Sparse Sliced Inverse Regression via Cholesky Matrix Penalization

Linh H. Nghiem, Francis K.C. Hui, Samuel Müller, and A.H. Welsh

Australian National University, University of Sydney and Macquarie University

Supplementary Material

S1 Additional Lemmas and Propositions

In order to prove the main theoretical results in the paper, we need additional technical definitions and auxillary results. First, we say that the sample covariance matrix of predictors $\hat{\boldsymbol{\Sigma}}$ satisfies the restricted eigenvalue condition over a set T with parameter (q, r) if and only if

$$\left\|\hat{\boldsymbol{L}}^\top v\right\|_2^2 = v^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{v} \geq r \left\|\boldsymbol{v}\right\|_2^2, \text{ for all } \boldsymbol{v} \in \mathcal{C}(T, q) = \{\boldsymbol{v} \in \mathbb{R}^p \mid \left\|\boldsymbol{v}_{T^c}\right\|_1 \leq q \left\|\boldsymbol{v}_T\right\|_1\}.$$

This condition is essential in obtaining the consistency of the Lasso estimator in the linear model (Wainwright, 2019). It is also essential for the consistency of the Cholesky matrix penalization (CHOMP) estimator as shown below.

We begin by obtaining the following bound of the difference between the CHOMP estimator and the pseudo-true parameter as defined in Section

4 of the main paper.

Lemma 1. *Assume the sample covariance matrix* $\hat{\mathbf{\Sigma}}$ *satisfies the restricted eigenvalue condition with parameter* $q = 3$ *and some positive constant* r. *Then, any solution of the equation (3.8) of the main paper with tuning parameter bounded below as* $\mu_j \geq 2 \left\| \hat{\boldsymbol{\eta}}_j - \hat{\mathbf{\Sigma}} \tilde{\boldsymbol{\beta}}_j \right\|_\infty$ *satisfies* $\left\| \hat{\boldsymbol{\beta}}_j - \tilde{\boldsymbol{\beta}}_j \right\|_2 \leq 3r^{-1}\mu_j s_j^{1/2}$, *for* $j = 1, \ldots, d$.

This result parallels the basic consistency result for the Lasso estimator in the linear model (Wainwright (2019)). The bound on the right hand side of (1) is inversely proportional to the restricted eigenvalue constant θ, which is expected because a higher θ implies a higher curvature around the optimal $\hat{\beta}_j$. Also, the bound scales with $s_j^{1/2}$; this is also natural because we are trying to estimate an unknown vector with s_j non-zero entries. We first prove Lemma 1.

S1.1 Proof of Lemma 1

As Lemma 1 holds for each dimension $j = 1, \ldots, d$, we remove the subscript j in the development below. First, we prove that $\delta \in C(S, 3)$ defined in the paper. By definition of $\hat{\boldsymbol{\beta}}$, we have

$$\frac{1}{2} \left\| \hat{\boldsymbol{L}}^\top \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\kappa}} \right\|_2^2 + \mu \left\| \hat{\boldsymbol{\beta}} \right\|_1 \leq \frac{1}{2} \left\| \hat{\boldsymbol{L}}^\top \tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\kappa}} \right\|_2^2 + \mu \left\| \tilde{\boldsymbol{\beta}} \right\|_1 .$$

Writing $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}} + \boldsymbol{\delta}$ we obtain

$$\frac{1}{2}\left\|\hat{\boldsymbol{L}}^\top\boldsymbol{\delta} - \left(\hat{\boldsymbol{\kappa}} - \hat{\boldsymbol{L}}^\top\tilde{\boldsymbol{\beta}}\right)\right\|_2^2 + \mu\left\|\hat{\boldsymbol{\beta}}\right\|_1 \leq \frac{1}{2}\left\|\hat{\boldsymbol{L}}^\top\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\kappa}}\right\|_2^2 + \mu\left\|\tilde{\boldsymbol{\beta}}\right\|_1. \quad \text{(S1.1)}$$

Expanding the first term on the left hand side of (S1.1), we have

$$\frac{1}{2}\left\|\hat{\boldsymbol{L}}^\top\boldsymbol{\delta} - \left(\hat{\boldsymbol{\kappa}} - \hat{\boldsymbol{L}}^\top\tilde{\boldsymbol{\beta}}\right)\right\|_2^2 = \frac{1}{2}\left\|\hat{\boldsymbol{L}}^\top\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\kappa}}\right\|_2^2 - \boldsymbol{\delta}^\top\hat{\boldsymbol{L}}(\hat{\boldsymbol{\kappa}} - \hat{\boldsymbol{L}}^\top\tilde{\boldsymbol{\beta}}) + \frac{1}{2}\left\|\hat{\boldsymbol{L}}^\top\boldsymbol{\delta}\right\|_2^2,$$

Hence,

$$\begin{aligned}
0 \leq \frac{1}{2}\left\|\boldsymbol{L}^\top\boldsymbol{\delta}\right\|_2^2 &\leq \boldsymbol{\delta}^\top\hat{\boldsymbol{L}}(\hat{\boldsymbol{\kappa}} - \hat{\boldsymbol{L}}^\top\tilde{\boldsymbol{\beta}}) + \mu\left(\left\|\tilde{\boldsymbol{\beta}}\right\|_1 - \left\|\hat{\boldsymbol{\beta}}\right\|_1\right) \\
&\overset{(i)}{\leq} \left\|\boldsymbol{\delta}\right\|_1\left\|\hat{\boldsymbol{L}}\hat{\boldsymbol{\kappa}} - \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^\top\tilde{\boldsymbol{\beta}}\right\|_\infty + \mu\left(\left\|\tilde{\boldsymbol{\beta}}\right\|_1 - \left\|\hat{\boldsymbol{\beta}}\right\|_1\right) \\
&\overset{(ii)}{=} \left\|\boldsymbol{\delta}\right\|_1\left\|\hat{\boldsymbol{\eta}} - \hat{\boldsymbol{\Sigma}}\tilde{\boldsymbol{\beta}}\right\|_\infty + \mu\left(\left\|\tilde{\boldsymbol{\beta}}\right\|_1 - \left\|\hat{\boldsymbol{\beta}}\right\|_1\right) \\
&\overset{(iii)}{\leq} \frac{1}{2}\mu\left\|\boldsymbol{\delta}\right\|_1 + \mu\left(\left\|\tilde{\boldsymbol{\beta}}\right\|_1 - \left\|\hat{\boldsymbol{\beta}}\right\|_1\right),
\end{aligned} \quad \text{(S1.2)}$$

where step (i) follows from Holder's inequality, step (ii) follows from the definitions $\hat{\boldsymbol{\kappa}} = \hat{\boldsymbol{L}}^{-1}\hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^\top$, and step (iii) follows from the condition $\mu \geq 2\left\|\hat{\boldsymbol{\eta}} - \hat{\boldsymbol{\Sigma}}\tilde{\boldsymbol{\beta}}\right\|_\infty$. Then, we have

$$\frac{1}{2}\left\|\boldsymbol{\delta}\right\|_1 + \left\|\tilde{\boldsymbol{\beta}}\right\|_1 - \left\|\hat{\boldsymbol{\beta}}\right\|_1 \geq 0. \quad \text{(S1.3)}$$

Because $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}} + \boldsymbol{\delta}$ and $\tilde{\boldsymbol{\beta}}_{T^c} = 0$, applying the (reverse) triangle inequality gives

$$\left\|\tilde{\boldsymbol{\beta}}\right\|_1 - \left\|\hat{\boldsymbol{\beta}}\right\|_1 = \left\|\tilde{\boldsymbol{\beta}}_S\right\|_1 - \left\|\tilde{\boldsymbol{\beta}}_S + \boldsymbol{\delta}_S\right\|_1 - \left\|\boldsymbol{\delta}_{S^c}\right\|_1 \leq \left\|\boldsymbol{\delta}_S\right\|_1 - \left\|\boldsymbol{\delta}_{S^c}\right\|_1,$$

Furthermore, we have $\|\boldsymbol{\delta}\|_1 = \|\boldsymbol{\delta}_S\|_1 + \|\boldsymbol{\delta}_{S^c}\|_1$. Therefore, equation (S1.3) gives

$$\begin{aligned} 0 &\leq \frac{1}{2}\|\boldsymbol{\delta}\|_1 + \left\|\tilde{\boldsymbol{\beta}}\right\|_1 - \left\|\hat{\boldsymbol{\beta}}\right\|_1 \\ &\overset{(iv)}{\leq} \frac{1}{2}\|\boldsymbol{\delta}_S\|_1 + \frac{1}{2}\|\boldsymbol{\delta}_{S^c}\|_1 + \|\boldsymbol{\delta}_S\|_1 - \|\boldsymbol{\delta}_{S^c}\|_1 = \frac{3}{2}\|\boldsymbol{\delta}_S\|_1 - \frac{1}{2}\|\boldsymbol{\delta}_{S^c}\|_1 \\ &\overset{(v)}{\leq} \frac{3}{2}\|\boldsymbol{\delta}_S\|_1 . \end{aligned}$$

It follows from step (iv) that $\|\boldsymbol{\delta}_{S^c}\|_1 \leq 3\|\boldsymbol{\delta}_S\|_1$, or $\boldsymbol{\delta} \in C(S,3)$. Finally, applying the restricted eigenvalue condition of the sample covariance matrix (defined in Section 4 of the main paper), we obtain

$$\begin{aligned} \frac{1}{2}\theta\|\boldsymbol{\delta}\|_2^2 \leq \frac{1}{2}\left\|\hat{\boldsymbol{L}}^\top\boldsymbol{\delta}\right\|_2^2 &\overset{(vi)}{\leq} \frac{1}{2}\mu\|\boldsymbol{\delta}\|_1 + \mu\left(\left\|\tilde{\boldsymbol{\beta}}\right\|_1 - \left\|\hat{\boldsymbol{\beta}}\right\|_1\right) \\ &\overset{(vii)}{\leq} \frac{3}{2}\mu\|\boldsymbol{\delta}_S\|_1 \overset{(viii)}{\leq} \frac{3}{2}\mu\sqrt{s}\|\boldsymbol{\delta}_S\|_2 \leq \frac{3}{2}\mu\sqrt{s}\|\boldsymbol{\delta}\|_2 , \end{aligned}$$

where step (vi) follows from (S1.2), step (vii) follows from step (v), and step $(viii)$ follows from the Cauchy-Schwartz inequality. Finally, we obtain

$$\left\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\right\|_2 = \|\boldsymbol{\delta}\|_2 \leq \frac{3}{\theta}\mu\sqrt{s}$$

as required.

S1.2 Additional Propositions

Next, we state the following results from Lin et al. (2019) which essentially imply that the conditions for Lemma 1 in the main paper hold with probability tending to one. We begin with the restricted eigenvalue condition for

the sample covariance matrix $\hat{\mathbf{\Sigma}}$.

