# Appendices

## A Proof of Theorem 2.1

First, a useful lemma is given.

**Lemma A.1.** Denote $\mathcal{F}_u = \{\int_{\Omega_{-u}} \left(f(x) - \sum_{v \subset u} f_v(x)\right) \mathrm{d}x_{-u} | f \in \mathcal{N}_\Phi, f_v \in \mathcal{F}_v\}$. Suppose $\Phi \in \Omega \times \Omega \to \mathbb{R}$ is a symmetric positive-definite kernel on $\Omega = [0,1]^d$ and $\Phi$ is a product kernel. Then,

$$f_u \in \mathcal{F}_u = \{f_v + g_u | g_u \in \mathcal{N}_{\Phi_u}, v \subset u, f_v \in \mathcal{F}_v\},$$

where $\Phi_u = \prod_{j \in u} \phi_j$.

*Proof.* Initially consider a finite element. The proof proceeds by induction. For $u = \emptyset$, we have that if $f \in \mathcal{N}_\Phi$, then

$$f_\emptyset = \int_\Omega f(x)\mathrm{d}x = \int_\Omega \sum_{y \in X} \beta_y \Phi(x,y)\mathrm{d}x = \sum_{y \in X} \beta_y \int_\Omega \Phi(x,y)\mathrm{d}x := \alpha \in \mathbb{R}.$$

This shows $f_\emptyset \in \mathcal{F}_\emptyset = \{f(\cdot) = \alpha | \alpha \in \mathbb{R}\}$.

Let $f_u \in \mathcal{F}_u$ for any $|u| \leq k$. Note that $\int_{\Omega_{-u}} \mathrm{d}x_{-u} = 1$ for any $u$, since $\Omega = [0,1]^d$. Thus, for $|u'| = k+1$,

$$\begin{aligned}
f_{u'}(x) &= \int_{\Omega_{-u'}} \left( f(x) - \sum_{v \subset u'} f_v(x) \right) \mathrm{d}x_{-u'} = \int_{\Omega_{-u'}} f(x)\mathrm{d}x_{-u'} - \sum_{v \subset u'} f_v(x) \\
&= \sum_{y \in X} \beta_y \int_{\Omega_{-u'}} \Phi(x,y)\mathrm{d}x_{-u'} - \sum_{v \subset u'} f_v(x) \\
&= \sum_{y \in X} \beta_y \int_{\Omega_{-u'}} \prod_{j=1}^{d} \phi(x_j, y_j)\mathrm{d}x_{-u'} - \sum_{v \subset u'} f_v(x) \\
&= \sum_{y \in X} \beta_y \prod_{j \in u'} \phi_j(x_j, y_j) \int_{\Omega_{-u'}} \prod_{j \notin u'} \phi_j(x_j, y_j)\mathrm{d}x_{-u'} - \sum_{v \subset u'} f_v(x) \\
&= \sum_{y \in X} \tilde{\beta}_y \prod_{j \in u'} \phi_i(x_i, y_i) - \sum_{v \subset u'} f_v(x),
\end{aligned}$$

where $\tilde{\beta}_y = \beta_y \int_{\Omega_{-u'}} \prod_{j \notin u'} \phi_j(x_j, y_j) \mathrm{d}x_{-u'}$. Hence, since $\sum_{y \in X} \tilde{\beta}_y \phi_{u'}(\cdot, y_i) \in \mathcal{N}_{\Phi_{u'}}$ and $f_v \in \mathcal{F}_v$ for any $|v| \leq k$, we have $f_{u'} \in \mathcal{F}_{u'} = \{f = f_v + g_{u'} | g_{u'} \in \mathcal{N}_{\Phi_{u'}}, v \subset u', f_v \in \mathcal{F}_v\}$. Therefore, by induction, $f_u \in \mathcal{F}_u = \{f_v + g_u | g_u \in \mathcal{N}_{\Phi_u}, v \subset u, f_v \in \mathcal{F}_v\}$ is true for any $u \subseteq D$.

Since any element of an RKHS is bounded (Aronszajn, 1950), we may use the dominated convergence theorem (Bartle, 1995) to interchange the integral and the limit of the finite sums to extend to an arbitrary element. □

By Lemma A.1, we have $f(x) = \sum_{u \subseteq D} f_u(x)$, where $f_u(x) \in \mathcal{F}_u = \{f_v + g_u | g_u \in \mathcal{N}_{\Phi_u}, v \subset u, f_v \in \mathcal{F}_v\}$. Thus, by the fact that $g_u^{(1)} + g_u^{(2)} \in \mathcal{N}_{\Phi_u}$ for $g_u^{(1)}, g_u^{(2)} \in \mathcal{N}_{\Phi_u}$, $f(x)$ can be represented as $f(x) = \sum_{u \subseteq D} f_u(x)$, where $f_u \in \mathcal{N}_{\Phi_u}$.

# B Algorithm for Estimation

1. Let $\mathcal{A}$ denote the set of active groups and $\mathcal{C}$ the set of candidate groups. Start with $\mathcal{A} = \emptyset$ and $\mathcal{C} = \{(u, r) | u = \{1\}, \ldots, \{d\}, r = 1\}$. Set an initial penalty $\lambda_{\max}$ and a small increment $\Delta$.

2. Set up an overlapping group lasso algorithm which minimizes the penalized likelihood function

$$\frac{1}{n} \sum_{i=1}^{n} \left( y_i - \sum_{(u,r) \in \mathcal{C}} \sum_{k=1}^{n_u(r)} \beta_u^{rk} \varphi_u^{rk}(x_{iu}) \right)^2 + \lambda \sum_{(u,r) \in \mathcal{C}} \sqrt{N_u(r) \sum_{v \subseteq u} \sum_{s \leq r} \sum_{k=1}^{n_v(s)} (\beta_v^{sk})^2}.$$

   Denote the input-output function as $\hat{\beta}_\lambda = \texttt{grplasso}(\lambda, \mathcal{C}, \hat{\beta}_{\lambda+\Delta})$. The inputs include a penalty value $\lambda$, the candidate set $\mathcal{C}$ and the estimated coefficient with penalty value $\lambda + \Delta$, and the output $\hat{\beta}_\lambda$ is the corresponding estimated coefficient by the algorithm. Start with $\lambda = \lambda_{\max}$ and $\hat{\beta}_{\lambda+\Delta} = 0$.

3. Do $\hat{\beta}_\lambda = \texttt{grplasso}(\lambda, \mathcal{C}, \hat{\beta}_{\lambda+\Delta})$ and obtain the set of active groups $\mathcal{A}' \subseteq \mathcal{C}$ based on $\hat{\beta}_\lambda$. Set $\lambda = \lambda - \Delta$. If $\mathcal{A}' \setminus \mathcal{A} \neq \emptyset$, then $\mathcal{A} \leftarrow \mathcal{A}'$ and $\mathcal{C} \leftarrow \mathcal{C} \cup \mathcal{C}'$, where $\mathcal{C}'$ contains the new candidate groups necessary to satisfy strong effects heredity given the updated $\mathcal{A}$.

4. Repeat step 3 until some convergence criterion is met.

## C Confidence Interval Algorithm

1. Let $\varphi^*$ denote the basis function evaluations at a particular predictive location $x^*$. Extend $\varphi^*$ to a basis of $\mathbb{R}^p$ and denote it as $A = (\varphi^*, c_2, \ldots, c_p)$. Compute $(\tilde{Z}_i, \tilde{Q}_i)^T = A^{-1}\varphi_i$ for $i = 1, \ldots, n$ and $(\hat{\eta}_1, \hat{\eta}_{(-1)}^T) = A^T\hat{\beta}_\lambda$, where $\hat{\beta}_\lambda$ is the estimated coefficient with penalty $\lambda$.

2. Compute the estimated decorrelated score function

$$\hat{S}(0, \hat{\eta}_{(-1)}) = -\frac{1}{n\hat{\sigma}^2}\sum_{i=1}^{n}(y_i - \hat{\eta}_{(-1)}^T\tilde{Q}_i)(\tilde{Z}_i - \hat{w}^T\tilde{Q}_i),$$

   where

$$\hat{w} = \arg\min \left\| \frac{1}{n}\sum_{i=1}^{n}\tilde{Q}_i(\tilde{Z}_i - w^T\tilde{Q}_i) \right\|_2 + \lambda''\|w\|_1,$$

   and $\hat{\sigma}^2$ is a consistent estimator of $\sigma^2$. For example, $\sigma^2$ can be estimated by $\hat{\sigma}^2 = \frac{1}{n-s}\sum_{i=1}^{n}(y_i - \hat{\beta}_\lambda^T\varphi_i)^2$, where $s$ the the number of non-zero elements in $\hat{\beta}_\lambda$. Another estimator is the cross-validation based variance estimator. Define the $K$ cross-validation folds as $\{D_1, \ldots, D_K\}$ and compute

$$\hat{\sigma}^2 = \min_\lambda \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in D_k}(y_i - (\hat{\beta}_\lambda^{(-k)})^T\varphi_i)^2,$$

   where $\hat{\beta}_\lambda^{(-k)}$ is the overlapping group lasso estimate at $\lambda$ over the data after the $k^{th}$ fold is omitted. This estimator has been used for the variance estimation in lasso regression problems. See Fan et al. (2012).

3. Compute the interval

$$[c_{\alpha/2}/b, c_{1-\alpha/2}/b],$$

   where $c_{\alpha/2} = -\hat{S}(0, \hat{\eta}_{(-1)}) + \sqrt{\frac{b}{n}}\Phi^{-1}(\alpha/2)$, $c_{1-\alpha/2} = -\hat{S}(0, \hat{\eta}_{(-1)}) + \sqrt{\frac{b}{n}}\Phi^{-1}(1-\alpha/2)$, $b = \frac{1}{n\hat{\sigma}^2}\sum_{i=1}^{n}\tilde{Z}_i(\tilde{Z}_i - \hat{w}^T\tilde{Q}_i)$. By some algebraic manipulation, one can show that this interval is same as the one in Corollary 5.1.

# D Confidence Interval Algorithm Modification for Large $n$

1. In Algorithm C, replace $\tilde{Q}_i$ by $\tilde{Q}_{*i}$ and $p$ by $p_*$, where the nuisance $\varphi_{ij}$, $j = 1, \ldots, p_*$ only contain basis functions in the candidate groups at the selected $\lambda$, say $\mathcal{C}_\lambda$.

2. Replace $\hat{w}$ by
$$\hat{w}_* = \left( \sum_{i=1}^{n} \tilde{Q}_{*i} \tilde{Q}_{*i}^T + \eta I_{p_*-1} \right)^{-1} \left( \sum_{i=1}^{n} \tilde{Q}_{*i} \tilde{Z}_i \right) \tag{D.1}$$
with a small positive $\eta$, where $I_{p_*-1}$ is a $(p_* - 1) \times (p_* - 1)$ identity matrix.

