Ultra high dimensional generalised additive model: Unified Theory and Methods

August 18, 2020

(Running title: High dimensional GAM)

Kaixu Yang

Department of Statistics and Probability

Michigan State University

USA

Tapabrata Maiti

Department of Statistics and Probability

Michigan State University

USA

Abstract

Generalised additive model is a powerful statistical learning and predictive modeling tool that has been applied in a wide range of applications. The need of high-dimensional additive modeling is eminent in the context of dealing with high through-put data such as genetic data analysis. In this article, we studied a two step selection and estimation method for ultra high dimensional generalised additive models. The first step applies group lasso on the expanded bases of the functions. With high probability this selects all nonzero functions without having too much over selection. The second step uses adaptive group lasso with any initial estimators, including the group lasso estimator, that satisfies some regular conditions. The adaptive group lasso estimator is shown to be selection consistent with improved convergence rates. Tuning parameter selection is also discussed and shown to select the true model consistently under GIC procedure. The theoretical properties are supported by extensive numerical study.

Keywords— Adaptive group lasso; Generalised additive model; High dimensional variable selection; Selection consistency; Tuning parameter selection.

1 Introduction

The main objective of this work is to establish theory driven high dimensional generalised additive modeling method with nonlinear links. The methodology includes convergence rate, variable selection consistency and tuning parameter selection consistency. Additive models play important roles in nonparametric statistical modeling and machine learning. Although this important statistical learning tool has been used in many important applications and there are free software available for implementing these models along with their variations, to our surprise, there is no literature that has studied the high-dimensional GAM with non-identity link systematically with theoretical founda-

tion. Generalised additive modeling allows nonlinear relationship between a response variable and a set of predictor variables. This general set up includes the special case, namely, the generalised linear models, by letting each additive component be a linear function. In general, let (y_i, \mathbf{X}_i) , i = 1, ..., n be independent observations, where y_i 's are response variables whose corresponding p-dimensional predictor vectors are \mathbf{X}_i 's. A generalised additive model (Hastie and Tibshirani, 1986) is defined as

$$\mu_i = E(y_i | \mathbf{X}_i) = g^{-1} \left(\sum_{j=1}^{p_n} f_j(X_{ij}) \right),$$
 (1)

where $g(\cdot)$ is a link function, f_j 's are unspecified smooth functions and X_{ij} is the jth component of vector \mathbf{X}_i . One of the functions could be a constant, which is the intercept term, but this is not necessary. The number of additive components is written as p_n , since it sometimes (usually in high dimensional set up) increases as n increases. A simple case that many people have studied is $p_n = p$, where the number of additive components is fixed and usually less than the sample size n. The choice of link function is as simple as in generalised linear models, where people prefer to choose link functions that make the distribution of the response variables belong to the popular exponential family. A widely used generalised additive model has the identity link function $g(\mu) = \mu$, which gives the classical additive model

$$y_i = \sum_{j=1}^{p_n} f_j(X_{ij}) + \epsilon_i, \tag{2}$$

where ϵ_i 's are i.i.d random variables with mean 0 and finite variance σ^2 .

On the other hand, high dimensional data analysis has become a part of many modern days scientific applications. Often the number of predictors p_n is much larger than the number of observations n, which is usually written as $p_n \gg n$. One of the most interesting scale is p_n increases exponentially as n increases, i.e. $\log p_n = O(n^{\rho})$ for some

constant $\rho > 0$. Fan and Lv (2011) called this as non-polynomial dimensionality or ultra high-dimensionality.

In this paper, we consider the generalised additive model in a high-dimensional set up. To avoid identification problems, the functions are assumed to be sparse, i.e. only a small proportion of the functions are non-zero and all others are exactly zero. A more generalised set up is that the number of nonzero functions, denoted s_n , also diverges as n increases. This case is also considered in this paper.

Many others have worked on generalised additive models. Common approaches use basis expansion to deal with the nonparametric functions, and perform variable selection and estimation methods on the bases. Meier et al. (2009) considered a simpler case (2), with a new sparsity-smoothness penalty and proved it's oracle property. They also performed a simulation study under logit link with their new penalty, however, no theoretical support was provided. Fan et al. (2011) proposed the nonparametric independence screening (NIS) method in screening the model (2). However, the selection consistency and the generalised link functions were not discussed. Marra and Wood (2011) discussed the practical variable selection in additive models, but not in the highdimensional set up. Liu et al. (2013) considered a two-step oracally efficient approach in generalised additive models in the low dimensional set up, but no variable selection in the high dimensional set up was done. Huang et al. (2010) focused on the variable selection of (2) with fixed number of nonzero functions and identity link function using a two step approach: first group lasso (Bakin, 1999; Yuan and Lin, 2006) on the bases to select the nonzero predictors and then use adaptive group lasso to estimate the bases coefficients. They then established the selection consistency and provided the rate of convergence of the estimation. Amato et al. (2016) reviewed several existing algorithms highlighting the connections between them, including the non-negative garrote, COSSO and adaptive shrinkage, and presented some computationally efficient algorithms for

fitting the additive models. Nandy et al. (2017) extended the consistency and rate of convergence of Huang et al. (2010) to spatial additive models. Fan and Zhong (2018) studied the GAM with identity link under the endogeneity setting. It worth mentioning that alternative methods to penalization have also been studied, for example, Tutz and Binder (2006) studied fitting GAM and perform variable selection implicitly through likelihood based boosting.

However, though widely used, no systematic theory about selection and estimation consistency and rate of convergence has been established for generalised additive models with non-identity link functions in the high-dimensional set up.

In this paper, we establish the theory part for generalised additive models with non-identity link functions in high dimensional set up. We develop a two-step selection approach, where in the first step we use group lasso to perform a screening, which, under mild assumptions, is able to select all nonzero functions and not over-select too much. In the second step, the adaptive group lasso procedure is used and is proved to select the true predictors consistently.

Another important practical issue in variable selection and penalised optimization problems is tuning parameter selection. Various cross validation (CV) techniques have been used in practice for a long time. Information criteria such as Akaike information criterion (AIC), AICc, Bayesian information criterion (BIC), Mallow's C_p and etc. have been used to select 'the best' model as well. Many equivalences among the tuning parameter selection methods have been shown in the Gaussian linear regression case. However, the consistency of these selection methods were not established. Later some variations of the information criteria such as modified BIC (Zhang and Siegmund, 2007; Wang et al., 2009) extended BIC (Chen and Chen, 2008) and generalised information criterion (GIC) (Fan and Tang, 2013) were proposed and shown to have good asymptotic properties in penalised linear models and penalised likelihoods. However, the results are

not useful for grouped variables in additive models, for which basis expansion technique is usually used and thus brings grouped selection.

In this paper, we generalise the result of generalised information criterion (GIC) by Fan and Tang (2013) to group-penalised likelihood problems and show that under some common conditions and with a good choice of the parameter in GIC, we are able to select the tuning parameter that corresponds to the true model.

In section 2, the model is specified and basic approach is discussed. Notations and basic assumptions are also introduced in this section. Section 3 gives the main results of the two steps selection and estimation procedure. Section 4 develops the tuning parameter selection. Extensive simulation study and real data example are presented in section 5 followed by a short discussion in section 6. The proofs of all theorems are deferred to supplementary materials.

2 Model

We consider the generalised additive model (1) with the link function corresponding to an exponential family distribution of the response. For each of the n independent observations, the density function is given as

$$f_{y_i}(y) = c(y) \exp\left[\frac{y\theta_i - b(\theta_i)}{\phi}\right], \ 1 \le i \le n, \ \theta_i \in \mathbb{R}.$$
 (3)

Without loss of generality, we assume that the dispersion parameter $0 < \phi < \infty$ is assumed to be a known constant. Specifically we assume $\phi = 1$. We consider a fixed-design throughout this paper, i.e., the design matrix \boldsymbol{X} is assumed to be fixed. However, we have shown in appendix A that the same theory works for a random design under simple assumptions on the distribution of \boldsymbol{X} . The additive relationship assumes that the densities of y_i 's depend on \boldsymbol{X}_i 's through the additive structure $\theta_i = \sum_{j=1}^{p_n} f_j(X_{ij})$.

This is the canonical link. If we use other link functions, for example, $A(\cdot)$, the theory also works as long as the functions $A(\cdot)$ satisfies the Lipschitz conditions for some order. Let $b^{(k)}(\cdot)$ be the k-th derivative of $b(\cdot)$, then by property of the exponential family, the expectation and variance matrix of $\mathbf{y} = (y_1, ..., y_n)^T$, under mild assumptions of $b(\cdot)$, is given by $\boldsymbol{\mu}(\boldsymbol{\theta})$ and $\phi \Sigma(\boldsymbol{\theta})$, where

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = (b^{(1)}(\theta_1), ..., b^{(1)}(\theta_n))^T \text{ and } \Sigma(\boldsymbol{\theta}) = \text{diag}\{b^{(2)}(\theta_1), ..., b^{(2)}(\theta_n)\}.$$
 (4)

The log-likelihood (ignoring the term c(y) which is not interesting to us in parameter estimation) can be written as

$$l = \sum_{i=1}^{n} \left[y_i \left(\sum_{j=1}^{p_n} f_j(X_{ij}) \right) - b \left(\sum_{j=1}^{p_n} f_j(X_{ij}) \right) \right].$$
 (5)

Assume that the additive components belong to the Sobolev space $W_2^d([a,b])$. According to Schumaker (1981), see pages 268-270, there exists B-spline approximation

$$f_{nj}(x) = \sum_{k=1}^{m_n} \beta_{jk} \phi_k(x), \qquad 1 \le j \le p.$$
 (6)

with $m_n = K_n + l$, where K_n is the number of internal knots and $l \ge d$ is the degree of the splines. Generally, it is recommended that d = 2 and l = 4, i.e., cubic splines.

Using the approximation above, Huang et al. (2010) proved that f_{nj} well approximates f_j in the sense of rate of convergence that

$$||f_j - f_{nj}||_2^2 = \int_a^b (f_j(x) - f_{nj}(x))^2 dx = O(m_n^{-2d}).$$
 (7)

Therefore, using the basis approximation, the log-likelihood (ignoring the term c(y)

which is not related to the parameters) can be written as

$$l = \sum_{i=1}^{n} \left[y_i \left(\sum_{j=1}^{p_n} \sum_{k=1}^{m_n} \beta_{jk}^0 \Phi_k(x_{ij}) \right) - b \left(\sum_{j=1}^{p_n} \sum_{k=1}^{m_n} \beta_{jk}^0 \Phi_k(x_{ij}) \right) \right] = \sum_{i=1}^{n} \left[y_i \left(\boldsymbol{\beta}^{0T} \Phi_i \right) - b \left(\boldsymbol{\beta}^{0T} \Phi_i \right) \right],$$
(8)

where $\boldsymbol{\beta}^0$ and Φ_i are the vector basis coefficients and bases defined below.

It's also worth noting that the number of bases m_n increases as n increases. This is necessary since Schumaker (1981) mentioned that one need to have sufficient partitions to well approximate f_j by f_{nj} . If we fix m_n , i.e. let $m_n = m_0$, though in the later part we will show the approach to estimate the basis coefficients can have better rate of convergence, the approximation error between the additive components and the spline functions $||f_j(x) - f_{nj}(x)||_2 = [\int_a^b (f_j(x) - f_{nj}(x))^2 dx]^{1/2} = O(1)$ will increase and lead to inconsistent estimations. Therefore, m_n , or more precisely, K_n , need to increase with n.

Our selection and estimation approach will be based on the bases approximated log likelihood (8). Before starting the methodology, we list the notations and state the assumptions we need in this paper.

Notations

The design matrix is $\boldsymbol{X}_{(n \times p_n)} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)^T$. The basis matrix is $\Phi_{(n \times m_n p_n)} = (\Phi_1, ..., \Phi_n)^T$, where $\Phi_i = (\phi_1(x_{i1}), ..., \phi_{m_n}(x_{i1}), ..., \phi_1(x_{ip_n}), ..., \phi_{m_n}(x_{ip_n}))^T$.

The true basis parameters are $\boldsymbol{\beta}^0 = (\beta^0_{11}, ..., \beta^0_{1m_n}, ..., \beta^0_{p_n1}, ..., \beta^0_{p_nm_n})^T \in \mathbb{R}^{m_np_n}$

We assume the functions $f_1,...,f_{p_n}$ are sparse, then $\boldsymbol{\beta}^0$ is block-wise sparse, i.e. the blocks $\boldsymbol{\beta}^0_{\ 1}=(\beta^0_{11},...,\beta^0_{1m_n})^T,...,\boldsymbol{\beta}^0_{\ p_n}=(\beta^0_{p_n1},...,\beta^0_{p_nm_n})^T$ are sparse.

Let μ_y be the expectation of \mathbf{y} based on the true basis parameters and $\varepsilon = \mathbf{y} - \mu_y$. Define the relationship $a_n \leq b_n$ as there exists a finite constant c such that $a_n \leq cb_n$. For any function f define $||f||_2 = [\int_a^b f^2(x) dx]^{1/2}$, whenever the integral exists. For any two collections of indices $S, \tilde{S} \subseteq \{1, ..., p_n\}$, the difference set is denoted $S - \tilde{S}$. The cardinality of S is denoted $\operatorname{card}(S)$. For any $\boldsymbol{\delta} \in \mathbb{R}^{m_n p_n}$, define $\boldsymbol{\delta}_1, ..., \boldsymbol{\delta}_{p_n}$ as its sub-blocks, where $\boldsymbol{\delta}_i \in \mathbb{R}^{m_n}$, and define the block-wise support

$$supp_B(\delta) = \{j \in \{1, ..., p_n\}; \delta_j \neq 0\}.$$

Define the block-wise cardinality $\operatorname{card}_B(\boldsymbol{\delta}) = \operatorname{card}(\operatorname{supp}_B(\boldsymbol{\delta})).$

For
$$S = \{s_1, ..., s_q\} \subseteq \{1, ..., p_n\}$$
, define sub-block vector $\boldsymbol{\delta}_S = (\boldsymbol{\delta}_{s_1}^T, ..., \boldsymbol{\delta}_{s_q}^T)^T$.

The number of additive components is denoted p_n , which is possible to grow faster than the sample size n. Let $T = \operatorname{supp}_B(\boldsymbol{\beta}^0)$ and T^c be the compliment set. Let $\operatorname{card}(T) = s_n$, where s_n is allowed to diverge slower than n.

For each $U \subseteq \{1, ..., p_n\}$ with $card(U - T) \leq m$ for some m, define

$$\mathcal{B}(U) = \{ \boldsymbol{\delta} \in \mathbb{R}^{m_n p_n}; \operatorname{supp}_B(\boldsymbol{\delta}) \subseteq U \},$$

$$\mathcal{B}(m) = \{\mathcal{B}(U); \text{ for any } U \subseteq \{1, ..., p_n\}; \operatorname{Card}(U - T) \le m\}.$$

Let q be an integer such that $q > s_n$ and q = o(n). Define

$$\mathcal{B}_1 = \{ \boldsymbol{\beta} \in \mathcal{B} : \operatorname{card}_{\mathcal{B}}(\boldsymbol{\beta}) \leq q \},$$

where \mathcal{B} is a sufficiently large, convex and compact set in \mathbb{R}^d .

Assumptions

Assumption 1 (On design matrix)

Using the normalised B-spline bases, the basis matrix Φ has each covariate vector Φ_j , $j=1,...,p_n$ bounded, i.e., $\exists c_{\Phi}$ such that $\|\Phi_j\|_2 \leq \sqrt{n}c_{\Phi}, \forall j=1,...,m_n \times p_n$.

Assumption 2 (Restricted Eigenvalues RE)

For a given sequence N_n , there exist γ_0 and γ_1 such that

$$\gamma_0 \gamma_2^{2s_n} m_n^{-1} \le \frac{\boldsymbol{\delta}^T \Phi^T \Phi \boldsymbol{\delta}}{n \|\boldsymbol{\delta}\|_2^2} \le \gamma_1 m_n^{-1}, \tag{9}$$

where γ_2 is a positive constant such that $0 < \gamma_2 < 0.5$, for all $\boldsymbol{\delta} \in \mathcal{C}$, where $\boldsymbol{\delta}^T = (\boldsymbol{\delta}_1^T, ..., \boldsymbol{\delta}_{p_n}^T)$ and

$$C = \{ \boldsymbol{\delta} \in \mathbb{R}^{p_n m_n} : \|\boldsymbol{\delta}\|_2 \neq 0, \|\boldsymbol{\delta}\|_2 \leq N_n \text{ and } \operatorname{card}_B(\boldsymbol{\delta}) = o(s_n) \}.$$
 (10)

Assumption 3 (On the exponential family distribution)

The function $b(\theta)$ is three times differentiable with $c_1 \leq b''(\theta) \leq c_1^{-1}$ and $|b'''(\theta)| \leq c_1^{-1}$ in its domain for some constant $c_1 > 0$. For unbounded and non-Gaussian distributed Y_i , there exists a diverging sequence $M_n = o(\sqrt{n})$ such that

$$\sup_{\boldsymbol{\beta} \in \mathcal{B}_1} \max_{1 \le i \le n} \left| b' \left(\left| \Phi_i^T \boldsymbol{\beta} \right| \right) \right| \le M_n. \tag{11}$$

Additionally the error term $\epsilon_i = y_i - \mu_{y_i}$'s follow the uniform sub-Gaussian distribution, i.e., there exist constants $c_2 > 0$ such that uniformly for all i = 1, ..., n, we have

$$P(|\epsilon_i| \ge t) \le 2\exp(-c_2 t^2) \text{ for any } t > 0.$$
(12)

Assumption 4 (On nonzero function coefficients)

There exist a sequence $c_{f,n}$ that may tend to zero as $n \to \infty$ such that for all $j \in T$, the true nonzero functions f_j satisfy

$$\min_{j \in T} ||f_j||_2 \ge c_{f,n}.$$

We note that Assumption 1 is a standard assumption in high dimensional models, where the design matrix needs to be bounded from above. Assumption 2 is a well-known condition in high-dimension set up on the empirical Gram matrix (Bickel et al., 2009). It is different than the regular eigenvalue condition, since when n < p, the $p \times p$ Gram matrix has rank less than p, thus it must have zero eigenvalues. Therefore, it is not realistic to bound the eigenvalues away from zero for all $\boldsymbol{\nu} \in \mathbb{R}^{p_n m_n}$, but we need to restrict to some space \mathcal{C} . In our set up, \mathcal{C} is the restricted sub-block eigenvalue condition on sub-blocks of the Gram matrix studied by Belloni and Chernozhukov (2013). Though the lower bound and upper bound are imposed on the fixed design matrix, we gave a derivation in supplementary materials that this condition holds when \boldsymbol{X} is drawn from a continuously differentiable density function which is bounded away from 0 and infinity on the domain of \boldsymbol{X} . This result is similar to the results in Huang et al. (2010).

Assumption 3 is a standard assumption to generalised models. (11) and (12) together controls the tail behavior of the responses, and as mentioned by Fan and Tang (2013), ensure a general and broad applicability of the method. Analogous assumptions to (11) can also be seen in Fan et al. (2010) and Bühlmann and van de Geer (2011). Specifically, for example, we have $b(\boldsymbol{\theta}) = \log(1 + \exp(\boldsymbol{\theta}))$. It's easy to verify that both its second and third derivatives have their absolute values all bounded from above by 1. For equation (11), observe that the first derivative is the mean of Bernoulli distribution, and thus it is also bounded. The error term is also bounded by 1, therefore, taking $c_2 = \log(2)$ will make equation (12) satisfy all logistic regression cases. Moreover, bounded second moment in logistic regression ensure that there exist ϵ such that the probability p_i of each observation satisfies $\epsilon .$

Assumption 4 appears often in variable selection methodologies, because intuitively a nonzero function or covariate has to contribute enough to the response in order to be considered nonzero.

Remark 2.1. In assumption 2, $\delta = \beta - \beta_0$ is the difference vector between a β and the true coefficients β_0 , thus we can view C as a restricted neighborhood of β_0 , i.e.,

$$\mathcal{N}_{\boldsymbol{\beta}_0}^{RE} = \{ \boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \le N_n, \ m_n \times \operatorname{card}_B(\boldsymbol{\delta}) \le n^* = o(n) \}.$$

If $\beta \in \mathcal{N}^{RE}_{\beta_0}$, then by assumption 2 we have

$$\frac{(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T \Phi^T \Phi(\boldsymbol{\beta} - \boldsymbol{\beta}_0)}{n \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2} \ge \gamma_0 \gamma_2^{2s_n} m_n^{-1}.$$

This, together with the bounded variance assumption in assumption 3, ensures the restricted strong convexity of the target function, i.e., for a $\beta^* \in \mathcal{N}^{RE}_{\beta_0}$, we have

$$\frac{(\boldsymbol{\beta}^* - \boldsymbol{\beta}_0)^T \Phi^T \boldsymbol{\Sigma}(\boldsymbol{\beta}) \Phi(\boldsymbol{\beta}^* - \boldsymbol{\beta}_0)}{n \|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\|_2^2} \ge \gamma_0 c_1 \gamma_2^{2s_n} m_n^{-1}, \ \forall \ \boldsymbol{\beta} \in \mathcal{N}_{\boldsymbol{\beta}_0}^{RE}.$$
(13)

3 Methodology & Theoretical Properties

We propose a two step procedure for selecting high dimensional additive models with generalised link that has improved convergence rates compared to single stage selection.

3.1 First step: model screening

The objective of this step is to recover the true support T of the additive components. Let \hat{T} be a random support given by a model selection procedure and $|\hat{T}|$ be the number of variables selected. A good model selection procedure should satisfy the common screening consistency conditions

$$T \subset \hat{T}, \ |\hat{T}| = O(s_n), \text{ w.p. converging to 1.}$$
 (14)

There have been many variable selection penalization (Fan et al., 2004; Van de Geer, 2008; Fan et al., 2010; Fan and Lv, 2011) in generalised linear models and (Huang et al., 2010) in linear additive models where this condition holds. Specifically, Fan et al. (2010) satisfies the requirements in (14) in generalised linear models and Huang et al. (2010) also satisfies (14) in additive models with identity link function. In this paper, we show that under mild conditions, by maximizing the log-likelihood with group lassolike penalization, we can select a model that satisfies (14). We also provide a rate of convergence of this first step selection.

Define the objective function to be

$$L(\boldsymbol{\beta}; \lambda_{n1}) = -\frac{1}{n} \sum_{i=1}^{n} \left[y_i \left(\boldsymbol{\beta}^T \boldsymbol{\Phi}_i \right) - b \left(\boldsymbol{\beta}^T \boldsymbol{\Phi}_i \right) \right] + \lambda_{n1} \sum_{j=1}^{p_n} \|\boldsymbol{\beta}_j\|_2.$$
 (15)

Let $\hat{\beta}$ be the optimiser for (15), i.e.

$$\hat{\boldsymbol{\beta}} = \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^{p_n m_n}} L(\boldsymbol{\beta}; \lambda_{n1}).$$

Let
$$\hat{T} = \operatorname{supp}_B(\hat{\boldsymbol{\beta}})$$
.

The objective function is the negative log-likelihood plus the group lasso penalization term, and the parameters are estimated as the minimisers of the objective function. Here the negative log likelihood function is averaged among the n observations to ensure that it is under the same scale as the penalization function.

With this group lasso type penalised log-likelihood, the selected model has the following properties.

Theorem 3.1. Consider the model \hat{T} obtained by minimizing (15). Under Assumptions 1-4, for some constant C and any diverging sequence $\gamma_n > 0$, choose the regularization

parameter

$$\lambda_{n1}^b = C\sqrt{m_n}\sqrt{\frac{\gamma_n + \log(p_n m_n)}{n}}$$

for bounded response (i.e., $|y_i| < c$), and the regularization parameter

$$\lambda_{n1}^{ub} = \sqrt{m_n} \gamma_n \sqrt{\frac{\log(p_n m_n)}{n}}$$

for unbounded sub-Gaussian response, as the sample size increases,

(i) With probability tending to 1,

$$|\hat{T}| = O(s_n)$$

(ii) With probability tending to 1,

$$\sum_{j=1}^{p_n} \left\| \boldsymbol{\beta}_j^0 - \hat{\boldsymbol{\beta}}_j \right\|_2^2 = O_P\left(s_n \gamma_2^{-2s_n} \frac{m_n^2 \log(p_n m_n)}{n} \right) + O(\lambda_{n1}^{b^{-2}} m_n^2 s_n \gamma_2^{-2s_n}) + O(s_n^2 m_n^{1-2d} \gamma_2^{-2s_n})$$

for the bounded response and

$$\sum_{j=1}^{p_n} \left\| \boldsymbol{\beta}_j^0 - \hat{\boldsymbol{\beta}}_j \right\|_2^2 = O_P \left(s_n \gamma_2^{-2s_n} \gamma_n \frac{m_n^2 \log(p_n m_n)}{n} \right) + O(\lambda_{n1}^{ub^2} m_n^2 s_n \gamma_2^{-2s_n}) + O(s_n^2 m_n^{1-2d} \gamma_2^{-2s_n})$$

for any diverging sequence γ_n and unbounded sub-Gaussian response.

(iii) If $s_n \gamma_2^{-2s_n} m_n^2 \log(p_n m_n)/n \ll c_{f,n}$ $(s_n \gamma_2^{-2s_n} m_n^2 \gamma_n \log(p_n m_n)/n \ll c_{f,n}$ in the unbounded case), $\gamma_2^{-2s_n} m_n^2 \lambda_{n1}^2 s_n/m_n \ll c_{f,n}$ and $s_n^2 m_n^{1-2d} \gamma_2^{-2s_n} \ll c_{f,n}$, with probability tending to 1, all nonzero coefficients are selected.

The proof of this theorem is given in supplementary materials.

Remark 3.1. To avoid estimability issues, here the constants C are selected to be large enough such that the number of parameters to be estimated, i.e., the number of selected nonzero functions $|\hat{T}|$ multiplied by the number of basis function m_n should be less than

or equal to n. Moreover, considering the multicollinearity in the design matrix, the constants are chosen such that $m_n \times |\hat{T}| = o(n)$.

Remark 3.2. The additional term γ_n in the convergence rate is due to unboundedness nature of the response variable rather than due to non-linear link function.

Remark 3.3. For the special case, linear (Gaussian) additive model, our results coincide with Huang et al. (2010). The difference is that we study a fixed design with assumptions on the eigenvalues of the design matrix and they studied a random design with assumption on the distribution of the design matrix. We have put further assumption on the eigenvalue due to the divergence of s_n , the number of nonzero variables. In the special case that s_n is fixed, our assumptions coincides with the assumptions in Huang et al. (2010). Another difference is that we include a diverging term γ_n that establishes the rate of convergence with probability converging to one.

There are three terms in the convergence rate: the first term comes from the regression itself, the second term comes from shrinkage, and the third term comes from the spline approximation error.