Proposition 1. *Assume Condition (C1) in the paper holds. For some universal constants a_1, a_2 and a_3, if the sample size n satisfies $n > a_1 s \log(p)$, then the sample covariance matrix $\hat{\mathbf{\Sigma}}$ satisfies the restricted eigenvalue condition with parameter $(q, r) = (3, \sqrt{C_{min}}/8)$ over any set T of cardinality s with probability at least $1 - a_2 \exp(-a_3 n)$.*

Next, one key condition in Lemma 1 is that the tuning parameter has to satisfy the lower bound $\mu \geq 2 \left\| \hat{\boldsymbol{\eta}} - \hat{\mathbf{\Sigma}} \tilde{\boldsymbol{\beta}} \right\|_\infty$. Proposition 2 implies that this lower bound is well-controlled.

Proposition 2. *Assume conditions (C1)-(C3) in the main paper hold. Then*

$$\left\| \hat{\boldsymbol{\eta}}_j - \hat{\mathbf{\Sigma}} \tilde{\boldsymbol{\beta}}_j \right\|_\infty = O_p \left\{ \frac{\log(p)^{1/2}}{(n\hat{\lambda}_j)^{1/2}} \right\}, \quad j = 1, \ldots, d.$$

Proposition 2 implies that if we set $\mu_j = M \left\{ \log(p)/(n\hat{\lambda}_j) \right\}^{1/2}$ for a sufficiently large constant M, then we have $\mu_j \geq 2 \left\| \hat{\boldsymbol{\eta}} - \hat{\mathbf{\Sigma}} \tilde{\boldsymbol{\beta}} \right\|_\infty$ with probability tending to one. When $n \to \infty$, the ratio $p/n \to 0$, and hence $\log(p)/n \to 0$. As long as the eigenvalue $\hat{\lambda}_j$ is bounded away from zero, $\left\| \hat{\boldsymbol{\eta}}_j - \hat{\mathbf{\Sigma}} \tilde{\boldsymbol{\beta}}_j \right\|_\infty \to 0$, then any positive tuning parameter μ_j will asymptotically satisfy the bound.

Proposition 3. *If $n\lambda = p^\nu$ for $\nu > 1/2$, then $\|\tilde{\boldsymbol{\beta}}_j\|_2 \geq C \left(\lambda_j / \hat{\lambda}_j \right)^{1/2}$ and*

$\left(\lambda_j/\hat{\lambda}_j\right)^{1/2} \leq C \left\|\mathcal{P}(\Lambda)\widehat{\boldsymbol{\eta}}_j\right\|_2$ *for* $j = 1, \ldots, d$ *with probability tending to one.*

Proposition 3 implies that the norm of the pseudo-true parameter $\left\|\tilde{\boldsymbol{\beta}}_j\right\|_2$ is bounded away from zero and that the ratio $\lambda_j/\hat{\lambda}_j$ is bounded for each $j = 1, \ldots, d$

S2 Proof of Main Theorems

S2.1 Proof of Theorem 1

When the sample size $n \to \infty$, Proposition 1 in the main paper implies that the sample covariance matrix $\hat{\boldsymbol{\Sigma}}$ satisfies the restricted eigenvalue condition with probability tending to one. Furthermore, the condition on the tuning parameter μ implies that we can apply Lemma 1 with $\theta = C_{\min}^{1/2}/8 > 0$. Hence for each dimension $j = 1, \ldots, d$, we then have

$$\|\boldsymbol{\delta}_j\|_2 = \left\|\hat{\boldsymbol{\beta}}_j - \tilde{\boldsymbol{\beta}}_j\right\|_2 \leq 24 C_{\min}^{-1/2} s_j^{1/2} \mu_j \leq C \left\{\frac{s_j \log(p)}{n\lambda_j}\right\}^{1/2} \leq C \left\{\frac{s \log(p)}{n\lambda_d}\right\}^{1/2}.$$

Proposition 3 and Condition (C2) in the main paper imply that the norm $\left\|\tilde{\boldsymbol{\beta}}_j\right\|_2$ is bounded away from zero. As a result,

$$\left\|\mathcal{P}(\widehat{\boldsymbol{\beta}}_j) - \mathcal{P}(\boldsymbol{\beta}_j)\right\|_F = \left\|\mathcal{P}(\widehat{\boldsymbol{\beta}}_j) - \mathcal{P}(\tilde{\boldsymbol{\beta}}_j)\right\|_F \leq 4\frac{\|\widehat{\boldsymbol{\beta}}_j - \widetilde{\boldsymbol{\beta}}_j\|_2}{\|\widetilde{\boldsymbol{\beta}}\|_2} = 4\frac{\|\boldsymbol{\delta}_j\|_2}{\|\widetilde{\boldsymbol{\beta}}_j\|_2} \leq C \left\{\frac{s \log(p)}{n\lambda_d}\right\}^{1/2}$$

for a sufficiently large constant C. Furthermore, Lin et al. (2019) shows that the lengths of each vector $\tilde{\boldsymbol{\beta}}_j,\ j = 1, \ldots, d$ are bounded below by $C(\lambda/\hat{\lambda}_j)^{1/2}$, and the angles between any two vectors of $\tilde{\boldsymbol{\beta}}_j,\ j = 1, \ldots, d$ are bounded below by a constant. The Gram-Schmidt process then implies

$$\left\| \mathcal{P}(\widehat{\mathbf{B}}) - \mathcal{P}(\mathbf{B}) \right\|_F \leq C \left\{ \frac{s \log(p)}{n\lambda} \right\}^{1/2}$$

as claimed.

S2.2 Proof of Theorem 2

It suffices to prove selection consistency for each dimension. In the proof below, the notations $\tilde{\boldsymbol{\beta}}$, $\bar{\boldsymbol{\beta}}$, and $\boldsymbol{\beta}^*$ denote the pseudo-true parameter (defined in the main paper), the initial consistent estimate, and the Adaptive Cholesky estimate for each dimension respectively; furthermore the set S is the true index set of non-zero components of $\boldsymbol{\beta}$. The subscript used in the proof, for example $\boldsymbol{\beta}_k$, denotes the kth component of $\boldsymbol{\beta}$, and $\boldsymbol{\beta}_S$, denotes the vector of components of $\boldsymbol{\beta}$ whose indices belong to S. For any matrix A and a set T, the notation $A_{,T}$ and $A_{T,}$ denotes the submatrix of A with column indices in T and the submatrix of A with row indices in T respectively, and $A_{T,T}$ denotes the submatrix with both row and column indices in T.

First, let $\hat{\boldsymbol{\Delta}} = \text{diag}(\bar{\boldsymbol{\beta}}_1^{\gamma}, \ldots, \bar{\boldsymbol{\beta}}_p^{\gamma})$, a diagonal matrix whose elements

correspond to the inverse of the weight vector ω. For ease of notation, consider the case of $\gamma = 1$. Due to consistency of the initial estimator, the matrix $\hat{\boldsymbol{\Delta}}_{S,S}$ is invertible with probability one. Furthermore, since the sample covariance matrix $\hat{\boldsymbol{\Sigma}}$ satisfies the restricted eigenvalue condition with probability tending to one (Proposition 1), the minimum eigenvalue of the matrix $\hat{\boldsymbol{\Sigma}}_{S,S}$ is bounded away from zero with probability tending to one as well. In that case, each component of the adaptive Cholesky matrix penalization estimator can be computed as $\hat{\boldsymbol{\beta}}_k^* = \bar{\boldsymbol{\beta}}_k \hat{u}_k$, $k = 1, \ldots, p$, where the vector $\hat{\boldsymbol{u}} = (\hat{u}_1, \ldots, \hat{u}_p)^\top$ solves the following minimization problem

$$\hat{\boldsymbol{u}} = \arg\min_{\boldsymbol{u}} \frac{1}{2} \|\boldsymbol{V}\boldsymbol{u} - \hat{\boldsymbol{\kappa}}\|_2^2 + \mu \|\boldsymbol{u}\|_1$$

with $\boldsymbol{V} = \hat{\boldsymbol{L}}^\top \hat{\boldsymbol{\Delta}}$. Therefore, if $\hat{\boldsymbol{u}}$ recovers the exact sparsity pattern, so does the adaptive Cholesky matrix penalization estimate. From the Karush-Kuhn-Tucker condition, the estimate $\hat{u}$ satisfies

$$\boldsymbol{V}^\top \boldsymbol{V}\hat{\boldsymbol{u}} - \boldsymbol{V}^\top \hat{\boldsymbol{\kappa}} + \mu w = 0, \tag{S2.4}$$

where $\boldsymbol{w} = (w_1, \ldots, w_p)$ with $w_k = \text{sign}(\hat{u}_k)$ if $\hat{u}_k \neq 0$ and $|w_k| \leq 1$ otherwise. Therefore, $\hat{\boldsymbol{u}}$ recovers the exact sparsity pattern of $\boldsymbol{\beta}^*$ if and only if $\boldsymbol{u}_S \neq 0, \boldsymbol{w}_S = \text{sign}(\boldsymbol{\beta}_S), u_{S^c} = 0, |w_{S^c}| \leq 1$. Furthermore, by definition of $\boldsymbol{\kappa}$, the quantity $\boldsymbol{V}^\top \hat{\boldsymbol{\kappa}} = \hat{\boldsymbol{\Delta}}^\top \hat{\boldsymbol{L}} \hat{\boldsymbol{\kappa}} = \hat{\boldsymbol{\Delta}} \hat{\boldsymbol{\eta}}$, and $\boldsymbol{V}\hat{\boldsymbol{u}} = \boldsymbol{V}_{,S} \hat{\boldsymbol{u}}_S$. Combining these with condition (S2.4) above, if $\hat{\boldsymbol{u}}$ recovers the exact sparsity pattern of $\boldsymbol{\beta}$,

we then have

$$\boldsymbol{V}_{S,}^{\top}\boldsymbol{V}_{,S}\hat{\boldsymbol{u}}_S - \hat{\boldsymbol{\Delta}}_{S,S}\hat{\boldsymbol{\eta}}_S + \mu\text{sign}(\boldsymbol{\beta}_S) = 0$$

$$\boldsymbol{V}_{S^c,}^{\top}\boldsymbol{V}_{,S}\hat{\boldsymbol{u}}_S - \hat{\boldsymbol{\Delta}}_{S^c,S^c}^{\top}\hat{\boldsymbol{\eta}}_{S^c} + \mu\boldsymbol{w}_{S^c} = 0.$$

