On the Support Recovery of Jointly Sparse Gaussian Sources Via Sparse Bayesian Learning

Saurabh Khanna, *Member, IEEE,* and Chandra R. Murthy, *Senior Member, IEEE*

Abstract—In this work, we provide non-asymptotic, probabilistic guarantees for successful recovery of the common nonzero support of jointly sparse Gaussian sources in the multiple measurement vector (MMV) problem. The support recovery problem is formulated as the marginalized maximum likelihood (or type-II ML) estimation of the variance hyperparameters of a joint sparsity inducing Gaussian prior on the source signals. We derive conditions under which the resulting nonconvex constrained optimization perfectly recovers the nonzero support of a joint-sparse Gaussian source ensemble with arbitrarily high probability. The support error probability decays exponentially with the number of MMVs at a rate that depends on the smallest restricted singular value and the nonnegative null space property of the self Khatri-Rao product of the sensing matrix. Our analysis confirms that nonzero supports of size as high as $O(m^2)$ are recoverable from m measurements per sparse vector. Our derived sufficient conditions for support consistency of the proposed constrained type-II ML solution also guarantee the support consistency of any global solution of the multiple sparse Bayesian learning (M-SBL) optimization whose nonzero coefficients lie inside a bounded interval. For the case of noiseless measurements, we further show that a single MMV is sufficient for perfect recovery of the k-sparse support by M-SBL, provided all subsets of $k + 1$ columns of the sensing matrix are linearly independent.

Index Terms—Compressive sensing, support recovery, sparse Bayesian Learning, joint sparsity, Restricted Isometry Property, Khatri-Rao product.

I. INTRODUCTION

JOINT sparsity has emerged as an important and versatile signal structure in the field of sparse signal processing. Two or more vectors are said to be jointly sparse if their nonzero coefficients belong to the same index set, i.e., the vectors share a common nonzero support. Joint sparsity arises naturally in multi-modal or multi-channel analysis of signals residing in low dimensional signal subspaces. The underlying joint sparsity of signals can prove useful in resolving ambiguities in the common support that may arise due to erroneous estimation of the support of the individual sparse signal vectors from noisy measurements. This viewpoint has been successfully exploited in several real-world applications such as MIMO channel estimation [1]–[3], distributed source coding [4], [5], multi-task compressive sensing [6], distributed

S. Khanna is with the Dept. of ECE, IIT Roorkee, 247667, India. He was with the Dept. of ECE, Indian Institute of Science, Bangalore, India, during the initial part of this work. Email: sakhanna@ece.iitr.ac.in. C. R. Murthy is with the Dept. of ECE, Indian Institute of Science, 560012, India. Email: cmurthy@iisc.ac.in.

event localization [7], array signal processing [8], cooperative spectrum sensing [9]–[11], and user activity detection in massive machine-type communications [12].

In the sparse signal recovery literature, the estimation of jointly sparse signals is commonly referred to as the *multiple measurement vector* (MMV) problem [13] where the signal of interest is a matrix $\mathbf{X} \in \mathbb{R}^{n \times L}$ whose columns are jointly sparse vectors in $\mathbb{R}^n$. As a result, $\mathbf{X}$ is a row sparse matrix with only a fraction of its rows containing nonzero elements and the rest of the rows made up entirely of zeros. In the MMV problem, the goal is to recover $\mathbf{X}$ from its noisy, linear measurements $\mathbf{Y} \in \mathbb{R}^{m \times L}$. The measurement matrix $\mathbf{Y}$ (each column is called a single measurement vector (SMV)) is generated as

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{W}, \qquad (1)$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}$ is a known sensing matrix and $\mathbf{W} \in \mathbb{R}^{m \times L}$ models the unknown noise in the measurements. For $m < n$, the above linear system is underdetermined and has infinitely many solutions for $\mathbf{X}$. However, if $\mathbf{A}$ satisfies certain restricted isometry properties, a unique row-sparse solution can still be guaranteed [13]–[16]. In many applications, one is primarily interested in identifying the nonzero rows of $\mathbf{X}$. This gives rise to the *joint-sparse support recovery (JSSR)* problem where the goal is to recover the row support of $\mathbf{X}$ given $\mathbf{Y}$ and $\mathbf{A}$. Interestingly, unlike the nonzero coefficients in a k-row-sparse $\mathbf{X}$ which can be robustly recovered only if $m \geq k$, its nonzero row support can be recovered correctly even from $m < k$ measurements.

The vast majority of the JSSR solvers [13], [15], [17]–[21] implicitly assume that the number of nonzero rows in $\mathbf{X}$ is less than the number of measurements per SMV, i.e., $k < m$. Recently, *correlation-aware* solvers such as Co-LASSO [22] and M-SBL [23], [24], and their extensions, e.g., RD-CMP [25], Co-LASSO-EXPGRD [26] and MRNNQP [27] have been empirically shown to be capable of recovering sparse supports of size $k > m$ when L is large. However, a complete theoretical understanding of when these methods are guaranteed to be successful is still lacking. Especially for M-SBL, due to its nonseparable and nonconvex log-marginalized likelihood objective, any support consistency guarantees about its global or local maxima are currently known only under restrictive assumptions, e.g., for noiseless measurements and row orthogonality of the signal matrix $\mathbf{X}$ in [23], [24], for the SMV case and row orthogonality of the measurement matrix in [28], and under the assumption of prior knowledge of the true support size and source signal power in [12], [29], [30].

The main goal of the present work is to identify sufficient conditions for exact support recovery in the JSSR problem by a *correlation-aware procedure based on the M-SBL optimization*. This is accomplished by analyzing the support consistency of global maxima of a constrained variant of M-SBL's type-II maximum-likelihood optimization. Our newly derived sufficient conditions for exact recovery of k-sized supports in the JSSR problem confirm that the support error probability vanishes even for $k > m$ when the number of MMVs is large enough.

A. Existing results on support recovery using MMVs

Early theoretical works focussed on obtaining guarantees for a unique joint-sparse solution to the canonical ℓ_0 norm minimization problem:

$$L_0: \quad \min_{\mathbf{X} \in \mathbb{R}^{n \times L}} \|\mathbf{X}\|_0 \quad \text{s.t. } \mathbf{A}\mathbf{X} = \mathbf{Y}, \qquad (2)$$

where $\|\mathbf{X}\|_0$ denotes the number of nonzero rows in $\mathbf{X}$. In [13], [14], the authors showed that the L_0 problem admits a unique k-sparse solution provided $k < \lceil (\text{spark}(\mathbf{A}) - 1 + \text{rank}(\mathbf{Y}))/2 \rceil$, where spark($\mathbf{A}$) denotes the smallest integer p such that there exist p linearly dependent columns in $\mathbf{A}$. This result establishes that the SMV bottleneck of $k < m/2$ for ℓ_0 norm based support recovery can be overcome by using multiple measurement vectors. Furthermore, the sparsity bound suggests that supports of size $k < m$ are uniquely recoverable.

