Model Selection With Covariance Matching Based Non-negative Lasso

Arash Owrang^{a,1,*}, Yoram Bresler^b, and Magnus Jansson^{a,1}

^a*Department of information science and engineering, School of EECS, KTH Royal Institute of Technology, Stockholm, Sweden*

^b*Department of ECE, University of Illinois, Urbana-Champaign, USA*

Abstract

We consider the problem of model selection for high-dimensional linear regressions in the context of support recovery with multiple measurement vectors available. Here, we assume that the regression coefficient vectors have a common support and the elements of the additive noise vector are potentially correlated. Accordingly, to estimate the support, we propose a non-negative Lasso estimator that is based on covariance matching techniques. We provide deterministic conditions under which the support estimate of our method is guaranteed to match the true support. Further, we use the extended Fisher information criterion to select the tuning parameter in our non-negative Lasso. We also prove that the extended Fisher information criterion can find the true support with probability one as the number of rows in the design matrix grows to infinity. The numerical simulations confirm that our support estimate is asymptotically consistent. Finally, the simulations also show that the proposed method is robust to high correlation between columns of the design matrix.

Keywords: Model selection, high-dimensional inference, covariance matching, non-negative Lasso, regularization, extended Bayesian information criterion, generalized least squares, sparse multiple measurement vector model

*Corresponding author

Email address: owrang@kth.se (Arash Owrang)

¹This work was supported in part by the Swedish Research Council under contract 2015-05484 and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 742648).

1. Introduction

Variable selection for high-dimensional linear regressions can be successfully accomplished by ℓ_1 minimization methods. Among these methods, Lasso has been popular thanks to its desirable characteristics [1]. In particular, Lasso can correctly estimate the underlying model if certain conditions are satisfied [2, 3, 4]; one of which concerns the signal to noise ratio (SNR). That is, to achieve ideal model selection with Lasso, the magnitudes of the coefficients of the true model are required to be significantly larger than the noise variance [3, 4]. However, this condition cannot be guaranteed in many real-life applications such as radio astronomy, magnetic resonance imaging and genome-wide association studies. Interestingly though, in such circumstances, it is often the case that the support of the contributing regression coefficients remains the same for a certain period of time. Intuitively, to circumvent the low SNR situation, one may acquire multiple measurement vectors and modify Lasso such that it can be applied on multiple linear regression equations [5, 6, 7]. Such modified estimators are generally capable of identifying larger models than what ordinary Lasso can; however, they are still vulnerable to strong noise.

As an alternative approach, some studies consider the temporal and spatial correlations of the acquired measurement vectors and propose relevant covariance-based estimators [8, 9]. Although, [8, 9] generally can deal with low SNR situations, they suffer from other drawbacks: [9] performs inadequately when the elements of the noise vector are not identically distributed (see e.g. Fig. 3), and [8] is tailored to uncorrelated noise. Further, the method in [8] tends to overestimate the order of the model; hence, one must post-process the estimated model to acquire a refined estimate of the true model.

In this paper, we study the task of model selection when there are multiple measurement vectors available. To devise a method that can cope with realistic scenarios, we allow the elements of the noise vector to be spatially correlated. Here, the noise is assumed to have an unknown covariance matrix, which is linearly parameterized in a known, low dimensional basis. We propose a non-

negative Lasso estimator that is inspired by covariance matching techniques. In the process of developing this estimator, some partial results have been previously published [10]. In this paper, we provide deterministic conditions under which the support estimate of our method is guaranteed to match the true support. Further, to tune the corresponding regularization parameter in the Lasso estimator, we use the extended Fisher information criterion (EFIC) [11]. We prove that minimizing EFIC yields the true support with probability one as the number of rows in the corresponding design matrix grows to infinity. We believe that these are the first proven results of this kind for this problem. The numerical simulations corroborate that our support estimator is asymptotically consistent. The simulations also show that the proposed method is robust to high correlation between columns of the design matrix.

2. Method

2.1. Data Model and Problem Formulation

Suppose that the m -dimensional complex measurement vector $\mathbf{y}(t) \in \mathbb{C}^m$ is described at time index $t \in \mathbb{N}_+$ as

$$\mathbf{y}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{\Sigma}^{1/2}\boldsymbol{\epsilon}(t).$$

Matrix $\mathbf{A} \in \mathbb{C}^{m \times N}$ is a known full row rank matrix, signal vector $\mathbf{x}(t) \in \mathbb{C}^N$ is a sparse vector with unknown but fixed support $\mathcal{S}$ of maximum cardinality $|\mathcal{S}| = K \ll m^2$. Denoting by $\mathbf{x}_{\mathcal{S}}(t)$ the vector with elements $\mathbf{x}_i(t)$ with $i \in \mathcal{S}$, and by $\mathcal{S}^c$ the relative complement of $\mathcal{S}$ in the set $[N] \triangleq \{1, 2, \dots, N\}$, we have that $\mathbf{x}_{\mathcal{S}^c}(t) = 0$ for all $t \in \mathbb{N}_+$. The $\mathbf{x}_{\mathcal{S}}(t)$'s are temporally i.i.d. zero mean complex circular Gaussian random vectors with fixed but unknown diagonal covariance matrix. The noise vectors, $\{\boldsymbol{\epsilon}(t)\}_{t \in \mathbb{N}_+} \in \mathbb{C}^m$, are temporally i.i.d. circular Gaussian random vectors with zero mean and $\mathbb{E}(\boldsymbol{\epsilon}(t)\boldsymbol{\epsilon}^H(t)) = \mathbf{I}$, where $\mathbb{E}(\cdot)$ denotes the expected value operator and the H superscript denotes the conjugate transpose. Moreover, $\boldsymbol{\epsilon}(t)$'s are considered to be uncorrelated with $\mathbf{x}_{\mathcal{S}}(t)$'s. With these assumptions, conditionally on $\mathbf{A}$, the measurements $\mathbf{y}(t)$'s

are independent and identically distributed (i.i.d.) complex circular-symmetric zero mean Gaussian random vectors. In this paper, although this is not a restriction, we are mostly interested in scenarios in which the dimensions of $\mathbf{A}$ obey $m^2 < N$. The nuisance positive definite matrix $\mathbf{\Sigma} \in \mathbb{C}^{m \times m}$ is unknown and $\mathbf{\Sigma}^{1/2}$ denotes the unique positive definite square root of $\mathbf{\Sigma}$. Further, it is assumed that $\mathbf{\Sigma}$ can be linearly parameterized by some unknown real parameter vector and the eigenvalues of $\mathbf{\Sigma}$, denoted by $\Lambda_i(\mathbf{\Sigma})$ for $i \in [m]$, obey $\Lambda_i(\mathbf{\Sigma}) = \mathcal{O}(1)$. The primary goal of this study is to recover the true unknown support $\mathcal{S}$ from L observations, $\{\mathbf{y}(t)\}_{t=1}^L$.

2.2. The Covariance Model

Let the unknown positive definite matrix $\mathbf{R} \in \mathbb{C}^{m \times m}$ denote the covariance of $\mathbf{y}(t)$. Then, observe that the statistical assumptions considered for $\mathbf{x}(t)$'s and $\mathbf{\epsilon}(t)$'s imply

$$\mathbf{R} = \mathbf{A}\mathbf{P}\mathbf{A}^H + \mathbf{\Sigma},$$

where $\mathbf{P} \in \mathbb{R}^{N \times N}$ is a diagonal matrix with $p_{i,i} = \mathbb{E}(x_i(t)\bar{x}_i(t))$ for $i \in \mathcal{S}$, $p_{i,i} = 0$ for $i \in \mathcal{S}^c$ and $\bar{x}_i(t)$ denotes the complex conjugate of $x_i(t)$. We assume that $\mathbf{R}$ is positive definite. A sufficient condition for this is that $\mathbf{\Sigma}$ is positive definite. This is not a restrictive assumption in practice, because typical measurements will contain some uncorrelated noise. Now, associate $\mathbf{R}$ with the vector $\text{vec}(\mathbf{R}) \in \mathbb{C}^{m^2}$ defined as

$$\mathbf{r} = \text{vec}(\mathbf{R}) \triangleq [\mathbf{r}_1^T \ \mathbf{r}_2^T \ \dots \ \mathbf{r}_m^T]^T$$

where the column vector $\mathbf{r}_i$ denotes the i -th column of $\mathbf{R}$. By vectorizing both sides of the preceding covariance matrix equation and using the properties of the vec operator, we get

$$\begin{aligned} \mathbf{r} &= (\bar{\mathbf{A}} \otimes \mathbf{A})\text{vec}(\mathbf{P}) + \text{vec}(\mathbf{\Sigma}) \\ &= (\bar{\mathbf{A}} \circ \mathbf{A})\mathbf{p} + \text{vec}(\mathbf{\Sigma}), \end{aligned} \tag{1}$$

where the elements of the matrix $\bar{\mathbf{A}}$ are the complex conjugates of those of $\mathbf{A}$, the vector $\mathbf{p} = \text{diag}(\mathbf{P})$ is a non-negative sparse vector with cardinality equal to

$|\mathcal{S}|$ in $\mathbb{R}_+^N$ that contains the diagonal elements of $\mathbf{P}$, $|\mathcal{S}|$ denotes the cardinality of the set $\mathcal{S}$, and $\otimes$ and $\circ$ denote the Kronecker and the Khatri-Rao product, respectively; specifically, the i -th column of the Khatri-Rao product is defined as $(\mathbf{A} \circ \mathbf{B})_i = \mathbf{a}_i \otimes \mathbf{b}_i$.

For the development that follows in the upcoming sections, it is crucial to ensure the identifiability of the covariance model from the knowledge of $\mathbf{r}$. In this regard, let us momentarily assume that $\mathcal{S}$ is given. Under this condition, one can see that (1) has m^2 equations but $m^2 + |\mathcal{S}|$ real-valued unknown parameters if $\mathbf{\Sigma}$ is considered to be an arbitrary positive definite matrix. Obviously, some structure must be imposed on $\mathbf{\Sigma}$ [12, 13]. Here, we assume that $\mathbf{\Sigma}$ can be linearly parameterized as $\mathbf{\Sigma} = \sum_{i=1}^{\kappa} h_i \mathbf{Q}^{(i)}$, where $\mathbf{Q}^{(i)}$'s are known Hermitian matrices, h_i 's are real-valued unknown parameters that are located in the upper triangular part of $\mathbf{\Sigma}$, and κ is assumed to satisfy $\kappa \leq m^2 - 2|\mathcal{S}|$. We assume that the $\mathbf{Q}^{(i)}$'s are linearly independent, so that the parametrization of $\mathbf{\Sigma}$ is non-redundant, and given $\mathbf{\Sigma}$ and the $\mathbf{Q}^{(i)}$'s, the h_i 's are uniquely determined. Knowing $\mathbf{Q}^{(i)}$ is equivalent to being aware of the structure of the spatial correlation of the additive noise. This is a realistic assumption in some applications [13]. Examples of well-known covariance matrix structures that admit such linear representation with relatively small κ include Toeplitz, Hankel, banded, block diagonal, low-rank, etc. Now we can rewrite (1) as

$$\mathbf{r} = (\bar{\mathbf{A}} \circ \mathbf{A})\mathbf{p} + \mathbf{Q}\mathbf{h}, \quad (2)$$

where the known matrix $\mathbf{Q} \triangleq \begin{bmatrix} \text{vec}(\mathbf{Q}^{(1)}) & \text{vec}(\mathbf{Q}^{(2)}) & \dots & \text{vec}(\mathbf{Q}^{(\kappa)}) \end{bmatrix} \in \mathbb{C}^{m^2 \times \kappa}$, and the real-valued unknown parameter vector $\mathbf{h} \triangleq \begin{bmatrix} h_1 & h_2 & \dots & h_{\kappa} \end{bmatrix}^T \in \mathbb{R}^{\kappa}$.