Solving this system of equations, we then have

$$\hat{\boldsymbol{u}}_S = (\boldsymbol{V}_{S,}^{\top}\boldsymbol{V}_{,S})^{-1}\left\{\hat{\boldsymbol{\Delta}}_{S,S}^{\top}\hat{\boldsymbol{\eta}}_S - \mu\text{sign}(\boldsymbol{\beta}_S)\right\}$$

$$-\mu\boldsymbol{w}_{S^c} = \boldsymbol{V}_{S^c,}^{\top}\boldsymbol{V}_{,S}(\boldsymbol{V}_{S,}^{\top}\boldsymbol{V}_{,S})^{-1}\left\{\hat{\boldsymbol{\Delta}}_{S,S}^{\top}\hat{\boldsymbol{\eta}}_S - \mu\text{sign}(\boldsymbol{\beta}_S)\right\} - \hat{\boldsymbol{\Delta}}_{S^c,S^c}\hat{\boldsymbol{\eta}}_{S^c}.$$

(1) With this in mind, we will show that probability of underselection goes to zero by showing that $\text{pr}(\hat{u}_S \neq 0) \to 1$. In fact,

$$\begin{aligned}(\boldsymbol{V}_{S,}^{\top}\boldsymbol{V}_{,S})^{-1}\hat{\boldsymbol{\Delta}}_{S,S}^{\top}\hat{\boldsymbol{\eta}}_S &= \left(\hat{\boldsymbol{\Delta}}_{S,S}\hat{\boldsymbol{L}}_{S,}\hat{\boldsymbol{L}}_{,S}^{\top}\hat{\boldsymbol{\Delta}}_{S,S}\right)^{-1}\hat{\boldsymbol{\Delta}}_{S,S}\hat{\boldsymbol{\eta}}_S = \hat{\boldsymbol{\Delta}}_{S,S}^{-1}\hat{\boldsymbol{\Sigma}}_{S,S}^{-1}\hat{\boldsymbol{\eta}}_S \\ &= \hat{\boldsymbol{\Delta}}_{S,S}^{-1}\hat{\boldsymbol{\Sigma}}_{S,S}^{-1}\left(\hat{\boldsymbol{\eta}}_S - \tilde{\boldsymbol{\eta}}_S\right) + \hat{\boldsymbol{\Delta}}_{S,S}^{-1}\hat{\boldsymbol{\Sigma}}_{S,S}^{-1}\tilde{\boldsymbol{\eta}}_S \\ &= \underbrace{\hat{\boldsymbol{\Delta}}_{S,S}^{-1}\hat{\boldsymbol{\Sigma}}_{S,S}^{-1}\left(\hat{\boldsymbol{\eta}}_S - \tilde{\boldsymbol{\eta}}_S\right)}_{I_1} + \underbrace{\hat{\boldsymbol{\Delta}}_{S,S}^{-1}\hat{\boldsymbol{\Sigma}}_{S,S}^{-1}\boldsymbol{\Sigma}_{S,S}\tilde{\boldsymbol{\beta}}_S}_{I_2},\end{aligned}$$

where the last inequality follows from the definition that $\tilde{\boldsymbol{\eta}} = \boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}}$ and the vector $\tilde{\boldsymbol{\beta}}$ is a sparse vector. Using Proposition 2, condition (C1) and (C4), we then have

$$\|I_1\|_\infty \leq \left\|\boldsymbol{\Delta}_{S,S}^{-1}\right\|_\infty \left\|\hat{\boldsymbol{\Sigma}}_{S,S}^{-1}\right\|_\infty \|\hat{\boldsymbol{\eta}}_S - \tilde{\boldsymbol{\eta}}_S\|_\infty \leq \frac{s^{1/2}}{\rho_n C_{\min}^{1/2}} O\left(\frac{\log(p)^{1/2}}{(n\lambda)^{1/2}}\right) = O_p\left\{\frac{s^{1/2}\log(p)^{1/2}}{\rho_n(n\lambda)^{1/2}}\right\} \to 0,$$

since $n^{-1}\lambda^{-1}\rho_n^{-2}s\log(p) \to 0$. Next,

$$\begin{aligned}I_2 = \hat{\boldsymbol{\Delta}}_{S,S}^{-1}\hat{\boldsymbol{\Sigma}}_{S,S}^{-1}\boldsymbol{\Sigma}_{S,S}\tilde{\boldsymbol{\beta}}_S &= \hat{\boldsymbol{\Delta}}_{S,S}^{-1}\hat{\boldsymbol{\Sigma}}_{S,S}^{-1}\hat{\boldsymbol{\Sigma}}_{S,S}\tilde{\boldsymbol{\beta}}_S + \hat{\boldsymbol{\Delta}}_{S,S}^{-1}\hat{\boldsymbol{\Sigma}}_{S,S}^{-1}\left(\boldsymbol{\Sigma}_{S,S} - \hat{\boldsymbol{\Sigma}}_{S,S}\right)\hat{\boldsymbol{\beta}}_S \\ &= \hat{\boldsymbol{\Delta}}_{S,S}^{-1}\tilde{\boldsymbol{\beta}}_S + \hat{\boldsymbol{\Delta}}_{S,S}^{-1}\hat{\boldsymbol{\Sigma}}_{S,S}^{-1}\left(\boldsymbol{\Sigma}_{S,S} - \hat{\boldsymbol{\Sigma}}_{S,S}\right)\hat{\boldsymbol{\beta}}_S = I_{21} + I_{22}.\end{aligned}$$

Due to the consistency of the initial estimator $\bar{\boldsymbol{\beta}}$, each element of the term I_{21} converges to a non-zero constant with probability 1 at a rate $O_p(\boldsymbol{\delta}_n)$ since $\boldsymbol{\delta}_n = o(\rho_n)$. Since $\left\|\boldsymbol{\Sigma}_{S,S} - \hat{\boldsymbol{\Sigma}}_{S,S}\right\|_2 = O\left\{(s/n)^{1/2}\right\}$ (Wainwright, 2019), we have $\left\|\boldsymbol{\Sigma}_{S,S} - \hat{\boldsymbol{\Sigma}}_{S,S}\right\|_\infty = O(sn^{-1/2})$ and

$$\|I_{22}\|_\infty \leq \left\|\hat{\boldsymbol{\Delta}}_{S,S}^{-1}\right\|_\infty \left\|\hat{\boldsymbol{\Sigma}}_{S,S}\right\|_\infty^{-1} \left\|\boldsymbol{\Sigma}_{S,S} - \hat{\boldsymbol{\Sigma}}_{S,S}\right\|_\infty \left\|\tilde{\boldsymbol{\beta}}_S\right\|_\infty = O_p\left\{\rho_n^{-1} s^{3/2} n^{-1/2}\right\} \to 0$$

since $\rho_n^{-1} s^{3/2} n^{-1/2} \to 0$. Finally, we consider $\mu(V_S^\top V_S)^{-1}\text{sign}(\boldsymbol{\beta}_S) = \mu\left(\hat{\boldsymbol{\Delta}}_{S,S}\hat{\boldsymbol{\Sigma}}_{S,S}\hat{\boldsymbol{\Delta}}_{S,S}\right)^{-1}\text{sign}(\boldsymbol{\beta}_S)$. We have

$$\left\|\mu\left(\hat{\boldsymbol{\Delta}}_{S,S}\hat{\boldsymbol{\Sigma}}_{S,S}\hat{\boldsymbol{\Delta}}_{S,S}\right)^{-1}\text{sign}(\boldsymbol{\beta}_S)\right\|_\infty \leq \mu O_p(s^{1/2})\left\|\hat{\boldsymbol{\Delta}}_{S,S}^{-1}\right\|_\infty^2 = O_p\left(\frac{\mu s^{1/2}}{\rho_n^2}\right)$$

so this term also goes to zero when $\mu = o(\rho_n^2/s^{1/2})$.

(2) We will show that the probability of overselection also goes to zero. Define the term

$$Q = \boldsymbol{V}_{S^c,}^\top \boldsymbol{V}_{,S}(\boldsymbol{V}_{S,}^\top \boldsymbol{V}_{,S})^{-1}\left\{\hat{\boldsymbol{\Delta}}_{S,S}^\top \hat{\boldsymbol{\eta}}_S - \mu\text{sign}(\boldsymbol{\beta}_S)\right\} - \hat{\boldsymbol{\Delta}}_{S^c,S^c}\hat{\boldsymbol{\eta}}_{S^c},$$

so there would be no over-selection if $\|Q\|_\infty \leq \mu$. By the triangle inequality and the fact that $\|\hat{\boldsymbol{\eta}}\|_\infty = |\text{sign}(\boldsymbol{\beta}_S)| \leq 1$, we have

$$\begin{aligned}
\|Q\|_\infty &\leq \left\|\hat{\boldsymbol{\Delta}}_{S^c,S^c}\hat{\boldsymbol{\eta}}_{S^c}\right\|_\infty + \left\|\boldsymbol{V}_{S^c,}^\top \boldsymbol{V}_{,S}(\boldsymbol{V}_{S,}^\top \boldsymbol{V}_{,S})^{-1}\hat{\boldsymbol{\Delta}}_{S,S}^\top \hat{\boldsymbol{\eta}}_S\right\|_\infty + \mu\left\|\boldsymbol{V}_{S^c,}^\top \boldsymbol{V}_{,S}(\boldsymbol{V}_{S,}^\top \boldsymbol{V}_{,S})^{-1}\text{sign}(\boldsymbol{\beta}_S)\right\|_\infty \\
&\leq \left\|\hat{\boldsymbol{\Delta}}_{S^c,S^c}\right\|_\infty + \left\|\hat{\boldsymbol{\Delta}}_{S^c,S^c}\hat{\boldsymbol{L}}_{S^c,}\hat{\boldsymbol{L}}_{,S}^\top\left(\hat{\boldsymbol{L}}_{S,}\hat{\boldsymbol{L}}_{,S}^\top\right)^{-1}\right\|_\infty + \mu\left\|\hat{\boldsymbol{\Delta}}_{S^c,S^c}\hat{\boldsymbol{L}}_{S^c,}\hat{\boldsymbol{L}}_{,S}^\top\left(\hat{L}_{S,}\hat{\boldsymbol{L}}_{,S}^\top\right)^{-1}\hat{\boldsymbol{\Delta}}_{S,S}^{-1}\right\|_\infty \\
&= O_p(\boldsymbol{\delta}_n) + O_p(\boldsymbol{\delta}_n) + O_p\left(\frac{\mu\,\boldsymbol{\delta}_n}{\rho_n}\right) \leq \mu
\end{aligned}$$

as long as $\boldsymbol{\delta}_n / \mu \to 0$, where the last equality follows from $\left\| \hat{\boldsymbol{L}}_{S^c,} \hat{\boldsymbol{L}}_{,S}^\top \left(\hat{\boldsymbol{L}}_{S,} \hat{\boldsymbol{L}}_{,S}^\top \right)^{-1} \right\|_\infty = \left\| \mathcal{X}_{S^c}^\top \mathcal{X}_S \left(\mathcal{X}_S^\top \mathcal{X}_S \right)^{-1} \right\|_\infty = O(1)$ by condition (C5).