To circumvent the combinatorial hardness of the L_0 problem, [17] proposes to minimize the $\ell_{p,q}$ mixed-norm of $\mathbf{X}$ instead of the ℓ_0 norm. The $\ell_{p,q}$ norm of $\mathbf{X}$ is evaluated as $\|\mathbf{X}\|_{p,q} \triangleq \left(\sum_{i=1}^{m} \|\mathbf{X}(i,:)\|_p^q \right)^{1/q}$. Variants of the $\ell_{p,q}$ norm minimization problem with different combinations of p and q have been investigated independently in [13], [14], [31], [32]. For $p \geq 1, q = 1$, [14] has shown that $\ell_{p,q}$ norm minimization problem has a unique k-sparse solution, provided $\mathbf{A}$ satisfies $\|\mathbf{A}_{\mathcal{S}}^\dagger \mathbf{a}_j\|_1 < 1$, for all $j \notin \mathcal{S}$ and for all $\mathcal{S} \subset [n], |\mathcal{S}| \leq k$, where $\mathbf{A}_{\mathcal{S}}^\dagger = \left(\mathbf{A}_{\mathcal{S}}^T \mathbf{A}_{\mathcal{S}} \right)^{-1} \mathbf{A}_{\mathcal{S}}^T$. This also serves as a sufficient condition for exact support recovery via simultaneous orthogonal matching pursuit (SOMP) [15], a greedy support recovery algorithm. In [18], support recovery performance of various correlation based greedy and iterative hard-thresholding type algorithms is studied in the noiseless MMV setup. The sufficient conditions for exact support recovery are specified in terms of the asymmetric restricted isometry constants of the sensing matrix.

A common limitation of the aforementioned support recovery techniques is that they are capable of uniquely recovering supports of size up to only $k < m/2$. In [33], rank aware OMP and rank aware order recursive matching pursuit are shown to perfectly recover any k-sized support from noiseless measurements as long as $k < \text{spark}(\mathbf{A}) - 1$ and $\text{rank}(\mathbf{X}) = k$. For the rank defective case, i.e., $\text{rank}(\mathbf{X}) < k$, compressed sensing MUSIC [20] and subspace-augmented MUSIC [19] are still capable of recovering any $k < \text{spark}(\mathbf{A}) - 1$ sized support as long as partial support of size $k - \text{rank}(\mathbf{X})$ can be estimated by some other support recovery algorithm.

In [29], the MMV support recovery problem is formulated as a multiple hypothesis testing problem. Necessary and sufficient conditions for perfect support recovery with high probability are derived under the assumption that the columns of $\mathbf{X}$ are i.i.d. $\mathcal{N}(0, \text{diag}(\mathbf{1}_{\mathcal{S}^*}))$, where $\mathcal{S}^*$ denotes the true support set of a known size k. An exponential decay rate of the support error probability is derived in closed form, which is however not conducive to general interpretation or deeper analysis. For the particular case of randomly constructed Gaussian sensing matrix $\mathbf{A}$ with $m = \Omega\left(k \log \frac{n}{k}\right)$ rows, it is shown that $L \gg \frac{\log n}{\log \log n}$ suffices for diminishing support error probability with increasing L. One of our contributions in the present work is to extend this result to a more general signal prior on $\mathbf{X}$ and show that the support error probability vanishes even if m scales sublinearly in the support size k.

In [34], the support recovery problem is analyzed as a single-input-multi-output MAC communication problem. For number of nonzero rows fixed to k, $m = \Omega(\frac{k \log n}{c(\mathbf{X})})$ is shown to be both necessary and sufficient for successful support recovery as the problem size tends to infinity. The quantity $c(\mathbf{X})$ is a capacity like term that depends on the elements of the nonzero rows in $\mathbf{X}$ and the noise power. Even fewer measurements $m = \Omega\left(\frac{k}{L} \log n\right)$ suffices when each measurement vector is generated using an independently drawn sensing matrix [35]. The $m < k$ regime has been studied in [36], and it was shown that $L = \Theta\left(\frac{k^2}{m^2} \log(k(n-k))\right)$ is both necessary and sufficient to exactly recover the row-support of $\mathbf{X}$.

Different from bounding the support error probability in the JSSR problem, [37]–[39] use replica analysis to obtain a variational characterization of the minimum mean squared error incurred in reconstructing joint-sparse vectors from noisy MMVs, as the dimension n scales to infinity while keeping the measurement-rate m/n fixed. In addition, [37] shows that the replica analysis approach can be extended to obtain tight variational lower bounds for the mean weighted support set error in sparse supports reconstructed via the message passing algorithm.

B. Correlation-aware support recovery

A key insight was propounded in [22], that there often exists a latent structure in the MMV problem: *the nonzero rows of $\mathbf{X}$ are uncorrelated*. This signal structure can be enforced by modeling each column of $\mathbf{X}$ to be i.i.d. $\mathcal{N}(0, \text{diag}(\boldsymbol{\gamma}))$, where $\boldsymbol{\gamma} \in \mathbb{R}_+^n$ is a nonnegative vector of variance parameters. Under this source model, identifying the nonzero rows of $\mathbf{X}$ amounts to detecting the support of $\boldsymbol{\gamma}$. In [22], Co-LASSO is proposed to recover $\boldsymbol{\gamma}$. Instead of directly working with the linear observations $\mathbf{Y}$, Co-LASSO uses their covariance form, $\frac{1}{L}\mathbf{Y}\mathbf{Y}^T$, as input, and estimates $\boldsymbol{\gamma}$ as a solution of the following constrained ℓ_1-norm minimization problem:

$$\min_{\boldsymbol{\gamma} \in \mathbb{R}_+^n} \|\boldsymbol{\gamma}\|_1 \quad \text{s.t. } (\mathbf{A} \odot \mathbf{A})\boldsymbol{\gamma} = \text{vec}\left(\frac{1}{L}\mathbf{Y}\mathbf{Y}^T\right), \qquad (3)$$

where $\mathbf{A} \odot \mathbf{A}$ denotes the Khatri-Rao product (i.e., the columnwise Kronecker product) of $\mathbf{A}$ with itself. The linear constraints in (3) depict the second-order moment-matching constraints, specifically, the covariance matching equation:

$\frac{1}{L}\mathbf{Y}\mathbf{Y}^T \approx \mathbf{A}\,\text{diag}(\boldsymbol{\gamma})\mathbf{A}^T$. Since (3) comprises up to $(m^2 + m)/2$ linearly independent equations in $\boldsymbol{\gamma}$, sparsity levels as high as $O(m^2)$ are potentially recoverable. To recover the maximum level of sparsity, $k = (m^2 + m)/2$, a necessary condition derived in [22], [40] dictates that the columnwise self Khatri-Rao product matrix $\mathbf{A} \odot \mathbf{A}$ must have full Kruskal rank,[1] i.e., $\text{Krank}(\mathbf{A} \odot \mathbf{A}) = (m^2 + m)/2$. Another popular MMV algorithm, M-SBL [23], also imposes a common Gaussian prior $\mathcal{N}(0, \text{diag}(\boldsymbol{\gamma}))$ on the columns of $\mathbf{X}$ and hence implicitly exploits the latent uncorrelatedness of the nonzero entries in $\mathbf{X}$. Interestingly, similar to Co-LASSO, the performance of Sparse Bayesian Learning (SBL)-based support recovery methods depend on the null-space structure of the self Khatri-Rao product $\mathbf{A} \odot \mathbf{A}$. Making this connection explicit is one of the goals of this work.

The sparsistency of different types of constrained solutions of the M-SBL optimization has been earlier investigated in [23] for $k < m$, and in [12], [24] for $k \geq m$. Here, by the sparsistency of a support recovery method (or support estimate), we refer to its propensity to be a consistent estimator (or estimate) of the common nonzero support of the unknown joint sparse vectors, i.e., the support error probability decays to zero as $L \to \infty$. In [12], it is assumed that support size and the source signal powers are known a priori, while both [23] and [24] assume that the measurements are noiseless, and the nonzero rows of $\mathbf{X}$ are orthogonal. For finite L, the row-orthogonality condition is too restrictive for a deterministic $\mathbf{X}$ and almost never true for a random $\mathbf{X}$ drawn from a continuous distribution. In [41], necessary and sufficient conditions for amplitude-wise consistency of the M-SBL solution have been identified for $L \to \infty$. In contrast, we study the non-asymptotic case while taking the measurement noise into consideration, and also dispense with the restrictive row orthogonality condition on $\mathbf{X}$. Moreover, unlike [12], our analysis does not require the source signal powers to be known upfront.