3. For the deterministic case (4),

   (i) Define $K$ cross-validation folds as $\{D_1, \ldots, D_K\}$ and partition the original samples $\{x_i, y_i\}_{i=1}^n$ via the $k$ folds.

   (ii) Regard $\hat{\sigma}^2$ in Algorithm C as an unknown parameter. Let $\hat{u}^{(-k)}(x^*, \hat{\sigma}^2)$ and $\hat{l}^{(-k)}(x^*, \hat{\sigma}^2)$ be the upper and lower limits at a predictive location $x^*$ by Algorithm C over the data after the $k^{th}$ fold is omitted, respectively.

   (iii) Replace $\hat{\sigma}^2$ by
$$\hat{\sigma}_*^2 = \arg\min_{\hat{\sigma}^2} \left| \left( \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in D_k} \mathbb{1}\{ y_i \in [\hat{l}^{(-k)}(x_i, \hat{\sigma}^2), \hat{u}^{(-k)}(x_i, \hat{\sigma}^2)] \} \right) - (1 - \alpha) \right|,$$
   where $\mathbb{1}\{A\}$ is an indicator function of the set $A$.

# E Proof of Theorem 4.1

## E.1 Notation and Reformulation

First, we introduce some additional notation. For a matrix $M = [M_{jk}]$, let $\|M\|_{\max} = \max_{j,k} |M_{jk}|$, $\|M\|_1 = \sum_{j,k} |M_{jk}|$, and $\|M\|_{l_\infty} = \max_j \sum_k |M_{jk}|$. For $v = (v_1, ..., v_p)^T \in \mathbb{R}^p$, and $1 \leqslant q < \infty$, define $\|v\|_q = (\sum_{i=1}^p |v_i|^q)^{1/q}$. Define $\|v\|_0 = |\{i : v_i \neq 0\}|$. For $S \subseteq \{1, ..., p\}$, let $v_S = \{v_j : j \in S\}$ and $\bar{S}$ be the complement of $S$. Given $a, d \in \mathbb{R}$, we use $a \vee b$ and $a \wedge b$ to denote the maximum and minimum of $a$ and $b$.

For convenience, we restate the loss function as follows. Consider groups $J_1, ..., J_{p_n}$, where $J_j \subseteq \{1, ..., p\}$, and $\bigcup_{j=1}^{p_n} J_j = \{1, ..., p\}$. Notice that we do not require $J_{j_1} \bigcap J_{j_2} = \emptyset$.

Define $C_k = \{j : k \in J_j\}$ and $c_k = |C_k|$. Thus, $C_k$ is the set of indices of the groups variable $k$ belongs to and $c_k$ is the number of groups that variable $k$ belongs to. We can also treat $c_k$ as replicates of index $k$. For notational simplicity, in the proof we write $\hat{\beta}_n$ and $\beta_n^*$ as $\hat{\beta}$ and $\beta^*$, respectively. We also write $\varphi_n(X_i)$ as $\varphi_i$ for simplicity. Define the vector of variable $k$ coefficients over all groups in which it appears $\beta_{kC_k}^Z = (\beta_{kj_{k1}}, \ldots, \beta_{kj_{kc_k}})^T$, where $j_{kl}$ denotes the index of variable $k$ within the $l^{\text{th}}$ group in which it appears, and the vector of all coefficients $\beta^Z = ((\beta_{1C_1}^Z)^T, \ldots, (\beta_{pC_p}^Z)^T)^T$. Let $\beta_{J_j} = (\beta_{kj})_{k \in J_j}^T$, where $\beta_{kj}$ is the coefficient of the $k^{th}$ variable and $k$ is in $j^{th}$ group. Let $d_j = |J_j|$. Consider the following optimization problem

$$\hat{\beta}^{Z,\lambda_n} = \arg\min_{\beta^Z} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \sum_{k=1}^{p} \left( \sum_{m=1}^{c_k} \beta_{kj_{km}} \right) \varphi_{ki})^2 + \lambda_n \sum_{j=1}^{p_n} \sqrt{d_j} \|\beta_{J_j}\|_2 \right\}, \tag{E.2}$$

where $\lambda_n$ is a positive number. We define the overlapping group lasso estimator as

$$\hat{\beta}^{\lambda_n} = \left( \sum_{k=1}^{c_1} \hat{\beta}_{1j_{1k}}^{\lambda_n}, ..., \sum_{k=1}^{c_p} \hat{\beta}_{pj_{pk}}^{\lambda_n} \right)^T, \tag{E.3}$$

in which we stress $\lambda_n$ since it will influence the solution of (E.2). Notice that by this definition, the least squares term becomes $\frac{1}{2n} \sum_{i=1}^{n} (y_i - \varphi_i^T \hat{\beta}^{\lambda_n})^2$, which is the same as in original group lasso case. We use $\frac{1}{2n}$ instead of $\frac{1}{n}$ for brevity of the Karush-Kuhn-Tucker (KKT) conditions, which are as following.

**Proposition E.1.** Let $\varphi$ be the matrix with rows $\varphi_i^T$, $i = 1, \ldots, n$. Let $\psi_j$ denote the $j^{th}$ column of $\varphi$, for $j = 1, \ldots, p$. Necessary and sufficient conditions for $\hat{\beta}^Z$ to be a solution to (E.2) are

$$\begin{aligned} -\frac{1}{n} \psi_j^T (y - \varphi \hat{\beta}^{\lambda_n}) + \frac{\lambda_n \sqrt{d_k} \hat{\beta}_{jk}^{\lambda_n}}{\|\hat{\beta}_{J_k}^{\lambda_n}\|_2} = 0, \qquad & \forall j \in J_k \text{ with } \hat{\beta}_{J_k}^{\lambda_n} \neq 0 \\ \| -\frac{1}{n} \psi_j^T (y - \varphi \hat{\beta}^{\lambda_n}) \|_2 \leqslant \lambda_n \sqrt{d_k}, \qquad & \forall j \in J_k \text{ with } \hat{\beta}_{J_k}^{\lambda_n} = 0. \end{aligned}$$

The following lemma Liu and Zhang (2009) states that at most $n$ groups can be nonzero.

**Lemma E.1.** Suppose $\lambda_n > 0$, a solution $\hat{\beta}^{Z,\lambda_n}$ exists such that the number of nonzero groups $|S(\hat{\beta}^{Z,\lambda_n})| \leqslant n$, the number of data points, where $S(\beta) = \{J_j : \hat{\beta}_{J_j} \neq 0\}$.

*Proof.* The proof of Lemma 1 in Liu and Zhang (2009) is also valid here. □

By Lemma E.1, for brevity, sometimes we say $\hat{\beta}^{\lambda_n}$ with $|S(\hat{\beta}^{Z,\lambda_n})| \leqslant n$, which is derived by combining (E.2) and (E.3), is the solution of (E.2). We will also write $\|y-\varphi\beta\|_2^2$ instead of $\sum_{i=1}^n \left( y_i - \sum_{k=1}^p \left( \sum_{m=1}^{c_k} \beta_{kj_{km}} \right) \varphi_{ki} \right)^2$. Let $\bar{c} = \max_j\{c_1, ..., c_p\}$ and $\bar{d} = \max_j\{d_1, ..., d_{p_n}\}$, the maximum number of groups a variable appears in and maximum group size, respectively. Let $s$ be the number of nonzero elements in $\beta^*$ and $p$ be the dimension of $\beta^*$. Notice that $s$ and $p$ (as well as $\bar{c}$ and $\bar{d}$) can depend $n$.

### E.2 Proof of Theorem 4.1

Our proof follows a similar line to Meinshausen and Yu (2009), but extends their results to the overlapping group lasso. We only need to show the stochastic case. The deterministic case is true because the proof is still valid by taking $\epsilon = 0$. A sketch of the proof is as follows. We first define the coefficients obtained from the de-noised model as a de-noised estimator. Then, by showing the difference between the de-noised estimator and true coefficients, and the difference between de-noised estimator and the estimator obtained via overlapping group lasso are both small, we obtain $l_2$ convergence. All the proofs of the lemmas in this section are in Appendix H.

Before we state and prove the main result, we introduce a definition which is useful in the proof.

**Definition E.1.** Denote $y(\xi) = \varphi\beta^* + \xi(\epsilon+\delta)$ as a de-noised model with level $\xi$ $(0 \leqslant \xi \leqslant 1)$, we define

$$\hat{\beta}^{\lambda,\xi} = \arg\min_{\beta} \frac{1}{2n}\|y(\xi) - \varphi\beta\|_2^2 + \lambda_n \sum_{j=1}^{p_n} \sqrt{d_j}\|\beta_{J_j}\|_2 \tag{E.4}$$

to be the de-noised estimator at noise level $\xi$, where $\hat{\beta}^{\lambda,\xi}$ is defined similarly as in (E.3).

In order to characterize the eigenvalues of a matrix under sparsity, we introduce the following definition, which can be found in Meinshausen and Yu (2009).

**Definition E.2.** The $m$-sparse minimum and maximum eigenvalue of a matrix $C = \frac{1}{n}\varphi^T\varphi$ are $\phi_{\min}(m) = \min_{\beta:\|\beta\|_0 \leqslant m} \frac{\beta^T C\beta}{\beta^T\beta}$ and $\phi_{\max}(m) = \max_{\beta:\|\beta\|_0 \leqslant m} \frac{\beta^T C\beta}{\beta^T\beta}$. Also, denote $\phi_{\max} = \phi_{\max}((s\bar{c}+n)\bar{d})$ where $s$, $\bar{c}$, and $\bar{d}_n$ are defined as in section E.1.

Now we introduce an assumption concerning $\phi_{\min}(\cdot)$ and $\phi_{\max}$. Detailed discussion has been shown in Meinshausen and Yu (2009).

**Assumption E.1.** There exist constants $0 < \kappa_{\min} \leqslant \kappa_{\max} < \infty$ such that $\liminf_{n\to\infty} \phi_{\min}(s\bar{c}\bar{d}\max\{\log n, \bar{c}\}) \geqslant \kappa_{\min}$ and $\limsup_{n\to\infty} \phi_{\max} \leqslant \kappa_{\max}$.

For continuity, we repeat Theorem 4.1 here.

**Theorem 4.1.** Under Assumption E.1, if $\lambda_n \asymp \sigma\sqrt{\frac{\log p}{n}}$, $\bar{d}^2 = o(\log n)$, and $\|y(\cdot) - \varphi(\cdot)^T\beta^*\|_\infty = O_p(\lambda_n)$, for the (overlapping) group lasso estimator constructed in (E.2) and (E.3), with probability tending to 1 for $n \to \infty$,

$$\|\hat{\beta}^{\lambda_n} - \beta^*\|_2^2 \lesssim \frac{\bar{c}^2 s\bar{d}\log p}{n}.$$

Let $\beta^{\lambda_n} = \hat{\beta}^{\lambda_n,0}$. The $l_2$-consistency can be obtained by bounding the bias and variance terms, i.e.

$$\|\hat{\beta}^{\lambda_n} - \beta^*\|_2^2 \leqslant 2\|\hat{\beta}^{\lambda_n} - \beta^{\lambda_n}\|_2^2 + 2\|\beta^{\lambda_n} - \beta^*\|_2^2.$$

**Remark E.1.** The condition $\|y(\cdot) - \varphi(\cdot)^T\beta^*\|_\infty = O_p(\lambda_n)$ implies $B_i = O_p(\lambda_n)$. In the proof of Theorem 4.1, the condition $B_i = O_p(\lambda_n)$ is sufficient.