(3) Finally, we show the bound on the error of the projection matrix associated with the Adaptive Cholesky Matrix $\hat{B}^*$. By the same argument as in the proof of Theorem 1 (Section S2.1), it suffices to show that $\left\| \hat{\boldsymbol{\beta}}^* - \tilde{\boldsymbol{\beta}} \right\|_2 \leq C s^{1/2} \log(p)^{1/2} / (n\lambda)$. In fact, due to variable selection consistency, it suffices to show the bound holds for $\left\| \hat{\boldsymbol{\beta}}_S^* - \tilde{\boldsymbol{\beta}}_S \right\|_2$. The first-order condition then implies

$$\boldsymbol{V}_{S,}^\top \boldsymbol{V}_{,S} \hat{\boldsymbol{u}}_S - \hat{\boldsymbol{\Delta}}_{S,S} \hat{\boldsymbol{\eta}}_S + \mu \text{sign}(\hat{\boldsymbol{u}}_S) = 0.$$

By definition, we have $\hat{\boldsymbol{u}}_S = \hat{\boldsymbol{\Delta}}_{S,S}^{-1} \hat{\boldsymbol{\beta}}_S^*$ and $\boldsymbol{V}_{,S} = \hat{\boldsymbol{L}}_{,S}^\top \hat{\boldsymbol{\Delta}}_{S,S}$, so substituting them into the above equation gives

$$\begin{aligned} & \hat{\boldsymbol{\Delta}}_{S,S} \hat{\boldsymbol{L}}_{S,} \hat{\boldsymbol{L}}_{,S}^\top \hat{\boldsymbol{\beta}}_S^* = \hat{\boldsymbol{\Delta}}_{S,S} \hat{\boldsymbol{\eta}}_S - \mu \text{sign}(\hat{\boldsymbol{u}}_S) \\ \text{or} \quad & \hat{\boldsymbol{\Delta}}_{S,S} \hat{\boldsymbol{\Sigma}}_{S,S} \hat{\boldsymbol{\beta}}_S^* = \hat{\boldsymbol{\Delta}}_{S,S} (\hat{\boldsymbol{\eta}}_S - \tilde{\boldsymbol{\eta}}_S) + \hat{\boldsymbol{\Delta}}_{S,S} \tilde{\boldsymbol{\eta}}_S - \mu \text{sign}(\hat{\boldsymbol{u}}_S). \end{aligned}$$

Also, by definition $\tilde{\boldsymbol{\eta}}_S = \boldsymbol{\Sigma}_{S,S}^{-1} \tilde{\boldsymbol{\beta}}_S$, so substituting it into the above equation and doing one algebraic manipulation gives

$$\begin{aligned} & \hat{\boldsymbol{\Delta}}_{S,S} \hat{\boldsymbol{\Sigma}}_{S,S} (\hat{\boldsymbol{\beta}}_S^* - \tilde{\boldsymbol{\beta}}_S) = \hat{\boldsymbol{\Delta}}_{S,S} (\hat{\boldsymbol{\eta}}_S - \tilde{\boldsymbol{\eta}}_S) + \hat{\boldsymbol{\Delta}}_{S,S} (\boldsymbol{\Sigma}_{S,S} - \hat{\boldsymbol{\Sigma}}_{S,S}) \tilde{\boldsymbol{\beta}}_S - \mu \text{sign}(\hat{\boldsymbol{u}}_S) \\ \text{or} \quad & (\hat{\boldsymbol{\beta}}_S^* - \tilde{\boldsymbol{\beta}}_S) = \hat{\boldsymbol{\Sigma}}_{S,S}^{-1} (\hat{\boldsymbol{\eta}}_S - \tilde{\boldsymbol{\eta}}_S) + \hat{\boldsymbol{\Sigma}}_{S,S}^{-1} (\boldsymbol{\Sigma}_{S,S} - \hat{\boldsymbol{\Sigma}}_{S,S}) \tilde{\boldsymbol{\beta}}_S - \mu \hat{\boldsymbol{\Sigma}}_{S,S}^{-1} \hat{\boldsymbol{\Delta}}_{S,S}^{-1} \text{sign}(\hat{\boldsymbol{u}}_S). \end{aligned}$$

Therefore, the triangular inequality and the fact that $\text{sign}(\hat{u}_S) = \pm 1$ gives

$$\begin{aligned}\left\|\hat{\boldsymbol{\beta}}^*_S - \tilde{\boldsymbol{\beta}}_S\right\|_\infty &\leq \left\|\hat{\boldsymbol{\Sigma}}^{-1}_{S,S}\right\|_2 \|\hat{\boldsymbol{\eta}}_S - \tilde{\boldsymbol{\eta}}_S\|_\infty + \left\|\hat{\boldsymbol{\Sigma}}^{-1}_{S,S}\right\|_2 \left\|\boldsymbol{\Sigma}_{S,S} - \hat{\boldsymbol{\Sigma}}_{S,S}\right\|_2 \left\|\tilde{\boldsymbol{\beta}}_S\right\|_\infty + \mu \left\|\hat{\boldsymbol{\Sigma}}^{-1}_{S,S}\right\|_\infty \left\|\hat{\boldsymbol{\Delta}}^{-1}_{S,S}\right\|_\infty \\ &= O_p\left\{\frac{\log(p)^{1/2}}{(n\lambda)^{1/2}}\right\} + O_p\left(\frac{s^{1/2}}{n^{1/2}}\right) + O_p(\mu s^{1/2}\rho_n^{-1}) = O_p\left\{\frac{\log(p)^{1/2}}{(n\lambda)^{1/2}}\right\},\end{aligned}$$

due to the condition of the tuning parameters as stated in the Theorem.

Finally, we have

$$\left\|\hat{\boldsymbol{\beta}}^*_S - \tilde{\boldsymbol{\beta}}_S\right\|_2 \leq s^{1/2}\left\|\hat{\boldsymbol{\beta}}^*_S - \tilde{\boldsymbol{\beta}}_S\right\|_\infty \leq C\left\{\frac{s\log(p)}{(n\lambda)}\right\}^{1/2},$$

for a sufficiently large constant C, as claimed.

S2.3 Proof of Theorem 3

Recall that for each dimension $j = 1, \ldots, d$, the set $S_j = \{k : \beta_{jk} \neq 0\}$ the set of indices corresponding to non-zero components of the true dimension $\boldsymbol{\beta}_j$. Any index set $\mathcal{S} \subset \{1, \ldots, p\}$ such that $\mathcal{S} \not\supset S_j$ is referred to as an underfitted index set, while any $\mathcal{S} \supsetneq S_j$ other than S_j itself is referred to as an overfitted index set. Correspondingly, we can partition the values of the tuning parameter μ_j into the underfitted, true, and overfitted ranges respectively,

$$\Omega_{j-} = \{\mu_j : \hat{S}(\mu_j) \not\supset S_j\},\ \Omega_{0j} = \{\mu_j : \hat{S}(\mu_j) = S_j\},\ \text{and}\ \Omega_{j+} = \{\mu_j : \hat{S}(\mu_j) \supsetneq S_j\}$$

where $\hat{S}(\mu_j) = \{k : \hat{\boldsymbol{\beta}}^*_{jk}(\mu_j) \neq 0\}$, the set of indices corresponding to the nonzero component of $\hat{\boldsymbol{\beta}}^*_j(\mu_j)$, the adaptive Cholesky matrix penalization

estimator at the tuning parameter μ_j. We will show that, for any μ_j that cannot identify the true model and the value of τ_j stated in the Theorem 3, the resulting $\mathrm{PIC}(\mu_j;\tau_j)$ is consistently larger than $\mathrm{PIC}(\mu_0;\tau_j)$ with $\mu_0 \in \Omega_{0j}$. To simplify the notation, we use $\mathrm{PIC}(\mu_j)$. We will treat two cases of overfitting and underfitting separately.

Overfitted range

For $\mu_j \in \Omega_{j+}$ (the overfitted range), we have,

$$\mathrm{PIC}(\mu_j)-\mathrm{PIC}(\mu_0) = \left\|\mathcal{P}\left\{\hat{\boldsymbol{\beta}}_j^*(\mu_j)\right\} - \mathcal{P}(\bar{\boldsymbol{\beta}}_j)\right\|_F^2 - \left\|\mathcal{P}\left\{\hat{\boldsymbol{\beta}}_j^*(\mu_0)\right\} - \mathcal{P}(\bar{\boldsymbol{\beta}}_j)\right\|_F^2 + \tau_j\Delta_j \tag{S2.5}$$

where $\Delta_j = \left\|\hat{\boldsymbol{\beta}}_j^*(\mu_j)\right\|_0 - \left\|\hat{\boldsymbol{\beta}}_j^*(\mu_0)\right\|_0 > 0$. By the triangle inequality, we obtain

$$\left\|\mathcal{P}\left\{\hat{\boldsymbol{\beta}}_j^*(\mu_0)\right\} - \mathcal{P}(\bar{\boldsymbol{\beta}}_j)\right\|_F \leq \left\|\mathcal{P}\left\{\hat{\boldsymbol{\beta}}_j^*(\mu_0)\right\} - \mathcal{P}(\boldsymbol{\beta}_j)\right\|_F + \left\|\mathcal{P}(\boldsymbol{\beta}_j) - \mathcal{P}(\bar{\boldsymbol{\beta}})\right\|_F, \tag{S2.6}$$

For the first term in the right hand side of (S2.6), the tuning parameter μ_0 satisfies the condition for the tuning parameter in Theorem 2, so $\left\|\mathcal{P}\left\{\hat{\boldsymbol{\beta}}_j^*(\mu_0)\right\} - \mathcal{P}(\boldsymbol{\beta}_j)\right\|_F = O_p\left[\{s\log(p)/(n\lambda)\}^{1/2}\right]$. The second term $\left\|\mathcal{P}(\boldsymbol{\beta}_j) - \mathcal{P}(\bar{\boldsymbol{\beta}}_j)\right\|_F = O_p(\sqrt{p/n})$. Since $s\log(p) = o(p)$, the rate of convergence of the right hand side of (S2.6) is dominated by the rate of convergence of the unpenalized estimator; i.e $\left\|\mathcal{P}\left\{\hat{\boldsymbol{\beta}}_j^*(\mu_0)\right\} - \mathcal{P}(\bar{\boldsymbol{\beta}}_j)\right\|_F^2 = O(p/n)$.

Finally, since $\Delta_j > 0$ and $\tau_j \overset{>}{\sim} p/n$, the right hand side of (S2.5) is asymptotically positive, i.e $\text{PIC}(\mu_j) > \text{PIC}(\mu_0)$ for every $\mu_j \in \Omega_+$ when $n \to \infty$.