The parametrization of $\mathbf{\Sigma}$ in (2) has reduced the number of unknowns, but for unknown $\mathcal{S}$ the number of unknowns is $N + \kappa$, and recalling that $m^2 < N$, it is seen that (2) is an underdetermined system of equations. Thus, to guarantee the identifiability of $(\bar{\mathbf{A}} \circ \mathbf{A})\mathbf{p} + \mathbf{Q}\mathbf{h}$ from $\mathbf{r}$, no other vectorized covariance model of smaller or equal number of parameters should describe $\mathbf{r}$ equally well. That

is,

$$\begin{bmatrix} \bar{\mathbf{A}}_{\mathcal{S}} \circ \tilde{\mathbf{A}}_{\mathcal{S}} & \tilde{\mathbf{Q}} \end{bmatrix} \begin{bmatrix} \mathbf{W}_{\mathbf{A}_{\mathcal{S}}}^{-1} \mathbf{p}_{\mathcal{S}}^* \\ \mathbf{W}_{\mathbf{Q}}^{-1} \mathbf{h}^* \end{bmatrix} \neq \begin{bmatrix} \bar{\mathbf{A}}_{\mathcal{I}} \circ \tilde{\mathbf{A}}_{\mathcal{I}} & \tilde{\mathbf{Q}} \end{bmatrix} \begin{bmatrix} \mathbf{W}_{\mathbf{A}_{\mathcal{I}}}^{-1} \mathbf{p}_{\mathcal{I}} \\ \mathbf{W}_{\mathbf{Q}}^{-1} \mathbf{h} \end{bmatrix}$$

90 for $\{\mathbf{p}_{\mathcal{I}} > \mathbf{0} \mid \mathcal{I} \subset [N], |\mathcal{I}| \leq |\mathcal{S}|, \mathcal{I} \neq \mathcal{S}\}$, where the columns of matrices $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{Q}}$ are the normalized versions of those of $\mathbf{A}$ and $\mathbf{Q}$, respectively. Here, $\mathbf{W}_{\mathbf{A}_{\mathcal{S}}}$ is a diagonal matrix such that $(w_{\mathbf{A}_{\mathcal{S}}})_{i,i} = \|\mathbf{a}_i\|_2$ for $i \in \mathcal{S}$ with an analogous definition for $\mathbf{W}_{\mathbf{Q}}$, $\|\cdot\|_2$ denotes the Euclidean norm and $\cdot^*$ denotes the true value of the corresponding parameter vector. Consequently, inspired by [14] and [4],
 95 we impose the following condition to uniformly ensure the identifiability of the true covariance model.

Restricted eigenvalue property: There exists a positive constant $C_{\min} > 0$ such that

$$\min_{|\mathcal{I}| \leq 2K} \Lambda_{\min} \left(\begin{bmatrix} (\bar{\mathbf{A}}_{\mathcal{I}} \circ \tilde{\mathbf{A}}_{\mathcal{I}})^H \\ \tilde{\mathbf{Q}}^H \end{bmatrix} \begin{bmatrix} \bar{\mathbf{A}}_{\mathcal{I}} \circ \tilde{\mathbf{A}}_{\mathcal{I}} & \tilde{\mathbf{Q}} \end{bmatrix} \right) \geq C_{\min}, \quad (3)$$

where $\Lambda_{\min}(\cdot)$ denotes the smallest eigenvalue of the corresponding matrix.

100 The constant $C_{\min}$ is essentially an indicator of how incoherent the columns of $\begin{bmatrix} \bar{\mathbf{A}}_{\mathcal{S}} \circ \tilde{\mathbf{A}}_{\mathcal{S}} & \tilde{\mathbf{Q}} \end{bmatrix}$ are. Typical examples of $\mathbf{A}$ for which the restricted eigenvalue property (3) is satisfied are certain submatrices of the discrete Fourier transform matrix (corresponding to sparse linear sensor arrays [9]), and realizations of a sub-Gaussian or Bernoulli ensemble. Finally, we would like to
 105 highlight that the appearance of the scalar factor two in $\kappa \leq m^2 - 2|\mathcal{S}|$ is due to the requirement of every submatrix $\bar{\mathbf{A}}_{\mathcal{I}} \circ \mathbf{A}_{\mathcal{I}}$ to be full-rank for any $\mathcal{I}$ such that $|\mathcal{I}| \leq 2K$.

2.3. Covariance Matching Based Non-negative Lasso

As a first step towards developing our method, we determine an estimate of $\mathbf{h}$. For this purpose, construct the real-valued vector $\mathbf{f} \in \mathbb{R}^{m^2}$ from the real and imaginary parts of the entries in the upper triangular part of $\mathbf{R}$ and note that the relation between $\mathbf{f}$ and $\mathbf{r}$ can be described by

$$\mathbf{f} = \mathbf{Tr}.$$

Here, $\mathbf{T} \in \mathbb{C}^{m^2 \times m^2}$ is an invertible matrix with elements zero, one, $1/2$ and $\pm j/2$. This transformation allows us to exploit established statistical methods that are developed for real-valued scenarios. Now, estimate $\mathbf{f}$ as $\hat{\mathbf{f}} = \mathbf{T}\hat{\mathbf{r}}$, where the vector $\hat{\mathbf{r}}$ is defined as $\hat{\mathbf{r}} = \text{vec}(\hat{\mathbf{R}})$ and $\hat{\mathbf{R}} = 1/L \sum_{t=1}^L \mathbf{y}(t)\mathbf{y}^H(t)$ is the sample covariance matrix. We assume that $L \geq m$, so that by the assumption $\mathbf{R} > 0$, we have $\hat{\mathbf{R}} > 0$ w.p.1. Consequently, the difference between $\mathbf{f}$ and its estimate can be expressed as

$$\hat{\mathbf{f}} - \mathbf{f} = \mathbf{T}(\hat{\mathbf{r}} - (\bar{\mathbf{A}} \circ \mathbf{A})\mathbf{p} - \mathbf{Q}\mathbf{h}).$$

Observe that the covariance of the residual vector $\hat{\mathbf{f}} - \mathbf{f}$ obeys

$$\begin{aligned} \text{cov}(\hat{\mathbf{f}} - \mathbf{f}) &= \mathbf{T} \text{cov}(\hat{\mathbf{r}} - \mathbf{r}) \mathbf{T}^H \\ &\stackrel{E1}{=} \frac{1}{L} \mathbf{T}(\mathbf{R}^T \otimes \mathbf{R}) \mathbf{T}^H, \end{aligned}$$

where the equality in E1 follows [13] from the complex circular Gaussian model for $\mathbf{y}(t)$, using the Issrelis Gaussian moment theorem. It is clear that the elements of $\hat{\mathbf{f}} - \mathbf{f}$ are correlated. As a result, a viable estimate of $\mathbf{h}$ can be obtained by utilizing the generalized least squares; this involves $\text{cov}(\hat{\mathbf{f}} - \mathbf{f})$ [15, 16]. However, as can be seen, $\text{cov}(\hat{\mathbf{f}} - \mathbf{f})$ depends on the unknown $\mathbf{R}$. A standard approach is to replace $\mathbf{R}$ with its consistent estimate, see e.g. [16, Chapter 10.5.2.]. Consequently, the estimate of $\mathbf{h}$ can be found by minimizing the Mahalanobis distance between $\mathbf{T}\hat{\mathbf{r}}$ and $\mathbf{T}\mathbf{r}$ as

$$\eta(\mathbf{p}, \mathbf{h}) = \|\hat{\mathbf{\Gamma}}^{-1/2} \mathbf{T}(\hat{\mathbf{r}} - (\bar{\mathbf{A}} \circ \mathbf{A})\mathbf{p} - \mathbf{Q}\mathbf{h})\|_2^2$$

with respect to $\mathbf{h}$, where the matrix $\hat{\mathbf{\Gamma}} = \mathbf{T}(\hat{\mathbf{R}}^T \otimes \hat{\mathbf{R}})\mathbf{T}^H$ is the consistent estimate of $L\text{cov}(\hat{\mathbf{f}} - \mathbf{f})$. Because $\mathbf{T}$ is invertible, and $\hat{\mathbf{R}} > 0$ w.p.1, it follows that $\hat{\mathbf{\Gamma}} > 0$ w.p.1, and $\hat{\mathbf{\Gamma}}^{-1/2}$ is well-defined. Solving for the unique minimizer of the weighted squared objective $\eta(\mathbf{p}, \mathbf{h})$ with respect to $\mathbf{h}$ yields

$$\hat{\mathbf{h}} = (\hat{\mathbf{\Gamma}}^{-1/2} \mathbf{T}\mathbf{Q})^\dagger \hat{\mathbf{\Gamma}}^{-1/2} \mathbf{T}(\hat{\mathbf{r}} - (\bar{\mathbf{A}} \circ \mathbf{A})\mathbf{p}),$$

where $(\cdot)^\dagger$ is the least-squares left inverse of the corresponding matrix, i.e. $\mathbf{Q}^\dagger = (\mathbf{Q}^H \mathbf{Q})^{-1} \mathbf{Q}^H$. Next, by inserting $\hat{\mathbf{h}}$ back into $\eta(\mathbf{p}, \mathbf{h})$, we obtain the minimum

generalized least squares error parameterized by $\mathbf{p}$ as

$$\eta_{\min}(\mathbf{p}) = \underbrace{\|\mathbf{\Pi}_{\hat{\mathbf{T}}^{-1/2}\mathbf{TQ}}^\perp \hat{\mathbf{T}}^{-1/2} \mathbf{T}[\hat{\mathbf{r}} - (\bar{\mathbf{A}} \circ \mathbf{A})\mathbf{p}]\|_2^2}_{\mathbf{B}}, \quad (4)$$

where the matrix $\mathbf{\Pi}_{\hat{\mathbf{T}}^{-1/2}\mathbf{TQ}}^\perp = \mathbf{I} - \hat{\mathbf{T}}^{-1/2} \mathbf{TQ}(\hat{\mathbf{T}}^{-1/2} \mathbf{TQ})^\dagger$ is the orthogonal projection matrix onto the orthogonal complement of the range space of $(\hat{\mathbf{T}}^{-1/2} \mathbf{TQ})$. This projector can be expressed as $\mathbf{\Pi}_{\hat{\mathbf{T}}^{-1/2}\mathbf{TQ}}^\perp = \mathbf{\Omega}\mathbf{\Omega}^T$, where the columns of the matrix $\mathbf{\Omega} \in \mathbb{R}^{m^2 \times m^2 - \kappa}$ are an orthobasis for the range space of $\mathbf{\Pi}_{\hat{\mathbf{T}}^{-1/2}\mathbf{TQ}}^\perp$. Consequently, we can rewrite $\eta_{\min}(\mathbf{p})$ as

$$\eta_{\min}(\mathbf{p}) = \underbrace{\|\mathbf{\Omega}^T \hat{\mathbf{T}}^{-1/2} \mathbf{T}[\hat{\mathbf{r}} - (\bar{\mathbf{A}} \circ \mathbf{A})\mathbf{p}]\|_2^2}_{\boldsymbol{\alpha}},$$