C. Our contributions

1) We interpret the M-SBL optimization as a Bregman matrix divergence minimization problem; opening up new avenues to exploit the vast literature on Bregman divergence minimization towards devising faster, more robust algorithms for support recovery.

2) For the JSSR problem, we analyze the consistency of the estimated nonzero support inferred from the variance hyperparameters of a zero mean, row-sparsity inducing Gaussian prior on $\mathbf{X}$. These hyperparameters are estimated via a constrained type-II maximum likelihood (ML) procedure, called cM-SBL, wherein the nonzero coefficients of the ML estimate are constrained to lie inside a bounded interval. We show that the support error probability decays exponentially with number of MMVs, and the error exponent is related to the null space and restricted singular value properties of $\mathbf{A} \odot \mathbf{A}$, the self Khatri-Rao product of the sensing matrix $\mathbf{A}$

with itself. Explicit bounds on the number of MMVs sufficient for vanishing support error probability for both noisy and nearly noiseless measurements are derived (in Theorem 4). The support consistency guarantees obtained for the constrained type-II ML solution also apply to any global solution of the M-SBL optimization whose nonzero coefficients lie inside a known interval (see Corollary 4).

3) In the special case where the sensing matrix $\mathbf{A}$ is constructed from i.i.d. $\mathcal{N}\left(0, \frac{1}{m}\right)$ entries, we bound the MMV complexity for which the cM-SBL optimization exactly recovers the true k-sparse support using $m = \Omega(\sqrt{k}\log n)$ measurements per MMV. In the case of noiseless measurements, we show that M-SBL exactly recovers the true k-sparse support from a single measurement vector, provided $k < \text{spark}(\mathbf{A}) - 1$.

A key aspect of our results is that our sufficient conditions are expressed in terms of number of MMVs and properties of the sensing matrix $\mathbf{A}$. This makes our results applicable to both random as well as deterministic constructions of $\mathbf{A}$. As part of our analysis, we present a new lower bound for the $\frac{1}{2}$-Rényi divergence between a pair of multivariate Gaussian densities (in Proposition 1), and an interesting null space property of Khatri-Rao product matrices (in Theorem 3), which may be of independent interest.

The remainder of the paper is organized as follows. In section II, we formulate the JSSR problem and introduce our source model for $\mathbf{X}$. We also review the M-SBL algorithm [23] and interpret the M-SBL cost function as a Bregman matrix divergence. In section III, we cover some preliminary concepts that are used while analyzing the support recovery performance of the constrained and unconstrained variant of the SBL procedure. In section IV, we derive an abstract upper bound for the support error probability, which is used in section V to derive our main result, namely, the sufficient conditions for vanishing support error probability in M-SBL. In section VI, we discuss the implications of the new results in the context of several interesting special cases. Our final conclusions are presented in section VII.

D. Notation

Throughout this paper, scalar variables are denoted by lowercase alphabets and vectors are denoted by boldface lowercase alphabets. Matrices are denoted by boldface uppercase alphabets and calligraphic uppercase alphabets denote sets.

Given a vector $\mathbf{x}$, $\mathbf{x}(i)$ represents its ith entry. $\text{supp}(\mathbf{x})$ denotes the support of $\mathbf{x}$, the set of indices corresponding to nonzero entries in $\mathbf{x}$. Likewise, $\mathcal{R}(\mathbf{X})$ denotes the set of indices of all nonzero rows in $\mathbf{X}$ and is called the row-support of $\mathbf{X}$. For any $n \in \mathbb{N}$, $[n] \triangleq \{1, 2, \ldots, N\}$. For any n dimensional vector $\mathbf{x}$ and index set $\mathcal{S} \subseteq [n]$, the vector $\mathbf{x}_{\mathcal{S}}$ is an $|\mathcal{S}| \times 1$ sized vector retaining only those entries of $\mathbf{x}$ that are indexed by elements of $\mathcal{S}$. Likewise, $\mathbf{A}_{\mathcal{S}}$ is a submatrix comprising the columns of $\mathbf{A}$ indexed by $\mathcal{S}$. $\text{Null}(\mathbf{A})$ and $\text{Col}(\mathbf{A})$ denote the null space and column space of the matrix $\mathbf{A}$, respectively. The spectral, Frobenius and maximum absolute row sum matrix norms of $\mathbf{A}$ are denoted

---

[1]The Kruskal rank of an $m \times n$ matrix $\mathbf{A}$ is the largest integer k such that any k columns of $\mathbf{A}$ are linearly independent.

by $\|\mathbf{A}\|_2$, $\|\mathbf{A}\|_F$, and $\|\mathbf{A}\|_\infty$, respectively. $\mathbb{P}(\mathcal{E})$ denotes the probability of event $\mathcal{E}$. $\mathcal{N}(\mu, \boldsymbol{\Sigma})$ denotes the Gaussian probability density with mean μ and covariance matrix $\boldsymbol{\Sigma}$. For any square matrix $\mathbf{C}$, $\text{tr}(\mathbf{C})$ and $|\mathbf{C}|$ denote its trace and determinant, respectively. S_{++}^n denotes the set of all $n \times n$ positive definite matrices.

Given positive sequences $\{f_n\}_{n=1}^\infty$ and $\{g_n\}_{n=1}^\infty$, $g_n = O(f_n)$ denotes that there exists an $N \in \mathbb{N}$ and a universal constant $C > 0$ such that $g_n \leq Cf_n$ for $n \geq N$. Similarly, $g_n = \Omega(f_n)$ denotes that $g_n \geq cf_n$ for $n \geq N$ and some constant $c > 0$. Lastly, $f_n = \Theta(g_n)$ implies that $cf_n \leq g_n \leq Cf_n$ for $n \geq N$.

II. SYSTEM MODEL AND THE M-SBL ALGORITHM

A. Joint-sparse support recovery (JSSR)

We now formally state the JSSR problem. Suppose $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L$ are L distinct joint-sparse vectors in $\mathbb{R}^n$ with a common nonzero support denoted by the index set $\mathcal{S}^* \subseteq [n]$. Let K be the maximum size of the common support, i.e., $|\mathcal{S}^*| \leq K$. In JSSR, we are interested in recovering $\mathcal{S}^*$ from noisy underdetermined linear measurements $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_L$ generated as

$$\mathbf{y}_j = \mathbf{A}\mathbf{x}_j + \mathbf{w}_j, \quad 1 \leq j \leq L. \quad (4)$$

We assume that the sensing matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ is a non-degenerate matrix with $m \leq n$, i.e., any m columns of $\mathbf{A}$ are linearly independent, or $\text{spark}(\mathbf{A}) = m + 1$. The noise vector $\mathbf{w} \in \mathbb{R}^m$ is zero mean Gaussian distributed with diagonal covariance matrix $\sigma^2 \mathbf{I}_m$. The linear measurement model in (4) can be rewritten in a compact MMV form as $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{W}$, where $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_L]$, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L]$ and $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_L]$ are the observation, signal and noise matrices, respectively. Since the columns of $\mathbf{X}$ are jointly sparse with common support $\mathcal{S}^*$, $\mathbf{X}$ is a row sparse matrix with row support $\mathcal{R}(\mathbf{X}) = \mathcal{S}^*$.