Underfitted range

For $\mu_j \in \Omega_{j-}$ (the underfitted range), we want to show

$$\text{PIC}(\mu_j) - \text{PIC}(\mu_0) = \left\|\mathcal{P}\left\{\hat{\boldsymbol{\beta}}_j^*(\mu_j)\right\} - \mathcal{P}(\bar{\boldsymbol{\beta}}_j)\right\|_F^2 - \left\|\mathcal{P}\left\{\hat{\boldsymbol{\beta}}_j^*(\mu_0)\right\} - \mathcal{P}(\bar{\boldsymbol{\beta}}_j)\right\|_F^2 + \tau_j\left(\left\|\hat{\boldsymbol{\beta}}_j^*(\mu_j)\right\|_0 - s_j\right) > 0 \tag{S2.7}$$

occurs with probability tending to one as $n \to \infty$. By the same argument as in the previous section for the overfitted range, the second term in the right hand side of (S2.7) $\left\|\mathcal{P}\left\{\hat{\boldsymbol{\beta}}_j^*(\mu_0)\right\} - \mathcal{P}(\bar{\boldsymbol{\beta}}_j)\right\|_F^2 = O(p/n)$. For the firm term in (S2.7), applying the triangle inequality again, we have

$$\left\|\mathcal{P}\left\{\hat{\boldsymbol{\beta}}_j^*(\mu_j)\right\} - \mathcal{P}(\bar{\boldsymbol{\beta}}_j)\right\|_F \geq \left\|\mathcal{P}\left\{\hat{\boldsymbol{\beta}}_j^*(\mu_j)\right\} - \mathcal{P}(\boldsymbol{\beta}_j)\right\|_F - \left\|\mathcal{P}(\boldsymbol{\beta}_j) - \mathcal{P}(\bar{\boldsymbol{\beta}}_j)\right\|_F \tag{S2.8}$$

First, regarding the second term on the right hand side of (S2.8), we have $\left\|\mathcal{P}(\boldsymbol{\beta}_j) - \mathcal{P}(\bar{\boldsymbol{\beta}}_j)\right\|_F = O(\sqrt{p/n})$. For the first term on the right hand side of (S2.8), let $\mathcal{K}_j$ be the index set of underfitted components, i.e for all $k \in \mathcal{K}_j$, we have $\hat{\beta}_{jk}^*(\mu_j) = 0$ while $\beta_{jk} \neq 0$. Hence, all the elements whose at least one of the column and row indices of the estimated projection matrix

$\mathcal{P}\left\{\hat{\boldsymbol{\beta}}_j^*(\mu_j)\right\}$ are zero. Therefore, we obtain

$$\left\|\mathcal{P}\left\{\hat{\boldsymbol{\beta}}_j^*(\mu_j)\right\} - \mathcal{P}(\boldsymbol{\beta}_j)\right\|_F^2 \geq 2\frac{\sum_{k\in\mathcal{K}_j}\boldsymbol{\beta}_{jk}^2}{\left\|\boldsymbol{\beta}_j\right\|_2^2} - \frac{\sum_{k\in\mathcal{K}_j}\boldsymbol{\beta}_{jk}^4}{\left\|\boldsymbol{\beta}_j\right\|_2^4} - 2\frac{\sum_{k,r\in\mathcal{K}_j}\beta_{jk}^2\beta_{jr}^2}{\left\|\boldsymbol{\beta}_j\right\|_2^4}$$
$$= 2\frac{\left\|\boldsymbol{\beta}_{j\mathcal{K}_j}\right\|_2^2}{\left\|\boldsymbol{\beta}_j\right\|_2^2} - \frac{\left\|\boldsymbol{\beta}_{j\mathcal{K}_j}\right\|_2^4}{\left\|\boldsymbol{\beta}_j\right\|_2^4}.$$

Let $\xi_j = \min_{k=1,\ldots,s_j}\left\{\dfrac{\boldsymbol{\beta}_{jk}^2}{\boldsymbol{\beta}_j^\top\boldsymbol{\beta}_j}\right\}$, so we have

$$|\mathcal{K}_j|\xi_j \leq \frac{\left\|\boldsymbol{\beta}_{j\mathcal{K}_j}\right\|_2^2}{\left\|\boldsymbol{\beta}_j\right\|_2^2} < 1,$$

where $|\mathcal{K}_j|$ denotes the cardinality of the set $\mathcal{K}_j$. Since the function $f(x) = x(2-x)$ is monotonic increasing on $[0,1]$, we obtain

$$\left\|\mathcal{P}\left\{\hat{\boldsymbol{\beta}}_j^*(\mu_j)\right\} - \mathcal{P}(\boldsymbol{\beta}_j)\right\|_F^2 \geq 2|\mathcal{K}_j|\xi_j - |\mathcal{K}_j|^2\xi_j^2.$$

Note that $\left\|\hat{\boldsymbol{\beta}}_j^*(\mu_j)\right\|_0 \geq s_j - |\mathcal{K}_j|$, so when $p/n \to 0$, equation (S2.7) is satisfied if for all $|\mathcal{K}_j| = 1,\ldots,s_j$, we have

$$2|\mathcal{K}_j|\xi_j - |\mathcal{K}_j|^2\xi_j^2 - \tau_j|\mathcal{K}_j| > 0, \text{ i.e. } \tau_j < 2\xi_j - |\mathcal{K}_j|\xi_j^2 = \xi_j(2-|\mathcal{K}_j|\xi_j). \quad \text{(S2.9)}$$

Since $|\mathcal{K}_j|\xi_j$ is smaller than 1, equation (S2.9) is satisfied if $\tau_j < \xi_j$. In other words, if $\tau_j = o(\xi_j)$, then $\text{PIC}(\mu_j) > \text{PIC}(\mu_0)$ as $n \to \infty$ as claimed.

S3 More details about the Lasso sliced inverse regression estimator

In this section, we briefly review the Lasso sliced inverse regression (SIR) estimator and establish the connection between it and the CHOMP estimator. Assume that a random sample $(\boldsymbol{x}_i^\top, y_i),\ i = 1, \ldots, n$ is generated from the single index model $y_i = f(\boldsymbol{x}_i^\top \boldsymbol{\beta}_0, \varepsilon_i),\ i = 1, \ldots, n$, with the outcome y_i, and covariate x_i follows a p-dimensional elliptical distribution with location zero and scale matrix $\boldsymbol{\Sigma}$. Let $\mathcal{X}$ denote the $n \times p$ design matrix. The sliced inverse regression estimate for $\boldsymbol{\beta}_0$ is based on the relationship

$$\boldsymbol{\Sigma}\, \boldsymbol{\beta}_0 \propto \boldsymbol{\eta}. \tag{S3.10}$$

The covariance matrix $\boldsymbol{\Sigma}$ is estimated by the sample covariance matrix $\hat{\boldsymbol{\Sigma}} = n^{-1}\mathcal{X}^\top\mathcal{X}$. Next, without loss of generality, assume the data $(\boldsymbol{x}_i, y_i)$ are arranged such that $y_1 \leq y_2 \leq \ldots \leq y_n$. Then the data are divided into H equal-sized slices, denoted by $J_1, \ldots, J_H$ based on the increasing order of y. For ease of notation and arguments, assume $n = cH$ with $c > 0$. Next, construct a $H \times n$ matrix $\boldsymbol{M} = \mathbf{I}_H \otimes \mathbf{1}_c^\top$, where $\mathbf{I}_H$ denotes the identity matrix of dimension H, $\mathbf{1}_c$ denotes the $c \times 1$ vector with all entries being one, and $\otimes$ denotes the outer product. Next, compute the averages of covariates within each slice, $\bar{\boldsymbol{x}}_h^\top = c^{-1}\sum_{i=1}^n \boldsymbol{x}_i^\top 1(y_i \in J_h)$, and form a $H \times p$ matrix

$\mathcal{X}_H$ with each row being $\bar{\boldsymbol{x}}_h^\top,\ h = 1, \ldots, H$. With this formulation, $\mathcal{X}_H = \boldsymbol{M}\mathcal{X}/c$, so the conditional expectation $\boldsymbol{\Lambda} = \mathrm{var}\{\mathbb{E}(\mathbf{X}\,|y)\}$ is estimated by

$$\hat{\boldsymbol{\Lambda}} = H^{-1}\sum_{h=1}^{H} \bar{\boldsymbol{x}}_h \bar{\boldsymbol{x}}_h^\top = \frac{1}{H}\mathcal{X}_H^\top \mathcal{X}_H = \frac{1}{nc}\mathcal{X}^\top \boldsymbol{M}^\top \boldsymbol{M}\mathcal{X}.$$

Let $\hat{\lambda}$ and $\hat{\boldsymbol{\eta}}$ be the largest eigenvalue and its corresponding eigenvector of length one of $\hat{\boldsymbol{\Lambda}}$. Then

$$\hat{\lambda}\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\eta}} = \frac{1}{nc}\mathcal{X}^\top \boldsymbol{M}^\top \boldsymbol{M}\mathcal{X}\hat{\boldsymbol{\eta}}.$$

Let $\tilde{\boldsymbol{y}} = (c\hat{\lambda})^{-1}\boldsymbol{M}^\top \boldsymbol{M}\mathcal{X}\hat{\boldsymbol{\eta}}$, then we have $\hat{\boldsymbol{\eta}} = n^{-1}\mathcal{X}^\top \tilde{\boldsymbol{y}}$. Therefore, the estimated version of equation (S3.10) can be written as $\mathcal{X}^\top \mathcal{X}\boldsymbol{\beta}_0 \propto \mathcal{X}^\top \tilde{\boldsymbol{y}}$ and the Lasso SIR estimate is defined as

$$\hat{\boldsymbol{\beta}}^{\mathrm{L}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2n}\|\tilde{\boldsymbol{y}} - \mathcal{X}\boldsymbol{\beta}\|_2^2 + \mu\|\boldsymbol{\beta}\|_1,$$

with μ being an appropriate tuning parameter. For the multiple index model $y_i = f(x_i^\top \boldsymbol{\beta}_1, \ldots, x_i^\top \boldsymbol{\beta}_d, \varepsilon_i)$, the Lasso SIR estimator for each dimension is defined as

$$\hat{\boldsymbol{\beta}}_j^{\mathrm{L}} = \arg\min_{\boldsymbol{\beta}_j} \frac{1}{2n}\left\|\widetilde{y}_j - \mathcal{X}\boldsymbol{\beta}_j\right\|_2^2 + \mu_j\|\boldsymbol{\beta}_j\|_1,\ j = 1, \ldots, d, \tag{S3.11}$$

where $\widetilde{\boldsymbol{y}}_j = (c\hat{\lambda}_j)^{-1}M^\top M\mathcal{X}\hat{\boldsymbol{\eta}}_j$, with $\hat{\lambda}_j$ and $\hat{\boldsymbol{\eta}}_j$ being the j^{th} largest eigenvalue and its corresponding eigenvector of $\hat{\boldsymbol{\Lambda}}$, and the μ_j are tuning parameters.