Remark 3.4. Let $\hat{f}_{nj}(x) = \sum_{k=1}^{m_n} \hat{\beta}_{jk} \phi_k(x)$. We can also state the results of the first selection step in terms of functions, which is a direct consequence of theorem 3.1. First, we have (i) $|\hat{T}| = O(s_n)$ with probability tending to 1, and (ii) if $s_n m_n \gamma_2^{-2s_n} \log(p_n m_n)/n \ll c_{f,n} (s_n m_n \gamma_2^{-2s_n} \gamma_n \log(p_n m_n)/n \ll c_{f,n}$ in the unbounded case), $\lambda_{n1}^2 s_n m_n \gamma_2^{-2s_n} \ll c_{f,n}$ and $s_n^2 m_n^{-2d} \gamma_2^{-2s_n} \ll c_{f,n}$, with probability tending to 1, all nonzero coefficients are selected.

Moreover, by the properties of spline in De Boor (2001), see for example Stone (1986) and Huang et al. (2010), there exist positive constants c_1 and c_2 such that

$$c_1 m_n^{-1} \|\hat{\boldsymbol{\beta}}_{nj} - \boldsymbol{\beta}_{nj}\|_2^2 \le \|\hat{f}_{nj} - f_{nj}\|_2^2 \le c_2 m_n^{-1} \|\hat{\boldsymbol{\beta}}_{nj} - \boldsymbol{\beta}_{nj}\|_2^2, \tag{16}$$

we have

$$\sum_{i=1}^{p_n} \left\| f_j - \hat{f}_{nj} \right\|_2^2 = O_P \left(s_n \gamma_2^{-2s_n} \frac{m_n \log(p_n m_n)}{n} \right) + O(\lambda_{n1}^{b^{-2}} m_n s_n \gamma_2^{-2s_n}) + O(s_n^2 m_n^{-2d} \gamma_2^{-2s_n})$$

for the bounded response case and

$$\sum_{i=1}^{p_n} \left\| f_j - \hat{f}_{nj} \right\|_2^2 = O_P \left(s_n \gamma_2^{-2s_n} \gamma_n \frac{m_n \log(p_n m_n)}{n} \right) + O(\lambda_{n1}^{ub^2} m_n s_n \gamma_2^{-2s_n}) + O(s_n^2 m_n^{-2d} \gamma_2^{-2s_n})$$

for the unbounded case, for any diverging sequence γ_n .

Remark 3.5. The theorem and its remark together tell us under Assumptions 1-4, by choosing proper γ_n , the functions selected by minimizing the first target function satisfy

$$T \subset \hat{T}$$
 and $|\hat{T}| = O(s_n)$

with probability converging to 1, i.e. we obtained screening consistency.

3.2 Second step: Post selection

After we have a "good" initial estimator, we use the adaptive group lasso to recover the true model (Huang et al., 2010) and we are able to achieve selection consistency in probability under some mild assumptions. The adaptive group lasso idea is similar to adaptive lasso (Zou, 2006) which enjoys better theoretical properties than simple lasso. Chatterjee and Lahiri (2013) and Das et al. (2017) studied rate of convergence and other asymptotic properties of the adaptive lasso estimator. Define the objective function to be

$$L_a(\boldsymbol{\beta}; \lambda_{n2}) = -\frac{1}{n} \sum_{i=1}^n \left[y_i \left(\boldsymbol{\beta}^T \boldsymbol{\Phi}_i \right) - b \left(\boldsymbol{\beta}^T \boldsymbol{\Phi}_i \right) \right] + \lambda_{n2} \sum_{j=1}^{p_n} w_{nj} \|\boldsymbol{\beta}_j\|_2, \tag{17}$$

where the weights depend on the screening stage group lasso estimator

$$w_{nj} = + \begin{cases} \|\hat{\boldsymbol{\beta}}_j\|_2^{-1}, & \text{if } \|\hat{\boldsymbol{\beta}}_j\|_2 > 0\\ \infty, & \text{if } \|\hat{\boldsymbol{\beta}}_j\|_2 = 0 \end{cases}$$
 (18)

Let $\hat{\boldsymbol{\beta}}_{AGL}$ be the optimiser for (17), i.e.

$$\hat{\boldsymbol{\beta}}_{AGL} = \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^{m_n p}} L_a(\boldsymbol{\beta}; \lambda_{n2}).$$

For the choice of weights, the first stage estimators need not to be necessarily the solution of group lasso, but could be more general estimators that satisfy following assumptions.

Assumption 5

The initial estimator $\hat{\boldsymbol{\beta}}$ is r_n consistent at zero, i.e.,

$$r_n \max_{j \in T^c} \|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^0\|_2 = O_P(1),$$
 (19)

and there exists a constant c_3 such that

$$\mathbb{P}\left(\min_{j\in T} \|\hat{\boldsymbol{\beta}}\|_{2} \ge c_{3}b_{n1}\right) \to 1,\tag{20}$$

where $b_{n1} = \min_{j \in T} \|\beta_j^0\|_2$.

Assumption 6

Let $s_n^* = p_n - s_n$ be the number of zero components. The tuning parameter λ_{n2} satisfies

$$\frac{\sqrt{\log(s_n^* m_n)}}{n^{1/2} \lambda_{n2} r_n} + \frac{s_n}{\lambda_{n2} r_n m_n^{d+1/2}} + \frac{\lambda_{n2} r_n}{\gamma_n \sqrt{s_n/n}} = o(1)$$
 (21)

for any diverging sequence γ_n .

Assumption 5 gives the restrictions on the initial estimator. We don't require our initial estimator to be the group lasso estimator. Any initial estimator satisfying assumption 5 will be able to make the adaptive group lasso estimator consistently selects and estimates the true nonzero components. However, the rate of convergence of the adaptive group lasso estimator depends on the rate of convergence of the initial estimator, which is assumed to be r_n in assumption 5. Moreover, the initial estimator mustn't have a 0 estimation for the nonzero components, otherwise it will mislead the results in the proceeding step. Assumption 6 put restrictions on the tuning parameter λ_{n2} in the adaptive group lasso step. The first two terms gives the upper bound for λ_{n2} and the third term gives the lower bound. Only with "appropriate" choice of λ_{n2} we can have the selection consistency and estimation consistency.

It worth noting that if we take the group lasso estimator as our initial estimator, assumptions 5 and 6 are automatically satisfied. Specifically, a trivial choice of r_n would be

$$r_n = O_P^{-1} \left(\sqrt{s_n} \gamma_2^{-s_n} \frac{m_n \sqrt{\log(p_n m_n)}}{\sqrt{n}} \right) + O^{-1} (\lambda_{n_1}^b m_n \sqrt{s_n} \gamma_2^{-s_n}) + O^{-1} (s_n m_n^{0.5 - d} \gamma_2^{-s_n})$$

for the bounded response and

$$r_n = O_P^{-1} \left(\sqrt{s_n} \gamma_2^{-s_n} \sqrt{\gamma_n} \frac{m_n \sqrt{\log(p_n m_n)}}{\sqrt{n}} \right) + O^{-1} (\lambda_{n1}^{ub} m_n \sqrt{s_n} \gamma_2^{-s_n}) + O^{-1} (s_n m_n^{0.5 - d} \gamma_2^{-s_n})$$

for the unbounded case and any diverging sequence γ_n , since we observe that for $j \in T^c$, $\hat{\beta}_j$ is either estimated as zero, or has a rate of convergence to β_j bounded by the rate of convergence in theorem (3.1). For equation (20), observe that the rate of convergence of the group lasso estimator is higher order infinitesimal of the minimal signal strength of nonzero coefficients, thus taking $c_3 = 0.5$ is sufficient. In assumption 6, with our trivial

choice of r_n , we are able to find a range of tuning parameters that satisfy equation (21). Therefore, it's reasonable to take the group lasso estimator as an initial estimator for the adaptive group lasso.

Let the notation $\hat{\boldsymbol{\beta}}_n \stackrel{0}{=} \boldsymbol{\beta}^0$ denote that the sign of each $\hat{\boldsymbol{\beta}}_j$ and $\boldsymbol{\beta}_j^0$ are either both zero or both nonzero. Then we have the following asymptotic properties for the adaptive group lasso estimator.

Theorem 3.2. Assume assumptions 1-6 hold, consider the estimator $\hat{\boldsymbol{\beta}}_{AGL}$ by minimizing (17), we have

(i) If $f_{c,n} >> \sqrt{s_n/n}$, the adaptive group lasso consistently selects the true active predictors with probability converging to 1, i.e.,

$$\mathbb{P}\left(\hat{\boldsymbol{\beta}}_{AGL} \stackrel{0}{=} \boldsymbol{\beta}^{0}\right) \to 1. \tag{22}$$

(ii) The rate of convergence of the adaptive group lasso estimator is given by

$$\sum_{j \in T} \|\hat{\boldsymbol{\beta}}_{AGLj} - \boldsymbol{\beta}_j^0\|_2^2 = O_p\left(s_n \gamma_2^{-2s_n} m_n^2 \frac{\log(s_n m_n)}{n}\right) + O(s_n^2 \gamma_2^{-2s_n} m_n^{1-2d}) + O(\lambda_{n2}^2 m_n^2 s_n \gamma_2^{-2s_n})$$

for the bounded response case and

$$\sum_{j \in T} \|\hat{\boldsymbol{\beta}}_{AGLj} - \boldsymbol{\beta}_j^0\|_2^2 = O_p\left(\gamma_n s_n \gamma_2^{-2s_n} m_n^2 \frac{\log(s_n m_n)}{n}\right) + O(s_n^2 \gamma_2^{-2s_n} m_n^{1-2d}) + O(\lambda_{n2}^2 m_n^2 s_n \gamma_2^{-2s_n})$$

for the unbounded response case, where γ_n is any diverging sequence.

The proof of this theorem is given in supplementary materials. It's interesting to compare the adaptive group lasso results with Wang and Tian (2019), who studied the asymptotic properties of the adaptive group lasso for generalized linear models. It worth noting that we considered a more general case by allowing the group size to diverge with

n, and the eigenvalue to be bounded by sequences that depending on n on a broader domain. In the special case that corresponds to their assumptions, our results (Theorem 3.2) coincides with their results.

Similar to the group lasso estimator, we also derive the results for the non-parametric function estimations, stated in the following remark.

Remark 3.6. Let $\hat{f}_{AGLj}(x) = \Phi_j(x)\hat{\boldsymbol{\beta}}_{AGLj}$. We can also state the results of the first selection step in terms of functions, which is a direct consequence of theorem 3.1. First, we have the true nonzero subset is recovered with probability tending to 1. Moreover, by the same properties of spline as in Remark 3.4, we have

$$\sum_{j \in T} \|\hat{f}_{AGLj} - f_j\|_2^2 = O_p \left(s_n \gamma_2^{-2s_n} m_n \frac{\log(s_n m_n)}{n} \right) + O(s_n^2 \gamma_2^{-2s_n} m_n^{-2d}) + O(\lambda_{n2}^2 m_n s_n \gamma_2^{-2s_n})$$

for the bounded response case and

$$\sum_{j \in T} \|\hat{f}_{AGLj} - f_j\|_2^2 = O_p\left(\gamma_n s_n \gamma_2^{-2s_n} m_n \frac{\log(s_n m_n)}{n}\right) + O(s_n^2 \gamma_2^{-2s_n} m_n^{-2d}) + O(\lambda_{n2}^2 m_n s_n \gamma_2^{-2s_n})$$

for the unbounded case, for any diverging sequence γ_n .

The convergence rate for the group lasso estimator is

$$\sum_{j=1}^{p_n} \left\| f_j - \hat{f}_{nj} \right\|_2^2 = O_P\left(s_n \gamma_2^{-2s_n} \frac{m_n \log(p_n m_n)}{n} \right) + O(\lambda_{n1}^{b^2} m_n s_n \gamma_2^{-2s_n}) + O(s_n^2 m_n^{-2d} \gamma_2^{-2s_n}),$$

while for the adaptive group lasso estimator is

$$\sum_{j \in T} \|\hat{f}_{AGLj} - f_j\|_2^2 = O_p\left(s_n \gamma_2^{-2s_n} m_n \frac{\log(s_n m_n)}{n}\right) + O(\lambda_{n2}^2 m_n s_n \gamma_2^{-2s_n}) + O(s_n^2 \gamma_2^{-2s_n} m_n^{-2d})$$

The regression term differs by the size of candidate set. The price we pay by not knowing the true set is $\log(pm_n)$ in the group lasso step, and becomes $\log(s_nm_n)$ in the adaptive

group lasso step, since the initial estimator have recovered a super set of the true set with cardinality $O(s_n)$. The penalty term's difference appears on the tuning parameter, where λ_{n2} is of a smaller order than λ_{n1} with a multiplier of r_n^{-1} . According to our choice of λ_{n2} , it has a trivial upper bound which is of order $O(\lambda_{n1}^2)$. Therefore, the tuning parameter part in the penalty convergence rate term becomes quadratic. The approximation error term is not affected by the adaptive group lasso step.

The adaptive group lasso is important in two reasons: first, with probability tending to 1, this is enable to select the true nonzero components accurately, which is not always the case in group lasso; second, the rate of convergence of the adaptive group lasso estimator is faster than the rate of convergence of the group lasso estimator. The difference in the leading terms are in the order of r_n^{-1} . This makes the adaptive group lasso estimator to achieve a better error with the same sample size, or the same error with a smaller sample size.

The theorem and remark in this section ensure that under mild assumptions, we are able to recover the true model with probability tending to 1 and achieve a rate of convergence better than the initial estimator. Particularly, if the restrictions of n, p_n, m_n and s_n in the previous section satisfy, the group lasso estimator is actually a good initial estimator. Therefore, this two step procedure actually is a complete procedure that gives us a way to do this model selection and estimation on any high-dimensional generalised additive model. However, the procedure is not practically complete without proper selection of the tuning parameter λ . Therefor, we propose a theoretically validated tuning parameter selection in the next section.

4 Tuning parameter selection

One important issue in penalised methods is choosing a proper tuning parameter. It is known that the selection results are sensitive to the choice of tuning parameters. The theoretical results only provide the order of the tuning parameter, which is not very useful in practice. The reason is that the order of a sequence describes the limit properties when n goes to infinity. In reality, our n is a fixed number, so we must have a practical instruction on selecting the tuning parameter.

Despite its importance, there isn't much development for tuning parameter selection in the high dimensional literature. The conventional tuning parameter selection criteria tend to select too many predictors, thus is hard to reach selection consistency. Another reason, especially in group lasso problems, is that the solution path of group lasso is piecewise nonlinear, which makes the testing procedure even harder. Here, we propose the generalised information criterion (GIC) (Zhang et al., 2010; Fan and Tang, 2013) that supports consistent model selection.

Let $\hat{\boldsymbol{\beta}}^{\lambda}$ be the adaptive group lasso solution with tuning parameter λ . The generalised information criterion is defined as

$$GIC(\lambda) = \frac{1}{n} \{ D(\hat{\mu}_{\lambda}; \mathbf{Y}) + a_n |\hat{T}_{\lambda}| \}, \tag{23}$$

where $D(\hat{\mu}_{\lambda}; \mathbf{Y}) = 2\{l(\mathbf{Y}; \mathbf{Y}) - l(\hat{\mu}_{\lambda}; \mathbf{Y})\}$. Here the $l(\boldsymbol{\mu}; \mathbf{Y})$ is the log-likelihood function in equation (3) expressed as a function of the expectation $\boldsymbol{\mu}$ and \mathbf{Y} . $l(\mathbf{Y}; \mathbf{Y})$ represents the saturated model with $\boldsymbol{\mu} = \mathbf{Y}$, and $\hat{\mu}_{\lambda} = b'(\sum_{i=1}^{p_n} \hat{f}_j^{\lambda}(x_{ij})) = b'(\phi \hat{\boldsymbol{\beta}}^{\lambda})$ is our estimated expectation when the tuning parameter is λ . The hyperparameter a_n is to penalise the size of the model. Using GIC, under proper choice of a_n , we are able to select all active predictors consistently.

The importance of the following consistency theorem is that the result in the previous

section guarantees that with probability converging to 1, there exists a λ_{n0} that will be able to identify the true model. Therefore, a good choice of a_n will be able to identify the true model with probability converging to 1. For a support $A \subset \{1, ..., p\}$ such that $|A| \leq q$, where $q \geq s_n$ and q = o(n), let

$$I(\boldsymbol{\beta}(A)) = E\left[\log(f^*/g_A)\right] = \sum_{i=1}^n \left[b'(\Phi_i \boldsymbol{\beta}^0) \Phi_i^T(\boldsymbol{\beta}^0 - \boldsymbol{\beta}(A)) - b(\Phi_i^T \boldsymbol{\beta}^0) + b(\Phi_i^T \boldsymbol{\beta}(A))\right]$$
(24)

be the Kullback-Leibler (KL) divergence between the true model and the selected model, where f^* is the density of the true model, and g_A is the density of the model with population parameter $\beta(A)$. Let $\beta^*(A)$ be the model with the smallest KL divergence over all models with support A, and let

$$\delta_n = \inf_{\substack{A \not\supset T \\ |A| \le q}} \frac{1}{n} I(\boldsymbol{\beta}^*(A)).$$

Here we note that if $T \subset A$, the minimizer is automatically β^0 and thus the KLdivergence is zero. For an underfitted models $T \not\subset A$, δ_n describes how easily one can distinguish the models from the true model by measuring the minimum distance from the true model to the "best estimated models". Later in the theorems we will need to assume lower bounds on δ_n so that we will be able to reach our consistency results. The following theorem proves that GIC works under mild conditions.

Theorem 4.1. Under assumptions 1-6, suppose that $\delta_n q^{-1} R_n^{-1} \to \infty$, $n\delta_n s_n^{-1} a_n^{-1} \to \infty$ and $a_n \psi^{-1} \to \infty$, where R_n and ψ_n are defined in lemma B.3 and lemma B.4, we have, as $n \to \infty$,

$$\mathbb{P}\{\inf_{\lambda \in \Omega_{-} \cup \Omega_{+}} GIC_{a_{n}}(\lambda) > GIC_{a_{n}}(\lambda_{n0})\} \to 1, \tag{25}$$

where

$$\Omega_{-} = \{ \lambda \in [\lambda_{min}, \lambda_{max}] : T_{\lambda} \not\supset T \},\,$$

$$\Omega_{+} = \{ \lambda \in [\lambda_{min}, \lambda_{max}] : T_{\lambda} \supset T \text{ and } T_{\lambda} \neq T \},$$

where T_{λ} is the set of predictors selected by tuning parameter λ . λ_{min} can be chosen as the smallest λ such that the selected model has size q that satisfies the theorem assumption, and λ_{max} simply corresponds to a model with no variables.

The proof of this theorem is given in supplementary materials. In practice, a choice of a_n is proposed to be $m_n \log(\log(n)) \log(p_n)$. We have

Corollary 4.1. Under assumptions 1-6, with choice of $a_n = m_n \log(\log(n)) \log(p_n)$, we have

$$\mathbb{P}\{\inf_{\lambda\in\Omega_{-}\cup\Omega_{+}}GIC_{a_{n}}(\lambda)>GIC_{a_{n}}(\lambda_{n0})\}\to 1.$$

In our two step procedure, there are two tuning parameters to be selected: λ_{n1} in the group lasso step and λ_{n2} in the adaptive group lasso step. The choice of λ_{n2} is of more importance, since λ_{n1} only serve as the parameter in screening. As long as we have a screening step that satisfies (14), we are ready for the adaptive group lasso step. To be simple, we propose to use GIC for selecting both λ_{n1} and λ_{n2} . As a result of the previous theorem, we are able to reach selection consistency.

5 Numerical Properties

In this section we conduct various empirical exercises to illustrate our theoretically guided method in practice. To optimize the group lasso problems, we apply the algorithm named groupwise-majorization-descent (GMD) by Yang and Zou (2015), which approximates the convex log-likelihood part with second order Taylor expansion and solve it with a quadratic function's closed form solution, wrapped in a block coordinate descent algorithm. We made the algorithm in GAM available as a python class, which is accessible at https://github.com/KaixuYang/PenalizedGAM.

As smoothness is a concern in practical GAM computations, we bring the P-spline (Eilers and Marx, 1996) penalty into the model while implementing the model numerically. The P-spline penalty controls the difference between coefficients of consecutive basis functions, and thus yields smoother spline functions.

Specifically, let $l(\boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{y})$ be the loss function in section 3, either the group lasso loss function or the adaptive group lasso loss function. The loss function with smoothness penalty is defined as

$$l_s(\boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{y}) = l(\boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{y}) + \lambda_s \sum_{j=1}^p \boldsymbol{\beta}_j^T \boldsymbol{D} \boldsymbol{\beta}_j,$$
 (26)

where

$$\boldsymbol{D} = \begin{bmatrix} 1 & -1 & 0 & . \\ -1 & 2 & -1 & . \\ 0 & -1 & 2 & . \\ . & . & . & . \end{bmatrix}$$

A slightly modified soft-thresholding function is used to handle the combination of group lasso penalty and the smoothness penalty.

5.1 Simulated Examples

Here we undertake extensive simulation study to see the performance of our proposed two step selection and estimation approach. We investigate the performance of both uncorrelated and correlated covariates and we consider different sample sizes and varying number of predictors in each case.

In this section, we consider three different types of generalized models: the logistic regression (Bernoulli distribution), the Poisson regression (Poisson distribution) and the Gamma regression (Gamma distribution). Through the whole subsection, we choose

l=4 which implies a cubic B-spline. We choose $m_n=9$ for most cases unless stated otherwise. The choice of l and m_n implies that there are $m_n-l=5$ inner knots, which are evenly placed over the empirical percentiles of the training data. In this subsection, we compare the performance of the two-step approach with the Lasso (Tibshirani, 1996), the GAMBoost (Tutz and Binder, 2006) and the GAMSEL (Chouldechova and Hastie, 2015). We implement our two-step approach with our own package mentioned above. The Lasso is implemented with the scikit-learn package in python. The GAMBoost and GAMSEL methods are implemented using their packages in R. In the group lasso step, we choose the tuning parameter corresponding to n_g variables, where n_g is the largest number such that $n_g \times m_n <= n$. This choice prevents estimation issues when we have too many parameters. The GIC procedure is applied in the adaptive group lasso step to select tuning parameters. In the GIC procedure, the tuning parameter selection criterion is defined as

$$GIC(\lambda) = \frac{1}{n} \{ D(\hat{\mu}_{\lambda}; \mathbf{Y}) + a_n |\hat{T}| \}.$$
 (27)

From our results in the previous section, we choose $a_n = (\log \log n)(\log p)m_n$.

5.1.1 Logistic Regression

First, we consider the logistic regression

$$y_i \sim Bernoulli(\theta_i), i = 1, ..., n,$$
 (28)

where $\theta_i = logit^{-1}[\alpha + \sum_{j=1}^p f_j(x_{ij})]$ and x_{ij} is the (i, j) - th element of the design matrix X.

Example 5.1. We first consider the logistic additive model on an independent design matrix case, where each predictor in X is independent of other predictors. Each element of the design matrix is generated from a Unif(-1,1) distribution. We consider 3 dif-

ferent cases with all n, p and s increasing, which coincides with our theory in section 3. Specifically, the three cases are: n=100, p=200 and s=3; n=200, p=500 and s=4; n=300, p=3000 and s=5. A testing sample of size 1000 is generated independently to measure the performance. For all three cases, we have nonzero functions $f_1(x)=5\sin(3x)$, $f_2(x)=-4x^4+9.33x^3+5x^2-8.33x$ and $f_3(x)=x(1-x^2)\exp(3x)-4$. These three general terms include a periodic term, a polynomial term and an exponential term. The last two cases have one more function of $f_4(x)=4x$, a linear term. Finally, the last case has an addition $f_5(x)=4\sin(-5\log(\sqrt{x+3}))$, a complicated composite function. Without loss of generality, the first s functions are set to be nonzero. The constants in the functions are to ensure similar signal strength and smoothness. The other functions $f_{s+1}(x)=...=f_p(x)=0$.

Our results focus on NV, the average number of variables being selected; TPR, the true positive rate (what percent of the truly nonzero variables are selected); FPR, the false positive rate (where percent of the zero variables are selected); and PE, the prediction error. In the logistic regression problem, our metric to measure the prediction error will be the misclassification rate, which is also the measurement in Chouldechova and Hastie (2015). The simulation results are averaged over 100 repetitions.

The simulation results are summarised in table 1 on page 27. Compared with the classical method Lasso and the existing GAM methods GAMSEL and GAMBoost, the two-step approach performs the best in terms of both variable selection and estimation in the high-dimensional set up. The two-step approach performs significantly better in prediction errors. In variable selection, the two-step approach selects the closest number of variables to the ground truth, while keeping the TPR high and FPR low. The existing GAM algorithms have similar TPR but includes too many false positives. The existing GAM algorithms were not intended for very high-dimensional data, and thus fails to handle the variable selection and prediction at the same time. As mentioned in Fan and

Li (2001), the tuning parameter in the Lasso for consistent variable selection is not the same as the tuning parameter for best prediction. We can see this may also be true for the group lasso case, since the estimated nonzero coefficients in the group lasso step are over-penalized. This also proves that an adaptive group lasso step is important, in terms of both variable selection and prediction.

Table 1: Simulation results for the two-step approach compared with the Lasso, GAMSEL and GAMBoost in the three cases of Example 5.1. NV, average number of the variables being selected; TPR, the true positive rate; FPR, the false positive rate; and PE, prediction error (here is the misclassification rate). Results are averaged over 100 repetitions. Enclosed in parentheses are the corresponding standard errors.

	n=100					n=200			n=300			
	p=200				p=500				p=3000			
	s=3			s=4				s=5				
	NV	TPR	FPR	PE	NV	TPR	FPR	PE	NV	TPR	FPR	PE
Two-step	3.56	.920	.004	.148	4.82	.989	.002	.128	4.92	.968	.000	.122
	(1.19)	(.146)	(.005)	(.027)	(1.02)	(.057)	(.002)	(.018)	(0.535)	(.086)	(.000)	(.018)
Lasso	30.0	.920	.138	.249	64.7	.978	.122	.229	85.2	.816	.027	.211
	(17.9)	(.144)	(.090)	(.041)	(19.2)	(.452)	(.039)	(.024)	(68.3)	(.243)	(.022)	(.024)
GAMSEL	10.1	.820	.039	.241	14.0	.943	.021	.214	33.9	.986	.010	.208
	(11.1)	(.209)	(.055)	(.035)	(12.6)	(.112)	(.025)	(.023)	(27.9)	(.065)	(.009)	(.016)
GAMBoost	44.7	.738	.213	.231	85.4	1.00	.164	.196	138	.996	.044	.186
	(4.84)	(.055)	(.025)	(.027)	(6.88)	(000.)	(.014)	(.018)	(9.64)	(.028)	(.003)	(.015)

In practice, the predictors are sometimes correlated to each other. It's interesting to see how well the procedure performs in correlated predictor cases. Therefore, we also perform the same comparison on correlated predictors.