Next, one should be mindful of the rank of $\bar{\mathbf{A}} \circ \mathbf{A}$. Indeed, depending on the
110 structure of $\mathbf{A}$ in the given application, $\bar{\mathbf{A}} \circ \mathbf{A}$ can be rank deficient. The rank deficiency of $\bar{\mathbf{A}} \circ \mathbf{A}$ implies that the *effective* number of row samples in $\mathbf{\Omega}^T \hat{\mathbf{T}}^{-1/2} \mathbf{T}\hat{\mathbf{r}}$ is equal to $l < m^2 - \kappa$. In such scenarios, it is computationally beneficial to represent $\boldsymbol{\alpha}$ in the corresponding l -dimensional space. Thus, we simply utilize the singular value decomposition $\mathbf{USV}^T = \mathbf{\Omega}^T \hat{\mathbf{T}}^{-1/2} \mathbf{T}(\bar{\mathbf{A}} \circ \mathbf{A})$
115 and reformulate $\eta_{\min}(\mathbf{p})$ as

$$\eta_{\min}(\mathbf{p}) = \|\underbrace{\mathbf{U}^T \mathbf{\Omega}^T \hat{\mathbf{T}}^{-1/2} \mathbf{T}\hat{\mathbf{r}}}_{\mathbf{z}} - \underbrace{\mathbf{SV}^T}_{\boldsymbol{\Phi}} \mathbf{p}\|_2^2, \quad (5)$$

where the matrices $\mathbf{U} \in \mathbb{R}^{(m^2 - \kappa) \times l}$ and $\mathbf{V} \in \mathbb{R}^{N \times l}$ have orthonormal columns and the elements of the diagonal matrix $\mathbf{S}$ are the singular values of $\mathbf{\Omega}^T \hat{\mathbf{T}}^{-1/2} \mathbf{T}(\bar{\mathbf{A}} \circ \mathbf{A})$ arranged in decreasing order. Now, let $\mathbf{z} - \boldsymbol{\Phi}\mathbf{p}$ be associated with the residual vector $\mathbf{e} \in \mathbb{R}^l$ as

$$\mathbf{z} = \boldsymbol{\Phi}\mathbf{p} + \mathbf{e}. \quad (6)$$

Recall that $\mathbf{p}^*$ is a sparse vector with $|\text{supp}(\mathbf{p}^*)| \leq K \ll m^2$; thus, the data description in (6) can be considered as an ideal setup to utilize ℓ_1 -minimization for estimating the support of $\mathbf{p}^*$. Hence, following [17] we construct the following non-negative (NN) Lasso, hereinafter referred to as covariance matching based

Lasso (COM-Lasso), to estimate $\mathcal{S}$

$$\begin{aligned}\hat{\mathbf{p}}(\lambda) &= \arg \min_{\tilde{\mathbf{p}} \geq 0} \left\{ \min_{\mathbf{h} \in \mathbb{R}^\kappa} \frac{1}{2} \eta(\mathbf{W}_{\Phi}^{-1} \tilde{\mathbf{p}}, \mathbf{h}) \right\} + \lambda \mathbf{1}^T \tilde{\mathbf{p}} \\ &= \arg \min_{\tilde{\mathbf{p}} \geq 0} \frac{1}{2} \|\mathbf{z} - \tilde{\Phi} \tilde{\mathbf{p}}\|_2^2 + \lambda \mathbf{1}^T \tilde{\mathbf{p}},\end{aligned}\tag{7}$$

where the matrix $\mathbf{W}_{\Phi}$ is a diagonal matrix with the diagonal elements $(w_{\Phi})_{i,i} = \|\phi_i\|_2$ for $i \in [N]$, the vector $\mathbf{1}$ represents $\mathbf{1} = [1, 1, \dots, 1]^T$, columns of the matrix $\tilde{\Phi}$ are related to those of Φ as $\tilde{\Phi} = \Phi \mathbf{W}_{\Phi}^{-1}$, vector $\tilde{\mathbf{p}} = \mathbf{W}_{\Phi} \mathbf{p}$, and the unknown tuning parameter $\lambda \geq 0$ controls the cardinality of $\hat{\mathcal{I}}(\lambda) = \text{supp}(\hat{\mathbf{p}}(\lambda))$. Normalizing the columns of the design matrix such that they have unit-norm lengths is a standard procedure in constructing the Lasso estimator. By doing so, it is more likely that any sparse covariate with relatively small cardinality be equally identifiable by Lasso.

We complete the argument by addressing the asymptotic distribution of $\mathbf{e}$. In this regard, observe that $\hat{\Gamma}$ converges in probability to Γ and $\sqrt{L}(\hat{\mathbf{f}} - \mathbf{f})$ converges in distribution to $\mathcal{N}(\mathbf{0}, \Gamma)$ as $L \rightarrow \infty$. Consequently, by Slutsky's Theorem, one can say that asymptotically $\sqrt{L} \hat{\Gamma}^{-1/2}(\hat{\mathbf{f}} - \mathbf{f}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ as $L \rightarrow \infty$. Finally, because $\mathbf{e} = \mathbf{U}^T \Omega^T \hat{\Gamma}^{-1/2}(\hat{\mathbf{f}} - \mathbf{f})$, and that Ω and $\mathbf{U}$ have orthonormal columns, we can conclude that $\sqrt{L} \mathbf{e} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ as $L \rightarrow \infty$.

2.4. Exact Support Recovery When L is Finite

Assuming that the covariance model (2) is identifiable, the important question is whether COM-Lasso can ensure the exact support recovery. Here, we present deterministic conditions under which the support of (7) satisfies $\hat{\mathcal{I}} = \mathcal{S}$.

Theorem 1. *Suppose that we are given $L \geq m$ corrupted measurements $\{\mathbf{y}(t)\}_{t=1}^L$. If there exists $\lambda \geq 0$ such that*

$$\begin{aligned}\lambda &> \max_{i \in \mathcal{S}^c} \tilde{\phi}_i^T (\Pi_{\mathcal{S}}^{\perp} \mathbf{e} + \lambda \tilde{\Phi}_{\mathcal{S}}^{\dagger} \mathbf{1}) \text{ and} \\ \min_{i \in \mathcal{S}} \tilde{p}_i^* &> \|\tilde{\Phi}_{\mathcal{S}}^{\dagger} \mathbf{e} - \lambda (\tilde{\Phi}_{\mathcal{S}}^T \tilde{\Phi}_{\mathcal{S}})^{-1} \mathbf{1}\|_{\infty},\end{aligned}$$

then the estimate of (7) using that λ will satisfies $\text{supp}(\hat{\mathbf{p}}) = \mathcal{S}$.

Proof. The vector $\hat{\mathbf{p}}$ is the solution of (7) if it satisfies the optimality conditions

$$-\tilde{\boldsymbol{\phi}}_i^T(\mathbf{z} - \tilde{\boldsymbol{\Phi}}\hat{\mathbf{p}}) + \lambda - \xi_i = 0, \quad (8)$$

$$\hat{p}_i \xi_i = 0$$

for $i \in [N]$, as well as $\boldsymbol{\xi} \geq \mathbf{0}$ and $\hat{\mathbf{p}} \geq \mathbf{0}$. Here, the vector $\boldsymbol{\xi} \in \mathbb{R}_+^N$ is the
140 Lagrangian multiplier associated with the non-negativity constraint in (7). Observe now, that since $\hat{p}_i \xi_i = 0$ for $i \in [N]$, the choice of λ as

$$\lambda > \max_{i \in \mathcal{S}^c} \tilde{\boldsymbol{\phi}}_i^T(\mathbf{z} - \tilde{\boldsymbol{\Phi}}\hat{\mathbf{p}}) \quad (9)$$

ensures $\hat{\mathbf{p}}_{\mathcal{S}^c} = \mathbf{0}$. As a result, we can say $\hat{\mathbf{p}} = \tilde{\mathbf{p}}^* + \boldsymbol{\nu}$ for some vector $\boldsymbol{\nu} \in \mathbb{R}^N$ such that $\boldsymbol{\nu}_{\mathcal{S}^c} = \mathbf{0}$. Clearly, if $\tilde{\mathbf{p}}^*$ obeys

$$\hat{\mathbf{p}}_{\mathcal{S}} = \tilde{\mathbf{p}}_{\mathcal{S}}^* + \boldsymbol{\nu}_{\mathcal{S}} > \mathbf{0} \quad (10)$$

then $\text{supp}(\hat{\mathbf{p}}) = \mathcal{S}$. This implies that $\boldsymbol{\xi}_{\mathcal{S}} = \mathbf{0}$; hence, the optimality condition (8) can be rewritten as

$$\tilde{\boldsymbol{\Phi}}_{\mathcal{S}}^T(\mathbf{z} - \tilde{\boldsymbol{\Phi}}_{\mathcal{S}}(\tilde{\mathbf{p}}_{\mathcal{S}}^* + \boldsymbol{\nu}_{\mathcal{S}})) = \lambda \mathbf{1}$$

for $i \in \mathcal{S}$. Next, by inserting $\mathbf{z} = \tilde{\boldsymbol{\Phi}}_{\mathcal{S}}\tilde{\mathbf{p}}_{\mathcal{S}}^* + \mathbf{e}$ into the preceding identity and then solving it for $\boldsymbol{\nu}_{\mathcal{S}}$, we obtain

$$\boldsymbol{\nu}_{\mathcal{S}} = \tilde{\boldsymbol{\Phi}}_{\mathcal{S}}^\dagger \mathbf{e} - \lambda(\tilde{\boldsymbol{\Phi}}_{\mathcal{S}}^T \tilde{\boldsymbol{\Phi}}_{\mathcal{S}})^{-1} \mathbf{1}.$$

Now insert $\boldsymbol{\nu}_{\mathcal{S}}$ into (10) to obtain the second condition in the statement of
145 Theorem 1. Finally, we complete the argument by evaluating $\hat{\mathbf{p}}_{\mathcal{S}}$ from $\tilde{\boldsymbol{\Phi}}_{\mathcal{S}}^T(\mathbf{z} - \tilde{\boldsymbol{\Phi}}_{\mathcal{S}}\hat{\mathbf{p}}_{\mathcal{S}}) = \lambda \mathbf{1}$ and then inserting it into (9). $\square$

Regarding Theorem 1, two comments are in order. First, for any λ that satisfies the conditions in Theorem 1, the NN-Lasso gives the true support. Since λ is a continuous parameter, this implies that Theorem 1 is in fact satisfied for an interval of λ rather than a single value. Second, the condition $L \geq m$ guarantees, as discussed earlier, that the inverse $\hat{\mathbf{\Gamma}}^{-1/2}$ exists, and therefore that $\tilde{\boldsymbol{\Phi}}$ is well-defined, w.p.1. In addition, as can be seen, whether or not the

conditions of Theorem 1 are satisfied depends on the behavior of $\tilde{\Phi}$ and $\mathbf{e}$. However, it follows from the first condition of Theorem 1 for $\mathbf{e} = \mathbf{0}$ that if $\mathbf{1}^T \tilde{\Phi}_{\mathcal{S}}^\dagger \tilde{\phi}_i < 1$ fails for some $i \notin \mathcal{S}$, then there is little hope for (7) to succeed in recovering $\mathcal{S}$. Motivated by this argument, we now impose the non-negative version of the irrepresentable condition (NIR) on $\tilde{\Phi}$ [2, 18].

NIR condition: The matrix $\tilde{\Phi}$ satisfies the NIR condition if there exists some $\zeta \in (0, 1]$ such that

$$\max_{i \in \mathcal{S}^c} \mathbf{1}^T \tilde{\Phi}_{\mathcal{S}}^\dagger \tilde{\phi}_i \leq 1 - \zeta.$$

It has been shown [3, 18] that, by imposing the NIR condition for $\mathcal{S} \subset [N]$ such that $|\mathcal{S}| \leq K$, program (7) can recover $\mathcal{S}$ uniformly with very high probability, if $\min_{i \in \mathcal{S}} \tilde{p}_i^*$ is large enough and λ is properly chosen. Of course, to have the
150 NIR condition uniformly satisfied, $\tilde{\Phi}_{\mathcal{S}}$ must also be uniformly injective over $\mathcal{S}$. Lemma 1 in the Appendix guarantees the injectivity of $\tilde{\Phi}_{\mathcal{S}}$ under the restricted eigenvalue property.