B. Gaussian source assumption

We assume that if the ith row of the unknown signal matrix $\mathbf{X}$ is nonzero, then it is a Gaussian ensemble of L i.i.d. zero mean random variables with a common variance $\gamma^*(i)$ which lies in the interval $[\gamma_{\min}, \gamma_{\max}]$. We refer to this as **Assumption (A1)**. An immediate consequence of **(A1)** is that there exists a bounded, nonnegative, and at most K sparse vector, $\gamma^* \in \mathbb{R}_+^n$, such that the columns $\mathbf{x}_j$ are i.i.d. $\mathcal{N}(0, \boldsymbol{\Gamma}^*)$ with $\boldsymbol{\Gamma}^* \triangleq \text{diag}(\gamma^*)$. Furthermore, $\mathcal{R}(\mathbf{X})$ and $\text{supp}(\gamma^*)$ are the same and both are equal to $\mathcal{S}^*$.

C. Multiple sparse Bayesian learning (M-SBL)

We now review M-SBL [23], a type-II maximum likelihood (ML) procedure for estimating joint-sparse signals from compressive linear measurements. M-SBL models the columns of $\mathbf{X}$ as i.i.d. $\mathcal{N}(0, \boldsymbol{\Gamma})$, where $\boldsymbol{\Gamma} = \text{diag}(\gamma)$, and $\gamma = [\gamma(1), \gamma(2), \ldots, \gamma(n)]^T$ is an $n \times 1$ vector of unknown nonnegative variance parameters. The elements of γ are collectively called hyperparameters as they represent the parameters of the signal prior. Since the hyperparameter $\gamma(i)$

models the common variance of the ith row of $\mathbf{X}$, if $\gamma(i) = 0$, it drives the posterior estimate of $\mathbf{x}_j(i)$ to zero for $1 \leq j \leq L$. Consequently, if γ is estimated to be a sparse vector, it induces joint sparsity in $\mathbf{X}$.

In M-SBL, the hyperparameter vector γ is chosen to maximize the Bayesian evidence $p(\mathbf{Y}; \gamma)$, which is tantamount to finding the ML estimate of γ. Let $\hat{\gamma}_{\text{ML}}$ denote the ML estimate of γ, i.e.,

$$\hat{\gamma} = \arg \max_{\gamma \in \mathbb{R}_+^n} \log p(\mathbf{Y}; \gamma). \quad (5)$$

The Gaussian prior on $\mathbf{x}_j$ combined with the linear measurement model induces Gaussian observations, i.e., $p(\mathbf{y}_j; \gamma) = \mathcal{N}(0, \sigma^2 \mathbf{I}_m + \mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^T)$. For a fixed γ, the MMVs $\mathbf{y}_j$ are mutually independent. Hence, it follows that

$$\begin{aligned} \log p(\mathbf{Y}; \gamma) &= \sum_{j=1}^L \log p(\mathbf{y}_j; \gamma) \\ &\propto -L \log |\boldsymbol{\Sigma}_\gamma| - \text{Tr}\left(\boldsymbol{\Sigma}_\gamma^{-1} \mathbf{Y}\mathbf{Y}^T\right), \quad (6) \end{aligned}$$

where $\boldsymbol{\Sigma}_\gamma = \sigma^2 \mathbf{I}_m + \mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^T$. The log likelihood $\log p(\mathbf{Y}; \gamma)$ in (6) is a nonconvex function of γ and its global maximizer $\hat{\gamma}$ is not available in closed form. However, its local maximizers can still be found via fixed point iterations or the Expectation-Maximization (EM) procedure. In [23], it is empirically shown that the EM procedure faithfully recovers the true support $\mathcal{S}^*$, provided m and L are sufficiently large.

D. The M-SBL objective is a Bregman matrix divergence

We now present an interesting interpretation of M-SBL's log-marginalized likelihood objective in (6) which facilitates a deeper understanding of what is accomplished by its maximization. We begin by introducing the Bregman matrix divergence $\mathcal{D}_\varphi(\mathbf{X}, \mathbf{Y})$ between any two $n \times n$ positive definite matrices $\mathbf{X}$ and $\mathbf{Y}$ as

$$\mathcal{D}_\varphi(\mathbf{X}, \mathbf{Y}) \triangleq \varphi(\mathbf{X}) - \varphi(\mathbf{Y}) - \langle \nabla\varphi(\mathbf{Y}), \mathbf{X} - \mathbf{Y} \rangle, \quad (7)$$

where $\varphi : S_{++}^n \to \mathbb{R}$ is a convex function with $\nabla\varphi(\mathbf{Y})$ as its first order derivative evaluated at $\mathbf{Y}$. In (7), the matrix inner product $\langle \mathbf{X}, \mathbf{Y} \rangle$ is evaluated as $\text{tr}\left(\mathbf{X}\mathbf{Y}^T\right)$. For the specific case of $\varphi(\cdot) = -\log|\cdot|$, a strongly convex function, we obtain the Bregman LogDet matrix divergence given by

$$\mathcal{D}_{\text{logdet}}(\mathbf{X}, \mathbf{Y}) = \text{tr}\left(\mathbf{X}\mathbf{Y}^{-1}\right) - \log\left|\mathbf{X}\mathbf{Y}^{-1}\right| - n. \quad (8)$$

By termwise comparison of (6) and (8), we observe that the negative log likelihood $-\log p(\mathbf{Y}; \gamma)$ and $\mathcal{D}_{\text{logdet}}(\mathbf{R}_\mathbf{Y}, \boldsymbol{\Sigma}_\gamma)$ are the same up to a constant. In fact, in [42, Theorem 6], it is shown that there is a one-to-one correspondence between every regular exponential family of probability distributions and a unique and distinct Bregman divergence.

In the divergence term $\mathcal{D}_{\text{logdet}}(\mathbf{R}_\mathbf{Y}, \boldsymbol{\Sigma}_\gamma)$, the first argument $\mathbf{R}_\mathbf{Y} \triangleq \frac{1}{L}\mathbf{Y}\mathbf{Y}^T$ is the sample covariance matrix of the observations $\mathbf{Y}$ and the second argument $\boldsymbol{\Sigma}_\gamma = \sigma^2 \mathbf{I} + \mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^T$ is the parameterized covariance matrix of $\mathbf{Y}$. This connection between M-SBL's log likelihood cost and the LogDet divergence reveals that by maximizing the M-SBL cost, we seek a γ that minimizes the distance between $\mathbf{R}_\mathbf{Y}$ and $\boldsymbol{\Sigma}_\gamma$, with point wise distances measured using the Bregman LogDet divergence.

Thus, the M-SBL algorithm, at its core, is essentially a *second order moment matching* or *covariance matching* procedure for finding γ such that the associated covariance matrix Σ_γ is closest to the sample covariance matrix, in the Bregman LogDet divergence sense.

This new interpretation of the M-SBL cost as a Bregman matrix divergence elicits two interesting questions:

i Are there other matrix divergences besides LogDet Bregman matrix divergence which are better suited for covariance matching?

ii How to exploit the structural similarities between the M-SBL cost and the Bregman (LogDet) matrix divergence to devise faster and more robust techniques for the type-II likelihood maximization?

It is our opinion that exploring the use of other matrix divergences for covariance matching is worth further investigation with the potential for new, improved algorithms for support recovery. Preliminary results in this direction have been quite encouraging. For example, in [25], an α-Rényi divergence objective is considered for covariance matching, and a fast greedy algorithm is developed for joint-sparse support recovery.

III. SOME PRELIMINARY CONCEPTS

In this section, we review a few key definitions and results which will be used in the later sections.