Next, we show that the Lasso SIR has the same estimating equation as the CHOMP estimator. From the definition (S3.11) and the first order

condition, each component of the Lasso SIR $\hat{\boldsymbol{\beta}}_j^L = (\hat{\boldsymbol{\beta}}_{j1}^L, \ldots, \hat{\boldsymbol{\beta}}_{jp}^L)^\top$ satisfies

$$n^{-1}\boldsymbol{x}_k^\top(\tilde{\boldsymbol{y}} - \mathcal{X}\hat{\boldsymbol{\beta}}_j) + \mu_j b_{jk}^L = 0,$$

where $b_{jk}^L = \text{sign}(\hat{\boldsymbol{\beta}}_{jk})$ if $\hat{\boldsymbol{\beta}}_{jk} \neq 0$ and $b_{jk}^L \in [-1, 1]$ otherwise. Also, $n^{-1}\boldsymbol{x}_k^\top\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{\eta}}_{jk}$, and $n^{-1}\boldsymbol{x}_k^\top\mathcal{X} = \hat{\boldsymbol{\Sigma}}_k$, the k^{th} row of the sample covariance matrix $\hat{\boldsymbol{\Sigma}}$. Hence, for any tuning parameter μ_j, the estimating equation of the Lasso SIR is

$$-\hat{\boldsymbol{\Sigma}}_k\hat{\boldsymbol{\beta}}_j + \hat{\boldsymbol{\eta}}_{jk} + \mu_j b_{jk}^L = 0$$

exactly the same as the estimating equation of the CHOMP estimator shown in the paper. As a result, it is not surprising that the CHOMP and the Lasso SIR estimator require the same theoretical value of tuning parameters to ensure estimation consistency and share the same convergence rate.

Similar to any regularization method, the performance of the Lasso SIR depends critically on the choice of the tuning parameters μ_j. In their simulation study, Lin et al. (2019) implemented (S3.11) as a Lasso problem with design matrix $\mathcal{X}$ and outcome $\tilde{\boldsymbol{y}}_j$ and used ten-fold cross-validation to choose the tuning parameters μ_j. We show via a small simulation below that the Lasso SIR estimator with this choice of tuning parameters has performance close to the Lasso sliced inverse regression estimator where tuning parameters are chosen optimally. This finding justifies our compari-

son of the Lasso SIR estimator with tuning parameter selected via ten-fold cross-validation with other estimators in the simulation study of the main paper.

For the simulation study, we generate independent and identically distributed data $(\boldsymbol{x}_i^\top, y_i)$ as in the single index model simulation in Section 5.1 of the main paper with $s = 5$ and $n = 500$. The number of slices is fixed at $H = 20$ and the number of indices $d = 1$ is assumed to be known. We compare the average estimation error across 1000 samples of the Lasso SIR estimator under two methods for choosing the parameter. For the first method, the tuning parameter is chosen through ten-fold cross-validation. For the second method, the tuning parameter is chosen to minimize the actual estimation error; this choice of tuning parameter is referred to as the optimal tuning parameter. For any tuning parameter μ, the estimation error is defined as Error $= \left\| \mathcal{P}(\hat{\boldsymbol{\beta}}_\mu^{\mathrm{L}}) - \mathcal{P}(\boldsymbol{\beta}_0) \right\|_F^2$, the squared Frobenius norm of the difference between the estimated projection matrix and the true projection matrix. Note that the optimal tuning parameter is not available in practice, because it requires knowledge of the true vector $\boldsymbol{\beta}_0$.

It can be seen that the Lasso SIR estimator with tuning parameter selected via cross-validation gives very similar performance to the same estimator with optimal tuning parameter, where the difference in estimation

Table 1: Estimation error of the Lasso SIR estimator with tuning parameter selected via ten-fold cross-validation and with optimal tuning parameter. Standard errors are in parentheses.

p	$\boldsymbol{\beta}_0$	Cross-validation	Optimal
40	Large	0.28 (0.06)	0.27 (0.06)
	Small	0.43 (0.10)	0.41 (0.09)
100	Large	0.36 (0.06)	0.34 (0.06)
	Small	0.55 (0.10)	0.54 (0.10)

error is negligible. This result is surprising given the pseudo response $\tilde{y}$ does not contain independent components, so investigating why cross-validation still works for the Lasso sliced inverse regression estimation can be a topic for future research.

S3.1 An adaptive Lasso SIR estimator

Similar to the CHOMP estimator, the Lasso SIR estimator can be made adaptive by penalizing each component of $\boldsymbol{\beta}_j$ in (S3.11) differently. Specifically, an adaptive version of the Lasso SIR estimator is given by

$$\tilde{\boldsymbol{\beta}}_j^{\mathrm{L}} = \arg\min_{\boldsymbol{\beta}_j} \frac{1}{2n} \left\| \widetilde{y}_j - \mathcal{X}\boldsymbol{\beta}_j \right\|_2^2 + \mu_j \sum_{k=1}^{p} \omega_{jk} |\beta_{jk}|, \ j = 1, \ldots, d, , \quad \text{(S3.12)}$$

where $\mu_j > 0$ is a tuning parameter. Similar to the adaptive CHOMP estimator, we set the weights ω_{jk} to be $|\bar{\beta}_{jk}|^{-\gamma}$, with $\bar{\beta}_{jk}$ being the kth

component of an initial consistent estimate $\bar{\boldsymbol{\beta}}_j$ and γ a positive constant. In the setting when $n > p$, we choose $\bar{\boldsymbol{\beta}}_j$ to be the unpenalized estimate $\bar{\boldsymbol{\beta}}_j = \hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\eta}}_j$.

We conduct simulation studies to compare the performance of the Lasso SIR, adaptive Lasso SIR, CHOMP, and adaptive CHOMP estimators in single and multiple index models. The simulation settings are given in Sections 5.1 and 5.2 of the main paper, except that we only consider the scenario when $\boldsymbol{\Sigma}$ has an autoregressive structure for the single index model simulation. Two choices of γ are considered for adaptive estimators, $\gamma \in \{1, 2\}$. For the CHOMP-type estimators, we choose the tuning parameters based on the proposed PIC, while for the Lasso SIR-type estimators, we choose the tuning parameters based on ten-fold cross-validation. We compare estimators using the same metrics as given in Sections 5.1 and 5.2 of the main paper.

Table 2: Performance of different estimators in the single index model simulation. Standard errors are in parentheses. The lowest estimation error in each setting is highlighted.

Model	p	Metric	CHOMP	Adaptive CHOMP		Lasso SIR	Adaptive Lasso SIR	
				$\gamma = 1$	$\gamma = 2$		$\gamma = 1$	$\gamma = 2$
(I)	100	Error	0.26 (0.12)	0.12 (0.06)	**0.10** (0.05)	0.19 (0.06)	0.19 (0.08)	0.20 (0.11)
		FPR	0.00 (0.01)	0.00 (0.00)	0.00 (0.00)	0.19 (0.09)	0.12 (0.10)	0.10 (0.11)
		FNR	0.00 (0.03)	0.00 (0.01)	0.00 (0.00)	0.00 (0.01)	0.00 (0.00)	0.00 (0.00)
	200	Error	0.29 (0.13)	0.13 (0.07)	**0.12** (0.06)	0.23 (0.08)	0.25 (0.10)	0.32 (0.13)
		FPR	0.00 (0.01)	0.00 (0.00)	0.00 (0.00)	0.12 (0.06)	0.12 (0.09)	0.14 (0.10)
		FNR	0.01 (0.04)	0.00 (0.00)	0.00 (0.00)	0.00 (0.01)	0.00 (0.00)	0.00 (0.00)
(II)	100	Error	0.07 (0.04)	**0.03** (0.01)	**0.03** (0.01)	0.06 (0.02)	**0.03** (0.02)	**0.03** (0.01)
		FPR	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.18 (0.09)	0.01 (0.04)	0.00 (0.00)
		FNR	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	200	Error	0.08 (0.04)	**0.03** (0.01)	**0.03** (0.01)	0.06 (0.02)	0.04 (0.03)	**0.03** (0.01)
		FPR	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.12 (0.06)	0.02 (0.05)	0.00 (0.00)
		FNR	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
(III)	100	Error	0.11 (0.05)	**0.04** (0.02)	**0.04** (0.02)	0.08 (0.02)	0.06 (0.03)	**0.04** (0.02)
		FPR	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.19 (0.09)	0.06 (0.09)	0.00 (0.00)
		FNR	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	200	Error	0.11 (0.06)	**0.04** (0.02)	**0.04** (0.02)	0.09 (0.03)	0.09 (0.05)	**0.04** (0.02)
		FPR	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.12 (0.06)	0.08 (0.10)	0.00 (0.00)
		FNR	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)

Table 3: Performance of different estimators in the multiple index model simulation. Standard errors are in parentheses. The lowest estimation error in each setting is highlighted.

p	Sparsity	Metric	CHOMP	Adaptive CHOMP		Lasso SIR	Adaptive Lasso SIR	
				$\gamma = 1$	$\gamma = 2$		$\gamma = 1$	$\gamma = 2$
100	Same	Error	0.31 (0.25)	0.22 (0.27)	0.21 (0.28)	0.28 (0.27)	0.23 (0.23)	**0.19** (0.25)
		FPR	0.00 (0.01)	0.00 (0.02)	0.01 (0.02)	0.32 (0.11)	0.13 (0.10)	0.02 (0.06)
		FNR	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.01)
	Different	Error	0.38 (0.13)	0.24 (0.10)	**0.22** (0.09)	0.26 (0.06)	0.24 (0.09)	0.30 (0.25)
		FPR	0.00 (0.01)	0.00 (0.00)	0.00 (0.00)	0.39 (0.11)	0.20 (0.14)	0.11 (0.09)
		FNR	0.00 (0.02)	0.00 (0.01)	0.00 (0.01)	0.00 (0.00)	0.00 (0.00)	0.03 (0.10)
200	Same	Error	0.32 (0.26)	0.21 (0.28)	0.22 (0.29)	0.29 (0.26)	0.24 (0.21)	**0.19** (0.22)
		FPR	0.00 (0.00)	0.00 (0.01)	0.00 (0.02)	0.21 (0.08)	0.10 (0.09)	0.03 (0.06)
		FNR	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.01)	0.00 (0.01)	0.00 (0.01)
	Different	Error	0.40 (0.13)	0.24 (0.09)	**0.22** (0.09)	0.30 (0.08)	0.29 (0.12)	0.33 (0.21)
		FPR	0.00 (0.01)	0.00 (0.00)	0.00 (0.00)	0.26 (0.09)	0.17 (0.12)	0.12 (0.09)
		FNR	0.00 (0.01)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.01)	0.02 (0.07)

In the single index model simulation, Table 2 demonstrates that the adaptive CHOMP estimators have a better performance than the adaptive Lasso SIR estimator for model (I), but they have similar performances for models (II) and (III). In the multiple index model simulation, Table 3 demonstrates that the adaptive Lasso SIR estimators tend to perform slighly better than the adaptive CHOMP estimators when the true two dimensions have the same sparsity pattern, while the reverse holds when the true two dimensions have different sparsity patterns. In both simulations, compared to the Lasso SIR, the adaptive Lasso SIR estimators have considerably smaller false positive rate, thus reducing the estimation error remarkably. However, compared to the adaptive CHOMP, the adaptive Lasso SIR estimators still tends to overfit. We conjecture that the adaptive Lasso SIR estimator may also have an oracle property similar to the adaptive CHOMP estimator, which is a topic of future research. However, similar to the Lasso SIR estimator, it is not obvious how to extend the adaptive Lasso SIR to other inverse regression methods, while it is straightforward to do so for the adaptive CHOMP estimator.