3. Choice of λ

Similar to Lasso, the support of the solution of NN-Lasso evolves as the value of λ changes [19, 20, 18]. In details, by considering the proof of Theorem 1, one can see that the support of $\hat{\mathbf{p}}(\lambda)$ satisfies $\hat{\mathcal{I}}(\lambda) = \text{supp}(\hat{\mathbf{p}}(\lambda)) = \emptyset$ for the choice of λ as $\lambda \geq \lambda_1 = \max_{i \in [N]} \tilde{\phi}_i^T \mathbf{z}$. Then, as λ gradually decreases from λ_1 , $\hat{\mathcal{I}}(\lambda)$ evolves at some pivotal λ_i 's, producing the solution set of NN-Lasso as $\{\hat{\mathcal{I}}(\lambda_i)\}_{i=1}^K$. However, assuming that $\mathcal{S} \in \{\hat{\mathcal{I}}(\lambda_i)\}_{i=1}^K$, we need to choose λ carefully in solving (7) to identify $\mathcal{S}$ correctly. Several studies have considered the choice of λ in the Lasso estimator [3, 4]. Adapting their result to our case would result in the choice of λ as

$$\lambda(\tau) = \frac{1}{\sqrt{L}}(1 + \tau)\sqrt{2 \ln N}$$

for some constant $\tau > 0$. Unfortunately, the preceding choice of λ introduces the new free-parameter τ that again must be tuned. Instead, notice that estimating the true unknown support $\mathcal{S}$ from the set $\{\hat{\mathcal{I}}(\lambda_i)\}_{i=1}^K$ can in fact be considered

a multiple hypotheses testing problem, where hypothesis $\mathcal{H}_{\mathcal{I}}$ is described as

$$\mathcal{H}_{\mathcal{I}} : \mathbf{z} = \Phi_{\mathcal{I}} \mathbf{p}_{\mathcal{I}} + \mathbf{e}, \quad \mathcal{I} \in \{\hat{\mathcal{I}}(\lambda_i)\}_{i=1}^K.$$

A standard approach is to find $\hat{\mathcal{S}}$ by minimizing a model selection criterion [21, 22, 23]. However, since the dimension of the parameter space N is much larger than l , the classical model selection methods, like the Bayesian information criterion (BIC), perform poorly and tend to overestimate the size of the support [24, 25]. Specifically, BIC is consistent if $N = O(l^{1/2})$ [25, 26].

The deficiency of BIC in our setting motivates us to consider the extended Fisher information criterion (EFIC), which is shown to be consistent in selecting the true support as the dimensions of the problem, both the effective number l of measurements in the covariance matching model and the dimension N of the sparse vector p of signal powers, grow to infinity [11]. Letting both l and N grow to infinity is of particular interest from both the mathematical and the application point of view, e.g. for high-resolution imaging. In this regard, in what follows we allow N to grow as fast as $N = l^d$ for some constant $d > 1$. Now, by recalling that $\mathbf{e}$ is approximately Gaussian distributed with $\mathbf{e} \stackrel{\text{a}}{\sim} \mathcal{N}(\mathbf{0}, \frac{1}{L} \mathbf{I})$, EFIC can be formulated as

$$-2 \ln p(\mathbf{z}; \hat{\mathbf{p}}_{\mathcal{I}}^{\text{ML}} | \mathcal{H}_{\mathcal{I}}) + \ln \det \mathbf{F}(\hat{\mathbf{p}}_{\mathcal{I}}^{\text{ML}}) + 2c|\mathcal{I}| \ln N, \quad (11)$$

where $p(\mathbf{z}; \mathbf{p}_{\mathcal{I}} | \mathcal{H}_{\mathcal{I}})$ represents the probability density function of $\mathbf{z}$ parameterized by $\mathbf{p}_{\mathcal{I}}$, the vector $\hat{\mathbf{p}}_{\mathcal{I}}^{\text{ML}}$ denotes the maximum likelihood (ML) estimate of $\mathbf{p}_{\mathcal{I}}$ and the Fisher information matrix $\mathbf{F}(\mathbf{p}_{\mathcal{I}})$ is defined by

$$\mathbf{F}(\mathbf{p}_{\mathcal{I}}) \triangleq -\mathbb{E} \left(\frac{\partial^2 \ln p(\mathbf{z}; \mathbf{p}_{\mathcal{I}} | \mathcal{H}_{\mathcal{I}})}{\partial \mathbf{p}_{\mathcal{I}} \partial \mathbf{p}_{\mathcal{I}}^T} \right),$$

where the expectation is taken with respect to $p(\mathbf{z}; \mathbf{p}_{\mathcal{I}} | \mathcal{H}_{\mathcal{I}})$. In (11), $c > 0$ is a constant that controls the penalty level in EFIC and its appropriate choice is discussed in Theorem 2.

To evaluate EFIC, first note that

$$\ln p(\mathbf{z}; \mathbf{p}_{\mathcal{I}} | \mathcal{H}_{\mathcal{I}}) = -\frac{l}{2} \ln \left(\frac{2\pi}{L} \right) - \frac{L}{2} \|\mathbf{z} - \Phi_{\mathcal{I}} \mathbf{p}_{\mathcal{I}}\|_2^2.$$

Maximizing the log-likelihood function with respect to $\mathbf{p}_{\mathcal{I}}$, we obtain the ML estimate of $\mathbf{p}_{\mathcal{I}}$ as

$$\hat{\mathbf{p}}_{\mathcal{I}}^{\text{ML}} = \Phi_{\mathcal{I}}^{\dagger} \mathbf{z}.$$

Inserting $\hat{\mathbf{p}}_{\mathcal{I}}^{\text{ML}}$ into the log-likelihood function yields

$$-2 \ln p(\mathbf{z}; \hat{\mathbf{p}}_{\mathcal{I}}^{\text{ML}} | \mathcal{H}_{\mathcal{I}}) = l \ln(2\pi/L) + L \|\Pi_{\Phi_{\mathcal{I}}}^{\perp} \mathbf{z}\|_2^2. \quad (12)$$

A straightforward calculation results in $\mathbf{F}(\mathbf{p}_{\mathcal{I}}) = L \Phi_{\mathcal{I}}^T \Phi_{\mathcal{I}}$. Next, using the properties of the determinant function,

$$\det \mathbf{F}(\hat{\mathbf{p}}_{\mathcal{I}}^{\text{ML}}) = L^{|\mathcal{I}|} \det(\Phi_{\mathcal{I}}^T \Phi_{\mathcal{I}}). \quad (13)$$

175 Substituting (13) and (12) into (11) and ignoring the constant terms we get

$$\hat{\mathcal{I}}^{\text{EFIC}} = \arg \min_{\mathcal{I} \in \mathcal{J}} g(\mathcal{I}), \quad (14)$$

where

$$g(\mathcal{I}) = L \|\Pi_{\Phi_{\mathcal{I}}}^{\perp} \mathbf{z}\|_2^2 + \ln \det(\Phi_{\mathcal{I}}^T \Phi_{\mathcal{I}}) + |\mathcal{I}| \ln L + 2c|\mathcal{I}| \ln N$$

and the set $\mathcal{J} = \bigcup_{k=1}^K \{\mathcal{I} \mid |\mathcal{I}| = k\}$ is the union of all combinatorial subsets up to cardinality $K \ll l$.

The important question is whether (14) is consistent in selecting the true support as $l \rightarrow \infty$. Moreover, what is the appropriate choice of c ? In Theorem 180 2 we show that minimizing $g(\mathcal{I})$ over $\mathcal{J}$ obeys $\hat{\mathcal{I}}^{\text{EFIC}} = \mathcal{S}$ with probability one as $l \rightarrow \infty$. Also, we provide a guideline for the choice of c .

Theorem 2. Suppose that $N = l^d$ for some constant $d > 1$ and the covariance data model (2) satisfies the necessary identifiability conditions. Specifically, $[\bar{\mathbf{A}} \circ \mathbf{A} \quad \mathbf{Q}]$ satisfies the restricted eigenvalue property of order $2K$ for 185 some constant $K \geq |\mathcal{S}|$. Additionally, columns of $\mathbf{A}$ obey

$$\min_{i \neq j} \frac{\|\mathbf{a}_i\|_2^2}{\|\mathbf{a}_j\|_2^2} = \Omega(m^{-r}) \quad (15)$$

for some constant $r \geq 0$. Then, the EFIC estimate with

$$c > 1 + \frac{1}{d} - \frac{\ln L}{2d \ln l} + \frac{r \ln m}{d \ln l} \quad (16)$$

satisfies $\hat{\mathcal{I}} = \mathcal{S}$ with probability one as l and L both grow to infinity.

Regarding the preceding theorem some comments are in order. First, recall that $l \leq m^2$, and therefore m grows with l . Second, condition (15), which
190 bounds the disparity in norm between different columns of $\mathbf{A}$, still allows these norms to have substantially different rate of growth with respect to m , showing that EFIC is robust to disparity of these norms. (Note though that the assumed restricted eigenvalue condition (3) can be shown to require that the norm ratio be bounded below by a constant). As can be seen, the larger the exponent r the
195 larger c is; note that the large value of c counterbalances the negative effect of large r on (20). Third, (16) can be replaced by $c > 1 + \frac{\gamma}{d} - \frac{\ln L}{2d \ln l} + \frac{r \ln m}{d \ln l}$, where $\gamma > 0$ influences $\Pr(\hat{\mathcal{I}} = \mathcal{S})$. For further discussion on the choice of γ , we refer to [11]. Finally, EFIC (with exhaustive search over the candidate supports) is theoretically more robust to the correlation between ϕ_i 's than Lasso. To see this,
200 note that although the EFIC criterion $g(\mathcal{I})$ does depend on Φ , per Theorem 2, to succeed, EFIC as it only requires restricted eigenvalue property, that is, that the smallest singular value of $\begin{bmatrix} \bar{\mathbf{A}}_{\mathcal{I}} \circ \mathbf{A}_{\mathcal{I}} & \mathbf{Q} \end{bmatrix}$ for $\{\mathcal{I} \subset [N] \mid |\mathcal{I}| \leq 2K\}$ be larger than some fixed (but arbitrarily small) positive constant. This restricted eigenvalue property condition can be satisfied even when the columns of $\begin{bmatrix} \bar{\mathbf{A}} \circ \mathbf{A} & \mathbf{Q} \end{bmatrix}$ have
205 a fair amount of correlation. Because Φ is defined in terms of $\begin{bmatrix} \bar{\mathbf{A}} \circ \mathbf{A} & \mathbf{Q} \end{bmatrix}$, this implies that EFIC can tolerate a fair amount of correlation between the ϕ_i 's. In contrast, as seen in Theorem 1, the conditions for perfect support recovery with Lasso depend on the correlation between the ϕ_i 's, which is made more explicit in the NIR condition. Therefore, Lasso demands ϕ_i 's to be highly uncorrelated.

210 Next, we consider the application of EFIC to our problem. We obtain the solution set of (7) as $\{\hat{\mathcal{I}}(\lambda_i)\}_{i=1}^K$ by using the modified-LARS algorithm [19, 20] and then we apply EFIC on this solution set to find which of the support estimates describes $\mathbf{z}$ the best; see Alg. 1.

Algorithm 1 Estimating the support of $\mathbf{p}^*$ by applying EFIC on the solution set of (7).

```

1: for  $j = 1$  to  $K$  do
2:    $\{\hat{\mathcal{I}}(\lambda_j)\} \leftarrow$  execute modified-LARS at step  $j$ 
3:   evaluate  $g(\hat{\mathcal{I}}(\lambda_j))$ 
4: end for
5:  $\hat{\mathcal{I}}^{\text{EFIC}} = \arg \min_{\mathcal{I} \in \{\hat{\mathcal{I}}(\lambda_j)\}_{j=1}^K} g(\mathcal{I})$ 

```

The computational cost of Alg. 1 is determined by those of $g(\mathcal{I})$ and modified-LARS. The computational cost associated with modified-LARS is $O(l^2)$ [19] and a straightforward calculation shows that $g(\mathcal{I})$'s cost is the same. In conclusion, the computational complexity of Alg. 1 is $O(l^2)$. Recall that $l \leq m^2 - \kappa$ is the effective number of covariance matching equations determined by $\text{rank}(\bar{\mathbf{A}} \circ \mathbf{A})$, which may be smaller than implied by the above inequality. In any event, because $K \ll m^2$ the computational cost of Algorithm 1 is at most $O(m^4)$.