Example 5.2. In this example, we study the case where the design matrix contains correlated predictors. We generate the data in the following way. First we generate each element of $X_{n\times p}$ independently from Unif(-1,1). Then we generate u from Unif(-1,1), independently from $X_{n\times p}$. Then all columns of X are transformed using $X_j = (X_j + tu)/\sqrt{1+t^2}$. This procedure controls the correlation among predictors through t such that $corr(x_{ik}, x_{ij}) = t^2/(1+t^2)$. Here the simulation is run on n = 100, p = 200 and s = 3. All other set-ups are kept same as example 5.1. In our example, we choose $t = \sqrt{3/7}$, where the correlation is 0.3 and $t = \sqrt{7/3}$, where the correlation is 0.7.

The results are summarised in table 2 on page 29. In the correlated cases, all four methods are influenced, more or less. In terms of variable selection, the two-step approach still has the closest number of selected variables. The methods behave differently in terms of TPR and FPR. GAMBoost tends to have greater numbers in both TPR and FPR, while GAMSEL tends to have both lower numbers. The two-step approach balances between those two methods, while maintaining the smallest FPR among all methods. In terms of the prediction error, the two-step approach significantly beats the other methods. The results show good performance of the two-step approach, and again emphasize that the adaptive group lasso step is necessary for better selection and estimation.

This underselection for correlated predictors has been an issue for the lasso and adaptive lasso methods. For nonparametric additive models, Huang et al. (2010) found the same issue when dealing with correlated predictors. Also the NIS proposed by Fan et al. (2011) did not perform well in correlated predictors compared to uncorrelated case. Our two-step approach is not affected too much with the correlation, in terms of both

Table 2: Simulation results for the two-step approach compared with the Lasso, GAM-SEL and GAMBoost in Example 5.2 with correlation 0.3 and 0.7 for n=100, p=200 and s=3. NV, average number of the variables being selected; TPR, the true positive rate; FPR, the false positive rate; and PE, prediction error (here is the misclassification rate). Results are averaged over 100 repetitions. Enclosed in parentheses are the corresponding standard errors.

		Cor	=0.3		Cor=0.7				
	NV	TPR	FPR	PE	NV	TPR	FPR	PE	
Two-step	2.82	.753	.003	.171	2.05	.557	.002	.174	
	(.994)	(.229)	(.004)	(.033)	(.829)	(.170)	(.003)	(.022)	
Lasso	37.0	.690	.176	.312	21.9	.327	.103	.288	
	(38.2)	(.259)	(.194)	(.069)	(37.9)	(.291)	(.193)	(.047)	
GAMSEL	15.4	.573	.069	.342	12.5	.397	.057	.264	
	(16.0)	(.285)	(.079)	(.065)	(9.15)	(.271)	(.044)	(.033)	
GAMBoost	44.2	.977	.209	.268	33.7	.860	.158	.203	
	(5.21)	(.085)	(.026)	(.033)	(4.52)	(.178)	(.014)	(.026)	

variable selection and prediction.

It also happens in the real world that the signal strength is low. Therefore, it is interesting to consider a case where we have lower signal strength than in example 5.1. Example 5.3. In this example, we reduce the signal strength of example 5.1 by a factor of 2, while all other assumptions are kept the same. The results are shown in Table 3 on page 30. From the table we see that minimal signal strength is an important factor to the performance of variable selection in the generalized models. The performance is impacted by the signal strength for all models. The two-step approach still have the closest number of nonzero variables to the ground truth. Though the true positive rate is lower than that of the Lasso or the GAMBoost, the latter two methods have too many false positives. The Lasso or GAMBoost selects too many variables and should not be considered as good variable selection methods. Moreover, the prediction error of the two-step approach remain the best among all four methods.

Table 3: Simulation results for the two-step approach compared with the Lasso, GAM-SEL and GAMBoost in Example 5.3, with n=100, p=200, s=3 and signal strength reduced. NV, average number of the variables being selected; TPR, the true positive rate; FPR, the false positive rate; and PE, prediction error (here is the misclassification rate). Results are averaged over 100 repetitions. Enclosed in parentheses are the corresponding standard errors.

	NV	TPR	FPR	PE
Two-step	3.91	.703	.009	.218
r wo-step	(2.05)	(.240)	(.009)	(.033)
Lasso	30.0	.770	.142	.258
Lasso	(30.5)	(.304)	(.154)	(.036)
GAMSEL	15.3	.510	.070	.377
GAMBEL	(18.0)	(.266)	(.090)	(.054)
GAMBoost	50.3	.980	.240	.308
GAMDOOSU	(5.11)	(.079)	(.026)	(.028)

5.1.2 Other link functions

In this subsection, we study the performance of the two-step approach numerically on the Poisson regression and Gamma regression. In the Poisson regression, we have

$$y_i \sim Poisson(\theta_i), i = 1, ..., n,$$
 (29)

where $\theta_i = \exp[\alpha + \sum_{j=1}^p f_j(x_{ij})]$ and x_{ij} is the (i,j) - th element of the design matrix X. In the Gamma regression, we have

$$y_i \sim Gamma(\theta_i, \phi), \ i = 1, ..., n, \tag{30}$$

where $\theta_i = \exp[\alpha + \sum_{j=1}^p f_j(x_{ij})]$ and x_{ij} is the (i,j) - th element of the design matrix X. The dispersion parameter ϕ is assumed to be known. Without loss of generality, we take $\phi = 1$.

Example 5.4. In this example, we keep the same set up as in example 5.1 to generate the design matrix, and use the Poisson distribution/Gamma distribution above to generate

response variables. All other parameters are kept the same as in example 5.1, but the signal strength is set to 1/4 of the original signal strength, and we set n=100, p=200 and s=3. We compare the two-step approach with generalized linear models (GLM) and the GAMBoost. Note that the GAMSEL only supports Gaussian and Binomial link, thus is not used as a comparison here. The GAMBoost only supports generalized models with canonical link. The canonical link for Gamma regression suffers from the risk that the mean might fall outside of its range, thus the canonical link is not useful in practice. Therefore, we only use GAMBoost in Poisson regression as a comparison. Our algorithm works for both Gamma regression and Poisson regression, and to the best of our knowledge, is the only publicly available algorithm that supports both in the high-dimensional settings. The GLMs are run with the scikit-learn package in python.

The results are provided in Table 4. We see the two-step approach works significantly better than the linear model, and than the GAMBoost in the Poisson regression case, except for the true positive rate. The GAMBoost has a perfect true positive rate, which is slightly better than that of our two-step approach. However, the same issue as before is that it selected too many variables and make the false positive rate much higher than tolerable. Moreover, the prediction performance on the two-step approach is also in the first place in both the cases.

5.2 Real data examples

In this section, we provide three real data examples to illustrate our procedure. In the first example, we consider the case n > p in the classification set up, in the second example, we consider the high-dimensional set up n < p in the classification set up, and in the third example, we consider a Gamma regression model.

Example 5.5. In this example, we use the data set in Example 1 of Friedman et al. (2001), the spam data as an example of the case n > p. The data set is available

Table 4: Simulation results for the two-step approach compared with the Lasso, GAM-SEL and GAMBoost in Example 5.4 for Poisson regression and Gamma regression with $n=100,\ p=200$ and s=3. NV, average number of the variables being selected; TPR, the true positive rate; FPR, the false positive rate; and PE, prediction error (here is the misclassification rate). Results are averaged over 100 repetitions. Enclosed in parentheses are the corresponding standard errors. The GAMBoost method does not support Gamma regression with non-canonical link function, while the canonical link falls outside of range, therefore it does not support Gamma regression.

	F	Poisson F	Regressio	n	Gamma Regression				
	NV	TPR	FPR	PE	NV	TPR	FPR	PE	
Two-step	4.30	.930	.008	2.34	3.57	.997	.003	14.4	
	(1.51)	(.172)	(.009)	(.703)	(0.98)	(.033)	(.005)	(19.5)	
Lasso	13.4	.867	.054	3.51	12.5	.887	.048	42.3	
	(9.79)	(.189)	(.050)	(.403)	(7.72)	(.196)	(.039)	(11.5)	
GAMBoost	82.1	1.00	.401	15.4	NA	NA	NA	NA	
	(4.27)	(000.)	(.022)	(2.12)	INA	INA	INA	INA	

at https://web.stanford.edu/ hastie/ElemStatLearn/data.html. This data set has been studied in many different contexts with the objective being to predict whether an email is a spam or not based on a few features of the emails. There are n=4601 observations, among which 1813 (39.4%) are spams. There are p=57 predictors, including 48 continuous real [0,100] attributes of the relative frequency of 48 'spam' words out of the total number of words in the email, 6 continuous real [0,100] attributes of the relative frequency of 6 'spam' characters out of the total number of characters in the email, 1 continuous real attribute of average length of uninterrupted sequences of capital letters, 1 continuous integer attribute of length of longest uninterrupted sequence of capital letters, and 1 continuous integer attribute of total number of capital letters in the e-mail. The data was first log transformed, since most of the predictors have long-tailed distribution, as mentioned in Friedman et al. (2001). They were then centered and standardised.

The data was split into a training data set with 3067 observations and a testing data set with 1534 observations. We choose order l = 4 which implies a cubis B-spline.

We choose $m_n = 15$, which implies there are 11 inner knots, evenly placed over the empirical percentiles of the data. We compare the result with the logistic regression with Lasso penalty, the support vector machine (SVM) with Lasso penalty, and the sparse group lasso neural network (SGLNN, Feng and Simon (2017), see also Yang and Maiti (2020)). The Lasso and SMV are implemented with the skikit-learn module in python, and the SGLNN is implemented with the algorithm in the paper in python. By changing the tuning parameter or stopping criterion, we get estimations with different sparsity levels. All results are averaged over 50 repetitions. The classification error with different level of sparsity is shown in Figure 1 on page 34. The two-step approach and the neural network perform better than the linear models, which indicates a nonlinear relationship. The two-step approach has maximum accuracy 0.944, while that for the neural network is 0.946. The neural network performs a little better than the two-step approach due to its ability to model the interactions among predictors, but this difference is not significant. However, neural network has no interpretation and takes longer to train. All four methods have performance increase as more predictors are included, which indicates that all predictors contributes to some effect to the prediction. However, we are able to reach more than 0.9 accuracy with only 15 predictors included. With the GIC criterion, the two-step approach selects 14.6 ± 1.52 predictors, with an average accuracy of 0.914 ± 0.015 . The most frequently selected functions are shown in Figure 2 on page 35, which also shows that these functions are truly non-linear. The plots are of the original functions, i.e., before the logarithm transformation. The estimated functions are close to the results in Friedman et al. (2001), Chapter 9, with slight scale difference due to different penalization. The results show that the additive model by the adaptive group lasso is more suitable for this data than linear models.

Example 5.6. For high-dimensional classification example, we use the prostate cancer gene expression data described in http://featureselection.asu.edu/datasets.php. The

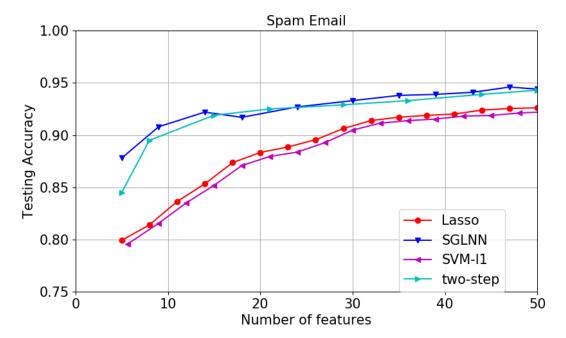


Figure 1: The classification accuracy against the number of nonzero variables measured on a testing set for Example 5.5 over 50 repetitions. The two-step approach, the logistic regression with Lasso, the l_1 norm penalized SVM and the sparse group lasso neural network are included in comparison.

data set has a binary response. 102 observations were studied on 5966 predictor variables, which indicates that the data set is really a high dimensional data set. The responses have values 1 (50 sample points) and 2 (52 sample points), where 1 indicates normal and 2 indicates tumor. All predictors are continuous predictors, with positive values.

To see the performance of our procedure, we ran 100 replications. In each replication, we randomly choose 76 of the observations as training data set and the rest 26 observations as testing data set. We choose order l = 4 which implies a cubis B-spline. We choose $m_n = 9$, which implies there are 5 inner knots, evenly placed over the empirical percentiles of the data. Similar to the last example, we compare the result with the logistic regression with Lasso penalty, the SVM with Lasso penalty, and SGLNN. The classification error with different level of sparsity is shown in Figure 3 on page 36. From the figure we see that compared with linear methods such as the logistic regression or

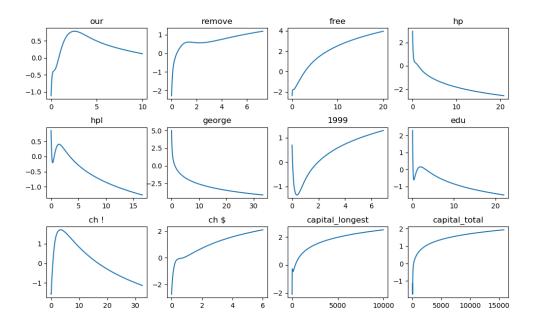


Figure 2: The estimated functions for the most frequently selected functions for Example 5.5.

support vector machine, the non-parametric approaches converges faster. The two-step approach reaches a testing accuracy of 0.945 when around 15 variables are included in the model, while the linear methods need over 30 variables to reach competitive results. Compared with neural network, the two-step approach is easier to implement with stabilized performances. A drawback of the non-parametric methods is to easily overfit for small sample, and that's the reason the performance drops as too many variables entered the into the model. With the GIC criterion, the two-step approach selects 3.25 ± 1.67 predictors, with an average accuracy of 0.914 ± 0.016 . To show the non-linear relationship, figure 4 on page 37 shows the estimated functions for the 6 most frequently selected variables.

Example 5.7. In this example, we investigate the performance of the two-step approach on Gamma regression. The data set is from National Oceanic and Atmospheric Administration (NOAA). We use the storm data, which includes the occurrence of storms in

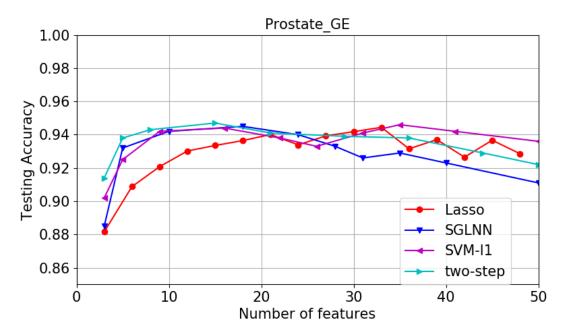


Figure 3: The classification accuracy against the number of nonzero variables measured on a testing set for Example 5.6 over 500 repetitions. The two-step approach, the logistic regression with Lasso, the l_1 norm penalized SVM and the sparse group lasso neural network are included in comparison.

the United States with the time, location, property damage, a narrative description and etc. Here we only take the data in Michigan from 2010 to 2018 and keep the narrative description as our predictor variable and the property damage as our response variable. The description is in text, therefore we applied wording embedding algorithm Word2vec (Mikolov et al., 2013) to transform each description into a numeric representation vector of length p=701, similar word embedding preprocessing can be found in Lee et al. (2020). The response variable property damage has a long tail distribution, thus we use a Gamma regression here. After removing outliers, the data set contains 3085 observations. In order to study the high-dimensional case, we randomly sample 10% of the observations as our training data (n=309) and the rest are used for validation. Moreover, the response is normalized with the location and scale parameters of gamma distribution.

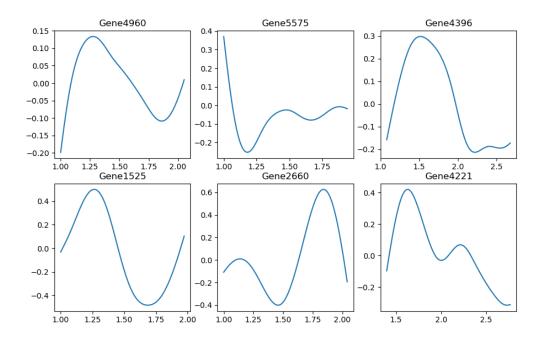


Figure 4: The estimated functions for the most frequently selected functions ordered by descending in frequency for Example 5.6.

To see the performance of our procedure, we ran 50 replications. We choose order l=4 which implies a cubis B-spline. We choose $m_n=9$, which implies there are 5 inner knots, evenly placed over the empirical percentiles of the data. Since there's limited libraries available for variable selection under high-dimensional gamma model, we compare the two-step approach with the linear regression with Lasso on a logarithm transformation on the response variable. The prediction error with different level of sparsity is shown in Figure 5 on page 38. With the GIC criterion, the two-step approach selects 34.45 ± 3.52 predictors, with an average MSE of 0.004334 ± 0.000115 . However, from the plot we see that the linear model was not able to reach this accuracy through the whole solution path, with the best accuracy of 0.004337 at around 80 nonzero variables. This example also shows the superior of the non-parametric model over linear models.

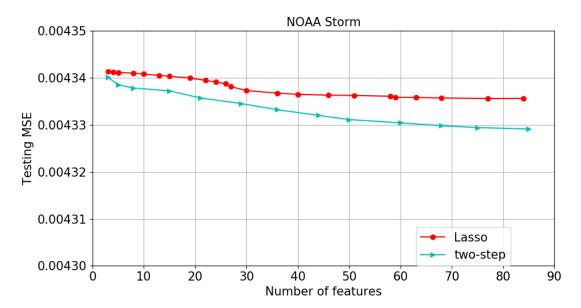


Figure 5: The testing MSE against the number of nonzero variables measured on a testing set for Example 5.7 over 50 repetitions. The two-step approach and logarithm transformation with the Lasso are included in comparison.

6 Discussion

In this paper, we considered ultra high-dimensional ($\log p_n = O(n^{\rho})$) generalised additive model with a diverging number of nonzero functions ($s_n \to 0$ as $n \to \infty$). After using basis expansion on the nonparametric functions, we used two step procedures—group lasso and adaptive group lasso to select the true model. We have proved the screening consistency of the group lasso estimator and the selection consistency of the adaptive group lasso estimator. The rates of convergence of both estimators were also derived, which proved that the adaptive group lasso does have an improvement on the estimator. The whole paper provides a solid foundation for the existing methods. Finally we proved that under this nonparametric set up, the generalised information criterion (GIC) is a good way to select the tuning parameter that consistently selects the true model.

In this paper, we used a fixed design on the data matrix X. A random design on X could be considered, i.e., X has a continuous distribution function $f_X(X)$ on its interval

[a,b], however, extra assumptions such as the boundedness of the density function are needed to reach the same result. Also we proved the selection consistency of the GIC procedure on the adaptive group lasso estimator, conditioning that the initial estimator satisfies (14), which is possessed by the group lasso procedure with probability tending to 1. However, the theory of screening consistency for the group lasso estimator is still to be established. This is a challenging problem, since there doesn't have to exist a tuning parameter that gives selection consistency in the group lasso procedure, but this is an interesting problem that deserves further investigation. We also discussed the subset selection and subset selection with shrinkage under our set up. The theoretical investigation suggests the other penalty functions may not have clear advantages over the proposed procedure.

Moreover, the heteroskedastic error case is also attracting in high-dimensional GAM. The square root Lasso (Belloni et al., 2011) has been proved to overcome this issue, however, it hasn't been extended to the non-parametric set up. It could be interesting to apply square root Lasso on the GAM to incorporate this case. This is a demanding topic that deserves further investigation as well.

References

Amato, U., Antoniadis, A., and De Feis, I. (2016). Additive model selection. *Statistical Methods & Applications*, 25(4):519–564.

Bakin, S. (1999). Adaptive regression and model selection in data mining problems. PhD thesis, School of Mathematical Sciences, Australian National University.

Barzilai, J. and Borwein, J. M. (1988). Two-point step size gradient methods. *IMA journal of numerical analysis*, 8(1):141–148.

- Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. bernoulli 19 521–547. *Mathematical Reviews (Math-SciNet): MR3037163 Digital Object Identifier: doi*, 10.
- Belloni, A., Chernozhukov, V., and Wang, L. (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806.
- Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732.
- Bühlmann, P. and van de Geer, S. (2011). Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer Publishing Company, Incorporated, 1st edition.
- Chatterjee, A. and Lahiri, S. (2013). Rates of convergence of the adaptive lasso estimators to the oracle distribution and higher order refinements by the bootstrap. *The Annals of Statistics*, 41(3):1232–1259.
- Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771.
- Chouldechova, A. and Hastie, T. (2015). Generalized additive model selection. arXiv preprint arXiv:1506.03850.
- Das, D., Gregory, K., and Lahiri, S. (2017). Perturbation bootstrap in adaptive lasso. arXiv preprint arXiv:1703.03165.
- De Boor, C. (2001). A practical guide to splines (revised ed.) springer. New York.
- Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. Statistical science, pages 89–102.

- Fan, J., Feng, Y., and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494):544–557.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.
- Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with np-dimensionality. IEEE Transactions on Information Theory, 57(8):5467–5484.
- Fan, J., Peng, H., et al. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961.
- Fan, J., Song, R., et al. (2010). Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics*, 38(6):3567–3604.
- Fan, Q. and Zhong, W. (2018). Nonparametric additive instrumental variable estimator:

 A group shrinkage estimation perspective. *Journal of Business & Economic Statistics*,

 36(3):388–399.
- Fan, Y. and Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 75(3):531–552.
- Feng, J. and Simon, N. (2017). Sparse-input neural networks for high-dimensional non-parametric regression and classification. arXiv preprint arXiv:1711.07592.
- Friedman, J., Hastie, T., and Tibshirani, R. (2001). The elements of statistical learning, volume 1. Springer series in statistics Springer, Berlin.
- Hastie, T. and Tibshirani, R. (1986). [generalized additive models]: Rejoinder. *Statist.* Sci., 1(3):314–318.

- Huang, J., Horowitz, J. L., and Wei, F. (2010). Variable selection in nonparametric additive models. *Annals of statistics*, 38(4):2282.
- Lee, G. Y., Manski, S., and Maiti, T. (2020). Actuarial applications of word embedding models. ASTIN Bulletin: The Journal of the IAA, 50(1):1–24.
- Liu, R., Yang, L., and Härdle, W. K. (2013). Oracally efficient two-step estimation of generalized additive model. *Journal of the American Statistical Association*, 108(502):619–631.
- Liu, Y. and Wu, Y. (2007). Variable selection via a combination of the 1 0 and 1 1 penalties. *Journal of Computational and Graphical Statistics*, 16(4):782–798.
- Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *The Annals of Statistics*, pages 3498–3528.
- Marra, G. and Wood, S. N. (2011). Practical variable selection for generalized additive models. *Computational Statistics & Data Analysis*, 55(7):2372–2387.
- Mazumder, R., Radchenko, P., and Dedieu, A. (2017). Subset selection with shrinkage: Sparse linear modeling when the snr is low. arXiv preprint arXiv:1708.03288.
- Meier, L., Van De Geer, S., and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71.
- Meier, L., Van de Geer, S., Bühlmann, P., et al. (2009). High-dimensional additive modeling. *The Annals of Statistics*, 37(6B):3779–3821.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

- Nandy, S., Lim, C. Y., and Maiti, T. (2017). Additive model building for spatial regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79(3):779–800.
- Schumaker, L. (1981). Spline functions: basic theory. 1981. John Wiley&Sons, New York.
- Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.
- Stone, C. J. (1985). Additive regression and other nonparametric models. *The annals of Statistics*, pages 689–705.
- Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. *The Annals of Statistics*, pages 590–606.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.
- Tutz, G. and Binder, H. (2006). Generalized additive modeling with implicit variable selection by likelihood-based boosting. *Biometrics*, 62(4):961–971.
- Van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso.

 The Annals of Statistics, pages 614–645.
- van der Vaart, A. and Wellner, J. (1996). Weak Convergence and Empirical Processes:

 With Applications to Statistics. Springer Series in Statistics. Springer.
- Wang, H., Li, B., and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B* (Statistical Methodology), 71(3):671–683.

- Wang, M. and Tian, G.-L. (2019). Adaptive group lasso for high-dimensional generalized linear models. *Statistical Papers*, 60(5):1469–1486.
- Wei, F. and Huang, J. (2010). Consistent group selection in high-dimensional linear regression. Bernoulli: official journal of the Bernoulli Society for Mathematical Statistics and Probability, 16(4):1369.
- Yang, K. and Maiti, T. (2020). Statistical aspects of high-dimensional sparse artificial neural network models. *Machine learning and knowledge extraction*, 2(1):1–19.
- Yang, Y. and Zou, H. (2015). A fast unified algorithm for solving group-lasso penalize learning problems. *Statistics and Computing*, 25(6):1129–1141.
- Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.
- Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, pages 1567–1594.
- Zhang, N. R. and Siegmund, D. O. (2007). A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63(1):22–32.
- Zhang, Y., Li, R., and Tsai, C.-L. (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, 105(489):312–323.
- Zhou, S., Shen, X., Wolfe, D., et al. (1998). Local asymptotics for regression splines and confidence regions. *The annals of statistics*, 26(5):1760–1782.

Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American statistical association, 101(476):1418-1429.

Kaixu Yang
Department of Statistics and Probability
Michigan State University
619 Red Cedar Rd. Room 507
East Lansing, MI, 48824
yangkaix@msu.edu

Tapabrata Maiti

Department of Statistics and Probability

Michigan State University

619 Red Cedar Rd. Room 424

East Lansing, MI, 48824

maiti@msu.edu

Appendix A Derivation of assumption 2

Though assumption 2 is imposed on the fixed design matrix, however, it holds if the design matrix X is drawn from a continuous density and the density g_j of X_j is bounded away from 0 and infinity by b and B, respectively, on the interval [a, b]. Let δ_A be the sub-vector of δ which include all nonzero entries. Without loss of generality, let $\delta_A = \{\delta_1, ..., \delta_k\}$, where $\delta_k \in \mathbb{R}^{m_n}$ and $k = O(s_n)$. Let Φ_A be the corresponding sub-matrix of Φ .

