–187.

[41] Zhang, W., Wan, Y., Allen, G. et al. (2013). Molecular pathway identification using biological network-regularized logistic models. *BMC Genomics* **14** S7.

[42] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101** 1418–1429. MR2279469