Note that in the application of EFIC in Alg. 1 to select between the candidate models generated by modified-LARS, there may exist some preferred growth rate of L with respect to m , which we leave for future research. In our numerical simulations, we have set $L = 4m \ln(m)$. When $L = O(m \ln(m))$, the Frobenius norm of $\hat{\mathbf{R}} - \mathbf{R}$ does not converge to zero (it may converges to zero if $L = \Omega(m \ln(m))$), but the eigenvalues of $\hat{\mathbf{R}}$ stay relatively close to those of $\mathbf{R}$ [27]. As the numerical experiments demonstrate, this mild growth of L with m is sufficient to provide good performance of Alg. 1.

4. Empirical Results

In this section, we provide some numerical results to evaluate the performance of COM-Lasso in recovering the true support. In this regard, we calculate the empirical probability of exact support recovery over 400 Monte Carlo trials. Moreover, to have a better understanding of the effectiveness of COM-

235 Lasso, we compare the performance of COM-Lasso with two existing methods.
 One of which is the SPICE method which is an iterative method that has been
 introduced in [8] and the other one is the complex version of what has been
 proposed in [9] as

$$\min_{\tilde{\mathbf{p}} \geq \mathbf{0}} \frac{1}{2} \|\mathbf{T}[\hat{\mathbf{r}} - (\tilde{\mathbf{A}} \circ \tilde{\mathbf{A}})\tilde{\mathbf{p}}]\|_2^2 + \lambda \|\tilde{\mathbf{p}}\|_1. \quad (17)$$

240 In our experience, the estimate of SPICE is far from being sparse when the
 number of iterations is not large; thus, to find the support estimate of SPICE,
 the experimenter should carefully apply a thresholding policy on the estimate
 of SPICE to obtain an estimate of the support. Here, we have *experimentally*
 set the threshold to a certain value for which the support of SPICE's estimate
 coincides with $\mathcal{S}$ for the majority of trials.

We now describe the setting for our numerical simulations. In each Monte
 Carlo trial, the true support $\mathcal{S}$ is chosen randomly from $\binom{N}{|\mathcal{S}|}$ possible combina-
 tions where the cardinality of $|\mathcal{S}|$ is fixed to 20. The restricted signal vectors,
 $\mathbf{x}_{\mathcal{S}}(t)$'s, are i.i.d. having complex Gaussian distribution with $\mathbf{x}_{\mathcal{S}}(t) \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$.
 The i.i.d. noise vectors $\boldsymbol{\epsilon}(t)$'s are drawn from a standard complex Gaussian
 distribution, independently from $\mathbf{x}_{\mathcal{S}}(t)$'s, with $\mathcal{CN}(\mathbf{0}, \mathbf{I})$. Further, the rows of
 the matrix $\mathbf{A}$ are chosen in an i.i.d. manner from a multivariate complex Gaus-
 sian distribution such that $\mathbf{a}_{*i} \sim \mathcal{CN}(\mathbf{0}, \mathbf{C})$, where the matrix $\mathbf{C} \in \mathbb{R}^{N \times N}$ is
 characterized as

$$\mathbf{C} = \begin{pmatrix} 1 & \mu & \mu & \dots & \mu \\ \mu & 1 & \mu & \mu & \dots & \mu \\ \mu & \mu & 1 & \mu & \dots & \mu \\ \vdots & \vdots & & \dots & \vdots \\ \mu & \mu & \dots & & 1 \end{pmatrix}.$$

245 Here, the constant $\mu \in [0, 1)$ controls the degree of correlation between columns
 of $\mathbf{A}$. Moreover, the dimension of the signal space is linked to the number
 of row measurements as well as the degrees of freedom in the noise space as
 $N = \lceil (m^2 - \kappa)^d \rceil$, where $\lceil \cdot \rceil$ is the ceiling function and $d = 1.2$. Note that
 columns of a Gaussian random $\mathbf{A}$ have equal weights with high probability;

Figure 1: The empirical probability of $\{\hat{\mathcal{T}} = \mathcal{S}\}$ versus m when $\mathbf{A}$ has an uncorrelated structure, i.e. $\mu = 0$, and $\Sigma = \sigma^2 \mathbf{I}$. Here, $|\mathcal{S}| = 20$, $\sigma^2 = 10$, $N = \lceil (m^2 - 1)^d \rceil$ for $d = 1.2$ and $L = 4m \ln m$.

hence, we set $r = 0$. Finally, we set $L = 4m \ln m$ and as recommended in [11], we let $\gamma = 2$ which results in $c = 1 + 2/d - \ln(4m \ln m)/(2d \ln l)$. In all figures, COM-Lasso-oracle and (17) show the performance of the corresponding methods when the cardinality of $\mathcal{S}$ is known. Also, COM-Lasso- λ illustrates the performance of COM-Lasso when $\lambda = \frac{1+\tau}{\sqrt{L}} \sqrt{2 \ln N}$. Remark that we choose τ as $\tau = 3$ in these numerical simulations to tune COM-Lasso- λ such that it performs at its best.

Figs. 1 and 2 plot the empirical probability of $\hat{\mathcal{T}} = \mathcal{S}$ versus m when $\Sigma = \sigma^2 \mathbf{I}$ and $\sigma^2 = 10$. As can be seen, Fig. 1 illustrates that COM-Lasso can potentially achieve perfect support recovery for $m \geq 26$ with fairly small number of vector observations in time when SNR is low. This indicates that even though $\tilde{\Phi}$ is exceedingly overcomplete, $\tilde{\Phi}$ still satisfies the NIR condition if m is large enough. Fig. 1 shows that there is a small performance gap between oracle and EFIC. This gap is most probably due to the fact that $\mathbf{e}$ is far from being

Figure 2: The empirical probability of $\{\hat{\mathcal{T}} = \mathcal{S}\}$ versus m when $\mathbf{A}$ has a correlated structure and $\mathbf{\Sigma} = \sigma^2 \mathbf{I}$. Here, $\mu = 0.5$, $|\mathcal{S}| = 20$, $\sigma^2 = 10$, $N = \lceil (m^2 - 1)^d \rceil$ for $d = 1.2$ and $L = 4m \ln m$.

Gaussian distributed with $\mathcal{N}(\mathbf{0}, 1/L)$ when m is small; that is, the calculations
 265 for the choice of c in EFIC is inaccurate when m is small. Fig. 1 also shows
 that, even though $\lambda(\tau)$ is tuned to perform at its best, it is still outperformed
 by EFIC. SPICE performs near-identical to COM-Lasso in the setup of Fig. 1.
 Expectedly, COM-Lasso-oracle outperforms (17)-oracle. Further, Fig. 2 shows
 that the performance of COM-Lasso-oracle is slightly shifted due to the large
 270 μ . This implies that support recovery by COM-Lasso is relatively robust to
 the correlation between $\mathbf{a}_i$'s. Moreover, Fig. 2 unveils that $\Pr\{\hat{\mathcal{T}} = \mathcal{S}\}$ for
 COM-Lasso-EFIC, COM-Lasso- λ and SPICE has slightly worsen due to large
 μ . Finally, (17)-oracle completely fails in identifying $\mathcal{S}$ when $\mathbf{a}_{*i}$'s are highly
 correlated.

275 Next, we assess the performance of support recovery when $\mathbf{\Sigma}$ is a diagonal
 matrix. Figs. 3 and 4 illustrate the empirical probability of support recovery
 versus m when the diagonal elements of $\mathbf{\Sigma}$ are normalized random positive
 integers such that $\text{tr}(\mathbf{\Sigma}) = 10m$. These numerical simulations show that the

Figure 3: The empirical probability of $\{\hat{\mathcal{T}} = \mathcal{S}\}$ versus m when $\mathbf{A}$ has an uncorrelated structure, i.e. $\mu = 0$, and $\mathbf{\Sigma}$ is a diagonal matrix. Here, $|\mathcal{S}| = 20$, $\text{tr}(\mathbf{\Sigma}) = 10m$, $N = \lceil (m^2 - m)^d \rceil$ for $d = 1.2$ and $L = 4m \ln m$.

performance of COM-Lasso-oracle is nearly identical for the same μ . This implies that COM-Lasso properly takes the nuisance parameter $\mathbf{h}$ into account. Moreover, when $\mu = 0$, the performance gap between EFIC and COM-Lasso-oracle is rather small. As can be seen, Fig. 3 shows that $\Pr\{\hat{\mathcal{T}} = \mathcal{S}\}$ worsens for (17)-oracle and SPICE when $\mathbf{\Sigma}$ has a diagonal structure. Continuing, Fig. 4 shows that SPICE performs poorly when $\mu = 0.5$. Again, (17)-oracle completely fails to estimate the true support when $\mathbf{a}_{*i}$'s are strongly correlated. Finally, COM-Lasso-EFIC outperforms COM-Lasso- λ both when $\mu = 0$ and when $\mu = 0.5$.

5. Conclusion

Model selection for high-dimensional linear regressions in the presence of strong and possibly correlated noise is a challenging task. To fulfill this task, by considering multiple number of measurement vectors, we propose a non-negative

Figure 4: The empirical probability of $\{\hat{\mathcal{T}} = \mathcal{S}\}$ versus m when $\mathbf{A}$ has a correlated structure and Σ is diagonal matrix. Here, $\mu = 0.5$, $|\mathcal{S}| = 20$, $\text{tr}(\Sigma) = 10m$, $N = \lceil (m^2 - m)^d \rceil$ for $d = 1.2$ and $L = 4m \ln m$.

Lasso estimator that exploits the covariance matching techniques. We provide deterministic conditions under which the proposed non-negative Lasso can identify the true model. Further, to choose Lasso's tuning parameter properly, we use EFIC. We prove that by minimizing EFIC one can find the true model with probability one as the number of rows in the design matrix grows to infinity. The simulation results indicate that our method is asymptotically consistent in estimating the true model. The numerical simulations also show that the proposed method is robust to high correlation between the columns of the design matrix. Although numerical simulations illustrate that our method performs desirably, additional analysis would be of interest and is left for future work. Specifically, it would be interesting to show that conditions in Theorem (1) hold as m grows to infinity, establish the link between the consistency of EFIC and COM-Lasso, and show the robustness of COM-Lasso to high correlation between columns of the design matrix.

6. Appendix

Lemma 1. Suppose that $L \geq m$ and $[\bar{\mathbf{A}} \circ \tilde{\mathbf{A}} \quad \tilde{\mathbf{Q}}]$ satisfies the restricted eigenvalue property of order $2K$. Then, $\tilde{\Phi}_{\mathcal{I}}$ is a full column-rank matrix with very high probability for any $\mathcal{I} \in [N]$ such that $|\mathcal{I}| \leq 2K$.