Let $T = \{t : \beta_i^* \neq 0, \beta_{it}^* \text{ is a component of } \beta^{Z*}\}$ represent the set of indices for all the groups with possibly nonzero coefficient vectors. Let $s_n = |T|$. Thus, $s_n \leqslant s\bar{c}$. The solution $\beta^{\lambda_n}$ can, for each value of $\lambda_n$, be written as $\beta^{\lambda_n} = \beta^* + \gamma^{\lambda_n}$, where $\gamma^{\lambda_n}$ is defined as the solution of the following optimization problem:

$$\begin{aligned} \arg\min_{\gamma} \quad & f(\gamma, \gamma^Z) \\ \text{s.t.} \quad & \sum_{k=1}^{c_i} \beta_{ik}^Z = \beta_i^*, \quad i = 1, ..., p; \\ & \sum_{k=1}^{c_i} \gamma_{ij_{ik}}^Z = \gamma_i, \quad i = 1, ..., p, \end{aligned} \tag{E.5}$$

where

$$f(\gamma, \gamma^Z) = n\gamma^T A\gamma + \lambda_n \sum_{t\in T^c} \sqrt{d_t}\|\gamma_t^Z\|_2 + \lambda_n \sum_{t\in T} \sqrt{d_t}(\|\gamma_t^Z + \beta_t^Z\|_2 - \|\beta_t^Z\|_2),$$

where $A = \frac{1}{n}\varphi^T\varphi$. This optimization problem is obtained by plugging $\beta^* + \gamma^{\lambda_n}$ into (E.4). Notice the arg min problem is with respect to $\gamma$ instead of $(\gamma, \gamma^Z)$.

Next, we state a lemma which bounds the $l_2$-norm of $\gamma^{\lambda_n}$. Its proof is provided in Appendix H.1.

**Lemma E.2.** Under Assumption E.1, with a positive constant $C$, the $l_2$-norm of $\gamma^{\lambda_n}$ is bounded for sufficiently large values of $n$ by $\|\gamma^{\lambda_n}\|_2 \leqslant \frac{\lambda_n\sqrt{\bar{c}s_n\bar{d}}}{n} \Big/ \left(\sqrt{\frac{\kappa_{\min}}{2}(1 - \frac{4\bar{d}}{\log n})} - \sqrt{\frac{2\kappa_{\max}\bar{d}^2}{\log n}}\right)$.

Now, we bound the variance term. For every subset $M \subset \{1, ..., p\}$ with $|M| \leqslant n$, denote $\hat{\theta}^M \in \mathbb{R}^{|M|}$ the restricted least square estimator of the noise $\epsilon$,

$$\hat{\theta}^M = (\varphi_M^T\varphi_M)^{-1}\varphi_M^T(\epsilon + B), \tag{E.6}$$

where $B = (B_1, .., B_n)^T$ and $\epsilon = (\epsilon_1, .., \epsilon_n)^T$. Now we state lemmas, which bound the $l_2$-norm of this estimator, and are also useful for the following parts of this development. First we define sub-exponential variables, sub-exponential norms, sub-Gaussian variables, and sub-Gaussian norms.

**Definition E.3.** (sub-exponential variable and sub-exponential norm) A random variable $X$ is called sub-exponential if there exists some positive constant $K_1$ such that $\mathbb{P}(|X| > t) \leqslant \exp(1 - t/K_1)$ for all $t \geqslant 0$. The sub-exponential norm of $X$ is defined as $\|X\|_{\psi_1} = \sup_{q\geqslant 1} q^{-1}(\mathbb{E}|X|^q)^{1/q}$.

**Definition E.4.** (sub-Gaussian variable and sub-Gaussian norm) A random variable $X$ is called sub-Gaussian if there exists some positive constant $K_2$ such that $\mathbb{P}(|X| > t) \leqslant \exp(1 - t^2/K_2)$ for all $t \geqslant 0$. The sub-Gaussian norm of $X$ is defined as $\|X\|_{\psi_2} = \sup_{q\geqslant 1} q^{-1/2}(\mathbb{E}|X|^q)^{1/q}$.

**Lemma E.3.** Let $\bar{m}_n$ be a sequence with $\bar{m}_n = o(n)$ and $\bar{m}_n \to \infty$ for $n \to \infty$

$$\max_{M:|M|\leqslant \bar{m}_n} \|\theta^M\|_2^2 \leqslant C^2 \frac{\bar{m}_n \log p}{n\phi^2_{\min}(\bar{d})}.$$

*Proof.* See Appendix H.2. □

Now define $A_{\lambda_n,\xi}$ to be

$$A_{\lambda_n,\xi} = \left\{ k : \lambda_n \frac{\sqrt{d_k}\hat{\beta}_{jk}}{\|\hat{\beta}_{J_k}\|_2} = \frac{1}{n}\psi_j^T(Y(\xi) - \varphi\hat{\beta}), \text{ with } j \in J_k \right\},$$

which represents the set of active groups for the de-noised problem.

**Lemma E.4.** If, for a fixed value of $\lambda_n$, the number of active variables of the de-noised estimators $\hat{\beta}^{\lambda_n,\xi}$ is for every $0 \leqslant \xi \leqslant 1$ bounded by $m'$, then

$$\|\hat{\beta}^{\lambda_n,0} - \hat{\beta}_n^{\lambda}\|_2^2 \leqslant C \max_{M:|M|\leqslant m'} \|\theta^M\|_2^2.$$

*Proof.* See Appendix H.3. □

The next lemma provides an asymptotic upper bound on the number of selected variables.

**Lemma E.5.** For $\lambda_n \geqslant \sqrt{\frac{\log p}{n}}$, the maximal number of selected variables, $\sup_{0\leqslant\xi\leqslant 1} \sum_{k\in A_{\lambda,\xi}} d_k$, is bounded, with probability tending to 1 for $n \to \infty$, by

$$\sup_{0\leqslant\xi\leqslant 1} \sum_{k\in A_{\lambda,\xi}} d_k \leqslant C_1 s_n \bar{d}\bar{c}.$$

*Proof.* See Appendix H.4. □

Now combining Lemmas E.3, E.4, and E.5, we have

$$\|\hat{\beta}^{\lambda_n,0} - \hat{\beta}^{\lambda_n}\|_2^2 \leqslant C \frac{s\bar{d}\bar{c}^2 \log p}{n\phi^2_{\min}(s\bar{d}\bar{c}^2)}.$$

Combining this and Lemma E.2, gives

$$\begin{aligned}\|\hat{\beta}^{\lambda_n}-\beta\|_2^2 &\leqslant C\frac{s\bar{d}\bar{c}^2\log p}{n\phi_{\min}^2(s\bar{d}\bar{c}^2)}+\frac{\lambda_n^2\bar{c}^2 s\bar{d}}{n^2}\Big/\left(\sqrt{\frac{\kappa_{\min}}{2}(1-\frac{4\bar{d}}{\log n})}-\sqrt{\frac{2\kappa_{\max}\bar{d}^2}{\log n}}\right)^2\\ &\leqslant C\frac{s\bar{d}\bar{c}^2\log p}{n}+C\frac{\bar{c}^2 s\bar{d}\log p}{n}\Big/\left(\sqrt{\frac{\kappa_{\min}}{2}(1-\frac{4\bar{d}}{\log n})}-\sqrt{\frac{2\kappa_{\max}\bar{d}^2}{\log n}}\right)^2\\ &\lesssim \frac{\bar{c}^2 s\bar{d}\log p}{n},\end{aligned}$$

which completes the proof of Theorem 4.1.

# F Proof of Corollary 4.1

Since $\beta^*$ satisfies (3),

$$\int_\Omega \varphi(x)(y(x)-\varphi(x)^T\beta^*)\mathrm{d}x=0.$$

Therefore, the oracle risk of $\hat{\beta}$ can be bounded by

$$\begin{aligned}&\int_\Omega (y(x)-\varphi(x)^T\hat{\beta})^2\mathrm{d}x-\int_\Omega (y(x)-\varphi(x)^T\beta^*)^2\mathrm{d}x\\ =&\int_\Omega (2y(x)-\varphi(x)^T\hat{\beta}-\varphi(x)^T\beta^*)(\varphi(x)^T(\beta^*-\hat{\beta}))\mathrm{d}x\\ =&\int_\Omega (2y(x)-2\varphi(x)^T\beta^*+\varphi(x)^T\beta^*-\varphi(x)^T\hat{\beta})(\varphi(x)^T(\beta^*-\hat{\beta}))\mathrm{d}x\\ =&\int_\Omega (\varphi(x)^T\beta^*-\varphi(x)^T\hat{\beta})(\varphi(x)^T(\beta^*-\hat{\beta}))\mathrm{d}x\\ =&\int_\Omega (\beta^*-\hat{\beta})^T\varphi(x)\varphi(x)^T(\beta^*-\hat{\beta})\mathrm{d}x\\ \leqslant& C\|\beta^*-\hat{\beta}\|_2^2,\end{aligned}$$

where the last inequality is because of Assumption E.1. Because $\|y(\cdot)-\varphi(\cdot)^T\beta^*\|_\infty=O_p(\lambda_n)$, we have $\int_\Omega(y(x)-\varphi(x)^T\beta^*)^2\mathrm{d}x=O_p(\lambda_n^2)$, which completes the proof.

# G Proof of Theorem 5.1

In this section we will prove Theorem 5.1. A sketch of proof is as follows, following the overall approach in Ning and Liu (2017). First, we introduce a decorrelated score function, and prove the decorrelated function converges weakly to a normal distribution under $l_2$-consistency, which is stated in Theorem G.1. The result is then applied to the overlapping group lasso model with known variance of error. Then by showing the difference between the decorrelated score function with known variance and decorrelated score function with estimated variance is small, we finish the proof of Theorem 5.1.