S4 Matrix Lasso estimator with different choices of tuning parameters

In this section, we demonstrate the difficulty of selecting tuning parameters for the Matrix Lasso estimator by examining its numerical performance with common methods of tuning parameter selection in practice. We simulate the data from the single index and multiple index models as given in Section 5.1 and 5.2 of the main paper, except that for the single index model, we only consider the case when the covariates have an autoregressive covariance structure. Recall that the Matrix Lasso estimator for the jth dimension of the central subspace is defined to be

$$\hat{\boldsymbol{\beta}}_j^{\mathrm{ML}} = \arg\min_{\boldsymbol{\beta}_j} \frac{1}{2} \left\| \hat{\boldsymbol{\eta}}_j - \hat{\boldsymbol{\Sigma}} \boldsymbol{\beta}_j \right\|_2^2 + \mu_j \left\| \boldsymbol{\beta}_j \right\|_1, \ j = 1, \ldots, d. \tag{S4.13}$$

where $\mu_j > 0$ is a tuning parameter. Computationally, equation (S4.13) is similar to the optimization problem corresponding to the Lasso estimator in the regular linear model with the design matrix to be $\hat{\boldsymbol{\Sigma}}$ and response vector $\hat{\boldsymbol{\eta}}_j$. In R, we can solve the problem using the `glmnet` package (Friedman et al., 2010). We consider the following methods of tuning parameter selection:

- Cross-validation (CV): We use the `cv.glmnet` command with all its default options and select the tuning parameter that either minimizes

the cross-validation score (CV min) or is the largest value whose cross-validation score is within one standard error of the minimum (CV 1se-rule). Noting that this cross-validation treats $\hat{\mathbf{\Sigma}}$ and $\hat{\boldsymbol{\eta}}_j$ as if they contained independent rows.

- PIC. We select the tuning parameter to minimize our proposed PIC given by

$$\text{PIC} = \left(\left\| \mathcal{P}\left\{ \hat{\beta}_j\left(\mu_j\right) \right\} - \mathcal{P}\left\{ \bar{\beta}_j\left(\mu_j\right) \right\} \right\|_F^2 + \frac{\log p}{p} \left\| \hat{\beta}_j(\mu_j) \right\|_0 \right) 1_{\hat{\beta}_j(\mu_j)\neq 0} + \infty 1_{\hat{\beta}_j(\mu_j)=0}.$$

- Optimal selection: We select the tuning parameter that minimizes the estimation error of the associated projection matrix, i.e for each dimension $j = 1, \ldots, d$, we select μ_j that minimizes $\left\| \mathcal{P}\left\{ \hat{\beta}_j\left(\mu_j\right) \right\} - \mathcal{P}(\boldsymbol{\beta}) \right\|_F^2$. Note that this optimal selection cannot be done in practice, since it depends on the true parameter $\boldsymbol{\beta}$. We include this in the simulation as a benchmark to compare the other tuning parameter methods.

Tables 4 and 5 demonstrate that with optimal tuning parameter selection, the Matrix Lasso estimator typically has a lot of false positives. Furthermore, using PIC to select the tuning parameter typically leads to larger estimation error than using cross-validation. The CV (min) rule tends to work better than the CV (1se) rule in terms of estimation error, although in some settings (for example single index model (II) with $p = 200$), both

rules lead to similar performance. However, notably, except in single index models (II) and (III), the performance of CV methods is relatively far from the performance under optimal tuning parameter selection, suggesting tuning parameter selection methods that are often used in practice do not guarantee good performance for the Matrix Lasso.

Table 4: Performance of Matrix Lasso estimators with different choices of tuning parameters in the single index model simulation. Standard errors are included in parentheses.

Model	p	Metric	CV (min)	CV (1se)	PIC	Optimal
(I)	100	Error	0.39 (0.16)	0.60 (0.35)	0.70 (0.24)	0.26 (0.07)
		FPR	0.66 (0.23)	0.22 (0.21)	0.03 (0.04)	0.44 (0.16)
		FNR	0.02 (0.10)	0.18 (0.26)	0.19 (0.17)	0.00 (0.00)
	200	Error	0.49 (0.09)	0.40 (0.14)	0.72 (0.22)	0.33 (0.09)
		FPR	0.72 (0.16)	0.43 (0.15)	0.02 (0.03)	0.36 (0.14)
		FNR	0.00 (0.01)	0.01 (0.05)	0.19 (0.17)	0.00 (0.01)
(II)	100	Error	0.10 (0.08)	0.19 (0.24)	0.37 (0.19)	0.09 (0.04)
		FPR	0.40 (0.15)	0.23 (0.14)	0.03 (0.03)	0.35 (0.12)
		FNR	0.00 (0.03)	0.03 (0.11)	0.04 (0.09)	0.00 (0.00)
	200	Error	0.11 (0.02)	0.11 (0.05)	0.35 (0.17)	0.09 (0.03)
		FPR	0.50 (0.12)	0.34 (0.10)	0.02 (0.02)	0.33 (0.12)
		FNR	0.00 (0.00)	0.00 (0.01)	0.03 (0.07)	0.00 (0.00)
(III)	100	Error	0.13 (0.10)	0.25 (0.28)	0.45 (0.21)	0.11 (0.04)
		FPR	0.51 (0.16)	0.27 (0.16)	0.03 (0.03)	0.40 (0.13)
		FNR	0.00 (0.05)	0.04 (0.13)	0.06 (0.11)	0.00 (0.00)
	200	Error	0.17 (0.03)	0.15 (0.05)	0.42 (0.19)	0.13 (0.04)
		FPR	0.62 (0.13)	0.41 (0.11)	0.03 (0.03)	0.34 (0.13)
		FNR	0.00 (0.00)	0.00 (0.00)	0.04 (0.09)	0.00 (0.00)

Table 5: Performance of Matrix Lasso estimators with different choices of tuning parameters in the multiple index model simulation. Standard errors are included in parentheses.

p	Sparsity	Metric	CV (min)	CV (1se-rule)	PIC	Optimal
100	Same	Error	0.49 (0.32)	0.81 (0.41)	0.74 (0.29)	0.33 (0.26)
		FPR	0.63 (0.16)	0.24 (0.17)	0.05 (0.05)	0.55 (0.19)
		FNR	0.02 (0.13)	0.13 (0.28)	0.04 (0.08)	0.01 (0.08)
	Different	Error	0.48 (0.26)	1.01 (0.35)	0.91 (0.23)	0.33 (0.07)
		FPR	0.63 (0.16)	0.19 (0.16)	0.06 (0.06)	0.69 (0.13)
		FNR	0.03 (0.14)	0.21 (0.33)	0.10 (0.12)	0.00 (0.00)
200	Same	Error	0.48 (0.24)	0.51 (0.28)	0.73 (0.29)	0.37 (0.26)
		FPR	0.66 (0.08)	0.39 (0.10)	0.03 (0.04)	0.44 (0.18)
		FNR	0.00 (0.00)	0.00 (0.03)	0.04 (0.09)	0.00 (0.05)
	Different	Error	0.46 (0.10)	0.60 (0.23)	0.93 (0.21)	0.41 (0.09)
		FPR	0.66 (0.08)	0.39 (0.11)	0.05 (0.04)	0.59 (0.13)
		FNR	0.00 (0.01)	0.02 (0.05)	0.10 (0.11)	0.00 (0.00)

S5 CHOMP for SIR in high dimensional settings

S5.1 Method

In this section, we demonstrate how the CHOMP technique can be extended to sufficient dimension methods such as the SIR in high dimensional settings. In such scenario, one particular challenge of implementing the CHOMP and the adaptive CHOMP estimators is to find a good estimator for the Cholesky factor $\boldsymbol{L}$ of the population covariance matrix $\boldsymbol{\Sigma}$ and its inverse. While this is hard in general, we can estimate $\boldsymbol{L}$ efficiently when the population covariance matrix has some special structure.

In this section, we consider a regression setting where the covariates have a natural order (for example when they are collected over time) and the population covariance (and correlation) matrix are banded, i.e $\sigma_{jk} = 0$ if $|j-k| > K$ with K known. Such covariance structure has been considered extensively in the literature of high-dimensional covariance estimation, see for example Pourahmadi (2013) and Khare et al. (2019). In this case, let $\boldsymbol{\Sigma} = \boldsymbol{C}\boldsymbol{D}\boldsymbol{C}^\top$ be the modified Cholesky decompositon of $\boldsymbol{\Sigma}$ such that $\boldsymbol{D}$ is a diagonal matrix and $\boldsymbol{C} = (c_{jk})$ is a lower triangular matrix with $c_{jj} = 1$ and $c_{jk} = 0$ if $|j - k| > K$. As suggested by Rothman et al. (2010), the off-diagonal elements of $\boldsymbol{C}$ and the diagonal elements of $\boldsymbol{D}$ can

be estimated sequentially by fitting a sequence of linear regressions. Let $\boldsymbol{x}^{(j)}, j = 1, \ldots, p$ denote the jth column of the design matrix $\mathcal{X}$. For the first variable, set $\boldsymbol{e}_1 = \boldsymbol{x}^{(1)}$. For $j = 2, \ldots, p$, let $\boldsymbol{c}_j^{(k)} = (c_{j,j-k}, \ldots, c_{j,j-1})^\top$ and $\boldsymbol{Z}_j^{(k)} = (\boldsymbol{e}_{j-k}, \ldots, \boldsymbol{e}_{j-1})$, where the index $j - k$ is understood to mean $\max(1, j - k)$, then we compute sequentially

$$\hat{\boldsymbol{c}}_j^{(k)} = \arg\min_{\boldsymbol{c}_j^{(k)}} \left\| \boldsymbol{x}^{(j)} - \boldsymbol{Z}_j^{(k)} \boldsymbol{c}_j^{(k)} \right\|_2^2, \boldsymbol{e}_j = \boldsymbol{x}^{(j)} - \boldsymbol{Z}_j^{(k)} \hat{\boldsymbol{c}}_j^{(k)}.$$

Finally the diagonal elements of $\boldsymbol{D}$ are estimated as $\hat{d}_{jj} = n^{-1} \left\| \boldsymbol{e}_j \right\|_2^2$, and the Cholesky factor $\boldsymbol{L}$ is estimated by $\hat{\boldsymbol{L}} = \widehat{\boldsymbol{C}} \widehat{\boldsymbol{D}}^{1/2}$, where $\widehat{\boldsymbol{D}} = \mathrm{diag}(\hat{d}_{jj})$ and $\widehat{\boldsymbol{C}} = (\hat{c}_{jk}),\ j, k = 1, \ldots, p$.