Proof. Because $\hat{\Gamma}^{-1/2} \mathbf{T}$ has full rank with very high probability and that $[\bar{\mathbf{A}} \circ \tilde{\mathbf{A}} \quad \tilde{\mathbf{Q}}]$ satisfies the restricted eigenvalue property, we can say that $\hat{\Gamma}^{-1/2} \mathbf{T} \times [\bar{\mathbf{A}}_{\mathcal{I}} \circ \mathbf{A}_{\mathcal{I}} \quad \mathbf{Q}]$ has full rank with very high probability. Thus,

$$\Xi = \begin{bmatrix} \Psi_{\mathcal{I}}^T \Psi_{\mathcal{I}} & \Psi_{\mathcal{I}}^T \mathbf{G} \\ \mathbf{G}^T \Psi_{\mathcal{I}} & \mathbf{G}^T \mathbf{G} \end{bmatrix}$$

is also a full rank matrix, where matrices $\Psi_{\mathcal{I}}$ and $\mathbf{G}$ are defined as $\Psi_{\mathcal{I}} = \hat{\Gamma}^{-1/2} \mathbf{T}(\bar{\mathbf{A}}_{\mathcal{I}} \circ \mathbf{A}_{\mathcal{I}})$ and $\mathbf{G} = \hat{\Gamma}^{-1/2} \mathbf{T} \mathbf{Q}$. Now, the Schur complement of $\mathbf{G}^T \mathbf{G}$ in the preceding block matrix is

$$\begin{aligned} \Psi_{\mathcal{I}}^T \Psi_{\mathcal{I}} - \Psi_{\mathcal{I}}^T \mathbf{G} (\mathbf{G}^T \mathbf{G})^{-1} \mathbf{G}^T \Psi_{\mathcal{I}} &= \Psi_{\mathcal{I}}^T \Pi_{\mathbf{G}}^{\perp} \Psi_{\mathcal{I}} \\ &= (\mathbf{B}(\bar{\mathbf{A}}_{\mathcal{I}} \circ \mathbf{A}_{\mathcal{I}}))^T \mathbf{B}(\bar{\mathbf{A}}_{\mathcal{I}} \circ \mathbf{A}_{\mathcal{I}}). \end{aligned}$$

310 By considering the properties of the Schur complement together with the fact that $\mathbf{G}^T \mathbf{G}$ and Ξ are full rank matrices, we can conclude that $\mathbf{B}(\bar{\mathbf{A}}_{\mathcal{I}} \circ \mathbf{A}_{\mathcal{I}})$ has full rank. Finally, observe that $\Omega \mathbf{U} \tilde{\Phi}_{\mathcal{I}} = \mathbf{B}(\bar{\mathbf{A}}_{\mathcal{I}} \circ \mathbf{A}_{\mathcal{I}}) \mathbf{W}_{\Phi_{\mathcal{I}}}^{-1}$; thus, by considering that $\Omega \mathbf{U}$ has orthonormal columns we can conclude that $\tilde{\Phi}_{\mathcal{I}}$ is a full rank matrix for $\{\mathcal{I} | \mathcal{I} \subset [N], |\mathcal{I}| \leq 2K\}$. $\square$

Lemma 2. Suppose that $[\bar{\mathbf{A}} \circ \tilde{\mathbf{A}} \quad \tilde{\mathbf{Q}}]$ satisfies the restricted eigenvalue property of order $2K$. Then, for sufficiently large L , the columns of Φ defined in (5) obey

$$\begin{aligned} i) \quad \|\phi_i\|_2^2 &\geq \frac{C_{\min}}{\left[\hat{C}_0 \max_{i \in \mathcal{S}} p_{i,i}\right]^2} \left(\frac{\min_{i \notin \mathcal{S}} \|\mathbf{a}_i\|_2^4}{\max_{i \in \mathcal{S}} \|\mathbf{a}_i\|_2^4} \right), & i \notin \mathcal{S}, \\ ii) \quad \|\phi_i\|_2^2 &\leq \frac{1}{p_i^2}, & i \in \mathcal{S}, \end{aligned}$$

with high probability for some constant $\dot{C}_0 > 0$. Further,

$$iii) \|\Pi_{\Phi_{\mathcal{I}}}^\perp \Phi_{\mathcal{S}} \mathbf{p}_{\mathcal{S}}\|_2^2 \geq \Lambda_{\min}(\tilde{\Phi}_{\dot{\mathcal{S}} \cup \mathcal{I}}^T \tilde{\Phi}_{\dot{\mathcal{S}} \cup \mathcal{I}}) (\min_{i \in \dot{\mathcal{S}}} p_i^2) \sum_{i \in \dot{\mathcal{S}}} \|\phi_i\|_2^2,$$

315 where the set $\dot{\mathcal{S}}$ is defined as $\dot{\mathcal{S}} \triangleq \mathcal{S} \setminus \mathcal{I}$.

Proof. For i) : First, recall the definition of Φ and $\mathbf{B}$ in equations (5) and (4), respectively. Then, note that

$$(w_\Phi)_{i,i}^2 \triangleq \|\phi_i\|_2^2 = \|\mathbf{U}\mathbf{S}(\mathbf{V}^T)_i\|_2^2 = \|\Omega^T \hat{\Gamma}^{-1/2}(\bar{\mathbf{a}}_i) \otimes \mathbf{a}_i\|_2^2 = \|\mathbf{B}(\bar{\mathbf{a}}_i \otimes \mathbf{a}_i)\|_2^2.$$

Next, by using some straightforward manipulations, it can be shown that

$$\|\mathbf{B}(\bar{\mathbf{a}}_i \otimes \mathbf{a}_i)\|_2^2 = \|\Pi_{(\hat{\mathbf{R}}^T \otimes \hat{\mathbf{R}})^{-1/2} \mathbf{Q}}^\perp (\hat{\mathbf{R}}^T \otimes \hat{\mathbf{R}})^{-1/2} (\bar{\mathbf{a}}_i \otimes \mathbf{a}_i)\|_2^2.$$

Then, utilize the closest point theorem to show that

$$\begin{aligned} & \|\Pi_{(\hat{\mathbf{R}}^T \otimes \hat{\mathbf{R}})^{-1/2} \mathbf{Q}}^\perp (\hat{\mathbf{R}}^T \otimes \hat{\mathbf{R}})^{-1/2} (\bar{\mathbf{a}}_i \otimes \mathbf{a}_i)\|_2^2 \\ &= \min_{\nu \in \mathbb{C}^\kappa} \|(\hat{\mathbf{R}}^T \otimes \hat{\mathbf{R}})^{-1/2} (\bar{\mathbf{a}}_i \otimes \mathbf{a}_i - \mathbf{Q}\nu)\|_2^2 \\ &\stackrel{E_2}{\geq} \Lambda_{\max}^{-1}(\hat{\mathbf{R}}^T \otimes \hat{\mathbf{R}}) \min_{\nu \in \mathbb{C}^\kappa} \|\bar{\mathbf{a}}_i \otimes \mathbf{a}_i - \mathbf{Q}\nu\|_2^2, \end{aligned}$$

where we have used the definition of the smallest eigenvalue to obtain the inequality in E_2 . It follows that

$$\begin{aligned} & \min_{i \notin \mathcal{S}} (w_\Phi)_{i,i}^2 \\ &\geq \Lambda_{\max}^{-1}(\hat{\mathbf{R}}^T \otimes \hat{\mathbf{R}}) \min_{i \notin \mathcal{S}} \min_{\nu \in \mathbb{C}^\kappa} \|\bar{\mathbf{a}}_i \otimes \mathbf{a}_i - \mathbf{Q}\nu\|_2^2 \\ &\stackrel{E_3}{=} \Lambda_{\max}^{-2}(\hat{\mathbf{R}}) \min_{i \notin \mathcal{S}} \min_{\nu \in \mathbb{C}^\kappa} \|\bar{\mathbf{a}}_i \otimes \mathbf{a}_i - \mathbf{Q}\nu\|_2^2 \\ &\stackrel{E_4}{=} \Lambda_{\max}^{-2}(\hat{\mathbf{R}}) \min_{i \notin \mathcal{S}} \min_{\nu \in \mathbb{C}^\kappa} \left\| \begin{bmatrix} \tilde{\bar{\mathbf{a}}}_i \otimes \tilde{\mathbf{a}}_i & \tilde{\mathbf{Q}} \end{bmatrix} \begin{bmatrix} \|\bar{\mathbf{a}}_i \otimes \mathbf{a}_i\|_2 \\ -\mathbf{W}_{\mathbf{Q}}\nu \end{bmatrix} \right\|_2^2 \\ &\stackrel{E_5}{\geq} \Lambda_{\max}^{-2}(\hat{\mathbf{R}}) \left(\min_{i \notin \mathcal{S}} \min_{\nu \in \mathbb{C}^\kappa} \left\| \begin{bmatrix} \|\bar{\mathbf{a}}_i \otimes \mathbf{a}_i\|_2 \\ -\mathbf{W}_{\mathbf{Q}}\nu \end{bmatrix} \right\|_2^2 \right) \\ &\quad \times \min_{i \notin \mathcal{S}} \Lambda_{\min} \left(\begin{bmatrix} \tilde{\bar{\mathbf{a}}}_i \otimes \tilde{\mathbf{a}}_i & \tilde{\mathbf{Q}} \end{bmatrix}^H \begin{bmatrix} \tilde{\bar{\mathbf{a}}}_i \otimes \tilde{\mathbf{a}}_i & \tilde{\mathbf{Q}} \end{bmatrix} \right) \\ &\geq C_{\min} \Lambda_{\max}^{-2}(\hat{\mathbf{R}}) \min_{i \notin \mathcal{S}} \|\mathbf{a}_i\|_2^4. \end{aligned}$$

Here, the identity in E_3 is due to the properties of the Kronecker product and the fact that $\hat{\mathbf{R}}$ is a positive definite matrix, the identity in E_4 is obtained by normalizing the columns of $\begin{bmatrix} \bar{\mathbf{a}}_i \otimes \mathbf{a}_i & \mathbf{Q} \end{bmatrix}$, the elements of the diagonal matrix $\mathbf{W}_{\mathbf{Q}}$ are $\|\mathbf{q}_i\|_2$'s, the inequality in E_5 is again because of the definition of the smallest eigenvalue and the final inequality is obtained by considering the properties of the Kronecker product together with the definition of the restricted eigenvalue property. To bound $\Lambda_{\max}(\hat{\mathbf{R}})$, one first should note that $\mathbf{R}$ has only $|\mathcal{S}|$ dominant eigenvalues. Thus, following [28, 29], we can say that eigenvalues of $\hat{\mathbf{R}}$ are close to those of $\mathbf{R}$ with high probability if $L = \Omega(m \ln m)$. Now observe that

$$\Lambda_{\max}(\mathbf{R}) \leq \Lambda_{\max}(\mathbf{A}_{\mathcal{S}} \mathbf{P}_{\mathcal{S}, \mathcal{S}} \mathbf{A}_{\mathcal{S}}^H) + \Lambda_{\max}(\boldsymbol{\Sigma}).$$

Moreover, by using the properties of eigenvalues, we can say

$$\Lambda_{\max}(\mathbf{A}_{\mathcal{S}} \mathbf{P}_{\mathcal{S}, \mathcal{S}} \mathbf{A}_{\mathcal{S}}^H) = \Lambda_{\max}(\mathbf{P}_{\mathcal{S}, \mathcal{S}}^{1/2} \mathbf{A}_{\mathcal{S}}^H \mathbf{A}_{\mathcal{S}} \mathbf{P}_{\mathcal{S}, \mathcal{S}}^{1/2}) \leq \max_{i \in \mathcal{S}} p_{i,i} \Lambda_{\max}(\mathbf{A}_{\mathcal{S}}^H \mathbf{A}_{\mathcal{S}}). \quad (18)$$

Next, normalize columns of $\mathbf{A}_{\mathcal{S}}$ as $\mathbf{A}_{\mathcal{S}} = \tilde{\mathbf{A}}_{\mathcal{S}} \mathbf{W}_{\mathbf{A}_{\mathcal{S}}}$ and again exploit the properties of eigenvalue, so that

$$\Lambda_{\max}(\mathbf{A}_{\mathcal{S}}^H \mathbf{A}_{\mathcal{S}}) = \Lambda_{\max}(\mathbf{W}_{\mathbf{A}_{\mathcal{S}}} \tilde{\mathbf{A}}_{\mathcal{S}}^H \tilde{\mathbf{A}}_{\mathcal{S}} \mathbf{W}_{\mathbf{A}_{\mathcal{S}}}) \leq \max_{i \in \mathcal{S}} \|\mathbf{a}_i\|_2^2 \Lambda_{\max}(\tilde{\mathbf{A}}_{\mathcal{S}}^H \tilde{\mathbf{A}}_{\mathcal{S}}).$$

Further, Gerschgorin's Theorem implies that $\Lambda_{\max}(\tilde{\mathbf{A}}_{\mathcal{S}}^H \tilde{\mathbf{A}}_{\mathcal{S}}) = \mathcal{O}(1)$; thus,

$$\Lambda_{\max}(\mathbf{A}_{\mathcal{S}}^H \mathbf{A}_{\mathcal{S}}) = \mathcal{O}(\max_{i \in \mathcal{S}} \|\mathbf{a}_i\|_2^2)$$

To summarize, insert the preceding identity into (18) and note that $\Lambda_i(\boldsymbol{\Sigma}) = \mathcal{O}(1)$, hence

$$\Lambda_{\max}(\mathbf{R}) \leq (\dot{C}_0 \max_{i \in \mathcal{S}} p_{i,i}) \max_{i \in \mathcal{S}} \|\mathbf{a}_i\|_2^2 (1 + o(1)).$$

The conclusion is straightforward.