## G.1 Hypothesis Test based on Decorrelated Function and $l_2$-Consistency

In this section, we will introduce a decorrelated score function, and prove several results similar to Ning and Liu (2017) but with $l_2$-consistency instead of $l_1$. Suppose we are given $n$ independently identically distributed $U_1, ..., U_n$, which come from the same probability distribution following from a high dimensional statistical model $\mathcal{P} = \{\mathbb{P}_\beta : \beta \in \Omega\}$, where $\beta$ is a $p$ dimensional unknown parameter and $\Omega$ is the parameter space. Let the true value of $\beta$ be $\beta^*$, which is sparse in the sense that the number of non-zero elements of $\beta$ is much smaller than $n$, order $\log n$. We consider the case in which we are interested in only one parameter. Suppose $\beta = (\beta_1, \beta_{-1})$, where $\beta_1 \in \mathbb{R}$ and $\beta_{-1} \in \mathbb{R}^{p-1}$. Let $\beta_1^*$ and $\beta_{-1}^*$ be the true value of $\beta_1$ and $\beta_{-1}$, respectively. For simplicity, we assume the null hypothesis is $H_0 : \beta_1^* = 0$, which can be generalized to the case $\beta_1^* = \beta_{1,0}$ in a straight forward manner. Suppose the negative log-likelihood function is

$$\ell(\beta_1, \beta_{-1}) = \frac{1}{n} \sum_{i=1}^{n} (-\log f(U_i; \beta_1, \beta_{-1})),$$

where $f$ is the p.d.f. corresponding to the model $\mathbb{P}_\beta$, which it will be assumed has at least two continuous derivatives with respect to $\beta$. The information matrix for $\beta$ is defined as $I = \mathbb{E}_\beta(\nabla^2 \ell(\beta))$, and the partial information matrix is $I_{\beta_1|\beta_{-1}} = I_{\beta_1\beta_1} - I_{\beta_1\beta_{-1}} I_{\beta_{-1}\beta_{-1}}^{-1} I_{\beta_{-1}\beta_1}$, where $I_{\beta_1\beta_1}$, $I_{\beta_1\beta_{-1}}$, $I_{\beta_{-1}\beta_{-1}}$, and $I_{\beta_{-1}\beta_1}$ are the corresponding partitions of $I$. Let $I^* = \mathbb{E}_{\beta^*}(\nabla^2 \ell(\beta^*))$.

In this paper, we are considering testing parameters for high dimensional models and,

as mentioned in Ning and Liu (2017), the traditional score function does not have a simple limiting distribution in the high dimensional setting. Thus, we use a decorrelated score function as mentioned in Ning and Liu (2017) defined as

$$S(\beta_1, \beta_{-1}) = \nabla_{\beta_1} \ell(\beta_1, \beta_{-1}) - w^T \nabla_{\beta_{-1}} \ell(\beta_1, \beta_{-1}),$$

where $w = I^{-1}_{\beta_{-1}\beta_{-1}} I_{\beta_{-1}\beta_1}$. Notice that $\mathbb{E}_\beta(S(\beta)\nabla_{\beta_{-1}}\ell(\beta)) = 0$. Suppose we are given the estimator $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_{-1})$ and tuning parameter $\lambda'$. We estimate $\hat{w}$ by solving

$$\hat{w} = \arg\min \|w\|_1, \text{ s.t. } \|\nabla^2_{\beta_1\beta_{-1}} \ell(\hat{\beta}) - w^T \nabla^2_{\beta_{-1}\beta_{-1}} \ell(\hat{\beta})\|_2 \leqslant \lambda'. \tag{G.7}$$

We use this method to estimate $w$ because since $w$ has dimension $d$ which is much greater than $n$, we need some sparsity of $w$, which is useful in the rest part of this paper. Thus, we can obtain estimated decorrelated score function $\hat{S}(\beta_1, \hat{\beta}_{-1}) = \nabla_{\beta_1} \ell(\beta_1, \hat{\beta}_{-1}) - \hat{w}^T \nabla_{\beta_{-1}} \ell(\beta_1, \hat{\beta}_{-1})$.

Along the same lines as Ning and Liu (2017), we need the following assumptions. Assumption G.1 states that the estimators $\hat{\beta}$ and $\hat{w}$ converge to zero. However, we assume $l_2$-consistency here, which is weaker than the condition in Ning and Liu (2017).

**Assumption G.1.** Assume that

$$\lim_{n\to\infty} \mathbb{P}_{\beta^*}(\|\hat{\beta}_{-1} - \beta^*_{-1}\|_2 \lesssim \eta_1(n)) = 1 \text{ and } \lim_{n\to\infty} \mathbb{P}_{\beta^*}(\|\hat{w} - w^*\|_1 \lesssim \eta_2(n)) = 1,$$

where $w^* = I^{*-1}_{\beta_{-1}\beta_{-1}} I^*_{\beta_{-1}\beta_1}$, and $\eta_1(n)$ and $\eta_2(n)$ converges to 0, as $n \to \infty$.

Assumption G.2 states that the derivative of log-likelihood function is near zero at the true parameters.

**Assumption G.2.** Assume that

$$\lim_{n\to\infty} \mathbb{P}_{\beta^*}(\|\nabla_{\beta_{-1}} l(0, \beta^*_{-1})\|_\infty \lesssim \eta_3(n)) = 1,$$

for some $\eta_3(n) \to 0$, as $n \to \infty$.

Assumption G.3 states that the Hessian matrix is relative smooth, so that we can use $\lambda'$ to control $\eta_4(n)$.

**Assumption G.3.** Assume that for $\beta_{-1,\nu} = \nu\beta^*_{-1} + (1-\nu)\hat{\beta}_{-1}$ with $\nu \in [0,1]$,

$$\lim_{n\to\infty} \mathbb{P}_{\beta^*}(\sup_{\nu\in[0,1]} \|\nabla^2_{\beta_1\beta_{-1}} l(0,\beta_{-1,\nu}) - \hat{w}^T \nabla^2_{\beta_{-1}\beta_{-1}} l(0,\beta_{-1,\nu})\|_2 \lesssim \eta_4(n)) = 1,$$

for some $\eta_4(n) \to 0$, as $n \to \infty$.

Assumption G.4 is the central limit theorem for a linear combination of the score functions.

**Assumption G.4.** For $v^* = (1, -w^{*T})^T$, it holds that

$$\frac{\sqrt{n} v^{*T} \nabla l(0,\beta^*_{-1})}{\sqrt{v^T I^* v}} \xrightarrow{\text{dist.}} N(0,1),$$

where $I^* = \mathbb{E}_{\beta^*}(\nabla^2 l(0,\beta^*_{-1}))$. Furthermore, assume that $C' \leqslant I^*_{\beta_1|\beta_{-1}} < \infty$, where $I^*_{\beta_1|\beta_{-1}} = I^*_{\beta_1\beta_1} - w^{*T} I^*_{\beta_{-1}\beta_1}$, and $C' > 0$ is a constant.

Assumption G.5 states that we can estimate the information matrix relatively accurately.

**Assumption G.5.** Assume

$$\lim_{n\to\infty} \mathbb{P}_{\beta^*}(\|\nabla^2 l(\hat{\beta}) - I^*\|_{\max} \lesssim \eta_5(n)) = 1$$

for some $\eta_5(n) \to 0$, as $n \to \infty$.

Now under Assumptions G.1 to G.5, we can prove a version of Theorem 3.5 in Ning and Liu (2017) which applies to the (potentially) overlapping group lasso.

**Theorem G.1.** Under Assumptions G.1 to G.5, with probability tending to one,

$$n^{1/2} |\hat{S}(0,\hat{\beta}_{-1}) - S(0,\beta^*_{-1})| \lesssim n^{1/2}(\eta_2(n)\eta_3(n) + \eta_1(n)\eta_4(n)). \tag{G.8}$$

If $n^{1/2}(\eta_2(n)\eta_3(n) + \eta_1(n)\eta_4(n)) = o(1)$, we have

$$n^{1/2} \hat{S}(0,\hat{\beta}_{-1}) I^{*-1/2}_{\beta_1|\beta_{-1}} \xrightarrow{\text{dist.}} N(0,1). \tag{G.9}$$

*Proof.* See Theorem 3.5 in Ning and Liu (2017). The only difference is under $l_2$-consistency,

$$|I_1| \leqslant \|\nabla^2_{\beta_1\beta_{-1}} l(0, \tilde{\beta}_{-1}) - \hat{w}^T \nabla^2_{\beta_{-1}\beta_{-1}} l(0, \tilde{\beta}_{-1})\|_2 \|\hat{\beta}_{-1} - \beta^*_{-1}\|_2 \lesssim \eta_1(n)\eta_4(n).$$

□

**Corollary G.1.** Assume that Assumptions G.1 to G.5 hold. It also holds that $\|w^*\|_1 \eta_5(n) = o(1)$, $\eta_2(n)\|I^*_{\beta_1\beta_{-1}}\|_\infty = o(1)$, and $n^{1/2}(\eta_2(n)\eta_3(n) + \eta_1(n)\eta_4(n)) = o(1)$. Under $H_0 : \beta^*_1 = 0$, we have for any $t \in \mathbb{R}$,

$$\lim_{n\to\infty} |\mathbb{P}_{\beta^*}(\hat{U}_n \leqslant t) - \Phi(t)| = 0, \tag{G.10}$$

where $\hat{U} = n^{1/2}\hat{S}(0, \hat{\beta}_{-1})\hat{I}^{-1/2}_{\beta_1|\beta_{-1}}$.

*Proof.* See the proof of Corollary 3.7 in Ning and Liu (2017). □

### G.2 Linear model and the corresponding decorrelated score function

Now we apply the consequences of the general results to the linear model as described in the previous section. In this section we first assume that the variance of noise is known. Consider the linear regression, $y_i = \beta^*_1 \varphi_{i1} + \beta^{*T}_{-1}\varphi_{i,-1} + B_i + \epsilon_i$, where $\varphi_{i1} \in \mathbb{R}$, $\varphi_{i,-1} \in \mathbb{R}^{p-1}$, $B_i \in \mathbb{R}$, and the error $\epsilon_i$ satisfies $\mathbb{E}(\epsilon_i) = 0$, $\mathbb{E}(\epsilon_i^2) = \sigma^2 > 0$ for $i = 1, ..., n$. Let $\varphi_i = (\varphi_{i1}, \varphi^T_{i,-1})^T$ denote the collection of all covariates for subject $i$. We first assume $\sigma^2$ is known.