A. ϵ-Cover, ϵ-net and covering number

Suppose $\mathcal{T}$ is a set equipped with a pseudo-metric d. For any set $\mathcal{A} \subseteq \mathcal{T}$, its ϵ-cover is defined as the coverage of $\mathcal{A}$ with open balls of radius ϵ and centers in $\mathcal{T}$. The set $\mathcal{A}^\epsilon$ comprising the centers of these covering balls is called an ϵ-net of $\mathcal{A}$. The minimum number of ϵ-balls that can cover $\mathcal{A}$ is called the ϵ-covering number of $\mathcal{A}$, and is given by

$$N_{\text{cov}}^\epsilon (\mathcal{A}, d) = \min \{|\mathcal{A}^\epsilon| : \mathcal{A}^\epsilon \text{ is an } \epsilon\text{-net of } \mathcal{A}\}.$$

In computational theory of learning, ϵ-net constructs are often useful in converting a union over the elements of a continuous set to a finite sized union.

B. Rényi divergence between multi-variate Gaussian distributions

For $p_1 = \mathcal{N}(0, \Sigma_1)$ and $p_2 = \mathcal{N}(0, \Sigma_2)$, the $\frac{1}{2}$-Rényi divergence $\mathcal{D}_{\frac{1}{2}}(p_1 || p_2)$ is available in closed form as [43]

$$\mathcal{D}_{\frac{1}{2}}(p_1, p_2) = \log \left| \frac{\Sigma_1 + \Sigma_2}{2} \right| - \frac{1}{2} \log |\Sigma_1| - \frac{1}{2} \log |\Sigma_2|. \quad (9)$$

Proposition 1. *Let p_1 and p_2 be two multivariate Gaussian distributions with zero mean and positive definite covariance matrices Σ_1 and Σ_2, respectively. Then, the $\frac{1}{2}$-Rényi divergence between p_1 and p_2 is bounded as*

$$\mathcal{D}_{\frac{1}{2}}(p_1, p_2) \geq \frac{1}{2} tr \left((\Sigma_1 - \Sigma_2)(\Sigma_1 + \Sigma_2)^{-1}(\Sigma_1 - \Sigma_2)(\Sigma_1 + \Sigma_2)^{-1} \right).$$

Proof. See Appendix A. □

C. Concentration of sample covariance matrix

Proposition 2 (Vershynin [44]). *Let $y_1, y_2, \ldots, y_L \in \mathbb{R}^m$ be L independent samples from $\mathcal{N}(0, \Sigma)$, and let $\Sigma_L = \frac{1}{L} \sum_{j=1}^L y_j y_j^T$ denote the sample covariance matrix. Then, for any $\epsilon > 0$,*

$$\||\Sigma_L - \Sigma\||_2 \leq \epsilon \||\Sigma\||_2, \quad (10)$$

holds with probability exceeding $1 - \delta$ provided $L \geq \frac{C}{\epsilon^2} \log \frac{2}{\delta}$, C being an absolute constant.

D. Spectral norm bounds for Gaussian matrices

Proposition 3 (Corollary 5.35 in [44]). *Let A be an $m \times n$ matrix whose entries are i.i.d. $\mathcal{N}(0, 1)$. Then for every $t \geq 0$,*

$$\||A\||_2 \leq \sqrt{m} + \sqrt{n} + t,$$

with probability at least $1 - 2e^{-t^2/2}$.

The following corollary presents a probabilistic bound for the spectral norm of a submatrix of an $m \times n$ sized Gaussian matrix obtained by sampling its columns.

Corollary 1. *Let A be an $m \times n$ sized matrix whose entries are i.i.d. $\mathcal{N}(0, 1)$. Then, for any $\mathcal{S} \subseteq [n], |\mathcal{S}| \leq k$, the submatrix $A_\mathcal{S}$ obtained by sampling the columns of A indexed by $\mathcal{S}$ satisfies*

$$\||A_\mathcal{S}\||_2 \leq \sqrt{m} + \sqrt{k} + \sqrt{6k \log n}.$$

with probability exceeding $1 - 2n^{-k}$.

Proof. See Appendix B. □

IV. SUPPORT ERROR ANALYSIS

Towards studying the sparsistency of the M-SBL solution, we first derive a Chernoff bound for the support error probability incurred by the solution of a constrained version of M-SBL optimization under assumption **A1**. We begin by introducing some of the frequently used notation in the table below.

$\mathcal{S}^*$	True row support of X.
γ^*	Principal diagonal of the common covariance matrix Γ^* of the i.i.d. columns in X. Consequently, $\text{supp}(\gamma^*) = \mathcal{S}^*$.
K	Maximum number of nonzero rows in X
$\mathcal{S}_k$	The collection of all support sets of k or lesser size, i.e., $\mathcal{S}_k = \{\mathcal{S} \subseteq [n], |\mathcal{S}| \leq k\}$
$\Theta(\mathcal{S})$	Bounded hyperparameter set associated with the support set $\mathcal{S}$, formally defined as $\Theta(\mathcal{S}) \triangleq \{\gamma \in \mathbb{R}_+^n : \text{supp}(\gamma) = \mathcal{S}, \gamma_{\min} \preceq \gamma_\mathcal{S} \preceq \gamma_{\max}\}$.
Θ_k	The collection of all k or less sparse vectors in $\mathbb{R}_+^n$ with nonzero coefficients lying in $[\gamma_{\min}, \gamma_{\max}]$. By definition, we have $\Theta_k = \bigcup_{\mathcal{S} \in \mathcal{S}_k} \Theta(\mathcal{S})$.

By assumption **A1** on X, γ^* belongs to the bounded parameter set Θ_n. Therefore, in order to estimate γ^* from the measurement vectors Y, we consider solving a constrained variant of the M-SBL optimization in (5), which we refer to as the cM-SBL problem:

$$\text{cM-SBL:} \quad \hat{\gamma} = \arg \max_{\gamma \in \Theta_n} \mathcal{L}(Y; \gamma). \quad (11)$$

The cM-SBL objective $\mathcal{L}(Y; \gamma)$ is the same as the M-SBL's log-marginalized likelihood $\log p(Y; \gamma)$ defined in (6). The

row support of $\mathbf{X}$ is estimated as the support of $\hat{\boldsymbol{\gamma}}$, where $\hat{\boldsymbol{\gamma}}$ is a solution of (11). Consider the set of "bad" MMVs,

$$\mathcal{E}_{\mathcal{S}^*} \triangleq \left\{ \mathbf{Y} \in \mathbb{R}^{m \times L} : \text{supp}\,(\hat{\boldsymbol{\gamma}}) \neq \mathcal{S}^* \right\}, \quad (12)$$

which result in erroneous estimation of $\mathcal{S}^*$ via cM-SBL. In other words, $\mathcal{E}_{\mathcal{S}^*}$ is the collection of undesired MMVs for which the M-SBL objective is globally maximized by some $\boldsymbol{\gamma} \in \Theta(\mathcal{S})$, $\mathcal{S} \neq \mathcal{S}^*$, i.e.,

$$\mathcal{E}_{\mathcal{S}^*} = \bigcup_{\mathcal{S} \in \mathcal{S}_n \setminus \{\mathcal{S}^*\}} \left\{ \mathbf{Y} : \max_{\boldsymbol{\gamma} \in \Theta(\mathcal{S})} \mathcal{L}(\mathbf{Y};\boldsymbol{\gamma}) \geq \max_{\boldsymbol{\gamma}' \in \Theta(\mathcal{S}^*)} \mathcal{L}(\mathbf{Y};\boldsymbol{\gamma}') \right\}. \quad (13)$$