By lemma 3 in Stone (1985), if the design matrix X is drawn from a continuous density and the density g_j of X_j is bounded away from 0 and infinity by b and B, respectively, on the interval [a, b], and $\operatorname{card}_B(\delta) = O(s_n)$, we have

$$\|\Phi_1 \delta_1 + ... + \Phi_k \delta_k\|_2 \ge \gamma_2^{k-1} (\|\Phi_1 \delta_1\|_2 + ... + \|\Phi_k \delta_k\|_2)$$

for some positive constant γ_2 such that $\delta_0 < 1 - 2\gamma_2^2 < 1$, where $\delta_0 = ((1 - bB^{-1})/2)$. Together with the triangle inequality, we have

$$\gamma_2^{k-1}(\|\Phi_1 \boldsymbol{\delta}_1\|_2 + \ldots + \|\Phi_k \boldsymbol{\delta}_k\|_2) \le \|\Phi_A \boldsymbol{\delta}_A\|_2 \le \|\Phi_1 \boldsymbol{\delta}_1\|_2 + \ldots + \|\Phi_k \boldsymbol{\delta}_k\|_2$$

By simple algebra, we have

$$\gamma_2^{2k-2}(\|\Phi_1\boldsymbol{\delta}_1\|_2^2 + \ldots + \|\Phi_k\boldsymbol{\delta}_k\|_2^2) \le \|\Phi_A\boldsymbol{\delta}_A\|_2^2 \le 2(\|\Phi_1\boldsymbol{\delta}_1\|_2^2 + \ldots + \|\Phi_k\boldsymbol{\delta}_k\|_2^2)$$

For any j = 1, ..., k, by lemma 6.2 in Zhou et al. (1998), we have

$$c_1 m_n^{-1} \le \lambda_{min} (n^{-1} \Phi_i^T \Phi_i) \le \lambda_{max} (n^{-1} \Phi_i^T \Phi_i) \le c_2 m_n^{-1}$$

for some c_1 and c_2 . Then we have

$$\begin{split} \frac{\boldsymbol{\delta}^T \Phi^T \Phi \boldsymbol{\delta}}{\|\boldsymbol{\delta}\|_2^2} &= \frac{\|\Phi_A \boldsymbol{\delta}_A\|_2^2}{\|\boldsymbol{\delta}_A\|_2^2} \\ &\geq \frac{\gamma_2^{2k-2} (\|\Phi_1 \boldsymbol{\delta}_1\|_2^2 + \ldots + \|\Phi_k \boldsymbol{\delta}_k\|_2^2)}{\|\boldsymbol{\delta}_A\|_2^2} \\ &= \gamma_2^{2k-2} \left(\frac{\|\Phi_1 \boldsymbol{\delta}_1\|_2^2}{\|\boldsymbol{\delta}_1\|_2^2} \frac{\|\boldsymbol{\delta}_1\|_2^2}{\|\boldsymbol{\delta}_A\|_2^2} + \ldots + \frac{\|\Phi_k \boldsymbol{\delta}_k\|_2^2}{\|\boldsymbol{\delta}_k\|_2^2} \frac{\|\boldsymbol{\delta}_k\|_2^2}{\|\boldsymbol{\delta}_A\|_2^2} \right) \\ &\geq \gamma_2^{2k-2} c_1 n m_n^{-1} \left(\frac{\|\boldsymbol{\delta}_1\|_2^2}{\|\boldsymbol{\delta}_A\|_2^2} + \ldots + \frac{\|\boldsymbol{\delta}_k\|_2^2}{\|\boldsymbol{\delta}_A\|_2^2} \right) \\ &= \gamma_2^{2k-2} c_1 n m_n^{-1} \end{split}$$

Let $\gamma_0 = \gamma_2^{-2} c_1$ and observe that $k = O(s_n)$, we have

$$\frac{\boldsymbol{\delta}^T \Phi^T \Phi \boldsymbol{\delta}}{n \|\boldsymbol{\delta}\|_2^2} \ge \gamma_0 \gamma_2^{2s_n} m_n^{-1}$$

Similarly, we have

$$\begin{split} \frac{\boldsymbol{\delta}^T \Phi^T \Phi \boldsymbol{\delta}}{\|\boldsymbol{\delta}\|_2^2} &= \frac{\|\Phi_A \boldsymbol{\delta}_A\|_2^2}{\|\boldsymbol{\delta}_A\|_2^2} \\ &\leq \frac{2(\|\Phi_1 \boldsymbol{\delta}_1\|_2^2 + \dots + \|\Phi_k \boldsymbol{\delta}_k\|_2^2)}{\|\boldsymbol{\delta}_A\|_2^2} \\ &= 2\left(\frac{\|\Phi_1 \boldsymbol{\delta}_1\|_2^2}{\|\boldsymbol{\delta}_1\|_2^2} \frac{\|\boldsymbol{\delta}_1\|_2^2}{\|\boldsymbol{\delta}_A\|_2^2} + \dots + \frac{\|\Phi_k \boldsymbol{\delta}_k\|_2^2}{\|\boldsymbol{\delta}_k\|_2^2} \frac{\|\boldsymbol{\delta}_k\|_2^2}{\|\boldsymbol{\delta}_A\|_2^2}\right) \\ &\leq 2c_2 n m_n^{-1} \left(\frac{\|\boldsymbol{\delta}_1\|_2^2}{\|\boldsymbol{\delta}_A\|_2^2} + \dots + \frac{\|\boldsymbol{\delta}_k\|_2^2}{\|\boldsymbol{\delta}_A\|_2^2}\right) \\ &= 2c_2 n m_n^{-1} \end{split}$$

Let $\gamma_1 = c_2$, we have

$$\frac{\boldsymbol{\delta}^T \Phi^T \Phi \boldsymbol{\delta}}{n \|\boldsymbol{\delta}\|_2^2} \leq \gamma_1 m_n^{-1}$$

Appendix B Proofs and lemma and theorems

The following lemmas are needed in proving theorems.

Lemma B.1. For any sequence $r_n > 0$, under assumption 1 and 3, we have for bounded response such that $|y_i| < c/2$ that

$$\mathbb{P}\left(\left\|\frac{\Phi^T\left(\mathbf{y}-\boldsymbol{\mu}_y\right)}{n}\right\|_{\infty} \le r_n\right) \ge 1 - 2p_n m_n \exp\left(-\frac{nr_n^2}{2c^2c_{\Phi}^2}\right) \tag{31}$$

Specifically, for a diverging sequence t_n , taking

$$r_n = \sqrt{2}cc_{\Phi}\sqrt{\frac{\log(p_n m_n) + t_n}{n}}$$

we have for response such that $|y_i| < c/2$ that

$$\mathbb{P}\left(\left\|\frac{\Phi^{T}\left(\boldsymbol{y}-\boldsymbol{\mu}_{y}\right)}{n}\right\|_{\infty} \leq r_{n}\right) \geq 1 - 2\exp(-t_{n})$$
(32)

Proof. Observe that

$$\frac{\Phi_j^T \left(\boldsymbol{y} - \boldsymbol{\mu}_y \right)}{n} = \sum_{i=1}^n \left(\frac{\phi_{ij} (y_i - \mu_{y_i})}{n} \right) := \sum_{i=1}^n \gamma(y_i)$$

It's easy to verify that $E\gamma(y_i)=0$ for i=1,...,n and $|\gamma(y_i)|=|\phi_{ij}(y_i-\mu_{y_i})/n|\leq cd_i$ for i=1,...,n. By assumption 1, we have $\sum_{i=1}^n d_i^2 \leq c_{\Phi}^2/n$ for i=1,...,n. Apply

Bonferroni's inequality and Hoeffding's inequality, we have

$$\mathbb{P}\left(\left\|\frac{\Phi^{T}\left(\mathbf{y}-\boldsymbol{\mu}_{y}\right)}{n}\right\|_{\infty} \leq r_{n}\right) = 1 - \mathbb{P}\left(\left\|\frac{\Phi^{T}\left(\mathbf{y}-\boldsymbol{\mu}_{y}\right)}{n}\right\|_{\infty} \geq r_{n}\right) \\
= 1 - \mathbb{P}\left(\left|\bigcup_{j=1}^{m_{n} \times p_{n}} \left\{\left|\frac{\Phi_{j}^{T}\left(\mathbf{y}-\boldsymbol{\mu}_{y}\right)}{n}\right| \geq r_{n}\right\}\right) \\
\geq 1 - \sum_{j=1}^{m_{n} \times p_{n}} \mathbb{P}\left(\left|\frac{\Phi_{j}^{T}\left(\mathbf{y}-\boldsymbol{\mu}_{y}\right)}{n}\right| \geq r_{n}\right) \\
\geq 1 - m_{n} \times p_{n} \times 2 \exp\left(-\frac{nr_{n}^{2}}{2u_{n}^{2}c_{\Phi}^{2}}\right) - c_{2}n^{1-c_{3}c_{4}^{2}}$$

with our choice of

$$r_n = \sqrt{2}cc_{\Phi}\sqrt{\frac{\log(p_n m_n) + t_n}{n}}$$

we have

$$\mathbb{P}\left(\left\|\frac{\Phi^{T}\left(\mathbf{y}-\boldsymbol{\mu}_{y}\right)}{n}\right\|_{\infty} \leq r_{n}\right) \geq 1 - m_{n} \times p_{n} \times 2 \exp\left(-\frac{nr_{n}^{2}}{8c^{2}c_{\Phi}^{2}}\right)$$

$$= 1 - m_{n} \times p_{n} \times 2 \exp\left(-\frac{n2c^{2}c_{\Phi}^{2}(\log(p_{n}m_{n}) + t_{n})}{2c^{2}c_{\Phi}^{2}n}\right)$$

$$= 1 - 2\exp(-t_{n})$$

Lemma B.2. In the unbounded response case, under assumptions 1 and 3, let $T_n = n^{-1} \|\Phi_j^T(\boldsymbol{y} - \boldsymbol{\mu}_y)\|_{\infty} = \max_{j=1,\dots,p_n m_n} n^{-1} |\Phi_j^T(\boldsymbol{y} - \boldsymbol{\mu}_y)|$, we have

$$ET_n = O(1)n^{-1/2}\sqrt{p_n m_n} (33)$$

and then for any diverging sequence a_n ,

$$\mathbb{P}\left(T_n \ge a_n \sqrt{\frac{\log(p_n m_n)}{n}}\right) \to 0 \text{ as } n \to 0$$
(34)

Proof. By the maximal inequality for sub-Gaussian random variables, for example, see Lemmas 2.2.1 and 2.2.2 in van der Vaart and Wellner (1996) and application see lemma 2 of Huang et al. (2010), we have

$$ET_n \le Cn^{-1}\sqrt{\log(p_n m_n)} \max_j \|\Phi_j\|_2$$

Then by assumption 1, we have

$$ET_n = O(1)n^{-1/2}\sqrt{p_n m_n}$$

Since $T_n \geq 0$, by Markov's inequality, we have

$$\mathbb{P}\left(T_n \ge a_n \sqrt{\frac{\log(p_n m_n)}{n}}\right) \le \frac{ET_n}{n^{-1/2} \sqrt{\log(p_n m_n)}} = \frac{C}{a_n} \to 0 \text{ as } n \to \infty$$
 (35)

Remark B.1. From the two lemmas we see that the difference between the bounded response case and the unbounded response case is the upper bound for the maximum of the random errors. For the bounded case, the error could be bounded by

$$r_n = C\sqrt{\frac{\log(p_n m_n) + t_n}{n}}$$

with any diverging sequence t_n . If we take $t_n = O(\log(p_n m_n))$, we have for a different

C, the bounded response errors to be bounded by

$$r_n = C\sqrt{\frac{\log(p_n m_n)}{n}}$$

with probability converging to 1. For the unbounded response case, with probability converging to 1, we need a diverging sequence a_n instead of a constant multiplied to the main term, i.e.,

$$r_n = a_n \sqrt{\frac{\log(p_n m_n)}{n}}$$

This difference is reflected on the choice of the tuning parameter λ .

Proof of theorem 3.1

Proof. First observe that due to the spline approximation, an error is bought into the model. Let $\theta = \sum_{j=1}^{p_n} f_j$ and $\theta^* = \sum_{j=1}^{p_n} f_{nj}$. By the proof of theorem 1 in Huang et al. (2010), we have

$$||f_j - f_{nj}||_{\infty} = O(m_n^{-d})$$

Therefore, we have

$$|\theta - \theta^*| \le \|\sum_{j=1}^{p_n} (f_j - f_{nj})\|_{\infty} \le \sum_{j=1}^{s_n} \|f_j - f_{nj}\|_{\infty} = O(s_n m_n^{-d})$$

Use Taylor expansion on $b'(\theta)$ around θ^* , we have

$$b'(\theta) - b'(\theta^*) = b''(\theta^{**})(\theta - \theta^*)$$

where θ^{**} lies between θ and θ^{*} . By assumption 3, we have

$$|\mu_{y_i} - \mu_{y_i}^*| = |b'(\theta) - b'(\theta^*)| \le c_1^{-1} |\theta - \theta^*| = O(s_n m_n^{-d}), \ i = 1, ..., n$$
(36)

where $\mu_{y_i}^*$ is the mean of the i^{th} observation evaluated at the spline approximated functions. Therefore, we have

$$\|\boldsymbol{\mu}_y - \boldsymbol{\mu}_y^*\|_{\infty} = O(s_n m_n^{-d})$$

As a direct result, we have

$$\frac{1}{n} \|\boldsymbol{\mu}_y - \boldsymbol{\mu}_y^*\|_2^2 = O(s_n^2 m_n^{-2d})$$
(37)

We start with part (i). The proof of this part is similar to the proof of part (i) of theorem 1 in Huang et al. (2010). But because of the non-identity link function, here we have to make some changes. By KKT conditions, a necessary and sufficient condition for $\hat{\beta}$ to be a minimiser of the target function is

$$\begin{cases}
\frac{1}{n} \Phi_k^T (\boldsymbol{y} - \hat{\boldsymbol{\mu}}_{\boldsymbol{y}}^*) = \frac{\lambda_{n1} \hat{\boldsymbol{\beta}}_k}{\|\hat{\boldsymbol{\beta}}_k\|_2}, \ \forall \ k \ s.t. \ \|\hat{\boldsymbol{\beta}}_k\|_2 > 0 \\
\frac{1}{n} \Phi_k^T (\boldsymbol{y} - \hat{\boldsymbol{\mu}}_{\boldsymbol{y}}^*) \in [-\lambda_{n1}, \lambda_{n1}], \ \forall \ k \ s.t. \ \|\hat{\boldsymbol{\beta}}_k\|_2 = 0
\end{cases}$$
(38)

where $\hat{\mu}_y^*$ is the mean of response approximated by splines and evaluated at the solution $\hat{\beta}$ and the second belonging relationship is element-wise. Let

$$s_k = \frac{\Phi_k^T(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_{\boldsymbol{y}}^*)}{n\lambda_{n1}}$$

Then we have

$$\begin{cases} \|s_k\|_2 = 1, \ \forall \ k \ s.t. \ \|\hat{\boldsymbol{\beta}}_k\|_2 > 0 \\ \|s_k\|_2 \le 1, \ \forall \ k \ s.t. \ \|\hat{\boldsymbol{\beta}}_k\|_2 = 0 \end{cases}$$
(39)

We consider the following subsets of $\{1,...,p\}$. Let A_1 be such that

$$\left\{k: \|\hat{\boldsymbol{\beta}}_k\|_2 > 0\right\} \subset A_1 \subset \left\{k: \frac{1}{n} \Phi_k^T (\boldsymbol{y} - \hat{\boldsymbol{\mu}}_{\boldsymbol{y}}^*) = \frac{\lambda_{n1} \hat{\boldsymbol{\beta}}_k}{\|\hat{\boldsymbol{\beta}}_k\|_2}\right\} \cup \{1, ..., s_n\}$$
(40)

Let $A_2 = \{1, ..., p\} \setminus A_1$, $A_3 = A_1 \setminus T$, $A_4 = A_1 \cap T^c$, $A_5 = A_2 \setminus T^c$ and $A_6 = A_2 \cap T^c$. Therefore, the relationships are

$$j \in T \quad j \in T^c$$

$$A_1 \text{: selected } j \text{ and some } j \in T \qquad A_3 \qquad A_4$$

$$A_2 \text{: } j \text{ not in } A_1 \text{ (includes unselected only)} \qquad A_5 \qquad A_6$$

Then we have

$$\Phi_{A_1}^T(\mathbf{y} - \hat{\boldsymbol{\mu}}_{A_1}^*) = S_{A_1} \tag{41}$$

where $S_{A_1} = (S_{K_1}^T, ..., S_{K_{q_1}}^T)^T$, $S_{K_i} = n\lambda_{n_1}s_{k_i}$ and $\hat{\boldsymbol{\mu}}_{A_1}^* = b'(\Phi_{A_1}\hat{\boldsymbol{\beta}}_{A_1})$. Also from the inequality in KKT, we have

$$-C_{A_2} \le \Phi_{A_2}^T (\boldsymbol{y} - \hat{\boldsymbol{\mu}}_{A_1}^*) \le C_{A_2}$$
(42)

where $C_{A_2} = (C_{K_1}^T, ..., C_{k_{q_2}}^T)^T$ and $C_{K_i} = n\lambda_{n_1}\mathbb{1}_{\{\|\hat{\boldsymbol{\beta}}_{K_i}\|_2 = 0\}} \cdot e_{m_n \times 1}$, where all the elements of e are 1. Let $\boldsymbol{\varepsilon}^* = \boldsymbol{y} - \boldsymbol{\mu}_{\boldsymbol{y}}^*$, then from (41) we have

$$\Phi_{A_1}^T(\boldsymbol{\mu}_{\boldsymbol{y}}^* + \boldsymbol{\varepsilon}^* - \hat{\boldsymbol{\mu}}_{A_1}^*) = S_{A_1}$$

use Taylor expansion on μ_y^* around $\hat{\mu}_{A_1}^*$, we have

$$\Phi_{A_1}^T \Sigma_1 \Phi_{A_1} (\boldsymbol{\beta}_{A_1} - \hat{\boldsymbol{\beta}}_{A_1}) + \Phi_{A_1}^T \Sigma_1 \Phi_{A_2} \boldsymbol{\beta}_{A_2} + \Phi_{A_1}^T \boldsymbol{\varepsilon}^* = S_{A_1}$$

where $\Sigma_1 = \Sigma(\boldsymbol{\theta}_1)$, $\boldsymbol{\theta}_1$ lies on the line segment joining $\Phi \boldsymbol{\beta}$ and $\Phi_{A_1} \hat{\boldsymbol{\beta}}_{A_1}$, and $\Sigma(\boldsymbol{\theta}) = \operatorname{diag}(b''(\theta_1), ..., b''(\theta_n))$ is the diagonal variance matrix evaluated at $\boldsymbol{\theta}$. From (42), we have

$$-C_{A_2} \leq \Phi_{A_2}^T \mathbf{\Sigma}_1 \Phi_{A_1} (\boldsymbol{\beta}_{A_1} - \hat{\boldsymbol{\beta}}_{A_1}) + \Phi_{A_2}^T \mathbf{\Sigma}_1 \Phi_{A_2} \boldsymbol{\beta}_{A_2} + \Phi_{A_2}^T \boldsymbol{\varepsilon}^* \leq C_{A_2}$$

Let $\Sigma_{ij} = \Phi_{A_i}^T \Sigma(\boldsymbol{\theta}_1) \Phi_{A_i} / n$, we have

$$\Sigma_{11}(\beta_{A_1} - \hat{\beta}_{A_1}) + \Sigma_{12}\beta_{A_2} = S_{A_1}$$

and

$$-C_{A_2} \leq \boldsymbol{\Sigma}_{21}(\boldsymbol{\beta}_{A_1} - \hat{\boldsymbol{\beta}}_{A_1}) + \boldsymbol{\Sigma}_{22}\boldsymbol{\beta}_{A_2} + \boldsymbol{\Phi}_{A_2}^T\boldsymbol{\varepsilon}^* \leq C_{A_2}$$

With our choice of λ_{n1} , the constants are sufficient large, by lemma 1 in Wei and Huang (2010), the eigenvalues of Σ_{11} are bounded from below. Thus without loss of generality, we assume Σ_{11} is invertible. Then we have

$$\frac{\boldsymbol{\Sigma}_{11}^{-1} S_{A_1}}{n} = \boldsymbol{\beta}_{A_1} - \hat{\boldsymbol{\beta}}_{A_1} + \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12} \boldsymbol{\beta}_{A_2} + \frac{\boldsymbol{\Sigma}_{11}^{-1}}{n} \boldsymbol{\Phi}_{A_1}^T \boldsymbol{\varepsilon}^*$$
 (43)

and

$$\frac{\|\mathbf{\Sigma}^{-1/2}(\boldsymbol{\mu}_{y} - \boldsymbol{\mu}_{y}^{*})\|_{2}}{n} + n\mathbf{\Sigma}_{22}\boldsymbol{\beta}_{A_{2}} - n\mathbf{\Sigma}_{21}\mathbf{\Sigma}11^{-1}\mathbf{\Sigma}_{12}\boldsymbol{\beta}_{A_{2}} \leq C_{A_{2}} - \Phi_{A_{2}}^{T}\boldsymbol{\varepsilon}^{*} - \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}S_{A_{1}} + \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\Phi_{A_{1}}^{T}\boldsymbol{\varepsilon}^{*}$$

$$(44)$$

Define

$$V_{1j} = \frac{1}{\sqrt{n}} \Sigma_{11}^{-1/2} Q_{A_j 1}^T S_{A_j}, \ j = 1, 3, 4$$

and

$$w_k = \Sigma_1^{1/2} (\boldsymbol{I} - \boldsymbol{P}_1) \Sigma_1^{1/2} \Phi_{A_k} \boldsymbol{\beta}_{A_k}, k = 2, ..., 6$$

where

$$m{P}_1 = m{\Sigma}^{1/2} \Phi_{A_1} (\Phi_{A_1}^T m{\Sigma} \Phi_{A_1})^{-1} \Phi_{A_1}^T m{\Sigma}^{1/2}$$

and Q_{A_jk} is the matrix representing the selection of variables in A_k from A_j .

Consider j=4. For any $k \in A_4$, we have $\|\hat{\boldsymbol{\beta}}_k\|_2 > 0$, then $\|s_k\|_2^2 = 1$. Then we have

 $||S_{A_4}||_2^2 = \sum_{k \in A_4} N(A_4)$, where $N(A_4)$ is the number of predictors in A_4 . Thus

$$||V_{14}||_{2}^{2} = \frac{1}{n} ||\Sigma_{11}^{-1/2} Q_{A_{4}1}^{T} S_{A_{4}}||_{2}^{2}$$

$$\geq \frac{1}{n} c_{1} ||Q_{A_{4}1}^{T} S_{A_{4}}||_{2}^{2}$$

$$= c_{1} n \sum_{k \in A_{4}} ||\lambda_{n_{1}} s_{k}||_{2}^{2}$$

$$\geq c_{1} n \lambda_{n_{1}}^{2} (q_{1} - s_{n})$$

That is

$$(q_1 - s_n)^+ \le \frac{\|V_{14}\|_2^2}{c_1 n \lambda_{n_1}^2} \tag{45}$$

Then, we need to find a bound for $||V_{14}||_2^2$ and $q_1 \leq (q_1 - s_n)^+ + s_n$ will be bounded. Using (43) and consider

$$\begin{split} V_{14}^T(V_{14} + V_{13}) &= S_{A_4}^T Q_{A_4 1} \frac{\boldsymbol{\Sigma}_{11}^{-1}}{n} S_{A_1} \\ &= S_{A_4}^T Q_{A_4 1} (\boldsymbol{\beta}_{A_1} - \hat{\boldsymbol{\beta}}_{A_1} + \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12} \boldsymbol{\beta}_{A_2} + \frac{\boldsymbol{\Sigma}_{11}^{-1}}{n} \boldsymbol{\Phi}_{A_1}^T \boldsymbol{\varepsilon}^*) \\ &= S_{A_4}^T Q_{A_4 1} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12} \boldsymbol{\beta}_{A_2} + \frac{S_{A_4}^T Q_{A_4 1} \boldsymbol{\Sigma}_{11}^{-1}}{n} \boldsymbol{\Phi}_{A_1}^T \boldsymbol{\varepsilon}^* + S_{A_4}^T (\boldsymbol{\beta}_{A_4} - \hat{\boldsymbol{\beta}}_{A_4}) \end{split}$$

Observe $\boldsymbol{\beta}_{A_4} = 0$, and

$$S_{A_4}^T \hat{\boldsymbol{\beta}}_{A_4} = \sum_{k \in A_4} \frac{\lambda_{n1} \hat{\boldsymbol{\beta}}_k^T \hat{\boldsymbol{\beta}}_k}{\|\hat{\boldsymbol{\beta}}_k\|_2} = \sum_{k \in A_4} \lambda_{n1} \|\hat{\boldsymbol{\beta}}_k\|_2 > 0$$

we have

$$V_{14}^{T}(V_{14} + V_{13}) \le S_{A_4}^{T}Q_{A_41}\Sigma_{11}^{-1}\Sigma_{12}\boldsymbol{\beta}_{A_2} + \frac{S_{A_4}^{T}Q_{A_41}\Sigma_{11}^{-1}}{n}\Phi_{A_1}^{T}\boldsymbol{\varepsilon}^*$$

On the other hand, by (44),

$$\begin{split} \|w_2\|_2^2 &= \boldsymbol{\beta}_{A_2}^T \boldsymbol{\Phi}_{A_2}^T \boldsymbol{\Sigma}_1^{1/2} (\boldsymbol{I} - \boldsymbol{P}_1) \boldsymbol{\Sigma}_1 (\boldsymbol{I} - \boldsymbol{P}_1) \boldsymbol{\Sigma}_1^{1/2} \boldsymbol{\Phi}_{A_2} \boldsymbol{\beta}_{A_2} \\ &\leq c_1^{-1} \boldsymbol{\beta}_{A_2}^T \boldsymbol{\Phi}_{A_2}^T \boldsymbol{\Sigma}_1^{1/2} (\boldsymbol{I} - \boldsymbol{P}_1) \boldsymbol{\Sigma}_1^{1/2} \boldsymbol{\Phi}_{A_2} \boldsymbol{\beta}_{A_2} \\ &= c_1^{-1} \boldsymbol{\beta}_{A_2}^T \boldsymbol{\Phi}_{A_2}^T \boldsymbol{\Sigma}_1 \boldsymbol{\Phi}_{A_2} \boldsymbol{\beta}_{A_2} + c_1^{-1} \boldsymbol{\beta}_{A_2}^T \boldsymbol{\Phi}_{A_2}^T \boldsymbol{\Sigma}_1 \boldsymbol{\Phi}_{A_1} (\boldsymbol{\Phi}_{A_1}^T \boldsymbol{\Sigma}_1 \boldsymbol{\Phi}_{A_1})^{-1} \boldsymbol{\Phi}_{A_1}^T \boldsymbol{\Sigma}_1 \boldsymbol{\Phi}_{A_2} \boldsymbol{\beta}_{A_2} \\ &= c_1^{-1} \boldsymbol{\beta}_{A_2}^T (n \boldsymbol{\Sigma}_{22} \boldsymbol{\beta}_{A_2} - n \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12} \boldsymbol{\beta}_{A_2}) \\ &\leq c_1^{-1} \boldsymbol{\beta}_{A_2}^T (C_{A_2} - \boldsymbol{\Phi}_{A_2}^T \boldsymbol{\varepsilon}^* - \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} S_{A_1} + \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Phi}_{A_1}^T \boldsymbol{\varepsilon}^*) \\ &= c_1^{-1} \boldsymbol{\beta}_{A_2}^T (C_{A_2} - c_1^{-1} \boldsymbol{\beta}_{A_2}^T (\boldsymbol{\Phi}_{A_2}^T - \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Phi}_{A_1}^T) \boldsymbol{\varepsilon}^* - c_1^{-1} \boldsymbol{\beta}_{A_2}^T \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} S_{A_1} \\ &= c_1^{-1} \boldsymbol{\beta}_{A_2}^T C_{A_2} - c_1^{-1} \boldsymbol{\beta}_{A_2}^T \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} S_{A_1} - c_1^{-1} w_2^T \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\varepsilon}^* \end{split}$$