Consider the overlapping group lasso estimator (E.3), the decorrelated score function is

$$S(\beta_1, \beta_{-1}) = -\frac{1}{n\sigma^2}\sum_{i=1}^n (y_i - \beta_1\varphi_{i1} - \beta^T_{-1}\varphi_{i,-1})(\varphi_{i1} - w^T\varphi_{i,-1}),$$

where $w = \mathbb{E}_\beta(\varphi_{i,-1}\varphi^T_{i,-1})^{-1}\mathbb{E}_\beta(\varphi_{i1}\varphi_{i,-1})$. Since the distribution of the design matrix does not depend on $\beta$, we can replace $\mathbb{E}_\beta(\cdot)$ by $\mathbb{E}(\cdot)$ for notation simplicity. Under the null hypothesis, $H_0 : \beta^*_1 = 0$, the decorrelated score function can be estimated by

$$\hat{S}(0, \hat{\beta}_{-1}) = -\frac{1}{n\sigma^2}\sum_{i=1}^n (y_i - \hat{\beta}^T_{-1}\varphi_{i,-1})(\varphi_{i1} - \hat{w}^T\varphi_{i,-1}),$$

where

$$\hat{w} = \arg\min \|w\|_1, \text{ s.t. } \left\| \frac{1}{n} \sum_{i=1}^{n} \varphi_{i,-1}(\varphi_{i1} - w^T \varphi_{i,-1}) \right\|_2 \leqslant \lambda'.$$

The (partial) information matrices are

$$I^* = \sigma^{-2}\mathbb{E}(\varphi_{i,-1}\varphi_{i,-1}^T), \text{ and } I^*_{\beta_1|\beta_{-1}} = \sigma^{-2}(\mathbb{E}(\varphi_{i1}^2) - \mathbb{E}(\varphi_{i1}\varphi_{i,-1}^T)\mathbb{E}(\varphi_{i,-1}\varphi_{i,-1}^T)^{-1}\mathbb{E}(\varphi_{i,-1}\varphi_{i1})),$$

which can be estimated by

$$\hat{I} = \frac{1}{n\sigma^2} \sum_{i=1}^{n} \varphi_{i,-1}\varphi_{i,-1}^T, \text{ and } \hat{I}_{\beta_1|\beta_{-1}} = \sigma^{-2}\left\{ \frac{1}{n} \sum_{i=1}^{n} \varphi_{i1}^2 - \hat{w}^T \left( \frac{1}{n} \sum_{i=1}^{n} \varphi_{i,-1}\varphi_{i1} \right) \right\},$$

respectively. Thus, the score test statistic is $\hat{U}_n = n^{1/2}\hat{S}(0, \hat{\beta}_{-1})\hat{I}^{-1/2}_{\beta_1|\beta_{-1}}$.

The following theorem states the asymptotic distribution $\hat{U}_n$ under null hypothesis.

**Theorem G.2.** Assume that

1. $\lambda_{\min}(\mathbb{E}(\varphi_i\varphi_i^T)) \geqslant 2\kappa_{\min}$ for some constant $\kappa_{\min} > 0$, and $\limsup_{n\to\infty} \phi_{\max} \leqslant \kappa_{\max}$, where $\phi_{\max}$ is defined in Definition E.2.

2. Let $S = \text{supp}(\beta^*)$ and $S' = \text{supp}(w^*)$ satisfy $|S| = s$ and $|S'| = s'$. Let $\bar{c}$ be the maximal number of replicates, $\bar{d}$ be the maximal number of group size. Assume $n^{-1/2}(s \vee s^*)\log p = o(1)$, $\bar{d}^2 = o(\log n)$ and $\frac{\bar{c}^2\bar{d}}{\log p} = o(1)$.

3. $\epsilon_i$, $w^{*T}\varphi_{i,-1}$, and $\varphi_{ij}$ are all sub-Gaussian with $\|\epsilon_i\|_{\Psi_2} \leqslant C$, $\|w^{*T}\varphi_{i,-1}\|_{\Psi_2} \leqslant C$, and $\|\varphi_{ij}\|_{\Psi_2} \leqslant C$, where $C$ is a positive constant.

4. $\lambda' \asymp \sqrt{\frac{\log p}{n}}$ and $\lambda \asymp \sigma\sqrt{\frac{\log p}{n}}$.

5. $B_i \lesssim \sqrt{\frac{\log p}{n}}$.

Then under $H_0 : \beta_1^* = 0$ for each $t \in \mathbb{R}$,

$$\lim_{n\to\infty} |\mathbb{P}_{\beta^*}(\hat{U}_n \leqslant t) - \Phi(t)| = 0.$$

*Proof.* Before the proof, we need the following lemmas in Ning and Liu (2017), which is used to ensure the assumptions of Theorem G.1 and Corollary G.1 hold. The proofs of Lemmas G.1, G.3, and G.4 can be found in Ning and Liu (2017). In the proof of Lemma G.4, one need to notice that $\varphi^T B$ can be bounded by assumption.

**Lemma G.1.** Under the conditions of Theorem G.2, with probability at least $1 - p^{-1}$, $\|\frac{1}{n}\sum_{i=1}^{n}(\varphi_{i1}\varphi_{i,-1} - \hat{w}^T\varphi_{i,-1}\varphi_{i,-1}^T)\|_\infty \leqslant C\sqrt{\frac{\log p}{n}}$, for some $C > 0$.

**Lemma G.2.** Under the conditions of Theorem G.2, with probability at least $1 - p^{-1}$,

$$\|\hat{\beta} - \beta^*\|_2^2 \leqslant C_1\frac{\bar{c}^2 s\bar{d}\log p}{n}, \text{ and } (\hat{\beta} - \beta^*)^T H_\varphi(\hat{\beta} - \beta^*) \leqslant C_1\kappa_{\max}\frac{\bar{c}^2 s\bar{d}\log p}{n},$$

where $H_\varphi = n^{-1}\sum_{i=1}^{n}\varphi_i\varphi_i^T$ and the constant $C_1 > 0$.

*Proof.* The first inequality is by Theorem 4.1. The second inequality is trivial. □

**Lemma G.3.** Under the conditions of Theorem G.2, with probability at least $1 - p^{-1}$,

$$\|\hat{w} - w^*\|_1 \leqslant 8C\kappa^{-1}s'\sqrt{\frac{\log p}{n}},$$

where $C > 0$ is a constant.

**Lemma G.4.** Under the conditions of Theorem G.2, it holds that $T^* \xrightarrow{\text{dist.}} N(0,1)$, and

$$\sup_{x\in\mathbb{R}}|\mathbb{P}_{\beta^*}(T^* \leqslant x) - \Phi(x)| \leqslant Cn^{-1/2},$$

where $T^* = n^{1/2}S(0,\beta_{-1}^*)/I_{\beta_1|\beta_{-1}}^{*1/2}$ and $C$ is a positive constant not depending on $\beta^*$.

Now we can check that the assumptions of Theorem G.1 and Corollary G.1 hold, which finishes the proof of Theorem G.2. □

Next we introduce some lemmas which give properties of sub-exponential variables and norms, as well as sub-Gaussian variables and norms, which will be used in the proof of Theorem 5.1.

**Lemma G.5.** (Bernstein Inequality) Let $X_1, ..., X_n$ be independent mean 0 sub-exponential random variables and let $K = \max_i \|X_i\|_{\Psi_1}$. Then for any $t > 0$,

$$\mathbb{P}_{\beta^*}\left(\frac{1}{n}\left|\sum_{i=1}^{n} X_i\right| \geqslant t\right) \leqslant 2\exp\left[-C\min\left(\frac{t^2}{K^2}, \frac{t}{K}\right)n\right],$$

where $C > 0$ is a constant.

**Lemma G.6.** Under the conditions of Theorem G.2 with probability at least $1 - p^{-1}$, $\|\frac{1}{n}\sum_{i=1}^{n} \varphi_i \epsilon_i\|_\infty \leqslant C\sqrt{\frac{\log p}{n}}$, for some $C > 0$.

The proofs of Lemmas G.5 and G.6 can be found in Ning and Liu (2017). Now, we can begin the proof of Theorem 5.1.

*Proof.* The proof is similar to Ning and Liu (2017) with a few changes. It is enough to show for any $\epsilon > 0$,

$$\lim_{n\to\infty} \sup_{\beta^* \in \Omega_0} \mathbb{P}_{\beta^*}(|\tilde{U}_n - \hat{U}_n| \geqslant \epsilon) = 0. \tag{G.11}$$

Notice that $|\tilde{U}_n - \hat{U}_n| = |\hat{U}_n||1 - \frac{\sigma^*}{\hat{\sigma}}|$. For a sequence of positive constants $t_n \to 0$ to be chosen later, we can show that $\lim_{n\to\infty} \sup_{\beta^* \in \Omega_0} \mathbb{P}_{\beta^*}(|\hat{U}_n| \geqslant t_n^{-1}) = 0$. It remains to show that

$$\lim_{n\to\infty} \sup_{\beta^* \in \Omega_0} \mathbb{P}_{\beta^*}\left(|1 - \frac{\sigma^*}{\hat{\sigma}}| \geqslant t_n\right) = 0. \tag{G.12}$$

Notice that

$$\begin{aligned}
\hat{\sigma}^2 - \sigma^{*2} &= \left(\frac{1}{n}\sum_{i=1}^{n}(B_i + \epsilon_i)^2 - \sigma^{*2}\right) + \hat{\Delta}^T H_\varphi \hat{\Delta} - 2\hat{\Delta}^T \frac{1}{n}\sum_{i=1}^{n}(\epsilon_i + B_i)\varphi_i \\
&= \left(\frac{1}{n}\sum_{i=1}^{n}(B_i + \epsilon_i)^2 - \sigma^{*2}\right) + \hat{\Delta}^T H_\varphi \hat{\Delta} - 2\hat{\Delta}^T \frac{1}{n}\sum_{i=1}^{n}\epsilon_i\varphi_i - 2\hat{\Delta}^T \frac{1}{n}\sum_{i=1}^{n} B_i\varphi_i \\
&= \left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 - \sigma^{*2}\right) + \hat{\Delta}^T H_\varphi \hat{\Delta} - 2\hat{\Delta}^T \frac{1}{n}\sum_{i=1}^{n}\epsilon_i\varphi_i + \frac{1}{n}\sum_{i=1}^{n} B_i^2 + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i B_i - 2\hat{\Delta}^T \frac{1}{n}\sum_{i=1}^{n} B_i\varphi_i.
\end{aligned} \tag{G.13}$$

where $\hat{\Delta} = \hat{\beta} - \beta^*$. Since $\|\epsilon_i^2\|_{\psi_1} \leqslant 2C^2$, by Lemma G.5, $|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 - \sigma^{*2}| \leqslant C\sqrt{\frac{\log n}{n}}$, for

some constant $C$, with probability tending to one. By Lemma G.2, we have $\Delta^T H_\varphi \Delta \leqslant C_1 \kappa_{\max} \frac{\bar{c}^2 s\bar{d} \log p}{n}$, for some constant $C_1$, with probability tending to one. By Lemma E.5 and Lemma G.2, we have

$$\begin{aligned}\|\hat{\Delta}\|_1 &\leqslant C_1 s\bar{d}\bar{c}^2 \|\hat{\Delta}\|_2 \\ &\leqslant C_2 s\bar{d}\bar{c}^2 \sqrt{\frac{\bar{c}^2 s\bar{d} \log p}{n}},\end{aligned}$$

for some constant $C_2 > 0$. By Lemma G.6, we have

$$\left\| \frac{1}{n} \sum_{i=1}^n \epsilon_i \varphi_i \right\|_\infty \leqslant C_3 \sqrt{\frac{\log p}{n}}.$$