We are interested in finding the conditions under which $\mathbb{P}(\mathcal{E}_{\mathcal{S}^*})$ can be made arbitrarily small. Since $\max\limits_{\boldsymbol{\gamma}' \in \Theta(\mathcal{S}^*)} \mathcal{L}(\mathbf{Y};\boldsymbol{\gamma}') \geq \mathcal{L}(\mathbf{Y};\boldsymbol{\gamma}^*)$, it follows that

$$\mathcal{E}_{\mathcal{S}^*} \subseteq \bigcup_{\mathcal{S} \in \mathcal{S}_n \setminus \{\mathcal{S}^*\}} \left\{ \mathbf{Y} : \max_{\boldsymbol{\gamma} \in \Theta(\mathcal{S})} \mathcal{L}(\mathbf{Y};\boldsymbol{\gamma}) \geq \mathcal{L}(\mathbf{Y};\boldsymbol{\gamma}^*) \right\}$$
$$= \bigcup_{\mathcal{S} \in \mathcal{S}_n \setminus \{\mathcal{S}^*\}} \bigcup_{\boldsymbol{\gamma} \in \Theta(\mathcal{S})} \left\{ \mathbf{Y} : \mathcal{L}(\mathbf{Y};\boldsymbol{\gamma}) - \mathcal{L}(\mathbf{Y};\boldsymbol{\gamma}^*) \geq 0 \right\}. \quad (14)$$

The continuous union over infinitely many elements of $\Theta(\mathcal{S})$ in (14) can be relaxed to a finite sized union by using the following ϵ-net argument. Consider $\Theta^\epsilon(\mathcal{S})$, a finite sized ϵ-net of the hyperparameter set $\Theta(\mathcal{S})$, such that for any $\boldsymbol{\gamma} \in \Theta(\mathcal{S})$, there exists an element $\boldsymbol{\gamma}' \in \Theta^\epsilon(\mathcal{S})$ satisfying $|\mathcal{L}(\mathbf{Y};\boldsymbol{\gamma}) - \mathcal{L}(\mathbf{Y};\boldsymbol{\gamma}')| \leq \epsilon$. Proposition 4 gives an upper bound on the size of such an ϵ-net.

Proposition 4. *Given a support set $\mathcal{S} \subseteq [n]$, there exists a finite set $\Theta^\epsilon(\mathcal{S}) \subset \Theta(\mathcal{S})$ such that it simultaneously satisfies*

(i) For any $\boldsymbol{\gamma} \in \Theta(\mathcal{S})$, there exists a $\boldsymbol{\gamma}' \in \Theta^\epsilon(\mathcal{S})$ such that $|\mathcal{L}(\mathbf{Y};\boldsymbol{\gamma}) - \mathcal{L}(\mathbf{Y};\boldsymbol{\gamma}')| \leq \epsilon$.

(ii) $|\Theta^\epsilon(\mathcal{S})| \leq \max\left\{ 1, \left(3C_{\mathcal{L},\mathcal{S}}(\gamma_{max} - \gamma_{min})\sqrt{|\mathcal{S}|}/\epsilon \right)^{|\mathcal{S}|} \right\}$, where $C_{\mathcal{L},\mathcal{S}}$ is the Lipschitz constant of $\mathcal{L}(\mathbf{Y};\boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$ in the bounded domain $\Theta(\mathcal{S})$.

The set $\Theta^\epsilon(\mathcal{S})$ is an ϵ-net of $\Theta(\mathcal{S})$.

Proof. See Appendix C. $\square$

From Proposition 4-(ii), we observe that both the construction as well as size of $\Theta^\epsilon(\mathcal{S})$ depends on the Lipschitz continuity of the log-likelihood $\mathcal{L}(\mathbf{Y};\boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$. By virtue of data-dependent nature of $\mathcal{L}(\mathbf{Y};\boldsymbol{\gamma})$, its Lipschitz constant $C_{\mathcal{L},\mathcal{S}}$ depends on the instantaneous value of $\mathbf{Y}$. To make the rest of the analysis independent of $\mathbf{Y}$, we introduce a new MMV set $\mathcal{G}$, conditioned on which, the Lipschitz constant $C_{\mathcal{L},\mathcal{S}}$ is uniformly bounded solely in terms of second-order statistics of $\mathbf{Y}$. A possible choice of $\mathcal{G}$ could be

$$\mathcal{G} \triangleq \left\{ \mathbf{Y} \subset \mathbb{R}^{m \times L} : \left\| \frac{1}{L}\mathbf{Y}\mathbf{Y}^T \right\|_2 \leq 2 \left\| \mathbb{E}\left[\mathbf{y}_1 \mathbf{y}_1^T \right] \right\|_2 \right\}. \quad (15)$$

By Proposition 9 in Appendix I, for $\mathbf{Y} \in \mathcal{G}$, $\mathcal{L}(\mathbf{Y};\boldsymbol{\gamma})$ is uniformly continuous with a Lipschitz constant that depends only on the spectral norm of $\mathbb{E}[\mathbf{y}_1 \mathbf{y}_1^T]$. Hence, the ϵ-net can now be constructed entirely independent of $\mathbf{Y}$. We denote this ϵ-net by $\Theta^\epsilon(\mathcal{S})|_{\mathcal{G}}$.

Since for arbitrary sets $\mathcal{A}$ and $\mathcal{B}$, $\mathcal{A} \subseteq (\mathcal{A} \cap \mathcal{B}) \cup \mathcal{B}^c$, the RHS in (14) relaxes as

$$\mathcal{E}_{\mathcal{S}^*} \subseteq \left\{ \bigcup_{\mathcal{S} \in \mathcal{S}_n \setminus \mathcal{S}^*} \bigcup_{\boldsymbol{\gamma} \in \Theta(\mathcal{S})} \left\{ \mathcal{L}(\mathbf{Y};\boldsymbol{\gamma}) - \mathcal{L}(\mathbf{Y};\boldsymbol{\gamma}^*) \geq 0 \right\} \cap \mathcal{G} \right\} \cup \mathcal{G}^c. \quad (16)$$

The continuous union over $\Theta(\mathcal{S})$ relaxes to a finite sized union over $\Theta^\epsilon(\mathcal{S})|_{\mathcal{G}}$ as shown below.

$$\mathcal{E}_{\mathcal{S}^*} \subseteq \left\{ \bigcup_{\mathcal{S} \in \mathcal{S}_n \setminus \mathcal{S}^*} \bigcup_{\boldsymbol{\gamma} \in \Theta^\epsilon(\mathcal{S})|_{\mathcal{G}}} \left\{ \mathcal{L}(\mathbf{Y};\boldsymbol{\gamma}) - \mathcal{L}(\mathbf{Y};\boldsymbol{\gamma}^*) \geq -\epsilon \right\} \cap \mathcal{G} \right\} \cup \mathcal{G}^c$$
$$\subseteq \left\{ \bigcup_{\mathcal{S} \in \mathcal{S}_n \setminus \mathcal{S}^*} \bigcup_{\boldsymbol{\gamma} \in \Theta^\epsilon(\mathcal{S})|_{\mathcal{G}}} \left\{ \mathcal{L}(\mathbf{Y};\boldsymbol{\gamma}) - \mathcal{L}(\mathbf{Y};\boldsymbol{\gamma}^*) \geq -\epsilon \right\} \right\} \cup \mathcal{G}^c.$$

By applying the union bound, we obtain

$$\mathbb{P}\left(\mathcal{E}_{\mathcal{S}^*}\right) \leq \sum_{\mathcal{S} \in \mathcal{S}_n \setminus \mathcal{S}^*} \sum_{\boldsymbol{\gamma} \in \Theta^\epsilon(\mathcal{S})|_{\mathcal{G}}} \mathbb{P}\left(\mathcal{L}(\mathbf{Y};\boldsymbol{\gamma}) - \mathcal{L}(\mathbf{Y};\boldsymbol{\gamma}^*) \geq -\epsilon \right)$$
$$+ \mathbb{P}\left(\mathcal{G}^c\right). \quad (17)$$

From (17), the support error probability $\mathbb{P}(\mathcal{E}_{\mathcal{S}^*})$ will be small when the summands, $\mathbb{P}\left(\mathcal{L}(\mathbf{Y};\boldsymbol{\gamma}) - \mathcal{L}(\mathbf{Y};\boldsymbol{\gamma}^*) \geq -\epsilon \right)$, $\boldsymbol{\gamma} \in \Theta^\epsilon(\mathcal{S})|_{\mathcal{G}}$, are individually sufficiently small so that their collective contribution remains small, and $\mathbb{P}(\mathcal{G}^c)$ is also small. In Theorem 1, we show that each summand corresponds to a large deviation event which occurs with an exponentially decaying probability.