Then we have

$$V_{14}^{T}(V_{14}+V_{13})+c_{1}\|w_{2}\|_{2}^{2} \leq \left(\frac{S_{A_{4}}^{T}Q_{A_{4}1}\boldsymbol{\Sigma}_{11}^{-1}}{n}\boldsymbol{\Phi}_{A_{1}}^{T}-w_{2}^{T}\boldsymbol{\Sigma}_{1}^{-1}\right)\boldsymbol{\varepsilon}^{*}-S_{A_{3}}^{T}Q_{A_{3}1}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\beta}_{A_{2}}+\boldsymbol{\beta}_{A_{2}}^{T}C_{A_{2}}$$

Define

$$u = \frac{\Phi_{A_1} \Sigma_{11}^{-1} Q_{A_4 1}^T S_{A_4} / n - \Sigma_1^{-1} w_2}{\|\Phi_{A_1} \Sigma_{11}^{-1} Q_{A_4 1}^T S_{A_4} / n - \Sigma_1^{-1} w_2\|_2}$$

Observe

$$\begin{split} &\|\Phi_{A_{1}}^{T} \boldsymbol{\Sigma}_{11}^{-1} Q_{A_{4}1}^{T} S_{A_{4}} / n - \boldsymbol{\Sigma}_{1}^{-1} w_{2} \|_{2} \\ &\leq 2 (\|\Phi_{A_{1}}^{T} \boldsymbol{\Sigma}_{11}^{-1} Q_{A_{4}1}^{T} S_{A_{4}} / n \|_{2}^{2} + \|\boldsymbol{\Sigma}_{1}^{-1} w_{2} \|_{2}^{2}) \\ &\leq 2 \|\Phi_{A_{1}}^{T} \boldsymbol{\Sigma}_{11}^{-1} Q_{A_{4}1}^{T} S_{A_{4}} / n \|_{2}^{2} + 2 c_{1}^{-2} \|w_{2}\|_{2}^{2} \\ &= 2 \|V_{14}\|_{2}^{2} + 2 c_{1}^{-2} \|w_{2}\|_{2}^{2} \end{split}$$

Observe $c_1 < c_1^{-1}$ implies $c_1 < 1$, then

$$||V_{14}||_{2}^{2} + c_{1}||w_{2}||_{2}^{2} + V_{14}^{T}V_{13} \leq (2c_{1}^{-2}||V_{14}||_{2}^{2} + 2c_{1}^{-2}||w_{2}||_{2}^{2})^{1/2}|u^{T}\boldsymbol{\varepsilon}^{*}|$$

$$+ \sqrt{n}||V_{13}||_{2}||\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\beta}_{A_{2}}||_{2} + \lambda_{n1}||\boldsymbol{\beta}_{A_{5}}||_{1}$$
 (46)

By (45), we have

$$||V_{13}||_{2}^{2} = \frac{1}{n} ||\mathbf{\Sigma}_{11}^{-1/2} Q_{A_{31}}^{T} S_{A_{3}}||_{2}^{2}$$

$$\leq c_{1}^{-1} \frac{||Q_{A_{31}} S_{A_{3}}||_{2}^{2}}{n}$$

$$= c_{1}^{-1} \sum_{k \in A_{3}} ||\lambda_{n_{1}} s_{k}||_{2}^{2}$$

$$\leq c_{1}^{-1} n \lambda_{n_{1}}^{2} N(A_{3})$$

By (46), we have

$$||V_{14}||_{2}^{2} + c_{1}||w_{2}||_{2}^{2}$$

$$\leq c_{1}^{-1}(2||V_{14}||_{2}^{2} + 2||w_{2}||_{2}^{2})^{1/2}|u^{T}\boldsymbol{\varepsilon}^{*}| + \sqrt{c_{1}^{-1}n\lambda_{n1}^{2}N(A_{3})}||V_{14}||_{2}$$

$$+ \sqrt{c_{1}^{-1}n\lambda_{n1}^{2}N(A_{3})}||\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\beta}_{A_{2}}||_{2} + \lambda_{n1}||\boldsymbol{\beta}_{A_{5}}||_{1}$$

$$(47)$$

Define

$$B_1 = \sqrt{c_1 n \lambda_{n1}^2 s_n}$$
 and $B_2 = \sqrt{c_1^{-1} n \lambda_{n1}^2 s_n}$

consider the event

$$\mathcal{E} = \left\{ |u^T \varepsilon^*|^2 \le \frac{(|A_1| \vee m_n) c_1^2 n \lambda_{n1}^2}{4m_n} = (|A_1| \vee m_n) \frac{c_1^3 B_1^2}{4s_n m_n} \right\}$$

later we will show that this event holds with probability tending to 1. On the event \mathcal{E} ,

by (45), we have

$$||V_{14}||_2^2 \ge \frac{q_1}{s_n} B_1^2 - B_1^2$$

then

$$|u^T \varepsilon^*|^2 \le \frac{c_1^3 q_1 m_n B_1^2}{4s_n m_n} \le \frac{c_1^3}{4} (\|V_{14}\|_2^2 + B_1^2)$$

and we have

$$\begin{split} c_1^{-1}(2\|V_{14}\|_2^2 + 2\|w_2\|_2^2)^{1/2}|u^T \boldsymbol{\varepsilon}^*| &\leq c_1^{-3}|u^T \boldsymbol{\varepsilon}^*|^2 + \frac{c_1^3}{4}c_1^{-2}(2\|V_{14}\|_2^2 + 2\|w_2\|_2^2) \\ &\leq \frac{1}{4}(\|V_{14}\|_2^2 + B_1^2) + \frac{c_1^3}{4}c_1^{-2}(2\|V_{14}\|_2^2 + 2\|w_2\|_2^2) \\ &\leq \frac{3}{4}\|V_{14}\|_2^2 + \frac{1}{4}B_1^2 + \frac{c_1}{2}\|w_2\|_2^2 \end{split}$$

Then we have

$$||V_{14}||_{2}^{2} + c_{1}||w_{2}||_{2}^{2} \leq \frac{3}{4}||V_{14}||_{2}^{2} + \frac{1}{4}B_{1}^{2} + \frac{c_{1}}{2}||w_{2}||_{2}^{2} + \sqrt{c_{1}^{-1}n\lambda_{n1}^{2}N(A_{3})}||V_{14}||_{2} + \sqrt{c_{1}^{-1}n\lambda_{n1}^{2}N(A_{3})}||\Sigma_{11}^{-1/2}\Sigma_{12}\boldsymbol{\beta}_{A_{2}}||_{2} + \lambda_{n1}||\boldsymbol{\beta}_{A_{5}}||_{1}$$

i.e.

$$||V_{14}||_{2}^{2} + 2c_{1}||w_{2}||_{2}^{2} \leq B_{1}^{2} + 4\sqrt{c_{1}^{-1}n\lambda_{n1}^{2}N(A_{3})}(||V_{14}||_{2} + ||\Sigma_{11}^{-1/2}\Sigma_{12}\boldsymbol{\beta}_{A_{2}}||_{2}) + \lambda_{n1}||\boldsymbol{\beta}_{A_{5}}||_{1}$$

Consider the set A_1 that contains all $\beta_k \neq 0$. We have $q_1 \geq s_n$ and

$$\left\{k: \|\hat{\boldsymbol{\beta}}_k\|_2 > 0 \text{ or } k \notin T^c\right\} \subset A_1 \subset \left\{k: \frac{1}{n} \Phi_k^T (\boldsymbol{y} - \hat{\boldsymbol{\mu}}_{\boldsymbol{y}}^*) = \frac{\lambda_{n1} \hat{\boldsymbol{\beta}}_k}{\|\hat{\boldsymbol{\beta}}_k\|_2} \text{ or } k \notin T^c\right\}$$
(48)

Then we have $A_5 = \emptyset$, $N(A_3) = s_n \le q_1$ and $\boldsymbol{\beta}_{A_2} = \mathbf{0}$. Then we have

$$||V_{14}||_2^2 \le B_1^2 + 4\sqrt{c_1^{-1}n\lambda_{n1}^2 s_n}||V_{14}||_2 = B_1^2 + 4B_2||V_{14}||_2$$

Use the truth that $x^2 \le c + 2bx$ implies $x^2 \le 2c + 4b^2$, we have

$$||V_{14}||_2^2 \le 2B_1^2 + 16B_2^2$$

Then we have from (45) that

$$(q_1 - s_n)^+ \le \frac{\|V_{14}\|_2^2}{c_1 n \lambda_{n1}^2} \le \frac{2B_1^2 + 16B_2^2}{c_1 n \lambda_{n1}^2} = c_5 s_n$$

where $c_5 = (2c_1^2 + 16)/c_1^2$, i.e.

$$(q_1 - s_n)^+ + s_n \le (c_5 + 1)s_n \tag{49}$$

We note that the constant c_5 only depends on c_1 and (48) simply requires larger A_1 , (49) holds for all A_1 satisfying (40). Note that (49) holds if

$$q_1 \le N(A_1 \cup A_5) \le \frac{n}{m_n} \text{ and } |u^T \varepsilon^*|^2 \le \frac{(|A_1| \lor m_n)c_1^2 n \lambda_{n_1}^2}{4m_n}$$
 (50)

So it remains to show that (50) holds with probability tending to 1. Define

$$x_{m}^{*} = \max_{|A|=m} \max_{\|U_{A_{k}}\|_{2}=1, k=1, \dots, m} \left| \boldsymbol{\varepsilon}^{*T} \right|$$

$$\frac{\Phi_{A}(\Phi_{A}^{T} \boldsymbol{\Sigma}_{A} \Phi_{A})^{-1} \bar{S}_{A} - \boldsymbol{\Sigma}_{A}^{-1/2} (\boldsymbol{I} - \boldsymbol{\Sigma}_{A}^{1/2} \Phi_{A} (\Phi_{A}^{T} \boldsymbol{\Sigma}_{A} \Phi_{A})^{-1} \Phi_{A}^{T} \boldsymbol{\Sigma}_{A}^{1/2}) \boldsymbol{\Sigma}_{A}^{1/2} \Phi \boldsymbol{\beta}}{\|\Phi_{A}(\Phi_{A}^{T} \boldsymbol{\Sigma}_{A} \Phi_{A})^{-1} \bar{S}_{A} - \boldsymbol{\Sigma}_{A}^{-1/2} (\boldsymbol{I} - \boldsymbol{\Sigma}_{A}^{1/2} \Phi_{A} (\Phi_{A}^{T} \boldsymbol{\Sigma}_{A} \Phi_{A})^{-1} \Phi_{A}^{T} \boldsymbol{\Sigma}_{A}^{1/2}) \boldsymbol{\Sigma}_{A}^{1/2} \Phi \boldsymbol{\beta}\|_{2}}$$

$$(51)$$

for $|A| = q_1 = m \ge 0$, $\bar{S}_A = (\bar{S}_{A_1}^T, ..., \bar{S}_{A_m}^T)^T$ where $\bar{S}_{A_k} = \lambda_{n1} U_{A_k}$, $||U_{A_k}||_2 = 1$ and Σ_A is the variance matrix evaluated at some θ corresponding to the remainder of the Taylor expansion when the subset A is considered. To simplify the notations, let $Q_A =$

 $\lambda_{n1}\Phi_A(\Phi_A^T\Sigma_A\Phi_A)^{-1}$ and $P_A=\Sigma_A^{1/2}\Phi_A(\Phi_A^T\Sigma_A\Phi_A)^{-1}\Phi_A^T\Sigma_A^{1/2}$, then we have

$$x_{m}^{*} = \max_{|A|=m} \max_{\|U_{A_{k}}\|_{2}=1, k=1,...,m} \left| \boldsymbol{\varepsilon}^{*T} \frac{Q_{A}U_{A} - \boldsymbol{\Sigma}_{A}^{-1/2} (\boldsymbol{I} - P_{A}) \boldsymbol{\Sigma}_{A}^{1/2} \boldsymbol{\Phi} \boldsymbol{\beta}}{\|Q_{A}U_{A} - \boldsymbol{\Sigma}_{A}^{-1/2} (\boldsymbol{I} - P_{A}) \boldsymbol{\Sigma}_{A}^{1/2} \boldsymbol{\Phi} \boldsymbol{\beta}\|_{2}} \right|$$
(52)

Define

$$\Omega_{m_0}^* = \{(U, \boldsymbol{\varepsilon}^*) : x_m^* \le C\sqrt{(|A| \lor 1)m_n \log(p_n m_n)}, \forall m = |A| \ge m_0\}$$

and

$$\Omega_{m_0} = \{(U, \boldsymbol{\varepsilon}) : x_m^{**} \le C\sqrt{(|A| \lor 1)m_n \log(p_n m_n)}, \forall m = |A| \ge m_0\}$$

for a large enough generic constant C, where

$$x_m^{**} = \max_{|A|=m} \max_{\|U_{A_k}\|_2 = 1, k = 1, \dots, m} \left| \boldsymbol{\varepsilon}^T \frac{Q_A U_A - \boldsymbol{\Sigma}_A^{-1/2} (\boldsymbol{I} - P_A) \boldsymbol{\Sigma}_A^{1/2} \boldsymbol{\Phi} \boldsymbol{\beta}}{\|Q_A U_A - \boldsymbol{\Sigma}_A^{-1/2} (\boldsymbol{I} - P_A) \boldsymbol{\Sigma}_A^{1/2} \boldsymbol{\Phi} \boldsymbol{\beta} \|_2} \right|$$

By triangle inequality and Cauchy-Schwarz inequality, we have

$$\left| \boldsymbol{\varepsilon}^{*T} \frac{Q_{A}U_{A} - \boldsymbol{\Sigma}_{A}^{-1/2} (\boldsymbol{I} - P_{A}) \boldsymbol{\Sigma}_{A}^{1/2} \boldsymbol{\Phi} \boldsymbol{\beta}}{\|Q_{A}U_{A} - \boldsymbol{\Sigma}_{A}^{-1/2} (\boldsymbol{I} - P_{A}) \boldsymbol{\Sigma}_{A}^{1/2} \boldsymbol{\Phi} \boldsymbol{\beta}\|_{2}} \right|$$

$$\leq \left| \boldsymbol{\varepsilon}^{T} \frac{Q_{A}U_{A} - \boldsymbol{\Sigma}_{A}^{-1/2} (\boldsymbol{I} - P_{A}) \boldsymbol{\Sigma}_{A}^{1/2} \boldsymbol{\Phi} \boldsymbol{\beta}}{\|Q_{A}U_{A} - \boldsymbol{\Sigma}_{A}^{-1/2} (\boldsymbol{I} - P_{A}) \boldsymbol{\Sigma}_{A}^{1/2} \boldsymbol{\Phi} \boldsymbol{\beta}\|_{2}} \right| + \|\boldsymbol{\theta}_{n}\|_{2}$$

$$\leq \left| \boldsymbol{\varepsilon}^{T} \frac{Q_{A}U_{A} - \boldsymbol{\Sigma}_{A}^{-1/2} (\boldsymbol{I} - P_{A}) \boldsymbol{\Sigma}_{A}^{1/2} \boldsymbol{\Phi} \boldsymbol{\beta}}{\|Q_{A}U_{A} - \boldsymbol{\Sigma}_{A}^{-1/2} (\boldsymbol{I} - P_{A}) \boldsymbol{\Sigma}_{A}^{1/2} \boldsymbol{\Phi} \boldsymbol{\beta}\|_{2}} \right| + Cn^{1/2} s_{n} m_{n}^{-d}$$

$$\leq \left| \boldsymbol{\varepsilon}^{T} \frac{Q_{A}U_{A} - \boldsymbol{\Sigma}_{A}^{-1/2} (\boldsymbol{I} - P_{A}) \boldsymbol{\Sigma}_{A}^{1/2} \boldsymbol{\Phi} \boldsymbol{\beta}}{\|Q_{A}U_{A} - \boldsymbol{\Sigma}_{A}^{-1/2} (\boldsymbol{I} - P_{A}) \boldsymbol{\Sigma}_{A}^{1/2} \boldsymbol{\Phi} \boldsymbol{\beta}} \right|_{2}} \right| + C\sqrt{(|A| \vee 1) m_{n} \log(p_{n} m_{n})}$$

Then we have

$$(U, \boldsymbol{\varepsilon}) \in \Omega_{m_0} \Rightarrow (U, \boldsymbol{\varepsilon}^*) \in \Omega_{m_0}^* \Rightarrow |u^T \boldsymbol{\varepsilon}^*|^2 \le |x_m^*|^2 \le \frac{(|A_1| \vee m_n)c_1^2 n \lambda_{n_1}^2}{4m_n} \text{ for } q_1 \ge m_0 \ge 0$$

Since ϵ_i 's are sub-Gaussian random variables by assumption 2, we have

$$1 - \mathbb{P}\left((U, \boldsymbol{\varepsilon}) \in \Omega_{q}\right)$$

$$= \mathbb{P}\left(x_{m}^{**} > C\sqrt{(m \vee 1)m_{n}\log(p_{n}m_{n})}, \forall m = |A| \geq m_{0}\right)$$

$$\leq \sum_{m=0}^{\infty} \mathbb{P}\left(x_{m}^{**} > C\sqrt{(m \vee 1)m_{n}\log(p_{n}m_{n})}\right)$$

$$\leq \sum_{m=0}^{\infty} \binom{p_{n}}{m} \mathbb{P}\left(\left|\boldsymbol{\varepsilon}^{T} \frac{Q_{A}U_{A} - \boldsymbol{\Sigma}_{A}^{-1/2}(\boldsymbol{I} - P_{A})\boldsymbol{\Sigma}_{A}^{1/2}\boldsymbol{\Phi}\boldsymbol{\beta}}{\|Q_{A}U_{A} - \boldsymbol{\Sigma}_{A}^{-1/2}(\boldsymbol{I} - P_{A})\boldsymbol{\Sigma}_{A}^{1/2}\boldsymbol{\Phi}\boldsymbol{\beta}\|_{2}}\right| > C\sqrt{(m \vee 1)m_{n}\log(p_{n}m_{n})}$$

$$\leq 2\sum_{m=0}^{\infty} \binom{p_{n}}{m} \exp\left(-C(m \vee 1)m_{n}\log(p_{n}m_{n})\right)$$

$$= 2(p_{n}m_{n})^{-Cm_{n}} + 2\sum_{m=1}^{\infty} \binom{p_{n}}{m}(p_{n}m_{n})^{-Cm_{n}}$$

$$\leq 2(p_{n}m_{n})^{-Cm_{n}} + 2\sum_{m=1}^{\infty} \frac{1}{m!} \left(\frac{p_{n}}{(p_{n}m_{n})^{Cm_{n}}}\right)^{m}$$

$$= 2(p_{n}m_{n})^{-Cm_{n}} + 2\exp\left(\frac{p_{n}}{(p_{n}m_{n})^{Cm_{n}}}\right) - 2 \to 0 \text{ as } n \to \infty$$

Therefore, the proof of part (i) is complete.

Then we prove part (ii). Consider the bounded response case. For a sequence N_n such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_2 \le N_n$, define $t = N_n/(N_n + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_2)$, then consider the convex combination $\boldsymbol{\beta}^* = t\hat{\boldsymbol{\beta}} + (1-t)\boldsymbol{\beta}^0$. We have $\boldsymbol{\beta}^* - \boldsymbol{\beta}^0 = t(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)$, which implies

$$\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^0\|_2 = t\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_2 = \frac{N_n\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_2}{N_n + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_2} \le N_n$$
 (53)

Recall the log likelihood function

$$l_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \left[y_i \left(\alpha + \boldsymbol{\beta}^T \Phi_i \right) - b \left(\alpha + \boldsymbol{\beta}^T \Phi_i \right) \right]$$

$$l_n(\boldsymbol{\beta}^*) = l_n(\boldsymbol{\beta}^0) + \frac{1}{n} \sum_{i=1}^n \left[y_i \Phi_i - \mu_{y_i}^* \Phi_i \right]^T (\boldsymbol{\beta}^* - \boldsymbol{\beta}^0)$$

$$- \frac{1}{2n} \sum_{i=1}^n (\boldsymbol{\beta}^* - \boldsymbol{\beta}^0)^T \Phi_i^T b'' (\alpha + \boldsymbol{\beta}^{**T} \Phi_i) \Phi_i (\boldsymbol{\beta}^* - \boldsymbol{\beta}^0)$$

$$= l_n(\boldsymbol{\beta}^0) + \frac{(\boldsymbol{\beta}^* - \boldsymbol{\beta}^0)^T \Phi^T (\boldsymbol{y} - \boldsymbol{\mu}_y^*)}{n} - \frac{1}{2n} (\boldsymbol{\beta}^* - \boldsymbol{\beta}^0)^T \Phi^T \boldsymbol{\Sigma} (\boldsymbol{\beta}^{**}) \Phi (\boldsymbol{\beta}^* - \boldsymbol{\beta}^0) \quad (54)$$

where $\boldsymbol{\beta}^{**}$ lines on the line joining $\boldsymbol{\beta}^{*}$ and $\boldsymbol{\beta}^{0}$, and

$$\Sigma(\boldsymbol{\beta}^{**}) = \operatorname{diag}\left(b''(\alpha + \boldsymbol{\beta}^{**T}\Phi_1), ...b''(\alpha + \boldsymbol{\beta}^{**T}\Phi_n)\right)$$

is the variance matrix of response when the coefficients take value on β^{**} . On the other hand, by convexity of the log likelihood function,

$$l_n(\boldsymbol{\beta}^*) = l_n(t\hat{\boldsymbol{\beta}} + (1-t)\boldsymbol{\beta}^0) \ge tl_n(\hat{\boldsymbol{\beta}}) + (1-t)l_n(\boldsymbol{\beta}^0)$$

by norm inequality, we have

$$\sum_{j=1}^{p_n} \|\boldsymbol{\beta}_j^*\|_2 = \sum_{j=1}^{p_n} \|t\hat{\boldsymbol{\beta}}_j + (1-t)\boldsymbol{\beta}_j^0\|_2 \le \sum_{j=1}^{p_n} (t\|\hat{\boldsymbol{\beta}}_j\|_2 + (1-t)\|\boldsymbol{\beta}_j^0\|_2)$$

joining the two inequalities above and by the definition of $\hat{\boldsymbol{\beta}}$ gives

$$l_n(\boldsymbol{\beta}^*) - \lambda_{n1} \sum_{j=1}^{p_n} \|\boldsymbol{\beta}_j^*\|_2 \ge t l_n(\hat{\boldsymbol{\beta}}) + (1-t) l_n(\boldsymbol{\beta}_j^0) - \lambda_{n1} \sum_{j=1}^{p_n} (t \|\hat{\boldsymbol{\beta}}_j\|_2 + (1-t) \|\boldsymbol{\beta}_j^0\|_2) \ge l_n(\boldsymbol{\beta}^0) - \lambda_{n1} \sum_{j=1}^{p_n} \|\boldsymbol{\beta}_j^0\|_2$$

which implies

$$l_n(\boldsymbol{\beta}^*) - l_n(\boldsymbol{\beta}^0) \ge \lambda_{n1} \sum_{j=1}^{p_n} \|\boldsymbol{\beta}_j^*\|_2 - \lambda_{n1} \sum_{j=1}^{p_n} \|\boldsymbol{\beta}_j^0\|_2$$
 (55)

By (54) and (55) together we have

$$\lambda_{n1} \sum_{j=1}^{p_n} \left(\|\boldsymbol{\beta}_j^*\|_2 - \|\boldsymbol{\beta}_j^0\|_2 \right) \leq \frac{(\boldsymbol{\beta}^* - \boldsymbol{\beta}^0)^T \Phi^T (\boldsymbol{y} - \boldsymbol{\mu}_{\boldsymbol{y}}^*)}{n} - \frac{1}{2n} (\boldsymbol{\beta}^* - \boldsymbol{\beta}^0)^T \Phi^T \boldsymbol{\Sigma} (\boldsymbol{\beta}^{**}) \Phi (\boldsymbol{\beta}^* - \boldsymbol{\beta}^0)$$

and move one term to the left hand side, we have

$$\frac{1}{2n} (\boldsymbol{\beta}^* - \boldsymbol{\beta}^0)^T \Phi^T \boldsymbol{\Sigma} (\boldsymbol{\beta}^{**}) \Phi (\boldsymbol{\beta}^* - \boldsymbol{\beta}^0)
\leq \frac{(\boldsymbol{\beta}^* - \boldsymbol{\beta}^0)^T \Phi^T (\boldsymbol{y} - \boldsymbol{\mu}_y^*)}{n} + \lambda_{n1} \sum_{j=1}^{p_n} (\|\boldsymbol{\beta}_j^0\|_2 - \|\boldsymbol{\beta}_j^*\|_2)
= \frac{(\boldsymbol{\beta}^* - \boldsymbol{\beta}^0)^T \Phi^T (\boldsymbol{y} - \boldsymbol{\mu}_y)}{n} + \frac{(\boldsymbol{\beta}^* - \boldsymbol{\beta}^0)^T \Phi^T (\boldsymbol{\mu}_y^* - \boldsymbol{\mu}_y)}{n} + \lambda_{n1} \sum_{j=1}^{p_n} (\|\boldsymbol{\beta}_j^0\|_2 - \|\boldsymbol{\beta}_j^*\|_2)
(56)$$