By Lemma G.5, $|\frac{1}{n} \sum_{i=1}^n \epsilon_i B_i| \lesssim \sqrt{1/n}$. By the assumptions of Theorem G.2, $\frac{1}{n} \sum_{i=1}^n B_i^2 \lesssim \frac{\log p}{n}$. Thus,

$$\begin{aligned}\left| \hat{\Delta}^T \frac{1}{n} \sum_{i=1}^n \epsilon_i \varphi_i \right| &\leqslant \|\hat{\Delta}\|_1 \left\| \frac{1}{n} \sum_{i=1}^n \epsilon_i \varphi_i \right\|_\infty \\ &\leqslant C_4 s\bar{d}\bar{c}^2 \sqrt{\bar{c}^2 s\bar{d}} \frac{\log p}{n},\end{aligned}$$

for some constant $C_4 > 0$. By assumption $B_i \lesssim \sqrt{\frac{\log p}{n}}$,

$$\begin{aligned}\left| \hat{\Delta}^T \frac{1}{n} \sum_{i=1}^n B_i \varphi_i \right| &\leqslant \|\hat{\Delta}\|_1 \left\| \frac{1}{n} \sum_{i=1}^n B_i \varphi_i \right\|_\infty \\ &\leqslant C_5 s\bar{d}\bar{c}^2 \sqrt{\bar{c}^2 s\bar{d}} \frac{\log p}{n},\end{aligned}$$

for some constant $C_5 > 0$. Thus, by (G.13), we have

$$|\hat{\sigma}^2 - \sigma^{*2}| \leqslant C_0 \sqrt{\frac{\log n}{n}} \vee (\bar{c}^2 s\bar{d})^{3/2} \frac{\log p}{n},$$

for some constant $C_0$, with probability tending to one. Thus,

$$|1 - \frac{\sigma^*}{\hat{\sigma}}| = \hat{\sigma}^{-2} |1 + \frac{\sigma^*}{\hat{\sigma}}| |\hat{\sigma}^2 - \sigma^{*2}| \lesssim |\hat{\sigma}^2 - \sigma^{*2}| \lesssim \sqrt{\frac{\log n}{n}} \vee (\bar{c}^2 s\bar{d})^{3/2} \frac{\log p}{n},$$

with probability tending to one, because $\sigma^{*2} > C^2$ and $\hat{\sigma}^2 = \sigma^{*2} + o_{\mathbb{P}}(1)$. Thus, if we choose $t_n \gtrsim \sqrt{\frac{\log n}{n}} \vee (\bar{c}^2 s\bar{d})^{3/2}\frac{\log p}{n}$, then (G.12) holds and (G.11) holds. Then by Theorem G.2, the result holds. □

# H Proofs of Lemmas

## H.1 Proof of Lemma E.2

*Proof.* For simplicity, we use $\lambda$ instead of $\lambda_n$, $\gamma$ instead of $\gamma^\lambda$, and $\gamma^Z$ instead of $\gamma^{Z,\lambda}$ in Appendix H. In this proof we will use $\gamma_t$ instead of $\gamma_{J_t}$ for brevity. Let $\gamma^Z(T)$ be the vector with elements $\gamma^Z_{ij_{ik}}(T) = \gamma^Z_{ij_{ik}} I_{\{\beta^*_i \neq 0\}}$. Similarly, $\gamma^Z_{ij_{ik}}(T^c) = \gamma^Z_{ij_{ik}} I_{\{\beta^*_i = 0\}}$. Thus, $\gamma^Z = \gamma^Z(T) + \gamma^Z(T^c)$. Notice $\{\beta^*_i \neq 0\} = \{i \in J_t, \text{ for some } t \in T\}$. Since $f(0,0) = 0$, and (E.5) is a minimizing problem, we have $f(\gamma, \gamma^Z) \leqslant 0$. Since $\gamma^T C \gamma \geqslant 0$ for any $\gamma$, and $\|\beta^Z_t\|_2 - \|\gamma^Z_t + \beta^Z_t\|_2 \leqslant \|\gamma^Z_t\|_2$ for any $t \in T$, combining $f(\gamma, \gamma^Z) \leqslant 0$, we have $\sum_{t\in T^c} \sqrt{d_t}\|\gamma^Z_t\|_2 \leqslant \sum_{t\in T} \sqrt{d_t}\|\gamma^Z_t\|_2$. Also, we have

$$\sum_{t\in T} \sqrt{d_t}\|\gamma^Z_t\|_2 \leqslant \sqrt{\sum_{t\in T} d_t}\|\gamma^Z(T)\|_2 \leqslant \sqrt{s_n \bar{d}}\|\gamma^Z\|_2. \tag{H.14}$$

The first inequality is true because of Cauchy's inequality, and the second inequality is true because $\bar{d} = \max\{d_1, ..., d_n\}$ and $s_n = |T|$.

For any $\beta^\lambda_{ij_{im_1}}$ and $\beta^\lambda_{ij_{im_2}}$, if they are both not zero, by KKT conditions, we have

$$-\frac{1}{n}\psi_i^T(y - \varphi\beta) + \frac{\lambda\sqrt{d_{j_{im_1}}}\beta^\lambda_{ij_{im_1}}}{\|\beta_{J_{j_{im_1}}}\|_2} = 0, \text{ and } -\frac{1}{n}\psi_i^T(y - \varphi\beta) + \frac{\lambda\sqrt{d_{j_{im_2}}}\beta^\lambda_{ij_{im_2}}}{\|\beta_{J_{j_{im_2}}}\|_2} = 0,$$

which indicates

$$\frac{\lambda\sqrt{d_{j_{im_1}}}\beta^\lambda_{ij_{im_1}}}{\|\beta_{J_{j_{im_1}}}\|_2} = \frac{\lambda\sqrt{d_{j_{im_2}}}\beta^\lambda_{ij_{im_2}}}{\|\beta_{J_{j_{im_2}}}\|_2}.$$

Since $\lambda > 0$, we have $\beta^\lambda_{ij_{im_1}}\beta^\lambda_{ij_{im_2}} \geqslant 0$. Notice if $\beta^\lambda_{ij_{im_1}}$ or $\beta^\lambda_{ij_{im_2}}$ is zero, $\beta^\lambda_{ij_{im_1}}\beta^\lambda_{ij_{im_2}} \geqslant 0$ still holds. Together with the constraints of optimization problem, we have $\gamma^\lambda_{ij_{im_1}}\gamma^\lambda_{ij_{im_2}} \geqslant 0$,

which indicates $\|\gamma^Z\|_2 \leqslant \|\gamma\|_2$. Thus, together with (H.14), we have

$$\sum_{t=1}^{p_n} \sqrt{d_t}\|\gamma_t^Z\|_2 \leqslant 2\sqrt{s_n\bar{d}}\|\gamma^Z\|_2 \leqslant 2\sqrt{s_n\bar{d}}\|\gamma\|_2. \tag{H.15}$$

Since $f(\gamma, \gamma^Z) \leqslant 0$, and ignoring the non-negative term $\lambda \sum_{t\in T^c} \sqrt{d_t}\|\gamma_t^Z\|_2$, it follows that

$$n\gamma^T C\gamma \leqslant \lambda\sqrt{s_n\bar{d}}\|\gamma^Z\|_2 \leqslant \lambda\sqrt{s_n\bar{d}}\|\gamma\|_2. \tag{H.16}$$

Next, we bound the term $n\gamma^T C\gamma$ from below. Pplugging the result into (H.16) will yield the desired upper bound on the $l_2$-norm of $\gamma$. Let $\|\gamma_{(1)}^Z\|_2 \geqslant \|\gamma_{(2)}^Z\|_2 \geqslant \cdots \geqslant \|\gamma_{(p_n)}^Z\|_2$ be the ordered block entries of $\gamma$. Let $\{u_n\}$ be a sequence of positive integers, such that $1 \leqslant u_n \leqslant p_n$ and define the set of $u_n$-largest groups as $U = \{k : \|\gamma_k^Z\|_2 \geqslant \|\gamma_{(u_n)}^Z\|_2\}$. Define analogously as before $\gamma^Z(U)$, $\gamma^Z(U^c)$, $\gamma(U)$, and $\gamma(U^c)$. Thus, $\gamma^T C\gamma = (\gamma(U) + \gamma(U^c))^T C(\gamma(U) + \gamma(U^c) = \|a + b\|_2^2$, where $a = \varphi\gamma(U)/\sqrt{n}$ and $b = \varphi\gamma(U^c)/\sqrt{n}$. Thus,

$$\gamma^T C\gamma = a^T a + 2b^T a + b^T b \geqslant (\|a\|_2 - \|b\|_2)^2. \tag{H.17}$$

Assume $l = \sum_{t=1}^{p_n} \|\gamma_t^Z\|_2$. Then for every $t = 1, ..., p_n$, $\|\gamma_{(t)}^Z\|_2 \leqslant l/t$, since $\gamma_{(t)}^Z$ is the $t^{th}$ largest group with respect to $\|\cdot\|_2$. Thus,

$$\|\gamma^Z(U^c)\|_2^2 = \sum_{t=u_n+1}^{p_n} \|\gamma_{(t)}^Z\|_2 \leqslant \left(\sum_{t=1}^{p_n} \|\gamma_t^Z\|_2^2\right)^2 \sum_{t=u_n+1}^{p_n} \frac{1}{t^2} \leqslant \left(\sum_{t=1}^{p_n} \sqrt{d_t}\|\gamma_t^Z\|_2\right)^2 \frac{1}{u_n}, \tag{H.18}$$

where the last inequality is because

$$\sum_{t=u_n+1}^{p_n} \frac{1}{t^2} \leqslant \int_{s=u_n}^{\infty} \frac{1}{s^2} ds = \frac{1}{u_n},$$

and $\sqrt{d_t} \geqslant 1$.