Theorem 1. *For $\boldsymbol{\gamma} \in \mathbb{R}_+^n$, let $p_{\boldsymbol{\gamma}}$ denote the marginal probability density of the columns of $\mathbf{Y}$ when the joint-sparse columns of $\mathbf{X}$ are drawn independently from $\mathcal{N}\left(0, \text{diag}(\boldsymbol{\gamma})\right)$. Then, the log-likelihood $\mathcal{L}(\mathbf{Y};\boldsymbol{\gamma}) = \sum_{j=1}^{L} \log p_{\boldsymbol{\gamma}}(\mathbf{y}_j)$ satisfies the following large deviation property.*

$$\mathbb{P}\left(\mathcal{L}(\mathbf{Y};\boldsymbol{\gamma}) - \mathcal{L}(\mathbf{Y};\boldsymbol{\gamma}^*) \geq -\epsilon \right) \leq \exp\left(-L\psi^*\left(-\frac{\epsilon}{L} \right) \right), \quad (18)$$

where $\psi^*(\cdot)$ is the Legendre transform[2] of $\psi(t) \triangleq (t - 1)\mathcal{D}_t(p_{\boldsymbol{\gamma}}, p_{\boldsymbol{\gamma}^*})$, and $\mathcal{D}_t$ is the t-Rényi divergence[3] (of order $t > 0$) between the probability densities $p_{\boldsymbol{\gamma}}$ and $p_{\boldsymbol{\gamma}^*}$.

Proof. See Appendix D. $\square$

Note that, when the measurement noise is Gaussian, the marginal density $p_{\boldsymbol{\gamma}}(\mathbf{y}_j)$ of the individual observations $\mathbf{y}_j$ is also Gaussian with zero mean and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} = \sigma^2 \mathbf{I}_m + \mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^T$. If $\sigma^2 > 0$, both marginals $p_{\boldsymbol{\gamma}}$ and $p_{\boldsymbol{\gamma}^*}$ are non-degenerate and hence the Rényi divergence $\mathcal{D}_t(p_{\boldsymbol{\gamma}}, p_{\boldsymbol{\gamma}^*})$ in Theorem 1 is well defined. We now restate Theorem 1 as

---

[2]For any convex function $f : \mathcal{X} \to \mathbb{R}$ on a convex set $\mathcal{X} \subseteq \mathbb{R}^n$, its Legendre transform is the function f^* defined by

$$f^*(\mathbf{z}) = \sup_{\mathbf{x} \in \mathcal{X}} \left(\langle \mathbf{z}, \mathbf{x} \rangle - f(\mathbf{x}) \right).$$

[3]Let $(\mathcal{X}, \mathcal{F})$ be a measurable space and P and Q be two probability measures on $\mathcal{F}$ with densities p and q, respectively, with respect to the dominating Lebesgue measure μ on $\mathcal{F}$. Then, for $\alpha \in \mathbb{R}^+ \setminus 1$, the *Rényi divergence* of order α between P and Q, denoted $\mathcal{D}_\alpha(p||q)$, is defined as

$$\mathcal{D}_\alpha(p||q) = \frac{1}{\alpha - 1} \log \int_{\mathcal{X}} p(x)^\alpha q(x)^{1-\alpha} \mu(dx).$$

Corollary 2, which is the final form of the large deviation result for $\mathcal{L}(\mathbf{Y}; \boldsymbol{\gamma})$ used later for bounding $\mathbb{P}(\mathcal{E}_{\mathcal{S}^*})$.

Corollary 2. *For any $\boldsymbol{\gamma} \in \mathbb{R}_+^n$, and the true variance parameters $\boldsymbol{\gamma}^*$, let the associated marginal densities $p_{\boldsymbol{\gamma}}$ and $p_{\boldsymbol{\gamma}^*}$ be defined as in Theorem 1, and suppose $\sigma^2 > 0$. Then, the log-likelihood $\mathcal{L}(\mathbf{Y}; \boldsymbol{\gamma})$ satisfies the large deviation property*

$$
\mathbb{P}\left(\mathcal{L}(\mathbf{Y};\boldsymbol{\gamma}) - \mathcal{L}(\mathbf{Y};\boldsymbol{\gamma}^*) \geq -\frac{L\mathcal{D}_{\frac{1}{2}}(p_{\boldsymbol{\gamma}}, p_{\boldsymbol{\gamma}^*})}{2}\right) \leq e^{-\frac{L\mathcal{D}_{\frac{1}{2}}(p_{\boldsymbol{\gamma}}, p_{\boldsymbol{\gamma}^*})}{4}}.
$$
(19)

Proof. The large deviation result is obtained by replacing $\psi^*\left(-\frac{\epsilon}{L}\right)$ in Theorem 1 by its lower bound $-\frac{t\epsilon}{L} - \psi(t)$, followed by setting $t = 1/2$ and $\epsilon = L\mathcal{D}_{\frac{1}{2}}(p_{\boldsymbol{\gamma}}, p_{\boldsymbol{\gamma}^*})/2$. $\square$

Note that, in the above, we have used the suboptimal choice $t = 1/2$ for the Chernoff parameter t, since its optimal value is not available in closed form. However, this suboptimal selection of t is inconsequential as it figures only as a multiplicative factor in the final MMV complexity. By using Corollary 2 in (17), we can bound $\mathbb{P}(\mathcal{E}_{\mathcal{S}^*})$ as

$$
\mathbb{P}(\mathcal{E}_{\mathcal{S}^*}) \leq \sum_{\mathcal{S} \in \mathcal{S}_n \setminus \mathcal{S}^*} |\Theta^\epsilon(\mathcal{S})|_{\mathcal{G}}| \exp\left(-\frac{L\mathcal{D}_{\mathcal{S}}^*}{4}\right) + \mathbb{P}(\mathcal{G}^c), \quad (20)
$$

with $\epsilon = \frac{L\mathcal{D}_{\mathcal{S}}^*}{2}$ and $\mathcal{D}_{\mathcal{S}}^*$ defined as

$$
\mathcal{D}_{\mathcal{S}}^* \triangleq \inf_{\boldsymbol{\gamma} \in \Theta(\mathcal{S})} \mathcal{D}_{\frac{1}{2}}(p_{\boldsymbol{\gamma}}, p_{\boldsymbol{\gamma}^*}). \quad (21)
$$

Suppose the support $\mathcal{S}$ differs from $\mathcal{S}^*$ in exactly $k_d^{\mathcal{S}, \mathcal{S}^*}$ locations, then

$$
\mathbb{P}(\mathcal{E}_{\mathcal{S}^*}) \leq \sum_{\mathcal{S} \in \mathcal{S}_n \setminus \mathcal{S}^*} \exp\left(-L k_d^{\mathcal{S}, \mathcal{S}^*}\left(\frac{\eta}{4} - \frac{\kappa_{\text{cov}}}{L}\right)\right) + \mathbb{P}(\mathcal{G}^c),
$$
(22)

where

$$
\eta \triangleq \min_{\mathcal{S} \in \mathcal{S}_n \setminus \mathcal{S}^*} \frac{\mathcal{D}_{\mathcal{S}}^*}{k_d^{\mathcal{S}, \mathcal{S}^*}}, \quad (23)
$$

$$
\kappa_{\text{cov}} \triangleq \max_{\mathcal{S} \in \mathcal{S}_n \setminus \mathcal{S}^*} \frac{\log\left|\Theta^\epsilon(\mathcal{S})|_{\mathcal{G}}\right|}{k_d^{\mathcal{S}, \mathcal{S}^*}}, \quad \epsilon = \frac{L\mathcal{D}_{\mathcal{S}}^*}{2}. \quad (24)
$$

Using the above, we can state the following theorem.