We have for the second term

$$\frac{(\boldsymbol{\beta}^{*} - \boldsymbol{\beta}^{0})^{T} \Phi^{T} (\boldsymbol{\mu}_{y}^{*} - \boldsymbol{\mu}_{y})}{n} \\
= \frac{(\boldsymbol{\beta}^{*} - \boldsymbol{\beta}^{0})^{T} \Phi^{T} \Sigma (\boldsymbol{\beta}^{**})^{1/2} \Sigma (\boldsymbol{\beta}^{**})^{-1/2} (\boldsymbol{\mu}_{y}^{*} - \boldsymbol{\mu}_{y})}{n} \\
\leq \frac{\|\Sigma (\boldsymbol{\beta}^{**})^{1/2} \Phi (\boldsymbol{\beta}^{*} - \boldsymbol{\beta}^{0})\|_{2} \|\Sigma (\boldsymbol{\beta}^{**})^{-1/2} (\boldsymbol{\mu}_{y}^{*} - \boldsymbol{\mu}_{y})\|_{2}}{n} \\
\leq \frac{(\boldsymbol{\beta}^{*} - \boldsymbol{\beta}^{0})^{T} \Phi^{T} \Sigma (\boldsymbol{\beta}^{**}) \Phi (\boldsymbol{\beta}^{*} - \boldsymbol{\beta}^{0})}{4n} + \frac{\|\Sigma (\boldsymbol{\beta}^{**})^{-1/2} (\boldsymbol{\mu}_{y}^{*} - \boldsymbol{\mu}_{y})\|_{2}^{2}}{n} \\
\leq \frac{(\boldsymbol{\beta}^{*} - \boldsymbol{\beta}^{0})^{T} \Phi^{T} \Sigma (\boldsymbol{\beta}^{**}) \Phi (\boldsymbol{\beta}^{*} - \boldsymbol{\beta}^{0})}{4n} + c_{1} d_{n} \tag{57}$$

where $d_n = O(s_n^2 m_n^{-2d})$, the first inequality follows from Cauchy-Schwarz inequality, the second inequality follows from the identity $uv \leq u^2/4 + v^2$, and the third inequality

follow from assumption 3 and (37). Then joining (56) and (57), we have

$$\frac{(\boldsymbol{\beta}^* - \boldsymbol{\beta}^0)^T \Phi^T \boldsymbol{\Sigma} (\boldsymbol{\beta}^{**}) \Phi (\boldsymbol{\beta}^* - \boldsymbol{\beta}^0)}{4n} \leq \frac{(\boldsymbol{\beta}^* - \boldsymbol{\beta}^0)^T \Phi^T (\boldsymbol{y} - \boldsymbol{\mu}_y)}{n} + \lambda_{n1} \sum_{j=1}^{p_n} (\|\boldsymbol{\beta}_j^0\|_2 - \|\boldsymbol{\beta}_j^*\|_2) + c_1 d_n \tag{58}$$

For the first term on the right hand side of (58), we have

$$\frac{\left(\boldsymbol{\beta}^{*} - \boldsymbol{\beta}^{0}\right)^{T} \Phi^{T} \left(\boldsymbol{y} - \boldsymbol{\mu}_{y}\right)}{n}$$

$$= \frac{\left(\boldsymbol{\beta}^{*} - \boldsymbol{\beta}^{0}\right)^{T} \Phi^{T} \boldsymbol{\Sigma} (\boldsymbol{\beta}^{**})^{1/2} \boldsymbol{\Sigma} (\boldsymbol{\beta}^{**})^{-1/2} \left(\boldsymbol{y} - \boldsymbol{\mu}_{y}\right)}{n}$$

$$\leq \frac{\left(\boldsymbol{\beta}^{*} - \boldsymbol{\beta}^{0}\right)^{T} \Phi^{T} \boldsymbol{\Sigma} (\boldsymbol{\beta}^{**}) \Phi (\boldsymbol{\beta}^{*} - \boldsymbol{\beta}^{0})}{8n} + \frac{2\|\boldsymbol{\Sigma} (\boldsymbol{\beta}^{**})^{-1/2} \left(\boldsymbol{y} - \boldsymbol{\mu}_{y}\right)\|_{2}^{2}}{n} \tag{59}$$

where the inequality is by the identity $a^Tb \leq ||a||_2^2/8 + 2||b||_2^2$. Joining (61) and (59), we have

$$\frac{(\boldsymbol{\beta}^* - \boldsymbol{\beta}^0)^T \Phi^T \boldsymbol{\Sigma} (\boldsymbol{\beta}^{**}) \Phi (\boldsymbol{\beta}^* - \boldsymbol{\beta}^0)}{8n} \le \frac{2 \|\boldsymbol{\Sigma} (\boldsymbol{\beta}^{**})^{-1/2} (\boldsymbol{y} - \boldsymbol{\mu}_y) \|_2^2}{n} + \lambda_{n1} \sum_{j=1}^{p_n} (\|\boldsymbol{\beta}_j^0\|_2 - \|\boldsymbol{\beta}_j^*\|_2) + c_1 d_n$$
(60)

By remark 2.1, we have

$$\frac{\gamma_0 c_1 \gamma_2^{2s_n} m_n^{-1}}{8} \|\boldsymbol{\beta}^* - \boldsymbol{\beta}^0\|_2^2 \le \frac{2 \|\boldsymbol{\Sigma}(\boldsymbol{\beta}^{**})^{-1/2} (\boldsymbol{y} - \boldsymbol{\mu}_y)\|_2^2}{n} + \lambda_{n1} \sum_{j=1}^{p_n} (\|\boldsymbol{\beta}_j^0\|_2 - \|\boldsymbol{\beta}_j^*\|_2) + c_1 d_n$$
(61)

Observe that

$$\begin{split} \| \boldsymbol{\Sigma} (\boldsymbol{\beta}^{**})^{-1/2} \left(\boldsymbol{y} - \boldsymbol{\mu}_{y} \right) \|_{2}^{2} &\leq c_{1}^{-1} \| \boldsymbol{y} - \boldsymbol{\mu}_{y} \|_{2}^{2} \\ &\leq \frac{c_{1}^{-1} m_{n}}{\gamma_{0} \gamma_{2}^{2s_{n}}} \| \Phi^{T} (\boldsymbol{y} - \boldsymbol{\mu}_{y}) \|_{2}^{2} \end{split}$$

Then by lemma B.1, we have

$$\frac{\gamma_0 c_1 \gamma_2^{2s_n} m_n^{-1}}{8} \|\boldsymbol{\beta}_{\{T \cup \hat{T}\}}^* - \boldsymbol{\beta}_{\{T \cup \hat{T}\}}^0\|_2^2 \le O_P\left(s_n m_n \frac{\log(p_n m_n)}{n \gamma_2^{2s_n}}\right) + \lambda_{n1} \sum_{j=1}^{p_n} \left(\|\boldsymbol{\beta}_j^0\|_2 - \|\boldsymbol{\beta}_j^*\|_2\right) + O(s_n^2 m_n^{-2d})$$
(62)

Observe that

$$\lambda_{n1} \sum_{j=1}^{p_{n}} \left(\|\boldsymbol{\beta}_{j}^{0}\|_{2} - \|\boldsymbol{\beta}_{j}^{*}\|_{2} \right) \\
\leq \lambda_{n1} \sum_{j \in T \cup \hat{T}} \|\boldsymbol{\beta}_{j}^{0} - \boldsymbol{\beta}_{j}^{*}\|_{2} \\
\leq \lambda_{n1} \sqrt{s_{n}} \|\boldsymbol{\beta}_{\{T \cup \hat{T}\}}^{*} - \boldsymbol{\beta}_{\{T \cup \hat{T}\}}^{0} \|_{2} \\
\leq \frac{\gamma_{0} c_{1} \gamma_{2}^{2s_{n}} m_{n}^{-1}}{16} \|\boldsymbol{\beta}_{\{T \cup \hat{T}\}}^{*} - \boldsymbol{\beta}_{\{T \cup \hat{T}\}}^{0} \|_{2}^{2} + \frac{4 \lambda_{n1}^{2} s_{n}}{\gamma_{0} c_{1} \gamma_{2}^{2s_{n}} m_{n}^{-1}} \tag{63}$$

where the first two inequalities are by norm inequality, and the third inequality is by the identity $a^Tb \leq ||a||_2^2 + ||b||_2^2/4$. Joining (62) and (63), we have

$$\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^0\|_2^2 = O_P\left(s_n \gamma_2^{-2s_n} \frac{m_n^2 \log(p_n m_n)}{n}\right) + O(\lambda_{n1}^2 m_n^2 s_n \gamma_2^{-2s_n}) + O(s_n^2 m_n^{1-2d} \gamma_2^{-2s_n})$$
(64)

For some N_n such that

$$\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^0\|_2 \le N_n/2$$

By definition of β^* , we have

$$\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^0\|_2 = \frac{N_n}{N_n + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_2} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_2 \le \frac{N_n}{2}$$

The inequality above implies

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_2 \le N_n$$

Therefore,

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_2^2 = O_P\left(s_n\gamma_2^{-2s_n}\frac{m_n^2\log(p_nm_n)}{n}\right) + O(\lambda_{n1}^2m_n^2s_n\gamma_2^{-2s_n}) + O(s_n^2m_n^{1-2d}\gamma_2^{-2s_n})$$

In the unbounded response case, the only difference that we have to make is in (59), we have

$$\frac{\left(\boldsymbol{\beta}^{*} - \boldsymbol{\beta}^{0}\right)^{T} \Phi^{T} \left(\boldsymbol{y} - \boldsymbol{\mu}_{y}\right)}{n} \\
\leq \frac{\gamma_{0} c_{1}}{8} \|\boldsymbol{\beta}_{\{T \cup \hat{T}\}}^{*} - \boldsymbol{\beta}_{\{T \cup \hat{T}\}}^{0}\|_{2}^{2} + O_{P}\left(s_{n} m_{n} a_{n} \frac{\log(p_{n} m_{n})}{n}\right) \tag{65}$$

where the convergence rate is by lemma B.2. Then with the choice of λ_{n1} for this case, we have

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_2^2 = O_P\left(s_n \gamma_2^{-2s_n} \frac{m_n^2 \log(p_n m_n)}{n}\right) + O(\lambda_{n1}^2 m_n^2 s_n \gamma_2^{-2s_n}) + O(s_n^2 m_n^{1-2d} \gamma_2^{-2s_n})$$

Part (iii) is a direct result of part (ii). By assumption 4, we have $||f_j||_2 \ge c_{f,n} > 0$, and we have

$$||f_{nj}||_2 \ge ||f_j||_2 - ||f_j - f_{nj}||_2 \ge c_{f,n} - O(m_n^{-d}) \ge \frac{1}{2}c_{f,n}$$

for large n. By the properties of spline in De Boor (2001), see for example Stone (1986) and Huang et al. (2010), there exist positive constants c_1 and c_2 such that

$$c_1 m_n^{-1} \|\boldsymbol{\beta}_j^0\|_2^2 \le \|f_{nj}\|_2 \le c_2 m_n^{-1} \|\boldsymbol{\beta}_j^0\|_2$$

Then we have $\|\boldsymbol{\beta}_{j}^{0}\|_{2}^{2} \geq c_{2}^{-1}m_{n}\|f_{nj}\|_{2}^{2} \geq 0.25c_{2}^{-1}m_{n}c_{f,n}^{2}$. Suppose there is a $j \in T$ such that $\|\hat{\boldsymbol{\beta}}_{j}\|_{2} = 0$, then we have

$$\|\boldsymbol{\beta}_{j}^{0}\|_{2} \ge 0.25c_{2}^{-1}m_{n}c_{f,n}^{2}$$

which is a contradiction to the result in (ii) and the theorem assumption. Therefore, part (iii) follows.

Proof of theorem 3.2

Proof. We start with part (i). To prove part (i), it's equivalent to prove that the selection is done as it is performed right on the active set, and none of the nonzero components are dropped with probability tending to 1. Let

$$\hat{\boldsymbol{\beta}}_{NZ} = \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^{p_n m_n}: \boldsymbol{\beta}_{T^c} = 0} L_a(\boldsymbol{\beta}; \lambda_{n2})$$

be the adaptive group lasso estimator restricted to the true nonzero components. First we show that with probability converging to 1, $\hat{\boldsymbol{\beta}}_{NZ}$ is the solution to minimizing (17), i.e., with probability converging to 1, the minimiser of (17) is $\hat{\boldsymbol{\beta}}_{NZ}$. Note that the adaptive group lasso is a convex optimization problem with affine constraints, therefore the KKT conditions are necessary and sufficient. The KKT conditions for a vector $\boldsymbol{\beta} \in \mathbb{R}^{p_n m_n}$ to be the solution of (17) is

$$\begin{cases}
\frac{1}{n}\Phi_j^T(\boldsymbol{y}-\boldsymbol{\mu}^*) = \lambda_{n2}w_{nj}\frac{\boldsymbol{\beta}_j}{\|\boldsymbol{\beta}_j\|_2}, & \text{if } \|\boldsymbol{\beta}_j\|_2 > 0 \\
\|\frac{1}{n}\Phi_j^T(\boldsymbol{y}-\boldsymbol{\mu}^*)\|_2 \le \lambda_{n2}w_{nj}, & \text{if } \|\boldsymbol{\beta}_j\|_2 = 0
\end{cases}$$
(66)

where $\mu^* = b'(\Phi \beta)$. It is sufficient to show that

$$\mathbb{P}\left(\hat{\boldsymbol{\beta}}_{NZ} \text{ satisfies (66)}\right) \to 1$$

Note that for any $j \in T$, we have the KKT conditions for $\hat{\boldsymbol{\beta}}_{NZ}$ that

$$\begin{cases}
\frac{1}{n} \Phi_{j}^{T}(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_{NZ}^{*}) = \lambda_{n2} w_{nj} \frac{\hat{\boldsymbol{\beta}}_{NZj}}{\|\hat{\boldsymbol{\beta}}_{NZj}\|_{2}}, & \text{if } \|\boldsymbol{\beta}_{j}\|_{2} > 0, \ j \in T \\
\|\frac{1}{n} \Phi_{j}^{T}(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_{NZ}^{*})\|_{2} \leq \lambda_{n2} w_{nj}, & \text{if } \|\boldsymbol{\beta}_{j}\|_{2} = 0, \ j \in T
\end{cases}$$
(67)

which are the equality condition in (66) and part of the inequality condition in (66). Therefore, it suffices to show that

$$\mathbb{P}\left(\|\frac{1}{n}\Phi_{j}^{T}(\boldsymbol{y}-\hat{\boldsymbol{\mu}}_{NZ}^{*})\|_{2} \leq \lambda_{n2}w_{nj}, \ \forall \ j \notin T\right) \to 1$$
(68)

This is equivalent to show that

$$\mathbb{P}\left(\|\frac{1}{n}\Phi_{j}^{T}(\boldsymbol{y}-\hat{\boldsymbol{\mu}}_{NZ}^{*})\|_{2} > \lambda_{n2}w_{nj}, \ \exists \ j \notin T\right) \to 0$$

$$(69)$$

Use Taylor expansion on $\frac{1}{n}\Phi_{j}^{T}(\boldsymbol{y}-\hat{\boldsymbol{\mu}}_{NZ}^{*})$, we have

$$\frac{1}{n}\Phi_j^T(\boldsymbol{y}-\hat{\boldsymbol{\mu}}_{NZ}^*) = \frac{1}{n}\Phi_j(\boldsymbol{y}-\boldsymbol{\mu}_y) + \frac{1}{n}\Phi_j^T(\boldsymbol{\mu}_y - b'(\Phi\boldsymbol{\beta}^0)) + \frac{1}{n}\Phi_j^T\boldsymbol{\Sigma}\Phi(\hat{\boldsymbol{\beta}}_{NZ} - \boldsymbol{\beta}^0)$$

where Σ is the variance matrix evaluated at some $\boldsymbol{\beta}^*$ located on the line segment joining $\boldsymbol{\beta}^0$ and $\hat{\boldsymbol{\beta}}_{NZ}$. Then we have

$$\mathbb{P}\left(\|\frac{1}{n}\Phi_{j}^{T}(\boldsymbol{y}-\hat{\boldsymbol{\mu}}_{NZ}^{*})\|_{2} > \lambda_{n2}w_{nj}, \exists j \notin T\right)$$

$$\leq \mathbb{P}\left(\|\frac{1}{n}\Phi_{j}(\boldsymbol{y}-\boldsymbol{\mu}_{y})\|_{2} > \frac{\lambda_{n2}w_{nj}}{3}, \exists j \notin T\right) + \mathbb{P}\left(\|\frac{1}{n}\Phi_{j}^{T}(\boldsymbol{\mu}_{y}-b'(\Phi\boldsymbol{\beta}^{0}))\|_{2} > \frac{\lambda_{n2}w_{nj}}{3}, \exists j \notin T\right)$$

$$+ \mathbb{P}\left(\|\frac{1}{n}\Phi_{j}^{T}\boldsymbol{\Sigma}\Phi(\hat{\boldsymbol{\beta}}_{NZ}-\boldsymbol{\beta}^{0})\|_{2} > \frac{\lambda_{n2}w_{nj}}{3}, \exists j \notin T\right)$$

$$\equiv P_{1} + P_{2} + P_{3}$$

Now let's consider P_1 . By assumption 3, the errors $y_i - \mu_{y_i}$'s are sub-Gaussian. For

bounded responses, we have by lemma B.1 and assumption 6 that

$$P_{1} = \mathbb{P}\left(\left\|\frac{1}{n}\Phi_{j}^{T}(\boldsymbol{y}-\hat{\boldsymbol{\mu}}_{NZ}^{*})\right\|_{2} > \lambda_{n2}w_{nj}, \ \exists \ j \notin T\right)$$

$$\leq \mathbb{P}\left(\left\|\frac{1}{n}\Phi_{j}^{T}(\boldsymbol{y}-\hat{\boldsymbol{\mu}}_{NZ}^{*})\right\|_{2} > C\lambda_{n2}r_{n}, \ \exists \ j \notin T\right) + o(1)$$

$$= \mathbb{P}\left(\max_{j \notin T}\left\|\frac{1}{n}\Phi_{j}^{T}(\boldsymbol{y}-\hat{\boldsymbol{\mu}}_{NZ}^{*})\right\|_{2} > C\lambda_{n2}r_{n}\right) + o(1)$$

$$\leq \mathbb{P}\left(\max_{j \notin T}\left\|\frac{1}{n}\Phi_{j}^{T}(\boldsymbol{y}-\hat{\boldsymbol{\mu}}_{NZ}^{*})\right\|_{2} > C\lambda_{n2}r_{n}\left\|\max_{j \notin T}\left\|\frac{1}{n}\Phi_{j}^{T}(\boldsymbol{y}-\hat{\boldsymbol{\mu}}_{NZ}^{*})\right\|_{2} \leq Cn^{-1/2}\sqrt{\log(s_{n}^{*}m_{n})}\right) + o(1)$$

$$\to 0 \ as \ n \to \infty$$

By lemma B.2, we have

$$E\left(\max_{j\notin T, k=1,\dots,m_n} \left\| \frac{1}{n} \Phi_{jk}^T (\boldsymbol{y} - \boldsymbol{\mu}_y) \right\|_2 \right) \le c_6 n^{-1/2} \sqrt{\log(s_n^* m_n)}$$
 (70)

for some constant c_6 . Observe that by assumption 5, we have $w_{nj} = O_P(r_n) \leq Cr_n$ for some general constant C. Then we have by Markov's inequality and assumption 6 that

$$P_{1} = \mathbb{P}\left(\left\|\frac{1}{n}\Phi_{j}^{T}(\boldsymbol{y}-\hat{\boldsymbol{\mu}}_{NZ}^{*})\right\|_{2} > \lambda_{n2}w_{nj}, \; \exists \; j \notin T\right)$$

$$\leq \mathbb{P}\left(\left\|\frac{1}{n}\Phi_{j}^{T}(\boldsymbol{y}-\hat{\boldsymbol{\mu}}_{NZ}^{*})\right\|_{2} > C\lambda_{n2}r_{n}, \; \exists \; j \notin T\right) + o(1)$$

$$= \mathbb{P}\left(\max_{j \notin T} \left\|\frac{1}{n}\Phi_{j}^{T}(\boldsymbol{y}-\hat{\boldsymbol{\mu}}_{NZ}^{*})\right\|_{2} > C\lambda_{n2}r_{n}\right) + o(1)$$

$$\leq \frac{E\left(\max_{j \notin T, k=1,\dots,m_{n}} \left\|\frac{1}{n}\Phi_{jk}^{T}(\boldsymbol{y}-\boldsymbol{\mu}_{y})\right\|_{2}\right)}{C\lambda_{n2}r_{n}} + o(1)$$

$$\leq \frac{c_{6}\sqrt{\log(s_{n}^{*}m_{n})}}{Cn^{1/2}\lambda_{n2}r_{n}} + o(1) \to 0 \; as \; n \to \infty$$

Then we consider P_2 . We have shown that

$$\frac{1}{n} \|\boldsymbol{\mu}_y - \boldsymbol{\mu}_y^*\|_2^2 = O(s_n^2 m_n^{-2d})$$

This implies that

$$\frac{1}{\sqrt{n}} \|\boldsymbol{\mu}_y - \boldsymbol{\mu}_y^*\|_2 = O(s_n m_n^{-d})$$

Then by assumption 1,

$$\max_{j \notin T} \left\| \frac{1}{n} \Phi_j (\boldsymbol{\mu}_y - \boldsymbol{\mu}_y^*) \right\|_2$$

$$\leq C m_n^{-1/2} \frac{1}{\sqrt{n}} \left\| \boldsymbol{\mu}_y - \boldsymbol{\mu}_y^* \right\|_2$$

$$= O(s_n m_n^{-d-1/2})$$

By assumption 6, we have $P_2 \to 0$ as $n \to \infty$. Next, we look at P_3 . By the definition of $\hat{\beta}_{NZ}$, we have by norm inequality

$$\frac{1}{n}\Phi_{j}^{T}\boldsymbol{\Sigma}\Phi(\hat{\boldsymbol{\beta}}_{NZ}-\boldsymbol{\beta}^{0})=\frac{1}{n}\Phi_{j}^{T}\boldsymbol{\Sigma}\Phi_{T}(\hat{\boldsymbol{\beta}}_{NZT}-\boldsymbol{\beta}_{T}^{0})$$

The MLE on the true nonzero set has a rate of convergence $\sqrt{s_n m_n/n}$. The penalised solution has been proved to be close to the MLE asymptotically (Zhang and Huang (2008); Fan and Li (2001); Lv and Fan (2009)). Knowing the true nonzero set, the rate of convergence of $\hat{\beta}_{NZ}$ is $\sqrt{s_n m_n/n}$. Then we have

$$P_{3} = \mathbb{P}\left(\left\|\frac{1}{n}\Phi_{j}^{T}\boldsymbol{\Sigma}\Phi_{T}(\hat{\boldsymbol{\beta}}_{NZT} - \boldsymbol{\beta}_{T}^{0})\right\|_{2} > \frac{\lambda_{n2}w_{nj}}{3}, \ \exists \ j \notin T\right)$$

$$\leq \mathbb{P}\left(\left\|\frac{1}{n}\Phi_{j}^{T}\boldsymbol{\Sigma}\Phi_{T}(\hat{\boldsymbol{\beta}}_{NZT} - \boldsymbol{\beta}_{T}^{0})\right\|_{2} > C\lambda_{n2}r_{n}, \ \exists \ j \notin T\right) + o(1)$$

$$\leq \mathbb{P}\left(\max_{j \notin T}\left\|\frac{1}{n}\Phi_{j}^{T}\boldsymbol{\Sigma}\Phi_{T}\right\|_{2} > \frac{C\lambda_{n2}r_{n}}{a_{n}\sqrt{s_{n}m_{n}/n}}\right) + \mathbb{P}\left(\left\|\hat{\boldsymbol{\beta}}_{NZT} - \boldsymbol{\beta}_{T}^{0}\right\|_{2} > a_{n}\sqrt{\frac{s_{n}m_{n}}{n}}\right) + o(1)$$

$$\to 0 \text{ as } n \to \infty$$

for any diverging sequence a_n , where the first probability in the last step goes to 0 by assumption 1 that the left hand side is of order $m_n^{-1/2}$ and assumption 6. The second

probability goes to 0 by the rate of convergence of $\hat{\beta}_{NZT}$.