Together with (H.15), we have $\|\gamma^Z(U^c)\|_2^2 \leqslant 4s_n\bar{d}\|\gamma^Z\|_2^2 \frac{1}{u_n}$. Since $\gamma(U)$ has at most

$\sum_{t\in U} d_t$ non-zero coefficients, and $\sum_{t\in U} d_t \leqslant u_n\bar{d}$,

$$\begin{aligned}\|a\|_2^2 &\geqslant \phi_{\min}\bigg(\sum_{t\in U} d_t\bigg)\|\gamma(U)\|_2^2 \geqslant \phi_{\min}\bigg(\sum_{t\in U} d_t\bigg)\|\gamma^Z(U)\|_2^2 \\ &= \phi_{\min}\bigg(\sum_{t\in U} d_t\bigg)(\|\gamma^Z\|_2^2 - \|\gamma^Z(U^c)\|_2^2) \geqslant \phi_{\min}\bigg(\sum_{t\in U} d_t\bigg)(1-\frac{4s_n\bar{d}}{u_n})\|\gamma^Z\|_2^2 \\ &\geqslant \phi_{\min}(u_n\bar{d})(1-\frac{4s_n\bar{d}}{u_n})\|\gamma^Z\|_2^2. \end{aligned} \tag{H.19}$$

The first inequality is true because of the definition of $\phi_{\min}(\cdot)$, and the equality is true because $\gamma^Z = \gamma^Z(U) + \gamma^Z(U^c)$. From Lemma E.1, $\gamma(U^c)$ has at most $n$ non-zero groups, which indicates

$$\|b\|_2^2 \leqslant \phi_{\max}(n\bar{d})\|\gamma(U^c)\|_2^2 \leqslant \phi_{\max}\|\gamma(U^c)\|_2^2 \leqslant \bar{d}\phi_{\max}\|\gamma^Z(U^c)\|_2^2 \leqslant \frac{4\phi_{\max}s_n\bar{d}^2}{u_n}\|\gamma^Z\|_2^2. \tag{H.20}$$

The first inequality is true because the definition of $\phi_{\max}(\cdot)$, the third inequality is true is because of Cauchy's inequality, and the last inequality is true because of (H.15) and (H.18). Thus, plugging (H.19) and (H.20) into (H.17), and combining with the facts $\sum_{t\in U} d_t \leqslant \bar{d}u_n$ and $\phi_{\max} \geqslant \phi_{\min}(u_n)$, under Assumption E.1, for sufficient large $n$, we have

$$\begin{aligned}\|a\|_2 - \|b\|_2 &\geqslant \bigg(\sqrt{\phi_{\min}(u_n\bar{d})(1-\frac{4s_n\bar{d}}{u_n})} - \sqrt{\frac{4\phi_{\max}s_n\bar{d}^2}{u_n}}\bigg)\|\gamma^Z\|_2 \\ &\geqslant \bigg(\sqrt{\phi_{\min}(u_n\bar{d})(1-\frac{4s_n\bar{d}}{u_n})} - \sqrt{\frac{2\kappa_{\max}s_n\bar{d}^2}{u_n}}\bigg)\|\gamma^Z\|_2\end{aligned}$$

Let $u_n = s_n\log n$, under Assumption E.1, for large $n$, we have

$$\|a\|_2 - \|b\|_2 \geqslant \bigg(\sqrt{\frac{\kappa_{\min}}{2}(1-\frac{4\bar{d}}{\log n})} - \sqrt{\frac{2\kappa_{\max}\bar{d}^2}{\log n}}\bigg)\|\gamma^Z\|_2.$$

Together with (H.16), we have

$$\frac{\lambda\sqrt{s_n\bar{d}}}{n}\|\gamma^Z\|_2 \geqslant \gamma^T C\gamma \geqslant \bigg(\sqrt{\frac{\kappa_{\min}}{2}(1-\frac{4\bar{d}}{\log n})} - \sqrt{\frac{2\kappa_{\max}\bar{d}^2}{\log n}}\bigg)^2\|\gamma^Z\|_2^2.$$

Since by Cauchy's inequality, we have $\|\gamma^Z\|_2^2 \geqslant \|\gamma\|_2^2/\bar{c}$. Thus,

$$\|\gamma\|_2^2 \leqslant \frac{\lambda^2 \bar{c} s_n \bar{d}}{n^2} \Big/ \left( \sqrt{\frac{\kappa_{\min}}{2}(1 - \frac{4\bar{d}}{\log n})} - \sqrt{\frac{2\kappa_{\max}\bar{d}^2}{\log n}} \right)^2,$$

which completes the proof. □

### H.2 Proof of Lemma E.3

*Proof.* From (E.6), for every $M$ with $|M| \leqslant \bar{m}_n$,

$$\|\theta^M\|_2^2 \leqslant \frac{1}{n^2 \phi_{\min}^2(\bar{m}_n)} \|\varphi_M^T(\epsilon + B)\|_2^2 \leqslant \frac{2}{n^2 \phi_{\min}^2(\bar{m}_n)} (\|\varphi_M^T \epsilon\|_2^2 + \|\varphi_M^T B\|_2^2) \tag{H.21}$$

By Lemma G.6, with probability at least $1 - d^{-1}$, $\| \sum_{i=1}^n \varphi_i \epsilon_i \|_\infty \leqslant C_1 \sqrt{n \log p}$. Thus,

$$\max_{M:|M| \leqslant \bar{m}_n} \|\varphi_M^T \epsilon\|_2^2 \leqslant \bar{m}_n \| \sum_{i=1}^n \varphi_i \epsilon_i \|_\infty^2 \leqslant \bar{m}_n C_1^2 n \log p,$$

where the first inequality is true because $\|\varphi_M^T \epsilon\|_2^2 \leqslant |M| \|\varphi_M^T \epsilon\|_\infty^2$, and $|M| \leqslant \bar{m}_n$.

By assumptions of Theorem 4.1,

$$\max_{M:|M| \leqslant \bar{m}_n} \|\varphi_M^T B\|_2^2 \leqslant \bar{m}_n \| \sum_{i=1}^n \varphi_i B_i \|_\infty^2 \leqslant \bar{m}_n C_2^2 n \log p.$$

Thus,

$$\max_{M:|M| \leqslant \bar{m}_n} \|\theta^M\|_2^2 \leqslant C^2 \frac{\bar{m}_n \log p}{n \phi_{\min}^2(\bar{m}_n)},$$

which finishes the proof. □

### H.3 Proof of Lemma E.4

*Proof.* Before the proof, we state a lemma.

**Lemma H.1.** For $x \in \mathbb{R}^q$, suppose $\hat{x}_1 = \arg\min_x f_1(x)$ and $\hat{x}_2 = \arg\min_x f_2(x)$ where $f_1(x) = \frac{1}{2} x^T A^T A x + b^T x$ with $A \in \mathbb{R}^{n \times q}$ which is full rank and $b \in \mathbb{R}^q$. Also, $f_2(x) = f_1(x) + c^T x$ with $c \in \mathbb{R}^q$. Let $A^Z$, $b^Z$ and $c^Z$ be defined in the same way as before. Let

$g_1(y^Z) = \frac{1}{2}\|A^Z y^Z\|_2^2 + (b^Z)^T y^Z + h(y^Z)$ and $g_2(y^Z) = \frac{1}{2}\|A^Z y^Z\|_2^2 + (b^Z)^T y^Z + (c^Z)^T y^Z + h(y^Z)$, where $h(y)$ is a convex function with respect to $y$ and everywhere sub-differentiable, and define $\hat{y}_1^Z = \arg\min_y^Z g_1(y^Z)$ and $\hat{y}_2^Z = \arg\min_y^Z g_1(y^Z)$. Then we have

$$\|\hat{y}_2 - \hat{y}_1\|_2 \leqslant \gamma \|\hat{x}_2 - \hat{x}_1\|_2.$$

*Proof.* Our proof is similar to Liu and Zhang (2009), with the only difference that $\|A^Z(\hat{y}_1^Z - \hat{y}_2^Z)\|_2^2 + (c^Z)^T(\hat{y}_1^Z - \hat{y}_2^Z) = \|A(\hat{y}_1 - \hat{y}_2)\|_2^2 + c^T(\hat{y}_1 - \hat{y}_2)$. □

Let $M(\xi) = A_{\lambda,\xi}$. Let $0 = \xi_1 < ... < \xi_{J+1} = 1$ be the points of discontinuity of $M(\xi)$. At these locations, variables either join the active set or are dropped from the active set. Fix some $j$ with $1 \leqslant j \leqslant J$. Denote by $M_j$ be the set of active groups $M(\xi)$ for any $\xi \in (\xi_j, \xi_{j+1})$. Assuming

$$\forall \xi \in (\xi_j, \xi_{j+1}) : \|\hat{\beta}^{\lambda,\xi} - \hat{\beta}^{\lambda,\xi_j}\|_2 \leqslant C(\xi - \xi_j)\|\hat{\theta}^{M_j}\|_2 \tag{H.22}$$

is true, where $\theta^{M_j}$ is the restricted OLS estimator of noise. Then

$$\begin{aligned}
\|\hat{\beta}^{\lambda,0} - \hat{\beta}^{\lambda}\|_2 &\leqslant \sum_{j=1}^{J} \|\hat{\beta}^{\lambda,\xi_j} - \hat{\beta}^{\lambda,\xi_{j+1}}\|_2 \\
&\leqslant C \max_{M:|M|\leqslant m} \|\theta^M\|_2 \sum_{j=1}^{J} (\xi_{j+1} - \xi_j) \\
&= C \max_{M:|M|\leqslant m} \|\theta^M\|_2.
\end{aligned}$$

By replacing $\hat{x}_1$, $\hat{x}_2$, $\hat{y}_1$ and $\hat{y}_2$ with $\xi\hat{\theta}^{M_j}$, $\xi_j\hat{\theta}^{M_j}$, $\hat{\beta}^{\lambda,\xi}$ and $\hat{\beta}^{\lambda,\xi_j}$ in Lemma H.1, respectively, we obtain (H.22). Hence, we complete the proof. □

### H.4 Proof of Lemma E.5

*Proof.* Our proof is similar to Meinshausen and Yu (2009). The only thing need to be noticed is that for (38) in Meinshausen and Yu (2009), we have

$$\begin{aligned}(\|(\varphi^Z_{A_{\lambda,\xi}})^T\varphi(\beta-\hat{\beta}^{\lambda,\xi})\|_2+\|(\varphi^Z_{A_{\lambda,\xi}})^T(\epsilon+B)\|_2)^2 &\leqslant 2(\|(\varphi^Z_{A_{\lambda,\xi}})^T\varphi(\beta-\hat{\beta}^{\lambda,\xi})\|_2^2+\|(\varphi^Z_{A_{\lambda,\xi}})^T(\epsilon+B)\|_2^2)\\ &\leqslant 2\bar{c}(\|\varphi^T_{A_{\lambda,\xi}}\varphi(\beta-\hat{\beta}^{\lambda,\xi})\|_2^2+\|\varphi^T_{A_{\lambda,\xi}}(\epsilon+B)\|_2^2).\end{aligned}$$

□

# I Stochastic Function

In this section, a stochastic function is considered. In particular, this example demonstrates tuning parameter selection. We consider the following function, which was used in Gramacy and Lee (2009),

$$f(x_1,x_2,x_3,x_4,x_5,x_6)=\exp\left\{\sin([0.9\times(x_1+0.48)]^{10})\right\}+x_2x_3+x_4+\epsilon, \tag{I.23}$$

where $\epsilon\sim\mathcal{N}(0,0.05^2)$ and $x_i\in[0,1], i=1,\ldots,6$. The function is nonlinear in $x_1, x_2$ and $x_3$, and linear in $x_4$. In $x_1$, it oscillates more quickly as it reaches the upper bound of the interval $[0,1]$. $x_5$ and $x_6$ are irrelevant variables.