Let $\hat{\boldsymbol{\kappa}}_j$ be calculated such as $\hat{\boldsymbol{L}} \hat{\boldsymbol{\kappa}}_j = \hat{\boldsymbol{\eta}}_j$, where $\boldsymbol{\eta}_j$ is calculated in the same way as outlined in Section 3.1 of the main paper. With the regression-based estimated Cholesky factor $\hat{\boldsymbol{L}}$, the CHOMP and adaptive CHOMP estimator for SIR are defined respectively as

$$\hat{\boldsymbol{\beta}}_j = \arg\min_{\boldsymbol{\beta}_j} \frac{1}{2} \left\| \hat{\boldsymbol{L}}^\top \boldsymbol{\beta}_j - \hat{\boldsymbol{\kappa}}_j \right\|_2^2 + \mu_j \left\| \boldsymbol{\beta}_j \right\|_1,\ j = 1, \ldots, d, \tag{S5.14}$$

and

$$\hat{\boldsymbol{\beta}}_j^* = \arg\min_{\boldsymbol{\beta}_j} \frac{1}{2} \left\| \hat{\boldsymbol{L}}^\top \boldsymbol{\beta}_j - \hat{\boldsymbol{\kappa}}_j \right\|_2^2 + \mu_j \sum_{k=1}^{p} \omega_{jk} |\beta_{jk}|,\ j = 1, \ldots, d., \tag{S5.15}$$

where $\omega_{jk} = |\bar{\beta}_{jk}|^{-\gamma}$, with $\bar{\beta}_{jk}$ being the kth component of an initial consistent estimate $\bar{\boldsymbol{\beta}}_j$ and γ a positive constant. Inhigh dimensional settings, the unpenalized sliced inverse regression estimator is not consistent (Lin

et al., 2018). Hence, for each dimension, we use the Lasso SIR estimator as the initial consistent estimator $\bar{\boldsymbol{\beta}}_j$ for computing the adaptive weight. Furthermore, to adjust for the convergence rate of the Lasso SIR estimator, we select the tuning parameters for the CHOMP and adaptive CHOMP from minimizing the following projection information criterion

$$\mathrm{PIC}(\mu_j;\tau_j) = \begin{cases} \left\| \mathcal{P}\left\{\hat{\boldsymbol{\beta}}(\mu_j)\right\} - \mathcal{P}(\bar{\boldsymbol{\beta}}_j) \right\|_F^2 + \frac{2}{p}\left\|\hat{\boldsymbol{\beta}}_j(\mu_j)\right\|_0, & \text{if } \hat{\boldsymbol{\beta}}(\mu_j) \neq 0 \\ \infty, & \text{if } \hat{\boldsymbol{\beta}}_j(\mu_j) = 0. \end{cases}$$

Estimation and selection consistency of the CHOMP-based estimators and of the PIC in high dimensional settings where the Cholesky factors are estimated based on regression will be topics of future research. Below, we will present a simulation study to demonstrate the empirical performance of this approach.

S5.2 Simulation

For the simulation, we generate data from the model (I) as in Section 5.1 of the main paper with the correlation matrix $\tilde{\boldsymbol{\Omega}}$ having off-diagonal elements $(\tilde{\omega})_{jk} = 1 - K^{-1}|j-k|$ if $|j-k| \leq K$ and 0 otherwise. We consider two values for K, namely $K \in \{3, 5\}$. The sample size is fixed at $n = 1000$ and the number of covariates varies over $p \in \{500, 1000, 1500\}$. We compute the CHOMP, adaptive CHOMP estimator with $\gamma = 1$ and $\gamma = 2$, and

the Lasso SIR estimators; then we compare them using the same metric as given in Section 5.1 of the main paper. Table 6 demonstrates that the adaptive CHOMP estimator with $\gamma = 2$ has the best performance in the considered settings. The Matrix Lasso and the CHOMP estimator tend to have approximately the same estimation error, which is usually higher than both the adaptive CHOMP and the Lasso SIR due to higher false negative rates. Compared to the Lasso SIR estimator, both the adaptive CHOMP estimators with $\gamma = 1$ and $\gamma = 2$ tend to reduce the false positive rates. However, the adaptive CHOMP with $\gamma = 1$ tends to underfit by having medium false negative rates (as seen in $p = 1500$). On the other hand, the adaptive CHOMP estimator with $\gamma = 2$ does not increase the false negative rate much and hence has the lowest estimation error.

S6 Additional Simulation Results

Table 6: Performance of the estimators in the simulation of single index model in high dimensional settings where the covariance of the covariates are banded. Standard errors are in parentheses. The lowest estimation error is highlighted for each setting.

p	K	Metric	CHOMP	Adaptive CHOMP		Lasso SIR	Mlasso
				$\gamma = 1$	$\gamma = 2$		
500	3	Error	0.89 (0.22)	0.45 (0.23)	**0.39** (0.17)	0.44 (0.15)	0.99 (0.22)
		FPR	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.10 (0.06)	0.01 (0.02)
		FNR	0.29 (0.18)	0.05 (0.12)	0.02 (0.07)	0.02 (0.05)	0.40 (0.23)
	5	Error	0.99 (0.21)	0.50 (0.28)	**0.43** (0.22)	0.49 (0.19)	1.06 (0.23)
		FPR	0.00 (0.00)	0.00 (0.01)	0.00 (0.01)	0.10 (0.06)	0.01 (0.01)
		FNR	0.42 (0.24)	0.12 (0.24)	0.08 (0.20)	0.03 (0.09)	0.51 (0.26)
1000	3	Error	1.06 (0.17)	0.53 (0.27)	**0.43** (0.19)	0.51 (0.17)	1.01 (0.19)
		FPR	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.07 (0.04)	0.01 (0.01)
		FNR	0.48 (0.21)	0.09 (0.17)	0.04 (0.09)	0.03 (0.08)	0.39 (0.22)
	5	Error	1.07 (0.18)	0.61 (0.31)	**0.52** (0.25)	0.57 (0.21)	1.06 (0.22)
		FPR	0.00 (0.00)	0.00 (0.00)	0.00 (0.01)	0.07 (0.04)	0.00 (0.01)
		FNR	0.53 (0.24)	0.18 (0.27)	0.09 (0.17)	0.05 (0.11)	0.52 (0.26)
1500	3	Error	1.05 (0.18)	0.77 (0.29)	**0.57** (0.25)	0.61 (0.24)	0.97 (0.25)
		FPR	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.04 (0.03)	0.00 (0.00)
		FNR	0.52 (0.24)	0.29 (0.27)	0.13 (0.18)	0.11 (0.17)	0.40 (0.24)
	5	Error	1.11 (0.16)	0.84 (0.30)	**0.62** (0.25)	0.64 (0.23)	1.03 (0.23)
		FPR	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.04 (0.03)	0.00 (0.00)
		FNR	0.59 (0.23)	0.36 (0.29)	0.15 (0.15)	0.13 (0.15)	0.48 (0.27)

Table 7: Performance of the estimators in the single index model simulation in Section 5.1 of the main paper with the correlation matrix $\tilde{\mathbf{\Omega}}$ having homogeneous structure. Standard errors are included in parentheses. The lowest estimation error in each setting is highlighted.

Model	p	Metric	CHOMP	Adaptive CHOMP		Lasso SIR	Mlasso
				$\gamma = 1$	$\gamma = 2$		
(I)	100	Error	0.24 (0.14)	**0.12** (0.07)	**0.12** (0.07)	0.23 (0.11)	0.43 (0.12)
		FPR	0.01 (0.01)	0.00 (0.00)	0.00 (0.00)	0.16 (0.08)	0.68 (0.30)
		FNR	0.01 (0.04)	0.00 (0.00)	0.00 (0.00)	0.00 (0.02)	0.02 (0.07)
	200	Error	0.25 (0.12)	**0.14** (0.08)	**0.14** (0.08)	0.26 (0.17)	0.57 (0.08)
		FPR	0.01 (0.01)	0.00 (0.00)	0.00 (0.01)	0.09 (0.05)	0.79 (0.21)
		FNR	0.01 (0.03)	0.00 (0.01)	0.00 (0.01)	0.01 (0.11)	0.00 (0.03)
(II)	100	Error	0.08 (0.06)	**0.03** (0.02)	**0.03** (0.01)	0.07 (0.03)	0.10 (0.09)
		FPR	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.16 (0.08)	0.34 (0.20)
		FNR	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.01 (0.03)
	200	Error	0.08 (0.06)	**0.03** (0.02)	**0.03** (0.02)	0.10 (0.13)	0.13 (0.11)
		FPR	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.10 (0.05)	0.39 (0.24)
		FNR	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.01 (0.07)	0.01 (0.04)
(III)	100	Error	0.10 (0.07)	**0.04** (0.02)	**0.04** (0.02)	0.09 (0.04)	0.14 (0.10)
		FPR	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.16 (0.08)	0.49 (0.22)
		FNR	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.01 (0.04)
	200	Error	0.11 (0.07)	**0.04** (0.02)	**0.04** (0.02)	0.11 (0.13)	0.20 (0.10)
		FPR	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.09 (0.05)	0.53 (0.24)
		FNR	0.00 (0.01)	0.00 (0.00)	0.00 (0.00)	0.00 (0.06)	0.01 (0.04)

Bibliography

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*, 1–22.

Khare, K., S.-Y. Oh, S. Rahman, and B. Rajaratnam (2019). A scalable sparse cholesky based approach for learning high-dimensional covariance matrices in ordered data. *Machine Learning 108*(12), 2061–2086.

Lin, Q., Z. Zhao, and J. S. Liu (2018). On consistency and sparsity for sliced inverse regression in high dimensions. *The Annals of Statistics 46*, 580–610.

Lin, Q., Z. Zhao, and J. S. Liu (2019). Sparse sliced inverse regression via lasso. *Journal of the American Statistical Association*, 1–33.

Pourahmadi, M. (2013). *High-dimensional Covariance Estimation*, Volume 882. John Wiley & Sons.

Rothman, A. J., E. Levina, and J. Zhu (2010). A new approach to cholesky-based covariance regularization in high dimensions. *Biometrika 97*(3), 539–550.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-asymptotic Viewpoint*. Cambridge University Press.