Therefore, we have that $\hat{\boldsymbol{\beta}}_{NZ}$ is our adaptive group lasso solution with probability converging to 1. The components selected by adaptive group lasso is asymptotically at most those which are actually nonzero. Then we want to prove that the true nonzero components are all selected with probability converging to 1. By our assumptions, we have

$$\min_{j \in T} \|\hat{\boldsymbol{\beta}}_{NZj}\|_{2} \ge \min_{j \in T} \|\boldsymbol{\beta}_{j}^{0}\|_{2} - \|\hat{\boldsymbol{\beta}}_{NZj} - \boldsymbol{\beta}_{j}^{0}\|_{2}$$

$$\ge c_{2}^{-1/2} m_{n}^{1/2} c_{f,n} - o_{P}(1)$$

$$> 0$$

Therefore, none of the true nonzero components are estimated as zero. Combining the two results above, we have that with probability converging to 1, the components selected by the adaptive group lasso are exactly the true nonzero components, i.e.,

$$\mathbb{P}\left(\hat{\boldsymbol{\beta}}_{AGL} \stackrel{0}{=} \boldsymbol{\beta}^{0}\right) \to 1 \text{ as } n \to \infty$$

Part (i) is proved. Then we look at part (ii), where based on the result in part (i), we only consider the high probability event that the selection of the adaptive group lasso estimator is perfect. Similar to part (ii) of theorem 3.1, we consider a convex combination of β^0 and $\hat{\beta}_{AGL}$

$$\boldsymbol{\beta}^* = t\hat{\boldsymbol{\beta}}_{AGL} + (1-t)\boldsymbol{\beta}^0$$

where $t = N_n/(N_n + \|\hat{\boldsymbol{\beta}}_{AGL} - \boldsymbol{\beta}^0\|_2)$ for some sequence N_n . Similar to (56), we have

$$\frac{1}{2n} (\boldsymbol{\beta}_{T}^{*} - \boldsymbol{\beta}_{T}^{0})^{T} \boldsymbol{\Phi}_{T}^{T} \boldsymbol{\Sigma} \boldsymbol{\Phi}_{T} (\boldsymbol{\beta}_{T}^{*} - \boldsymbol{\beta}_{T}^{0}) \leq \frac{(\boldsymbol{\beta}_{T}^{*} - \boldsymbol{\beta}_{T}^{0})^{T} \boldsymbol{\Phi}_{T}^{T} (\boldsymbol{y} - \boldsymbol{\mu}_{y})}{n} + \frac{(\boldsymbol{\beta}_{T}^{*} - \boldsymbol{\beta}_{T}^{0})^{T} \boldsymbol{\Phi}_{T}^{T} (\boldsymbol{\mu}_{y} - \boldsymbol{\mu}_{y}^{*})}{n} + \lambda_{n2} \sum_{j=1}^{s_{n}} w_{nj} (\|\boldsymbol{\beta}^{0}\|_{2} - \|\boldsymbol{\beta}^{*}\|_{2})$$
(71)

Then by the fact that $|a^Tb| \le ||a||_2^2 + ||b||_2^2/4$, we have

$$\frac{1}{2n} (\boldsymbol{\beta}_{T}^{*} - \boldsymbol{\beta}_{T}^{0})^{T} \Phi_{T}^{T} \boldsymbol{\Sigma} \Phi_{T} (\boldsymbol{\beta}_{T}^{*} - \boldsymbol{\beta}_{T}^{0})
\leq \frac{(\boldsymbol{\beta}_{T}^{*} - \boldsymbol{\beta}_{T}^{0})^{T} \Phi_{T}^{T} (\boldsymbol{y} - \boldsymbol{\mu}_{y})}{n} + \frac{1}{4n} (\boldsymbol{\beta}_{T}^{*} - \boldsymbol{\beta}_{T}^{0})^{T} \Phi_{T}^{T} \boldsymbol{\Sigma} \Phi_{T} (\boldsymbol{\beta}_{T}^{*} - \boldsymbol{\beta}_{T}^{0})
+ \frac{\|\boldsymbol{\Sigma}^{-1/2} (\boldsymbol{\mu}_{y} - \boldsymbol{\mu}_{y}^{*})\|_{2}^{2}}{n} + \lambda_{n2} \sum_{j=1}^{s_{n}} w_{nj} (\|\boldsymbol{\beta}^{0}\|_{2} - \|\boldsymbol{\beta}^{*}\|_{2})$$

Then by (37),

$$\frac{1}{4n}(\boldsymbol{\beta}_{T}^{*} - \boldsymbol{\beta}_{T}^{0})^{T} \boldsymbol{\Phi}_{T}^{T} \boldsymbol{\Sigma} \boldsymbol{\Phi}_{T}(\boldsymbol{\beta}_{T}^{*} - \boldsymbol{\beta}_{T}^{0}) \leq \frac{(\boldsymbol{\beta}_{T}^{*} - \boldsymbol{\beta}_{T}^{0})^{T} \boldsymbol{\Phi}_{T}^{T} (\boldsymbol{y} - \boldsymbol{\mu}_{y})}{n} + O(s_{n}^{2} m_{n}^{-2d}) + \lambda_{n2} \sum_{j=1}^{s_{n}} w_{nj} (\|\boldsymbol{\beta}^{0}\|_{2} - \|\boldsymbol{\beta}^{*}\|_{2})$$

By (13), the fact that $|a^T b| \leq ||a||_2^2 + ||b||_2^2/4$ and norm inequality, we have

$$\frac{\gamma_{0}c_{1}\gamma_{2}^{2s_{n}}m_{n}^{-1}}{4}\|\boldsymbol{\beta}^{*}-\boldsymbol{\beta}^{0}\|_{2}^{2} \leq \frac{(\boldsymbol{\beta}_{T}^{*}-\boldsymbol{\beta}_{T}^{0})^{T}\Phi_{T}^{T}(\boldsymbol{y}-\boldsymbol{\mu}_{y})}{n} + O(s_{n}^{2}m_{n}^{-2d}) + \frac{2(\max_{j\in T}w_{nj})^{2}}{\gamma_{0}c_{1}}\lambda_{n2}^{2}s_{n} + \frac{\gamma_{0}c_{1}\gamma_{2}^{2s_{n}}m_{n}^{-1}}{8}\|\boldsymbol{\beta}^{*}-\boldsymbol{\beta}^{0}\|_{2}^{2}$$

Then by assumption 6,

$$\frac{\gamma_0 c_1 \gamma_2^{2s_n} m_n^{-1}}{8} \|\boldsymbol{\beta}_T^* - \boldsymbol{\beta}_T^0\|_2^2 \le \frac{(\boldsymbol{\beta}_T^* - \boldsymbol{\beta}_T^0)^T \Phi_T^T (\boldsymbol{y} - \boldsymbol{\mu}_y)}{n} + O(s_n^2 m_n^{-2d}) + O(\lambda_{n2}^2 s_n)$$

Use the fact that $|a^Tb| \leq ||a||_2^2 + ||b||_2^2/4$ on the first term of the right hand side, we have

$$\frac{\gamma_{0}c_{1}\gamma_{2}^{2s_{n}}m_{n}^{-1}}{8}\|\boldsymbol{\beta}_{T}^{*}-\boldsymbol{\beta}_{T}^{0}\|_{2}^{2} \leq \frac{\gamma_{0}c_{1}\gamma_{2}^{2s_{n}}m_{n}^{-1}}{16}\|\boldsymbol{\beta}_{T}^{*}-\boldsymbol{\beta}_{T}^{0}\|_{2}^{2} \\
+ \frac{4}{\gamma_{0}c_{1}\gamma_{2}^{2s_{n}}m_{n}^{-1}n^{2}}\|\boldsymbol{\Phi}_{T}^{T}(\boldsymbol{y}-\boldsymbol{\mu}_{y})\|_{2}^{2} + O(s_{n}^{2}m_{n}^{-2d}) + O(\lambda_{n2}^{2}s_{n})$$

By norm inequality and lemma B.1, we have

$$\frac{4}{\gamma_0 c_1 \gamma_2^{2s_n} m_n^{-1} n^2} \|\Phi_T^T(\boldsymbol{y} - \boldsymbol{\mu}_y)\|_2^2 \le \frac{4}{\gamma_0 c_1 \gamma_2^{2s_n} m_n^{-1} n^2} s_n m_n \|\Phi_T^T(\boldsymbol{y} - \boldsymbol{\mu}_y)\|_{\infty} = O_P\left(s_n \gamma_2^{-2s_n} m_n \frac{\log(s_n m_n)}{n}\right)$$

Combine the last two results, we have with probability converging to 1,

$$\|\boldsymbol{\beta}_{T}^{*} - \boldsymbol{\beta}_{T}^{0}\|_{2}^{2} = O_{p}\left(s_{n}\gamma_{2}^{-2s_{n}}m_{n}^{2}\frac{\log(s_{n}m_{n})}{n}\right) + O(s_{n}^{2}\gamma_{2}^{-2s_{n}}m_{n}^{1-2d}) + O(\lambda_{n2}^{2}m_{n}^{2}s_{n}\gamma_{2}^{-2s_{n}})$$

Then similar to the argument in the proof of part (ii) of theorem 3.1, we have

$$\sum_{j \in T} \|\hat{\boldsymbol{\beta}}_{AGLj} - \boldsymbol{\beta}_j^0\|_2^2 = O_p\left(s_n \gamma_2^{-2s_n} m_n^2 \frac{\log(s_n m_n)}{n}\right) + O(s_n^2 \gamma_2^{-2s_n} m_n^{1-2d}) + O(\lambda_{n2}^2 m_n^2 s_n \gamma_2^{-2s_n})$$

In the unbounded response case, we replace lemma B.1 with lemma B.2 and get

$$\sum_{j \in T} \|\hat{\boldsymbol{\beta}}_{AGLj} - \boldsymbol{\beta}_j^0\|_2^2 = O_p\left(s_n \gamma_2^{-2s_n} m_n^2 a_n \frac{\log(s_n m_n)}{n}\right) + O(s_n^2 \gamma_2^{-2s_n} m_n^{1-2d}) + O(\lambda_{n2}^2 m_n^2 s_n \gamma_2^{-2s_n})$$

for any diverging sequence a_n . Part (ii) is proved.

Proof of theorem 4.1

Proof. The idea of the proof is similar to the proofs in Fan and Tang (2013), but due to the group penalization structure, some changes have to be made. First, the GIC criterion has the solution of adaptive group lasso, which is not easy to study. So we use a proxy, the MLE on the nonzero components selected by the adaptive group lasso

estimator. Let

$$\hat{\boldsymbol{\beta}}^{*}(A) = \underset{\{\boldsymbol{\beta} \in \mathbb{R}^{p_n m_n} : \text{supp}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = A\}}{\text{arg max}} \frac{1}{n} \sum_{i=1}^{n} \left[y_i \left(\boldsymbol{\beta}^T \boldsymbol{\Phi}_i \right) - b \left(\boldsymbol{\beta}^T \boldsymbol{\Phi}_i \right) \right]$$
(72)

for a given $A \subset \{1,...,p\}$, and the proxy of GIC is defined as

$$GIC_{a_n}^*(A) = \frac{1}{n} \{ D(\hat{\mu}_A^*; \mathbf{Y}) + a_n |A| \}$$
 (73)

where $\hat{\mu}_A^* = b'(\Phi \hat{\boldsymbol{\beta}}^*(A))$. The first result is that the proxy $GIC_{a_n}^*(T)$ well approximates $GIC_{a_n}(\lambda_0)$. To prove this, observe by the definition of $\hat{\boldsymbol{\beta}}_0 = \hat{\boldsymbol{\beta}}^*(T)$, we have the first order necessary condition

$$\frac{\partial}{\partial \boldsymbol{\beta}} l_n(\hat{\boldsymbol{\beta}}_0) = \mathbf{0} \tag{74}$$

Use Taylor expansion and by assumptions 1 and 2, we have

$$0 \geq GIC_{a_n}^*(T) - GIC_{a_n}(\lambda_0)$$

$$= \frac{1}{n} \left(l_n(\hat{\boldsymbol{\beta}}(\lambda_{n0})) - l_n(\hat{\boldsymbol{\beta}}_0) \right)$$

$$= -\frac{1}{n} \left(\hat{\boldsymbol{\beta}}(\lambda_{n0}) - \hat{\boldsymbol{\beta}}_0 \right)^T \Phi^T \boldsymbol{\Sigma}(\boldsymbol{\beta}^*) \Phi \left(\hat{\boldsymbol{\beta}}(\lambda_{n0}) - \hat{\boldsymbol{\beta}}_0 \right)$$

$$\geq -c_1 \gamma_0 \left\| \hat{\boldsymbol{\beta}}(\lambda_{n0}) - \hat{\boldsymbol{\beta}}_0 \right\|_2^2$$

$$(75)$$

where $\boldsymbol{\beta}^*$ lies on the line segment joining $\hat{\boldsymbol{\beta}}(\lambda_{n0})$ and $\hat{\boldsymbol{\beta}}_0$. Then we need to bound $\|\hat{\boldsymbol{\beta}}(\lambda_{n0}) - \hat{\boldsymbol{\beta}}_0\|_2^2$. By the definition of $\hat{\boldsymbol{\beta}}(\lambda_{n0})$, we have

$$\Phi_T^T \left(\mathbf{y} - b'(\Phi_T \hat{\boldsymbol{\beta}}_T(\lambda_{n0})) \right) + n\lambda_{n0} \boldsymbol{\nu}_T = \mathbf{0}$$
 (76)

where the elements of $\boldsymbol{\nu}_T$ are $w_{nj}\hat{\boldsymbol{\beta}}_j(\lambda_{n0})/\|\hat{\boldsymbol{\beta}}_j(\lambda_{n0})\|_2$ for $j \in T$. On the other hand, by

the definition of $\hat{\boldsymbol{\beta}}_0$, we have

$$\Phi_T^T \left(\mathbf{y} - b'(\Phi_T \hat{\boldsymbol{\beta}}_{0T}) \right) = \mathbf{0}$$
 (77)

Together we have

$$\Phi_T^T \left(b'(\Phi_T \hat{\boldsymbol{\beta}}_{0T}) - b'(\Phi_T \hat{\boldsymbol{\beta}}_T(\lambda_{n0})) \right) + n\lambda_{n0} \boldsymbol{\nu}_T = \mathbf{0}$$
 (78)

Use Taylor expansion on the left hand side of the equation, we have

$$\Phi_T^T \mathbf{\Sigma}(\boldsymbol{\beta}^{**}) \Phi_T \left(\hat{\boldsymbol{\beta}}_T(\lambda_{n0}) - \hat{\boldsymbol{\beta}}_{0T} \right) = n \lambda_{n0} \boldsymbol{\nu}_T$$
 (79)

where $\boldsymbol{\beta}^{**}$ lies on the line segment joining $\hat{\boldsymbol{\beta}}_T(\lambda_{n0})$ and $\hat{\boldsymbol{\beta}}_{0T}$. Taking 2 norm and together with assumptions 1 and 2 and the results in theorem 3.1, we have

$$\left\|\hat{\boldsymbol{\beta}}_{T}(\lambda_{n0}) - \hat{\boldsymbol{\beta}}_{0T}\right\|_{2} \leq C\lambda_{n0}\|\boldsymbol{w}_{T}\|_{2} \leq C\lambda_{n0}\sqrt{s_{n}}\|\boldsymbol{w}_{T}\|_{\infty}$$
(80)

where $\boldsymbol{w}_T = (w_{nj}, j \in T)'$. Then we have

$$\|\hat{\boldsymbol{\beta}}(\lambda_{n0}) - \hat{\boldsymbol{\beta}}_0\|_2 = O(\lambda_{n0}\sqrt{s_n}) \tag{81}$$

Choose a_n to be any diverging sequence, then we have

$$\|\hat{\boldsymbol{\beta}}(\lambda_{n0}) - \hat{\boldsymbol{\beta}}_0\|_2 = o(\lambda_{n0}\sqrt{s_n a_n}) \tag{82}$$

Then by (75), we have

$$GIC_{a_n}(\lambda_0) - GIC_{a_n}^*(T) = o(\lambda_{n0}\sqrt{s_n a_n})$$
(83)

As a direct result,

$$GIC_{a_{n}}(\lambda) - GIC_{a_{n}}(\lambda_{n0}) \ge (GIC_{a_{n}}^{*}(\alpha_{\lambda}) - GIC_{a_{n}}^{*}(T)) + (GIC_{a_{n}}^{*}(T) - GIC_{a_{n}}(\lambda_{n0}))$$

$$= (GIC_{a_{n}}^{*}(\alpha_{\lambda}) - GIC_{a_{n}}^{*}(T)) + o_{p}(\lambda_{n0}\sqrt{s_{n}a_{n}})$$
(84)

The using this proxy, next we prove that the proxy GIC^* is able to detect the distance between a selected model and the true model. Since the GIC^* depends only on the MLE and has nothing to do with the penalization, this is the same as the generalised linear model, but with the spline line approximation error being considered.

Due to the estimation problem, we are only interested in the models A such that $|A| \leq K$ where $Km_n = o(n)$. As the proof in Fan and Tang (2013), we consider the underfitted model and overfitted model (defined in their paper). Briefly, the underfitted models are A such that $A \not\supset T$ and the overfitted models are A such that $A \supsetneq T$. Also in the result of theorem 3.1, the model size $|A| = O(s_n) = o(n)$ and thus the KL divergence has a unique minimiser for every such model A, as discussed in Fan and Tang (2013).

Lemma B.3 implies that for all underfitted models

$$GIC_{A_n}^*(A) - GIC_{a_n}^*(T) = 2|A|I(\boldsymbol{\beta}^*(A)) + (|A| - |T|)a_n n^{-1} + |A|O_P(R_n)$$

$$\geq \delta_n - s_n a_n n^{-1} - O_P(KR_n)$$

$$\geq \frac{\delta_n}{2}$$

if $\delta_n K^{-1} R_n^{-1} \to \infty$ and $a_n = o(\delta_n s_n^{-1} n)$. This result states that there is a negligible increment on the GIC^* if one of the nonzero component is missed, when the parameters satisfy the conditions. Lemma B.4 implies that for all overfitted models

$$GIC_{a_n}^*(A) - GIC_{a_n}^*(T) = \frac{|A| - |T|}{n} [a_n - O_P(\psi_n)] > \frac{a_n}{2n}$$

if $a_n \psi_n \to \infty$. This result states that there is a negligible increment on the GIC^* if one of the zero component is selected along with the true model, when the parameters satisfy the conditions. Therefore,

$$\mathbb{P}\left(\inf_{A \not\supset T} GIC_{a_n}^*(A) - GIC_{a_n}^*(T) > \frac{\delta_n}{2} \text{ and } \inf_{A \supsetneq T} GIC_{a_n}^*(A) - GIC_{a_n}^*(T) > \frac{a_n}{2n}\right) \to 1$$
(85)

Combine this result with (84) and theorem assumptions, we have

$$\mathbb{P}\{\inf_{\lambda\in\Omega_{-}\cup\Omega_{+}}GIC_{a_{n}}(\lambda)>GIC_{a_{n}}(\lambda_{n0})\}\to 1$$

Lemma B.3. Under assumptions 2 and 3, as $n \to \infty$, we have

$$\sup_{\substack{|A| \le K \\ A \subset \{1, \dots, p_n\}}} \frac{1}{n|A|} |D(\hat{\boldsymbol{\mu}}_A^*; \boldsymbol{Y}) - D(\hat{\boldsymbol{\mu}}_0^*; \boldsymbol{Y}) - 2I(\boldsymbol{\beta}^*(A))| = O_P(R_n)$$

where either a) the responses are bounded or Gaussian distributed, $R_n = \sqrt{\gamma_n m_n \log(p_n)/n}$, and $m_n \log(p_n) = o(n)$; or b) the responses are unbounded and non-Gaussian distributed, $R_n = \sqrt{\gamma_n m_n \log(p_n)/n} + \gamma_n^2 m_n M_n^2 \log(p_n)/n \text{ and } \log(p) = o(\min\{n(\log n)^{-1}K^{-2}m_n^{-1}\gamma_n^{-1}, nM_n^{-2}\}).$

Proof. lemma B.3 is a direct result from lemma B.7 and lemma B.8. \Box

Lemma B.4. Under assumption 1, 2 and 3, and suppose $\log p = O(n^{\kappa})$ for some $0 < \kappa < 1$, as $n \to \infty$, we have

$$\frac{1}{|A|-|T|}\left(D(\hat{\boldsymbol{\mu}}_A^*;\boldsymbol{Y})-D(\hat{\boldsymbol{\mu}}_0^*;\boldsymbol{Y})\right)=O_P(\psi_n)$$

uniformly for all $A \supseteq T$ with |A| < K and either a) $\psi_n = m_n \sqrt{\gamma_n \log(p_n)}$ when the responses are bounded, $K = O(\min\{n^{(1-2\kappa)/6}, n^{(1-3\kappa)/8}\})$ and $\kappa \le 1/2$; or b) $\psi_n = 0$

 $m_n \gamma_n \log(p_n)$ when the responses are Gaussian bounded; or when the response are unbounded and non-Gaussian distributed, and the last three terms in lemma B.10 are dominated by $m_n \gamma_n \log p_n$.

Proof. lemma B.4 is a direct result from lemma B.9 and B.10.

Lemma B.5. Under assumptions 2-3, let γ_n be a slowly diverging sequence, if $\gamma_n L_n \sqrt{Km_n \log P_n/n} \rightarrow 0$ as $n \rightarrow \infty$, where $L_n = O(1)$ for the bounded case and $L_n = O(M_n + \sqrt{\log n})$ for the unbounded case, then we have

$$\sup_{|A| \le K} \frac{1}{|A|} Z_A \left(\gamma_n L_n \sqrt{|A| m_n \frac{\log p_n}{n}} \right) = O_P \left(\gamma_n^2 L_n^2 \frac{m_n \log p_n}{n} \right)$$

where

$$Z_A(N) = \sup_{\boldsymbol{\beta} \in \mathcal{B}_A(N)} \frac{1}{n} \left| l_n(\boldsymbol{\beta}) - l_n(\boldsymbol{\beta}^*(A)) - E\left[l_n(\boldsymbol{\beta}) - l_n(\boldsymbol{\beta}^*(A)) \right] \right|$$

and

$$\mathcal{B}_A(N) = \left\{ \boldsymbol{\beta} \in \mathbb{R}^P : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*(A)\|_2 \le N, supp_B(\boldsymbol{\beta}) = A \right\} \cup \left\{ \boldsymbol{\beta}^*(A) \right\}$$

Proof. Define

$$\Omega_n = \{ \| \boldsymbol{\varepsilon} \|_{\infty} \le \tilde{L_n} \}$$

If we take $\tilde{L}_n = C\sqrt{\log n}$, Fan and Tang (2013) has showed that $\mathbb{P}(\Omega_n) \to 1$. Let

$$\tilde{Z}_A(N) = \sup_{\boldsymbol{\beta} \in \mathcal{B}_A(N)} \frac{1}{n} |l_n(\boldsymbol{\beta}) - l_n(\boldsymbol{\beta}^*(A)) - E[l_n(\boldsymbol{\beta}) - l_n(\boldsymbol{\beta}^*(A))|\Omega_n]|$$

Then we have

$$\sup_{|A| \le K)} \frac{1}{|A|} Z_A(N) \le \sup_{|A| \le K)} \frac{1}{|A|} \tilde{Z}_A(N) + \sup_{|A| \le K, \beta \in \mathcal{B}_A(N)} \frac{1}{|A|} R_A(\beta)$$

where

$$R_A(\boldsymbol{\beta}) = \frac{1}{n} \left| E\left[l_n(\boldsymbol{\beta}) - l_n(\boldsymbol{\beta}^*(A)) \right] - E\left[l_n(\boldsymbol{\beta}) - l_n(\boldsymbol{\beta}^*(A)) | \Omega_n \right] \right|$$

By the definition of l_n , we have

$$R_{A}(\boldsymbol{\beta}) = \frac{1}{n} \left| E[\boldsymbol{\varepsilon} | \Omega_{n}]^{T} \Phi(\boldsymbol{\beta} - \boldsymbol{\beta}^{*}(A)) \right|$$

$$\leq \frac{1}{n} \left\| E[\boldsymbol{\varepsilon} | \Omega_{n}] \right\|_{2} \left\| \Phi(\boldsymbol{\beta} - \boldsymbol{\beta}^{*}(A)) \right\|_{2}$$

$$= \sqrt{\frac{1}{n} \sum_{i=1}^{n} (E[\epsilon_{i} | \Omega_{n}])^{2} \cdot \frac{1}{\sqrt{n}} \left\| \Phi(\boldsymbol{\beta} - \boldsymbol{\beta}^{*}(A)) \right\|_{2}}$$

$$\leq C \tilde{L}_{n} \exp(-C \tilde{L}_{n}) \|\boldsymbol{\beta} - \boldsymbol{\beta}^{*}(A) \|_{2}$$

where the first inequality is Cauchy-Schwartz inequality, and the second inequality is lemma 1 in Fan and Tang (2013) and assumption 1. Then we have

$$\sup_{|A| \le K, \boldsymbol{\beta} \in \mathcal{B}_A(N)} \frac{1}{|A|} R_A(\boldsymbol{\beta}) = C\tilde{L}_n \exp(-C\tilde{L}_n) N$$

Taking $\tilde{L}_n = C\sqrt{\log n}$, $N = \gamma_n L_n \sqrt{|A| \log(p_n m_n)/n}$ and under the lemma assumption, we have

$$\sup_{|A| \le K, \boldsymbol{\beta} \in \mathcal{B}_A(N)} \frac{1}{|A|} R_A(\boldsymbol{\beta}) = o(\log(p_n m_n)/n)$$
(86)

Then let's consider $\tilde{Z}_A(N)$. For any $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathcal{B}_A(N)$, by the mean value theorem, we have $b(\Phi_i^T\boldsymbol{\beta}_1) - B(\Phi_i^T\boldsymbol{\beta}_2) = b'(\Phi_i^T\tilde{\boldsymbol{\beta}})\Phi_i^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)$, where $\tilde{\boldsymbol{\beta}}$ lies on the line segment joining $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. We have the likelihood function

$$|-y_i \Phi_i^T \boldsymbol{\beta}_1 + b(\Phi_i^T \boldsymbol{\beta}_1) - (-y_i \Phi_i^T \boldsymbol{\beta}_2 + b(\Phi_i^T \boldsymbol{\beta}_2))|$$

$$= |(-y_i + b'(\Phi_i^T \tilde{\boldsymbol{\beta}}))|\Phi_i^T \boldsymbol{\beta}_1 - \Phi_i^T \boldsymbol{\beta}_2|$$

$$\leq (\tilde{L}_n + 2M_n)|\Phi_i^T (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)|$$

to be Lipschitz continuous. Let $w_1, ..., w_n$ be a Rademacher sequence independent of ε . By the symmetrization theorem and the concentration inequality, see chapter 14 of Bühlmann and van de Geer (2011), we have

$$E[\tilde{Z}_{A}(N)|\Omega_{n}] \leq 2E \left[\sup_{\boldsymbol{\beta} \in \mathcal{B}_{A}(N)} \frac{1}{n} \left| \sum_{i=1}^{n} w_{i} \left[-y_{i} \Phi_{i}^{T} \boldsymbol{\beta} + b(\Phi_{i}^{T} \boldsymbol{\beta}) - \left(-y_{i} \Phi_{i}^{T} \boldsymbol{\beta}^{*}(A) + b(\Phi_{i}^{T} \boldsymbol{\beta}^{*}(A))\right) \right] |\Omega_{n}| \right]$$

$$\leq 4L_{n}E \left[\sup_{\boldsymbol{\beta} \in \mathcal{B}_{A}(N)} \frac{1}{n} \left| \sum_{i=1}^{n} w_{i} \left[\Phi_{i}(\boldsymbol{\beta} - \boldsymbol{\beta}^{*}(A)) \right] |\Omega_{n}| \right] \right]$$

$$\leq 4L_{n}E \left[\left(\sup_{\boldsymbol{\beta} \in \mathcal{B}_{A}(N)} \|\boldsymbol{\beta} - \boldsymbol{\beta}^{*}(A)\|_{2} \right) \left(\sum_{j \in A} \sum_{i=1}^{n} \sum_{k=1}^{m_{n}} \left| \frac{1}{n^{2}} (w_{i} \phi_{ijk})^{2} \right| \right)^{1/2} \right]$$

$$\leq 4L_{n}N \sqrt{\frac{|A|m_{n}}{n}}$$

where the second last inequality is by Cauchy-Schwartz inequality, and the last inequality is by the definition of $\mathcal{B}_A(N)$ and w_i . Then since

$$\frac{1}{n} \sum_{i=1}^{n} (L_n \Phi_i^T (\beta(A) - \beta^0))^2 \le C L_n^2 N^2$$

Apply Massart's inequality, see theorem 14.2 in Bühlmann and van de Geer (2011), we have

$$\mathbb{P}\left(\tilde{Z}_A(N) \ge E[\tilde{Z}_A(N)|\Omega_n] + t\right) \le \exp\left(-\frac{1}{CL_n^2N^2}\frac{nt^2}{2}\right)$$

Take $t = 4L_n Nu \sqrt{|A|m_n/n}$ with u > 0, $N = L_n \sqrt{|A|m_n/n}(1+u)$, $u = \gamma_n \sqrt{\log p_n}$ and

observe that $\binom{p_n}{k} \leq (pe/k)^k$, we have

$$\mathbb{P}\left(\sup_{|A| \le K} \frac{1}{|A|} \tilde{Z}_A(N) \ge 4L_n^2 \frac{m_n}{n} (1+u)^2 |\Omega_n\right)$$

$$\le \sum_{|A| \le K} \mathbb{P}\left(\tilde{Z}_A(N) \ge 4|A|L_n^2 \frac{m_n}{n} (1+u)^2 |\Omega_n\right)$$

$$\le \sum_{k \le K} \left(\frac{pe}{k}\right)^k \exp(-CKm_n u^2)$$

$$\le \sum_{k \le K} \left(\frac{pe}{k}\right)^k \exp(-CKm_n \gamma_n \log p_n) \to 0$$

Then we have

$$\mathbb{P}\left(\sup_{|A| \le K} \frac{1}{|A|} \tilde{Z}_A(N) \ge \gamma_n^2 L_n^2 \frac{m_n}{n} \log p_n\right) = o(1) + \mathbb{P}(\Omega^c) \to 0$$