For *ii*) : By considering the properties of the projection matrix, we can deduce from (??) that

$$\|\phi_i\|_2^2 \leq \|(\hat{\mathbf{R}}^T \otimes \hat{\mathbf{R}})^{-1/2} (\bar{\mathbf{a}}_i \otimes \mathbf{a}_i)\|_2^2.$$

Representing the right-hand side Euclidean-norm as a quadratic form and using the properties of the Kronecker product gives

$$\|\phi_i\|_2^2 \leq (\mathbf{a}_i^H \hat{\mathbf{R}}^{-1} \mathbf{a}_i)^2.$$

Next, rewrite $\hat{\mathbf{R}}$ as

$$\hat{\mathbf{R}} = p_i \mathbf{a}_i \mathbf{a}_i^H + \underbrace{\sum_{j \in \mathcal{S} \setminus i} p_j \mathbf{a}_j \mathbf{a}_j^H}_{\mathbf{D}} + \mathbf{\Sigma} + \mathbf{E},$$

where the residual matrix $\mathbf{E} = \hat{\mathbf{R}} - \mathbf{R}$. Since $\mathbf{D}$ is a positive definite matrix with a very high probability for large enough L , by considering the Woodbury matrix identity, we can reformulate $\hat{\mathbf{R}}^{-1}$ as

$$\hat{\mathbf{R}}^{-1} = \mathbf{D}^{-1} - \frac{1}{p_i^{-1} + \beta} \mathbf{D}^{-1} \mathbf{a}_i \mathbf{a}_i^H \mathbf{D}^{-1},$$

where β represents $\beta = \mathbf{a}_i^H \mathbf{D}^{-1} \mathbf{a}_i$. Inserting the preceding identity into $\mathbf{a}_i^H \hat{\mathbf{R}}^{-1} \mathbf{a}_i$ yields

$$\mathbf{a}_i^H \hat{\mathbf{R}}^{-1} \mathbf{a}_i = \frac{\beta}{1 + p_i \beta} \leq \frac{1}{p_i},$$

which is what we needed to prove.

For *iii*) : split $\mathcal{S}$ as $\mathcal{S} = \dot{\mathcal{S}} \cup \{\mathcal{I} \cap \mathcal{S}\}$ and observe that

$$\|\Pi_{\Phi_{\mathcal{I}}}^{\perp} \Phi_{\mathcal{S}} \mathbf{p}_{\mathcal{S}}\|_2^2 = \left\| \Pi_{\Phi_{\mathcal{I}}}^{\perp} \begin{bmatrix} \Phi_{\mathcal{I} \cap \mathcal{S}} & \Phi_{\dot{\mathcal{S}}} \end{bmatrix} \begin{bmatrix} \mathbf{p}_{\mathcal{I} \cap \mathcal{S}} \\ \mathbf{p}_{\dot{\mathcal{S}}} \end{bmatrix} \right\|_2^2 = \|\Pi_{\Phi_{\mathcal{I}}}^{\perp} \Phi_{\dot{\mathcal{S}}} \mathbf{p}_{\dot{\mathcal{S}}}\|_2^2.$$

Next, reusing the similar techniques that have been established in Lemma 2-(i) yields

$$\|\Pi_{\Phi_{\mathcal{I}}}^{\perp} \Phi_{\mathcal{S}} \mathbf{p}_{\mathcal{S}}\|_2^2 \geq \Lambda_{\min}(\tilde{\Phi}_{\dot{\mathcal{S}} \cup \mathcal{I}}^T \tilde{\Phi}_{\dot{\mathcal{S}} \cup \mathcal{I}}) (\min_{i \in \dot{\mathcal{S}}} p_i^2) \sum_{i \in \dot{\mathcal{S}}} \|\phi_i\|_2^2.$$

□

Proof of Theorem 2. We would like to highlight that the main steps that should be taken to prove Theorem 2 in this paper are similar to those of Theorem 2 in [11]; thus, to condense this proof, we adopt the presentation in [11, Theorem 2]. However, we carefully express the main differences in detail. The proof consists of three main sections. First, we express two deterministic optimality conditions for EFIC. Then, we show that each of these conditions hold as $l \rightarrow \infty$.

The optimality conditions for EFIC: The strict minimizer of

$$\hat{\mathcal{I}}^{\text{EFIC}} = \arg \min_{\mathcal{I} \in \mathcal{J}} g(\mathcal{I}) \quad (19)$$

satisfies $\hat{\mathcal{I}}^{\text{EFIC}} = \mathcal{S}$ if $g(\mathcal{S}) < g(\mathcal{I})$ for $\mathcal{I} \in \mathcal{J} \setminus \mathcal{S}$. Observe, now, that $\mathcal{J} \setminus \mathcal{S}$ can be split into the union of two subsets as

$$\mathcal{J} \setminus \mathcal{S} = \left\{ \bigcup_{k=|\mathcal{S}|+1}^K \hat{\mathcal{I}}^k \right\} \cup \left\{ \bigcup_{k=0}^K \check{\mathcal{I}}^k \right\},$$

where

$$\begin{aligned} \hat{\mathcal{I}}^k &= \{\mathcal{I} | \mathcal{S} \subset \mathcal{I}, |\mathcal{I}| = k\}, \\ \check{\mathcal{I}}^k &= \{\mathcal{I} | \mathcal{S} \not\subset \mathcal{I}, |\mathcal{I}| = k\}. \end{aligned}$$

Having this in mind, by simplifying $g(\mathcal{S}) < g(\mathcal{I})$ for $\mathcal{I} \in \hat{\mathcal{I}}^k$ and $\mathcal{I} \in \check{\mathcal{I}}^k$, we can say that $\hat{\mathcal{I}}^{\text{EFIC}} = \mathcal{S}$ if

$$\begin{aligned} \mathcal{J}_1 : \|\mathbf{\Pi}_{\Phi_{\mathcal{S}}}^\perp \underline{\mathbf{e}}\|_2^2 - \|\mathbf{\Pi}_{\Phi_{\mathcal{I}}}^\perp \underline{\mathbf{e}}\|_2^2 &< \Upsilon(\mathcal{I}), & \mathcal{I} \in \bigcup_{k=|\mathcal{S}|+1}^K \hat{\mathcal{I}}^k, \\ \mathcal{J}_2 : \|\mathbf{\Pi}_{\Phi_{\mathcal{S}}}^\perp \underline{\mathbf{e}}\|_2^2 - L\|\mathbf{\Pi}_{\Phi_{\mathcal{I}}}^\perp \mathbf{z}\|_2^2 &< \Upsilon(\mathcal{I}), & \mathcal{I} \in \bigcup_{k=0}^K \check{\mathcal{I}}^k. \end{aligned}$$

Here, the vector $\underline{\mathbf{e}} = \sqrt{L}\mathbf{e}$ is the normalized version of $\mathbf{e}$ and

$$\Upsilon(\mathcal{I}) = \ln \frac{\det(\Phi_{\mathcal{I}}^T \Phi_{\mathcal{I}})}{\det(\Phi_{\mathcal{S}}^T \Phi_{\mathcal{S}})} + (|\mathcal{I}| - |\mathcal{S}|)(\ln L + 2c \ln N).$$

We continue this section of the proof by evaluating the asymptotic behavior of $\Upsilon(\mathcal{I})$. In this regard, using the result produced in [11, Theorem 2], we have

$$\ln \det(\Phi_{\mathcal{I}}^T \Phi_{\mathcal{I}}) = \sum_{i \in \mathcal{I}} \ln \|\phi_i\|_2^2 + C_0.$$

Next, by inserting the preceding identity into $\Upsilon(\mathcal{I})$ and considering Lemma 2-(i,ii) together with (15), we can conclude that for $\mathcal{I} \in \hat{\mathcal{I}}^k$

$$\begin{aligned} \Upsilon(\mathcal{I}) &= \sum_{i \in \mathcal{I} \setminus \mathcal{S}} \ln \|\phi_i\|_2^2 + C_1 + \Delta(\ln L + 2c \ln N) \\ &\geq 2\Delta \ln \left(\frac{\min_{i \notin \mathcal{S}} \|\mathbf{a}_i\|_2^2}{\max_{i \in \mathcal{S}} \|\mathbf{a}_i\|_2^2} \right) + \Delta(\ln L + 2c \ln N) + \mathcal{O}(1) \\ &\geq \Delta[-2r \ln m + \ln L + 2c \ln N](1 + o(1)) \end{aligned} \tag{20}$$

and for $\mathcal{I} \in \check{\mathcal{I}}^k$

$$\begin{aligned}
\Upsilon(\mathcal{I}) &= \sum_{i \in \mathcal{I} \setminus \mathcal{S}} \ln \|\phi_i\|_2^2 - \sum_{i \in \mathcal{S} \setminus \mathcal{I}} \ln \|\phi_i\|_2^2 + C_2 + (|\mathcal{I}| - |\mathcal{S}|) (\ln L + 2c \ln N) \\
&\geq 2(|\mathcal{I}| - |\mathcal{S}|) \ln \left(\frac{\min_{i \notin \mathcal{S}} \|\mathbf{a}_i\|_2^2}{\max_{i \in \mathcal{S}} \|\mathbf{a}_i\|_2^2} \right) + (|\mathcal{I}| - |\mathcal{S}|) (\ln L + 2c \ln N) + O(1) \\
&\geq (|\mathcal{I}| - |\mathcal{S}|) [-2r \ln m + \ln L + 2c \ln N] (1 + o(1))
\end{aligned} \tag{21}$$

325 where $\Delta = ||\mathcal{I}| - |\mathcal{S}||$.