Here, we consider 5 replicates at each unique training location, $n=5m$, as indicated in Wang and Haaland (2018), along with $n_{\text{test}}=10,000$ unique predictive locations randomly generated from a uniform distribution on $[0,1]^d$. Since the choice of tuning parameter $\lambda$ in (2) can be particularly crucial in stochastic function emulation, we consider AIC, BIC and 10-fold CV as selection criteria. For the implementation of 10-fold CV, 10 CPUs are requested for parallel computing. Table 1 shows the performance of traditional Gaussian process, local Gaussian process, and MRFA with these three selection criterion based on designs of increasing size $n$. It can be seen that, similar to the results in the previous subsections, traditional Gaussian process is only feasible at $n=1,000$, while MRFA is feasible and accurate for large problems. Even when traditional Gaussian process is feasible, MRFA is much faster in terms of fitting and prediction, and more accurate with any tuning parameter selection method. Local Gaussian process fitting is feasible for large problems,

but less accurate than MRFA and traditional Gaussian process. Among the three criteria, it can be seen that AIC, BIC and CV have relatively small differences in terms of prediction accuracy. Computationally, the tuning parameters can be chosen within 2 seconds using AIC or BIC, while the computational costs of CV can be considerable.

This example also illustrates the flexibility of the proposed method. From (I.23), the function appears not to satisfy the strong effect heredity conditions, because the main effects of $x_2$ and $x_3$ are not present. On the other hand, the function can be easily re-expressed in a form that does satisfy strong effect heredity. For example,

$$f(x_1, \ldots, x_6) = -1 + \exp\left\{\sin([0.9 \times (x_1 + 0.48)]^{10})\right\} + x_2 + x_3 + (x_2 - 1)(x_3 - 1) + x_4 + \epsilon,$$

which satisfies the strong effect heredity assumption because main effect functions of $x_2$ and $x_3$ appear in the function in addition to the interaction function $(x_2 - 1)(x_3 - 1)$.

| | $n$ | Fitting Time (sec.) | Prediction Time (sec.) | | Selection Time (sec.) | RMSE ($\times 10^{-1}$) |
|---|---|---|---|---|---|---|
| `mlegp` | 1,000 | 2524 | 88 | | | 1.64 |
| `laGP` | 1,000 | - | 394 | | | 7.30 |
| | 10,000 | - | 439 | | | 6.07 |
| | 100,000 | - | 457 | | | 4.70 |
| | 1,000,000 | - | 433 | | | 3.85 |
| `MRFA` | 1,000 | 96 | 8 | AIC | 1 | 1.36 |
| | | | | BIC | 1 | 1.36 |
| | | | | CV | 92 | 1.32 |
| | 10,000 | 443 | 23 | AIC | 1 | 0.18 |
| | | | | BIC | 1 | 0.19 |
| | | | | CV | 423 | 0.26 |
| | 100,000 | 2999 | 34 | AIC | 1 | 0.14 |
| | | | | BIC | 1 | 0.14 |
| | | | | CV | 2213 | 0.14 |
| | 1,000,000 | 61504 | 103 | AIC | 1 | 0.01 |
| | | | | BIC | 1 | 0.01 |
| | | | | CV | 55849 | 0.05 |

**Table 1:** *The 6-dimensional stochastic function example with* $n_{\text{test}} = 10,000$ *random predictive locations.*

## J Other Functions

In this section, we present three more example functions in comparison with `laGP` and `mlegp`, the 3-dimensional bending function (Plumlee and Apley, 2017), the 6-dimensional OTL circuit function (Ben-Ari and Steinberg, 2007), and the 10-dimensional wing weight function (Forrester et al., 2008). The details of these examples and their input ranges are given in Appendix K.

The comparison results are shown in Table 2. Similar to the results in the previous subsections, the results indicate the MRFA outperforms the traditional Gaussian process in terms of prediction accuracy, except for the wing function at $n = 1,000$ where the traditional Gaussian process fitting has better accuracy. The reason might be that the underlying wing weight function contains high-order interaction functions making it not particularly well-suited to low-order representation. See (K.24) in Appendix K. Nevertheless, even when the traditional Gaussian process fitting is feasible (at $n = 1,000$), the MRFA is much faster than traditional Gaussian process fitting. Local Gaussian process fitting is feasible for large problems and has better accuracy in the low-dimensional example (see Table 2(a)), but it is less accurate in the other two examples and in some cases slower than the MRFA.

## K Description of Functions in Section J

- The amount of deflection of a bending function is given by

$$D_e = \frac{4}{10^9}\frac{L^3}{bh^3},$$

  where the 3 inputs are $L, b$, and $h$.

- The midpoint voltage of a transformerless OTL circuit function is given by

$$V_m = \frac{(V_{b1} + 0.74)B(R_{c2} + 9)}{B(R_{c2} + 9) + R_f} + \frac{11.35R_f}{B(R_{c2} + 9) + R_f} + \frac{0.74R_f\beta(R_{c2} + 9)}{(B(R_{c2} + 9) + R_f)R_{c1}},$$

  where $V_{b1} = 12R_{b2}/(R_{b1} + R_{b2})$, and the 6 inputs are $R_{b1}, R_{b2}, R_f, R_{c1}, R_{c2}$, and $B$.

- The wing weight function models a light aircraft wing, where the wing's weight is

| $d = 3$ | $n$ | Fitting time (sec.) | Prediction time (sec.) | RMSE ($\times 10^{-5}$) |
|---|---|---|---|---|
| `mlegp` | 1,000 | 1807 | 140 | 5.64 |
| `laGP` | 1,000 | - | 310 | 0.66 |
| | 10,000 | - | 312 | 0.21 |
| | 100,000 | - | 311 | 0.08 |
| | 1,000,000 | - | 316 | 0.04 |
| `MRFA` | 1,000 | 49 | 8 | 2.16 |
| | 10,000 | 293 | 14 | 0.46 |
| | 100,000 | 3311 | 25 | 0.20 |
| | 1,000,000 | 113279 | 159 | 0.14* |

**(a)** *Performance of the 3-dimensional bending function. *Note that due to memory limits, in the cases $R_{max} = 3$ and $D_{max} = 3$ are considered instead.*

| $d = 6$ | $n$ | Fitting time (sec.) | Prediction time (sec.) | RMSE ($\times 10^{-4}$) |
|---|---|---|---|---|
| `mlegp` | 1,000 | 3976 | 173 | 13.70 |
| `laGP` | 1,000 | - | 314 | 102.71 |
| | 10,000 | - | 301 | 27.01 |
| | 100,000 | - | 323 | 11.43 |
| | 1,000,000 | - | 328 | 4.80 |
| `MRFA` | 1,000 | 294 | 19 | 7.81 |
| | 10,000 | 798 | 17 | 2.05 |
| | 100,000 | 6688 | 82 | 1.42 |
| | 1,000,000 | 122075 | 133 | 1.18* |

**(b)** *Performance of the 6-dimensional OTL circuit function. *Note that due to memory limits, in the cases $R_{max} = 3$ and $D_{max} = 3$ are considered instead.*

| $d = 10$ | $n$ | Fitting time (sec.) | Prediction time (sec.) | RMSE ($\times 10^{-1}$) |
|---|---|---|---|---|
| `mlegp` | 1,000 | 2922 | 228 | 1.56 |
| `laGP` | 1,000 | - | 327 | 19.74 |
| | 10,000 | - | 325 | 10.72 |
| | 100,000 | - | 329 | 5.04 |
| | 1,000,000 | - | 347 | 2.22 |
| `MRFA` | 1,000 | 1319 | 28 | 7.77 |
| | 10,000 | 1633 | 21 | 1.52 |
| | 100,000 | 12289 | 84 | 1.39 |
| | 1,000,000 | 168854 | 148 | 1.18* |

**(c)** *Performance of the 10-dimensional wing weight function. *Note that due to memory limits, in the cases $R_{max} = 1$ and $D_{max} = 3$ are considered instead.*

**Table 2:** *Performance of the bending, OTL circuit, and wing weight functions with $n_{\text{test}} = 10,000$ random predictive locations.*

given by

$$W = 0.036 S_w^{0.758} W_{fw}^{0.0035} \left(\frac{A}{\cos^2(\Lambda)}\right)^{0.6} q^{0.006} R^{0.04} \left(\frac{100 t_c}{\cos(\Lambda)}\right)^{-0.3} (N_z W_{dg})^{0.49} + S_w W_p, \tag{K.24}$$

where the 10 inputs are $S_w, W_{fw}, A, \Lambda, q, R, t_c, N_z, W_{dg}$, and $W_p$.

The input ranges are given in Table 3.

| Bending | | OTL circuit | | Wing weight | |
|---|---|---|---|---|---|
| $L$ | $\in [10, 20]$ | $R_{b1}$ | $\in [50, 150]$ | $S_w$ | $\in [150, 200]$ |
| $b$ | $\in [1, 2]$ | $R_{b2}$ | $\in [25, 70]$ | $W_{fw}$ | $\in [220, 300]$ |
| $h$ | $\in [0.1, 0.2]$ | $R_f$ | $\in [0.5, 3]$ | $A$ | $\in [6, 10]$ |
| | | $R_{c1}$ | $\in [1.2, 2.5]$ | $\Lambda$ | $\in [-10, 10]$ |
| | | $R_{c2}$ | $\in [0.25, 1.2]$ | $q$ | $\in [16, 45]$ |
| | | $\beta$ | $\in [50, 300]$ | $R$ | $\in [0.5, 1]$ |
| | | | | $t_c$ | $\in [0.08, 0.18]$ |
| | | | | $N_z$ | $\in [2.5, 6]$ |
| | | | | $W_{dg}$ | $\in [1700, 2500]$ |
| | | | | $W_p$ | $\in [0.025, 0.08]$ |

**Table 3:** *Input ranges of the OTL circuit function, the piston simulation function, and the wing weight function.*

# References


Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404.

Bartle, R. G. (1995). *The Elements of Integration and Lebesgue Measure*. John Wiley & Sons, New York.

Ben-Ari, E. N. and Steinberg, D. M. (2007). Modeling data from computer experiments: an empirical comparison of kriging with mars and projection pursuit regression. *Quality Engineering*, 19(4):327–338.

Fan, J., Guo, S., and Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B*, 74(1):37–65.

Forrester, A. I. J., Sobester, A., and Keane, A. J. (2008). *Engineering Design via Surrogate Modelling: a Practical Guide*. John Wiley & Sons, Chichester, UK.

Gramacy, R. B. and Lee, H. K. (2009). Adaptive design and analysis of supercomputer experiments. *Technometrics*, 51(2):130–145.

Liu, H. and Zhang, J. (2009). Estimation consistency of the group lasso and its applications. In *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 376–383.

Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1):246–270.

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

Plumlee, M. and Apley, D. W. (2017). Lifted Brownian kriging models. *Technometrics*, 59(2):165–177.

Wang, W. and Haaland, B. (2018). Controlling sources of inaccuracy in stochastic kriging. *Technometrics*, to appear.