Lemma B.6. Under assumptions 1-3, we have

$$\sup_{|A| \le K} \frac{1}{\sqrt{|A|}} \|\hat{\boldsymbol{\beta}}^*(A) - \boldsymbol{\beta}^*(A)\|_2 = O_P\left(\gamma_n L_n \sqrt{\frac{m_n \log p_n}{n}}\right)$$

Proof. Define the convex combination of $\hat{\boldsymbol{\beta}}^*(A)$ and $\boldsymbol{\beta}^*(A)$ to be the same way as we did in proving theorem 3.1 as $\hat{\boldsymbol{\beta}}_u(A)$. Then is remains to show

$$\sup_{|A| \le K} \frac{1}{\sqrt{|A|}} \|\hat{\boldsymbol{\beta}}_u(A) - \boldsymbol{\beta}^*(A)\|_2 = O_P\left(\gamma_n L_n \sqrt{\frac{m_n \log p_n}{n}}\right)$$

By the definition of $\hat{\boldsymbol{\beta}}^*(A)$ and the concavity of the likelihood function, we have

$$l_n(\hat{\boldsymbol{\beta}}_n(A)) - l_n(\boldsymbol{\beta}^*(A)) \ge 0$$

By the definition of $\beta^*(A)$, we have

$$E[l_n(\boldsymbol{\beta}^*(A) - l_n(\hat{\boldsymbol{\beta}}_u(A))] \ge 0$$

Combine the two inequalities above, we have

$$0 \le E[l_n(\boldsymbol{\beta}^*(A) - l_n(\hat{\boldsymbol{\beta}}_u(A))] \le l_n(\hat{\boldsymbol{\beta}}_u(A)) - l_n(\boldsymbol{\beta}^*(A)) - E[l_n(\hat{\boldsymbol{\beta}}_u(A) - l_n(\boldsymbol{\beta}^*(A))] \le nZ_A(N)$$
(87)

On the other hand, for any $\beta_A \in \mathbb{B}_A(N)$, we have

$$E[l_n(\boldsymbol{\beta}_A) - l_n(\boldsymbol{\beta}^*(A))] = E[\boldsymbol{y}^T \boldsymbol{\Phi} \boldsymbol{\beta}_A - \boldsymbol{1}^T b(\boldsymbol{\Phi} \boldsymbol{\beta}_A) - \boldsymbol{y}^T \boldsymbol{\Phi} \boldsymbol{\beta}^*(A) + \boldsymbol{1}^T b(\boldsymbol{\Phi} \boldsymbol{\beta}^*(A))]$$

$$= b'(\sum_{j=1}^{p_n} f_j)^T \boldsymbol{\Phi}[\boldsymbol{\beta}_A - \boldsymbol{\beta}^*(A)] - \boldsymbol{1}^T [b(\boldsymbol{\Phi} \boldsymbol{\beta}_A) - b(\boldsymbol{\Phi} \boldsymbol{\beta}^*(A))]$$

Observe that by the definition of $\beta^*(A)$, we have

$$\Phi[b'(\sum_{j=1}^{p_n} f_j) - b'(\Phi \beta^*(A))] = \mathbf{0}$$

use Taylor expansion, we have

$$E[l_n(\boldsymbol{\beta}_A) - l_n(\boldsymbol{\beta}^*(A))] = b'(\boldsymbol{\Phi}\boldsymbol{\beta}^*(A))^T \boldsymbol{\Phi}[\boldsymbol{\beta}_A - \boldsymbol{\beta}^*(A)] - \mathbf{1}^T [b(\boldsymbol{\Phi}\boldsymbol{\beta}_A) - b(\boldsymbol{\Phi}\boldsymbol{\beta}^*(A))]$$

$$= -\frac{1}{2} (\boldsymbol{\beta}_A - \boldsymbol{\beta}^*(A))^T \boldsymbol{\Phi}_A^T \tilde{\boldsymbol{\Sigma}} \boldsymbol{\Phi}_A (\boldsymbol{\beta}_A - \boldsymbol{\beta}^*(A))$$

$$\leq Cn \|\boldsymbol{\beta}_A - \boldsymbol{\beta}^*(A)\|_2^2$$

where the last inequality is by assumptions 1 and 2. Then we have

$$\|\boldsymbol{\beta}_A - \boldsymbol{\beta}^*(A)\|_2^2 \le CZ_A(N)$$

Take $N = \gamma_n L_n \sqrt{|A| m_n \log p_n / n}$ and by lemma B.5, we have

$$\sup_{|A| \le K} \frac{1}{\sqrt{|A|}} \|\hat{\boldsymbol{\beta}}_u(A) - \boldsymbol{\beta}^*(A)\|_2 = O_P\left(\gamma_n L_n \sqrt{\frac{m_n \log p_n}{n}}\right)$$

Then lemma B.6 follows.

Lemma B.7. Under assumptions 1-3, we have

$$\sup_{|A| \le K} \frac{1}{n|A|} \left(l_n(\hat{\boldsymbol{\beta}}^*(A)) - l_n(\boldsymbol{\beta}^*(A)) \right) \le \frac{\gamma_n^2 L_n^2 m_n \log p_n}{n}$$

Proof. Define the event

$$\mathcal{E} = \left\{ \sup_{|A| \le K} \frac{1}{\sqrt{|A|}} \|\hat{\boldsymbol{\beta}}^*(A) - \boldsymbol{\beta}^*(A)\|_2 = O_P\left(\gamma_n L_n \sqrt{\frac{m_n \log p_n}{n}}\right) \right\}$$

By lemma B.6, we have $\mathbb{P}(\mathbb{E}) \to 1$. Using the same argument as in (87) in proving lemma B.6, we have

$$0 \leq l_n(\hat{\boldsymbol{\beta}}^*(A)) - l_n(\boldsymbol{\beta}^*(A)) \leq l_n(\hat{\boldsymbol{\beta}}_u(A)) - l_n(\boldsymbol{\beta}^*(A)) - E[l_n(\hat{\boldsymbol{\beta}}_u(A)) - l_n(\boldsymbol{\beta}^*(A))] \leq nZ_A(N)$$

By lemma B.5, conditioning on \mathcal{E} , we have

$$l_n(\hat{\boldsymbol{\beta}}^*(A)) - l_n(\boldsymbol{\beta}^*(A)) \le nO_P\left(\gamma_n^2 L_n^2 \frac{|A| m_n \log p_n}{n}\right)$$

Then the lemma follow from $\mathbb{P}(A) \leq \mathbb{P}(A|\mathcal{E}) + \mathbb{P}(\mathcal{E}^c)$.

Lemma B.8. Under assumption 1-3, we have

$$\sup_{|A| \le K} \frac{1}{n|A|} |l_n(\boldsymbol{\beta}^*(A)) - E[l_n(\boldsymbol{\beta}^*(A))]| = O_P\left(\sqrt{\frac{\gamma_n m_n \log p_n}{n}}\right)$$

where $\log p_n = o(n)$ for bounded response and $\gamma_n m_n K^2 \log p_n = o(n)$ for unbounded

response.

Proof. By the definition, we have $l_n(\boldsymbol{\beta}^*(A)) - E[l_n(\boldsymbol{\beta}^*(A))] = \boldsymbol{\varepsilon}^T \Phi \boldsymbol{\beta}^*(A)$. For bounded response, by Hoeffding's inequality, we have

$$\mathbb{P}(|\boldsymbol{\varepsilon}^T \boldsymbol{\Phi} \boldsymbol{\beta}^*(A)| \ge t) \le C \exp\left(\frac{Ct^2}{\sum_{i=1}^n (\boldsymbol{\Phi}_i^T \boldsymbol{\beta}^*(A))^2}\right)$$
$$\le C \exp\left(-\frac{Ct^2}{n|A|m_n}\right)$$

Take $t = |A| \sqrt{n \gamma_n m_n \log p_n}$, we have

$$\mathbb{P}(|\boldsymbol{\varepsilon}^T \Phi \boldsymbol{\beta}^*(A)| \ge |A| \sqrt{n \gamma_n m_n \log p_n}) \le C \exp(-C|A| \gamma_n \log p_n)$$

Then we have

$$\sup_{|A| \le K} \frac{1}{n|A|} |l_n(\boldsymbol{\beta}^*(A)) - E[l_n(\boldsymbol{\beta}^*(A))]| = O_P\left(\sqrt{\frac{\gamma_n m_n \log p_n}{n}}\right)$$

If the responses are unbounded, we use Bernstein's inequality. First check the condition

$$E[|\Phi_{i}\boldsymbol{\beta}^{*}(A)\epsilon_{i}|^{m}] = m \int_{0}^{\infty} x^{m-1} \mathbb{P}(|\Phi_{i}\boldsymbol{\beta}^{*}(A)\epsilon_{i} \geq x) dx$$

$$= m|\Phi_{i}^{T}\boldsymbol{\beta}^{*}(A)|^{m} \int_{0}^{\infty} \left(\frac{x}{|\Phi_{i}^{T}\boldsymbol{\beta}^{*}(A)|}\right)^{m-1} \mathbb{P}\left(|\epsilon_{i}| \geq \frac{x}{|\Phi_{i}^{T}\boldsymbol{\beta}^{*}(A)|}\right) d\frac{x}{|\Phi_{i}^{T}\boldsymbol{\beta}^{*}(A)|}$$

$$\leq m|\Phi_{i}^{T}\boldsymbol{\beta}^{*}(A)|^{m} \int_{0}^{\infty} t^{m-1}C \exp(-Ct^{2}) dt)$$

$$\leq m|\Phi_{i}^{T}\boldsymbol{\beta}^{*}(A)|^{m} (\|\Phi\boldsymbol{\beta}^{*}(A)\|_{\infty}C)^{m-2} \frac{m!}{2}$$

Then by Bernstein's inequality, we have

$$\mathbb{P}(|\boldsymbol{\varepsilon}^T \boldsymbol{\Phi} \boldsymbol{\beta}^*(A)| \ge \sqrt{n}t) \le 2 \exp\left(-\frac{1}{2} \frac{nt^2}{C\|\boldsymbol{\Phi}_A \boldsymbol{\beta}^*(A)\|_2^2 + C\sqrt{n}\|\boldsymbol{\Phi}_A \boldsymbol{\beta}^*(A)\|_{\infty}t}\right)$$

Taking $t = |A|\sqrt{\gamma_n m_n \log p_n}$, we have

$$\mathbb{P}(|\boldsymbol{\varepsilon}^T \boldsymbol{\Phi} \boldsymbol{\beta}^*(A)| \ge \sqrt{n} |A| \sqrt{\gamma_n m_n \log p_n})$$

$$\le 2 \exp\left(-\frac{1}{2} \frac{n|A|^2 \gamma_n m_n \log p_n}{C \|\boldsymbol{\Phi}_A \boldsymbol{\beta}^*(A)\|_2^2 + C\sqrt{n} \|\boldsymbol{\Phi}_A \boldsymbol{\beta}^*(A)\|_{\infty} |A| \sqrt{\gamma_n m_n \log p_n}}\right)$$

$$\to 0$$

if
$$K^2 \gamma_n m_n \log p_n / n \to 0$$
.

Lemma B.9. Under assumptions 1-3, we have

$$\sup_{\substack{A\supset T\\|A|\leq K}}\frac{1}{|A|-|T|}(\boldsymbol{y}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1/2}\boldsymbol{B}_A\boldsymbol{\Sigma}_0^{-1/2}(\boldsymbol{y}-\boldsymbol{\mu}_0)=O_P(m_n(\gamma_n\log p_n)^\xi)$$

where

$$oldsymbol{B}_A = oldsymbol{\Sigma}_0^{1/2} \Phi_A (\Phi_A^T oldsymbol{\Sigma}_0 \Phi_A)^{-1} \Phi_A^T oldsymbol{\Sigma}_0^{1/2}$$

and $\xi = 1/2$ for bounded response and $\xi = 1$ for unbounded response.

Proof. Let k = |A| - |T| and $\mathbf{P}_A = \mathbf{B}_A - \mathbf{B}_T$. It's easy to verify that \mathbf{P}_A is a projection matrix, thus we have $tr(\mathbf{P}) = km_n$, $\sum_{i=1}^n P_{ii} = km_n$ and $\sum_{i,j} P_{ij} = km_n$. Let

$$ilde{oldsymbol{y}} = oldsymbol{\Sigma}_0^{-1/2} (oldsymbol{y} - oldsymbol{\mu}_0)$$

We have the decomposition

$$\frac{1}{m_n k} \tilde{\boldsymbol{y}}^T \boldsymbol{P}_A \tilde{\boldsymbol{y}} = \frac{1}{m_n k} \sum_{i=1}^n P_{ii} \tilde{y}_i^2 + \frac{1}{m_n k} \sum_{i \neq j} P_{ij} \tilde{Y}_i \tilde{Y}_j \equiv I_1(A) + I_2(A)$$

Let \tilde{y}_i^* be independent copies of \tilde{y}_i , then by the decoupling inequality, there exists a

constant C > 0 such that

$$\mathbb{P}\left(\frac{1}{m_n k} | \sum_{i \neq j} P_{ij} \tilde{y}_i \tilde{y}_j| \ge t\right) \le C \mathbb{P}\left(\frac{1}{m_n k} | \sum_{i \neq j} P_{ij} \tilde{Y}_i \tilde{Y}_j^*| \ge C^{-1} t\right)$$

For bounded response, apply Hoeffding's inequality, we have

$$\mathbb{P}(I_1(A) \ge 1 + x) \le 2 \exp\left(2\frac{Cx^2}{\sum_{i=1}^n (m_n k)^{-2} P_{ii}^2}\right) \le 2 \exp(-Cm_n kx^2)$$

Taking $x = \sqrt{\gamma \log p_n}$, use the inequality $\binom{p}{k} \leq (pe/k)^k$ and use the same technique as we used in proving lemma B.5, we have

$$\mathbb{P}\left(\sup_{|A| \le K} I_1(A) \ge 1 + \sqrt{\gamma_n \log p_n}\right) \le 2C \sum_{k=1}^K \left(\frac{(p_n - s_n)e}{k}\right)^k \exp(-Cm_n k \gamma_n \log p_n) \to 0$$

Then observe $\sum_{i\neq j} P_{ij}^2 = \sum_i (P_{ii} - P_{ii}^2) \leq m_n k$, we have following the decoupling inequality that

$$\mathbb{P}(|I_2(A)| \ge t) \le C \mathbb{P}\left(\frac{1}{m_n k} | \sum_{i \ne j} P_{ij} \tilde{Y}_i \tilde{Y}_j^*| \ge C^{-1} t\right)$$

$$\le C \exp\left(-\frac{C^{-2} (m_n k)^2 t^2}{\sum_{i \ne j} P_{ij}^2}\right)$$

$$\le C \exp(-C m_n k t^2)$$

Taking $t = \sqrt{\gamma_n \log p_n}$ and use the same technique as in the previous step, we have

$$\mathbb{P}\left(\sup_{|A| \le K} I_2(A) \ge 1 + \sqrt{\gamma_n \log p_n}\right) \to 0$$

In the unbounded case, we apply the Bernstein's inequality. In the same way as we did

in proving lemma B.8, we check the condition

$$E|P_{ii}\tilde{Y}_{i}^{2}|^{m} \le m!C^{m-2}\frac{P_{ii}^{2}}{2}$$

By Bernstein's inequality, we have

$$\mathbb{P}(I_1(A) \ge x^2) \le 2\exp(-Cm_nkx^2)$$

Taking $x = \sqrt{\gamma_n \log p_n}$, we have

$$\sup_{|A| \le K} I_1(A) = O_P(\gamma_n \log p_n)$$

For $I_2(A)$, we have

$$\sum_{i \neq j} |P_{ij}|^m E[|\tilde{\boldsymbol{y}}_i \tilde{\boldsymbol{y}}_j^*|^m] \le m! C^{m-2} \frac{P_{ij}^2}{2}$$

Then by Berstein's inequality and taking $x = \sqrt{\gamma_n \log p_n}$, we have

$$\mathbb{P}(|I_2(A)| \ge \gamma_n \log p_n) \to 0$$

Lemma B.10. Under assumptions 1-3, for all $A \supset T$ and $|A| \leq K$, we have

$$l_n(\hat{\boldsymbol{\beta}}^*(A)) - l_n(\boldsymbol{\beta}^(A)) = \frac{1}{2} (\boldsymbol{y} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_0^{-1/2} \boldsymbol{B}_A \boldsymbol{\Sigma}_0^{-1/2} (\boldsymbol{y} - \boldsymbol{\mu}_0) + |A|^{5/2} O_P \left(m_n^{5/2} \gamma_n^{5/2} L_n^2 \frac{(\log p_n)^{1+\xi/2}}{\sqrt{n}} \right) + |A|^4 O_P \left(m_n^4 \gamma_n^4 L_n^4 \frac{(\log p_n)^2}{n} \right) + |A|^3 O_P \left(m_n^3 \gamma_n^3 L_n^3 \frac{(\log p_n)^{3/2}}{\sqrt{n}} \right)$$

Proof. Use Taylor's expansion, we have

$$l_n(\hat{\boldsymbol{\beta}}^*(A)) - l_n(\boldsymbol{\beta}^*(A))$$

$$= (\hat{\boldsymbol{\beta}}^*(A) - \boldsymbol{\beta}^*(A))^T \Phi^T(\boldsymbol{y} - b'(\Phi \boldsymbol{\beta}^*(A)) - \frac{1}{2}(\hat{\boldsymbol{\beta}}^*(A) - \boldsymbol{\beta}^*(A))^T \Phi^T \boldsymbol{\Sigma}_0 \Phi(\hat{\boldsymbol{\beta}}^*(A) - \boldsymbol{\beta}^*(A)) + \text{Remainder}$$

$$\equiv I_1(A) + I_2(A) + I_3(A)$$

First, by the definition of $\hat{\boldsymbol{\beta}}^*(A)$, we have

$$\Phi_A^T[\boldsymbol{y} - b'(\Phi \hat{\boldsymbol{\beta}}^*(A))] = 0$$

Then by Taylor expansion, we have

$$\Phi_A^T \boldsymbol{y} = \Phi_A^T b'(\Phi \hat{\boldsymbol{\beta}}^*(A))$$

$$= \Phi_A^T b'(\Phi \boldsymbol{\beta}^*(A)) + \Phi_A^T \boldsymbol{\Sigma}_0 \Phi(\hat{\boldsymbol{\beta}}^*(A) - \boldsymbol{\beta}^*(A)) + \Phi_A^T \boldsymbol{\nu}_A$$

where $\nu_{Ai} = b'''(\Phi_i^T \hat{\boldsymbol{\beta}}^*(A))(\Phi_i^T (\hat{\boldsymbol{\beta}}^*(A) - \boldsymbol{\beta}^*(A)))^2/2$ and $\tilde{\boldsymbol{\beta}}^*(A)$ lies on the line segment joining $\hat{\boldsymbol{\beta}}^*(A)$ and $\boldsymbol{\beta}^*(A)$. By the definition of $\boldsymbol{\beta}^*(A)$, we have

$$\Phi_A^T[b'(\sum_{j=1}^{p_n} f_j) - b'(\Phi_A \beta^*(A))] = 0$$

we have

$$\hat{\boldsymbol{\beta}}^*(A) - \boldsymbol{\beta}^*(A) = (\Phi_A^T \boldsymbol{\Sigma}_0 \Phi_A)^{-1} \Phi_A^T (\boldsymbol{y} - b'(\sum_{j=1}^{p_n} -\boldsymbol{\nu}_A))$$

Therefore, we have

$$I_1(A) = (\boldsymbol{y} - \boldsymbol{\mu}_y)^T \boldsymbol{\Sigma}_0^{-1/2} \boldsymbol{B}_A \boldsymbol{\Sigma}_0^{-1/2} (\boldsymbol{y} - \boldsymbol{\mu}_y) + R_{1,A}$$

where $R_{1,A} = -\boldsymbol{\mu}_A^T \boldsymbol{\Sigma}_0^{-1/2} \boldsymbol{B}_A \boldsymbol{\Sigma}_0^{-1/2} \boldsymbol{\varepsilon}$. By Cauchy-Schwartz inequality, we have

$$|R_{1,A}| \le \|\boldsymbol{B}_A \boldsymbol{\Sigma}_0^{-1/2} \boldsymbol{\varepsilon}\|_2 \|\boldsymbol{\Sigma}_0^{-1/2} \boldsymbol{\nu}_A\|_2$$

 $\le (\|\boldsymbol{B}_T \boldsymbol{\Sigma}_0^{-1/2} \boldsymbol{\varepsilon}\|_2 + \|\tilde{R}_{1,A}\|_2) \|\boldsymbol{\Sigma}_0^{-1/2} \boldsymbol{\nu}_A\|_2$

where $\tilde{R}_{1,A} = (\boldsymbol{B}_A - \boldsymbol{B}_0) \boldsymbol{\Sigma}_0^{-1/2} \boldsymbol{\varepsilon}$. Observe that $\boldsymbol{\Sigma}_0 = E[\epsilon \epsilon^T]$ and $tr(\boldsymbol{B}_T \boldsymbol{B}_T) = m_n s_n$, take $\gamma_n \to \infty$, by Markov's inequality, we have

$$\mathbb{P}\left(\|\boldsymbol{B}_{T}\boldsymbol{\Sigma}_{0}^{-1/2}\boldsymbol{\varepsilon}\|_{2} \geq \sqrt{m_{n}s_{n}\gamma_{n}}\right) \leq \frac{1}{m_{n}s_{n}\gamma_{n}}E[\|\boldsymbol{B}_{T}\boldsymbol{\Sigma}_{0}^{-1/2}\boldsymbol{\varepsilon}\|_{2}^{2}]$$

$$= \frac{1}{m_{n}s_{n}\gamma_{n}}tr\{\boldsymbol{B}_{T}\boldsymbol{\Sigma}_{0}^{-1/2}E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{T}]\boldsymbol{\Sigma}_{0}^{-1/2}\boldsymbol{B}_{T}\}$$

$$= \frac{1}{\gamma_{n}} \to 0$$

Then we have

$$\|\boldsymbol{B}_T \boldsymbol{\Sigma}_0^{-1/2} \boldsymbol{\varepsilon}\|_2 = O_P(\sqrt{m_n s_n \gamma_n})$$
(88)

By lemma B.9, we have

$$(|A| - |T|)^{-1/2} \|\tilde{R}_{1,A}\|_2 = O_P(m_n^{1/2}(\gamma_n \log p_n)^{\xi})$$
(89)

Finally, we have

$$\|\boldsymbol{\Sigma}_{0}^{-1/2}\boldsymbol{\nu}_{A}\|_{2} \leq C\|\boldsymbol{\nu}_{A}\|_{2}$$

$$\leq C\left(\sum_{i=1}^{n}|\Phi_{i}^{T}(\hat{\boldsymbol{\beta}}^{*}(A)-\boldsymbol{\beta})^{*}(A))|^{4}\right)^{1/2}$$

$$\leq C\left(\sum_{i=1}^{n}\|\Phi_{iA}\|_{2}^{4}\|\hat{\boldsymbol{\beta}}^{*}(A)-\boldsymbol{\beta})^{*}(A)\|_{2}^{4}\right)^{1/2}$$

$$\leq Cm_{n}|A|n^{1/2}\|\hat{\boldsymbol{\beta}}^{*}(A)-\boldsymbol{\beta})^{*}(A)\|_{2}^{2}$$

$$= m_{n}^{2}|A|^{2}O_{P}\left(\gamma_{n}^{2}L_{n}^{2}\frac{\log p_{n}}{\sqrt{n}}\right)$$
(90)

Combining (88), (89) and (90), we have

$$I_1(A) = (\boldsymbol{y} - \boldsymbol{\mu}_y)^T \boldsymbol{\Sigma}_0^{-1/2} \boldsymbol{B}_A \boldsymbol{\Sigma}_0^{-1/2} (\boldsymbol{y} - \boldsymbol{\mu}_y) + O_P \left(|A|^{5/2} m_n^{5/2} \gamma_n^{5/2} L_n^2 \frac{(\log p_n)^{1+\xi/2}}{\sqrt{n}} \right)$$

Then we look at $I_2(A)$. We have

$$I_{2}(A) = \frac{1}{2} (\hat{\boldsymbol{\beta}}^{*}(A) - \boldsymbol{\beta}^{*}(A))^{T} \Phi^{T} \boldsymbol{\Sigma}_{0} \Phi (\hat{\boldsymbol{\beta}}^{*}(A) - \boldsymbol{\beta}^{*}(A))$$
$$= \frac{1}{2} (\boldsymbol{y} - \boldsymbol{\mu}_{y})^{T} \boldsymbol{\Sigma}_{0}^{-1/2} \boldsymbol{B}_{A} \boldsymbol{\Sigma}_{0}^{-1/2} (\boldsymbol{y} - \boldsymbol{\mu}_{y}) + \frac{1}{2} R_{2,A} - R_{1,A}$$

where

$$R_{2,A} = \nu_{A} \Sigma_{0}^{-1/2} \boldsymbol{B}_{A} \Sigma_{0}^{-1/2} \boldsymbol{\mu}_{A}$$

$$\leq C \|\nu_{A}\|_{2}^{2}$$

$$\leq C m_{n}^{2} |A|^{2} n \|\hat{\boldsymbol{\beta}}^{*}(A) - \boldsymbol{\beta}^{*}(A)\|_{2}^{4}$$

$$= O\left(m_{n}^{2} |A|^{4} \gamma_{n}^{4} L_{n}^{4} \frac{m_{n}^{2} (\log p_{n})^{2}}{n^{2}} n\right)$$

$$= O\left(|A|^{4} m_{n}^{4} \gamma_{n}^{4} L_{n}^{4} \frac{(\log p_{n})^{2}}{n}\right)$$

Therefore,

$$\begin{split} I_2(A) = & \frac{1}{2} (\boldsymbol{y} - \boldsymbol{\mu}_y)^T \boldsymbol{\Sigma}_0^{-1/2} \boldsymbol{B}_A \boldsymbol{\Sigma}_0^{-1/2} (\boldsymbol{y} - \boldsymbol{\mu}_y) + O_P \left(|A|^{5/2} m_n^{5/2} \gamma_n^{5/2} L_n^2 \frac{(\log p_n)^{1+\xi/2}}{\sqrt{n}} \right) \\ & + O \left(|A|^4 m_n^4 \gamma_n^4 L_n^4 \frac{(\log p_n)^2}{n} \right) \end{split}$$

Finally, we have for $I_3(A)$ that

$$|I_3(A)| \le Cn|A|^{3/2} m_n^{3/2} ||\hat{\boldsymbol{\beta}}^*(A) - \boldsymbol{\beta}^*(A)||_2^3$$
$$= O_P \left(|A|^3 m_n^3 \gamma_n^3 L_n^3 \frac{(\log p_n)^{3/2}}{\sqrt{n}} \right)$$

Combining the three results for $I_1(A)$, $I_2(A)$ and $I_3(A)$, we get the desired result. \square