Establishing $\Pr\{\mathcal{J}_1\}$ as $l \rightarrow \infty$: In this regard, first use the definition of the projection matrix, so that

$$\|\mathbf{\Pi}_{\Phi_{\mathcal{S}}}^\perp \underline{\epsilon}\|_2^2 - \|\mathbf{\Pi}_{\Phi_{\mathcal{I}}}^\perp \underline{\epsilon}\|_2^2 = \|\mathring{\mathbf{\Pi}}_{\Phi_{\mathcal{I} \setminus \mathcal{S}}}\|_2^2,$$

where $\mathring{\mathbf{\Pi}}_{\Phi_{\mathcal{I} \setminus \mathcal{S}}}$ is defined as $\mathring{\mathbf{\Pi}}_{\Phi_{\mathcal{I} \setminus \mathcal{S}}} = \mathbf{\Pi}_{\Phi_{\mathcal{I}}} - \mathbf{\Pi}_{\Phi_{\mathcal{S}}}$. Then, let $\mathring{E}_k$ be the event

$$\mathring{E}_k : \max_{\mathcal{I} \in \check{\mathcal{I}}^k} \|\mathring{\mathbf{\Pi}}_{\Phi_{\mathcal{I} \setminus \mathcal{S}}}\|_2^2 < \min_{\mathcal{I} \in \check{\mathcal{I}}^k} \Upsilon(\mathcal{I})$$

and observe that $\mathcal{J}_1$ is fulfilled if $\mathring{E}_k$ occurs for all $k \in \{|\mathcal{S}| + 1, \dots, K\}$. Observe, now, since $\underline{\epsilon}$ is approximately a standard Gaussian random variable, $\max_{\mathcal{I} \in \check{\mathcal{I}}^k} \|\mathring{\mathbf{\Pi}}_{\Phi_{\mathcal{I} \setminus \mathcal{S}}}\|_2^2 = O_p(\ln l)$ [11, Theorem 2]. In details,

$$\Pr \left\{ \max_{\mathcal{I} \in \check{\mathcal{I}}^k} \|\mathring{\mathbf{\Pi}}_{\Phi_{\mathcal{I} \setminus \mathcal{S}}}\|_2^2 \leq 2(\Delta \ln N + \gamma \ln l) \right\} \geq 1 - l^{-\gamma},$$

where $\gamma > 0$ is a positive constant. Consequently, using the Borel-Cantelli lemma provides that

$$\max_{\mathcal{I} \in \check{\mathcal{I}}^k} \|\mathring{\mathbf{\Pi}}_{\Phi_{\mathcal{I} \setminus \mathcal{S}}}\|_2^2 \leq 2(\Delta \ln N + \gamma \ln l) \tag{22}$$

with probability one as $m \rightarrow \infty$ for the choice of $\gamma > 1$. To conclude, using (22) and (20) implies that the event $\mathring{E}_k$ happens with probability one if

$$2(\Delta \ln N + \gamma \ln l) < \Delta(-2r \ln m + \ln L + 2c \ln N)$$

for $\Delta = 1, 2, \dots, K - |\mathcal{S}|$. A straightforward calculation yields

$$c > 1 + \gamma \frac{\ln l}{\ln N} - \frac{\ln L}{2 \ln N} + \frac{r \ln m}{\ln N}.$$

Eventually, because

$$\Pr \mathcal{J}_1 \geq \Pr \left\{ \bigcap_{k=|S|+1}^K \check{E}_k \right\},$$

we can say that $\mathcal{J}_1$ holds with probability one.

Establish $\Pr\{\mathcal{J}_2\}$ as $l \rightarrow \infty$: Expand $\|\Pi_{\Phi_{\mathcal{I}}}^\perp \mathbf{z}\|_2^2$ and observe that on the event

$$\check{E}_k : \min_{\mathcal{I} \in \check{\mathcal{I}}^k} \{A_1 + A_2 + \Upsilon(\mathcal{I})\} > 0 \quad (23)$$

for all $k \in [K]$, $\mathcal{J}_2$ is satisfied, where the terms A_1 and A_2 are defined as

$$\begin{aligned} A_1 &= L \|\Pi_{\Phi_{\mathcal{I}}}^\perp \Phi_S \mathbf{p}_S\|_2^2 + 2\sqrt{L} \boldsymbol{\epsilon}^T \Pi_{\Phi_{\mathcal{I}}}^\perp \Phi_S \mathbf{p}_S \text{ and} \\ A_2 &= \|\Pi_{\Phi_{\mathcal{I}}}^\perp \boldsymbol{\epsilon}\|_2^2 - \|\Pi_{\Phi_S}^\perp \boldsymbol{\epsilon}\|_2^2 = \|\Pi_{\Phi_S}^\perp \boldsymbol{\epsilon}\|_2^2 - \|\Pi_{\Phi_{\mathcal{I}}}^\perp \boldsymbol{\epsilon}\|_2^2. \end{aligned}$$

We evaluate each term individually. First,

$$\begin{aligned} A_1 &= L \|\Pi_{\Phi_{\mathcal{I}}}^\perp \Phi_S \mathbf{p}_S\|_2^2 \left(1 + \frac{2}{\sqrt{L}} \frac{\boldsymbol{\epsilon}^T \Pi_{\Phi_{\mathcal{I}}}^\perp \Phi_S \mathbf{p}_S}{\|\Pi_{\Phi_{\mathcal{I}}}^\perp \Phi_S \mathbf{p}_S\|_2} \right) \\ &\geq L \|\Pi_{\Phi_{\mathcal{I}}}^\perp \Phi_S \mathbf{p}_S\|_2^2 \times \left(1 - \underbrace{\frac{2}{\sqrt{L} \|\Pi_{\Phi_{\mathcal{I}}}^\perp \Phi_S \mathbf{p}_S\|_2} \max_{\mathcal{I} \in \check{\mathcal{I}}^k} \left| \boldsymbol{\epsilon}^T \frac{\Pi_{\Phi_{\mathcal{I}}}^\perp \Phi_S \mathbf{p}_S}{\|\Pi_{\Phi_{\mathcal{I}}}^\perp \Phi_S \mathbf{p}_S\|_2} \right|}_{A_3} \right). \end{aligned}$$

However, $A_3 \leq 2\sqrt{\ln \left(\frac{N}{|\mathcal{I}|}\right)}$ with probability one as $l \rightarrow \infty$ [11, Theorem 2].

Applying parts (ii) and (iii) in Lemma 2 gives

$$\frac{A_3}{\sqrt{L} \|\Pi_{\Phi_{\mathcal{I}}}^\perp \Phi_S \mathbf{p}_S\|_2} \leq C_3 \sqrt{\frac{\ln N}{L}}$$

with probability one as $m \rightarrow \infty$. Hence,

$$A_1 \geq L \|\Pi_{\Phi_{\mathcal{I}}}^\perp \Phi_S \mathbf{p}_S\|_2^2 (1 + o(1)).$$

Second, $A_2 = O_p(\ln l)$, cf. (22). To summarize, use (21) together with the bounds provided for A_1 and A_2 , thus

$$\begin{aligned} \min_{\mathcal{I} \in \check{\mathcal{I}}^k} \{A_1 + A_2 + \Upsilon(\mathcal{I})\} &\geq L \|\Pi_{\Phi_{\mathcal{I}}}^\perp \Phi_S \mathbf{p}_S\|_2^2 \\ &\quad + 2(|\mathcal{I}| - |\mathcal{S}|) \ln \left(\frac{\min_{i \notin \mathcal{S}} \|\mathbf{a}_i\|_2^2}{\max_{i \in \mathcal{S}} \|\mathbf{a}_i\|_2^2} \right) + O(\ln N) + O(\ln L). \end{aligned}$$

As can be seen, $L\|\Pi_{\Phi_T}^\perp \Phi_S \mathbf{p}_S\|_2^2$ is the dominant positive term. As a result, the event $\check{E}_k$ holds with probability one as $l \rightarrow \infty$. Eventually, we conclude that $\mathcal{J}_2$ holds with probability one as $l \rightarrow \infty$ since

$$\Pr \mathcal{J}_2 = \Pr \left\{ \bigcap_{k=0}^K \check{E}_k \right\}.$$

□

References

- [1] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Series B (Methodol.) 58 (1) (1996) 267–288.
- [2] P. Zhao, B. Yu, On model selection consistency of lasso, J. Mach. Learn. Res. 7 (2006) 2541–2563.
- 335 [3] E. Candes, Y. Plan, Near-ideal model selection by ℓ_1 minimization, Ann. of Statist. 37 (5A) (2009) 2145–2177. doi:10.1214/08-A0S653.
- [4] M. J. Wainwright, Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (lasso), IEEE Trans. Inf. Theory 55 (5) (2009) 2183–2202. doi:10.1109/TIT.2009.2016018.
- 340 [5] S. Negahban, M. J. Wainwright, Simultaneous support recovery in high dimensions: Benefits and perils of block ℓ_1/ℓ_∞ -regularization, IEEE Trans. on Inf. Theory 57 (6) (2011) 3841–3863.
- [6] G. Obozinski, M. J. Wainwright, M. Jordan, Support union recovery in high-dimensional multivariate regression, The Ann. of Statist. 39 (1) (2011) 1–47.
- 345 [7] K. Lounici, M. Pontil, A. B. Tsybakov, S. A. Van De Geer, Taking advantage of sparsity in multi-task learning, Conference on Computat. Learn. Theory (COLT).

- [8] P. Stoica, P. Babu, J. Li, Spice: A sparse covariance-based estimation
 350 method for array processing, *IEEE Tran. Signal Process.* 59 (2) (2011)
 629–638. doi:10.1109/TSP.2010.2090525.
- [9] P. Pal, P. P. Vaidyanathan, Pushing the limits of sparse support recovery
 using correlation information, *IEEE Tran. Signal Process.* 63 (3) (2015)
 711–726. doi:10.1109/TSP.2014.2385033.
- [10] A. Owrang, M. Jansson, Weighted covariance matching based square root
 355 lasso, in: 2015 IEEE International Conference on Acoustics, Speech and
 Signal Processing (ICASSP), IEEE, 2015, pp. 3751–3755.
- [11] A. Owrang, M. Jansson, A model selection criterion for high-dimensional
 linear regression, *IEEE Transactions on Signal Processing* 66 (13) (2018)
 360 3436–3446. doi:10.1109/TSP.2018.2821628.
- [12] P. Stoica, T. Soderstrom, On array signal processing in spatially correlated
 noise fields, *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.*
 39 (12) (1992) 879–882.
- [13] B. Ottersten, P. Stoica, R. Roy, Covariance matching estimation techniques
 365 for array signal processing applications, *Digit. Signal Process.* 8 (3) (1998)
 185 – 210. doi:http://dx.doi.org/10.1006/dspr.1998.0316.
- [14] S. Foucart, H. Rauhut, A mathematical introduction to compressive sens-
 ing, Vol. 1, Birkhäuser Basel, 2013.
- [15] A. C. Aitken, On least squares and linear combination of observations,
 370 *Proc. of the Roy. Soc. of Edinburgh* 55 (1936) 42–48.
- [16] W. H. Greene, *Econometric analysis*, Pearson Education India, 2003.
- [17] D. L. Donoho, J. Tanner, Sparse nonnegative solutions of underdetermined
 linear equations by linear programming (2005) 9446–9451.
- [18] L. Wu, Y. Yang, H. Liu, Nonnegative-lasso and application in index track-
 375 ing, *Computat. Statist. & Data Anal.* 70 (2014) 116–126.

- [19] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, *Ann. Statist.* 32 (2) (2004) 407–499. doi:10.1214/009053604000000067.
- [20] S. Rosset, J. Zhu, Piecewise linear regularized solution paths, *The Ann. of Statist.* 35 (3) (2007) 1012–1030.
- 380 [21] H. Akaike, A new look at the statistical model identification, *IEEE Trans. Autom. Control* 19 (6) (1974) 716–723. doi:10.1109/TAC.1974.1100705.
- [22] G. Schwarz, Estimating the dimension of a model, *Ann. Statist.* 6 (2) (1978) 461–464. doi:10.1214/aos/1176344136.
- [23] H. Bozdogan, Model selection and akaike’s information criterion (aic): The
385 general theory and its analytical extensions, *Psychometrika* 52 (3) (1987) 345–370. doi:10.1007/BF02294361.
- [24] K. W. Broman, T. P. Speed, A model selection approach for the identification of quantitative trait loci in experimental crosses, *J. R. Stat. Soc.: Series B (Stat. Methodol.)* 64 (4) (2002) 641–656.
- 390 [25] J. Chen, Z. Chen, Extended bayesian information criteria for model selection with large model spaces, *Biometrika* 95 (3) (2008) 759–771.
- [26] A. Owrang, In pursuit of ideal model selection for high-dimensional linear regression, Ph.D. thesis, KTH Royal Institute of Technology (2019).
- [27] J. A. Tropp, An introduction to matrix concentration inequalities, *Foundations and Trends® in Machine Learning* 8 (1-2) (2015) 1–230.
395
- [28] J. A. Tropp, User-friendly tail bounds for sums of random matrices, *Foundations of Computat. Mathematics* 12 (4) (2012) 389–434. doi:10.1007/s10208-011-9099-z.
- [29] R. Vershynin, How close is the sample covariance matrix to the actual
400 covariance matrix?, *J. of Theor. Probab.* 25 (3) (2012) 655–686. doi:10.1007/s10959-010-0338-z.