Theorem 2. *Suppose $\mathcal{S}^*$ is the true row support of the unknown $\mathbf{X}$ satisfying assumption A1 and $|\mathcal{S}^*| \leq K$. Then, for any $\delta \in (0, 1)$, $\mathbb{P}(\mathcal{E}_{\mathcal{S}^*}) \leq 2\delta$, if*

$$
L \geq \max\left\{\frac{8}{\eta}\log\left(\frac{6enK}{\delta}\right), \frac{8\kappa_{cov}}{\eta}, C\log\frac{2}{\delta}\right\}.
$$

Here, η and κ_{cov} are as defined in (23) and (24), respectively, and $C > 0$ is a universal numerical constant.

Proof. See Appendix E. $\square$

In Theorem 2, we finally have an abstract bound on the sufficient number of MMVs, L, which guarantees vanishing support error probability in cM-SBL, given that the true support is $\mathcal{S}^*$. However, the MMV bound is meaningful only when η (23) is strictly positive. We now proceed to deduce the conditions for which

1) $\eta > 0$,

2) η and κ_{cov} scale favorably with the MMV problem dimensions.

A. Bounds for η and κ_{cov}

To understand how small η in the MMV bound in Theorem 2 can be, we first derive a lower bound on $\mathcal{D}_{\mathcal{S}}^*$ for any $\mathcal{S} \subseteq [n]$, in the following proposition.

Proposition 5. *Let $p_{\boldsymbol{\gamma}}$ denote the parameterized multivariate Gaussian density with zero mean and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} = \sigma^2 \mathbf{I} + \mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^T$, $\boldsymbol{\Gamma} = \text{diag}(\boldsymbol{\gamma})$. For any pair $\boldsymbol{\gamma}, \boldsymbol{\gamma}^* \in \mathbb{R}_+^n$ such that $\mathcal{S} = supp(\boldsymbol{\gamma})$ and $\mathcal{S}^* = supp(\boldsymbol{\gamma}^*)$, the $\frac{1}{2}$-Rényi divergence between $p_{\boldsymbol{\gamma}}$ and $p_{\boldsymbol{\gamma}^*}$ satisfies*

$$
\mathcal{D}_{\frac{1}{2}}(p_{\boldsymbol{\gamma}}, p_{\boldsymbol{\gamma}^*}) \geq \frac{\|(\mathbf{A} \odot \mathbf{A})(\boldsymbol{\gamma} - \boldsymbol{\gamma}^*)\|_2^2}{4\left(\sigma^2 + \gamma_{max}\sigma_{max}^2(\mathbf{A}_{\mathcal{S} \cup \mathcal{S}^*})\right)^2},
$$

where $\mathbf{A} \odot \mathbf{A}$ denotes the columnwise Khatri-Rao product of $\mathbf{A}$ with itself and $\sigma_{max}(\cdot)$ denotes maximum singular value of the input matrix.

Proof. See Appendix F. $\square$

From Proposition 5, it can be observed that as long as the null space of $\mathbf{A} \odot \mathbf{A}$ is devoid of any vectors of the form $\boldsymbol{\gamma} - \boldsymbol{\gamma}^*$ (i.e., difference of a nonnegative vector and a nonnegative K-sparse vector), then η as defined in (23) is always strictly positive. This condition for strictly positive η can be formalized as the *nonnegative restricted null space property* of $\mathbf{A} \odot \mathbf{A}$, defined next.

Definition 1. *A matrix is said to satisfy the nonnegative restricted null space property (NN-RNSP) of order k if its null space does not contain any vectors that are expressible as the difference between a k (or lesser) sparse nonnegative vector and an arbitrary nonnegative vector.*

The requirement that $\mathbf{A} \odot \mathbf{A}$ satisfies a restricted null-space property similar to Definition 1 has been highlighted in [8] in the context of MMV-based support recovery using correlation-aware priors. In [8], it is shown that $\mathbf{A} \odot \mathbf{A}$ exhibits the NN-RNSP for appropriate sparse-array designs of the sensing matrix $\mathbf{A}$ provided the k-sparse nonnegative vector in Definition 1 satisfies a certain *support-separability condition*. In the theorem below, we present an interesting robust null-space property of $\mathbf{A} \odot \mathbf{A}$ which is satisfied under the mild condition that the columns of $\mathbf{A}$ are approximately normalized. This property will be crucial in establishing $\mathbf{A} \odot \mathbf{A}$'s NN-RNSP compliance towards ensuring the positivity of η, and also dispenses with the restrictive support-separability condition required in [8].

Theorem 3 (*Strong Robust Null Space Property of Self Khatri-Rao Products*). *Let $\mathbf{A}$ be an $m \times n$ sized real matrix with columns $\mathbf{a}_i$ satisfying $\|\mathbf{a}_i\|_2^2 \in [1 - \alpha, 1 + \alpha]$ for some $\alpha \in (0, 1)$ for $i \in [n]$. Then, the self Khatri-Rao product $\mathbf{A} \odot \mathbf{A}$ satisfies the following robust null space property:*

$$
\|(\mathbf{A} \odot \mathbf{A})\mathbf{v}\|_2^2 \geq \frac{(1 - \alpha)^2}{2m}\left(\|\mathbf{v}_+\|_1^2 + \|\mathbf{v}_-\|_1^2\right)
$$

for all $\mathbf{v} \in \mathbb{R}^n$ such that $\|\mathbf{v}_+\|_1 \geq 4\left(\frac{1+\alpha}{1-\alpha}\right)^2 \|\mathbf{v}_-\|_1$. Here, $\mathbf{v}_+$ and $\mathbf{v}_-$ are nonnegative vectors containing the absolute

values of the positive and negative elements of $\mathbf{v}$*, respectively, such that* $\mathbf{v} = \mathbf{v}_+ - \mathbf{v}_-$.

Proof. See Appendix G. $\qquad\square$

An interesting consequence of Theorem 3 is that as long as $\mathbf{A}$ is approximately column normalized, the null space of $\mathbf{A} \odot \mathbf{A}$ does not contain any vectors of the form $\boldsymbol{\gamma} - \boldsymbol{\gamma}^*$ when $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^*$ have widely different nonzero supports, particularly if $\|\boldsymbol{\gamma}\|_0 \geq 4 \left(\frac{\gamma_{\max}}{\gamma_{\min}}\right) \left(\frac{1+\alpha}{1-\alpha}\right)^2 \|\boldsymbol{\gamma}^*\|_0$. However, when the support sizes of $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^*$ are comparable, it is not as straightforward to ascertain whether $\boldsymbol{\gamma} - \boldsymbol{\gamma}^*$ lies in the null space of $\mathbf{A} \odot \mathbf{A}$ or not. In Proposition 6, we state verifiable sufficient conditions that subsume the NN-RNSP condition for $\mathbf{A} \odot \mathbf{A}$, which in turn guarantees that η in (23) is always strictly positive.

Proposition 6. *Let the sensing matrix* $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n]$ *satisfy the following two properties*[4]

(P1). $\exists \alpha \in (0, 1)$ *such that* $\|\mathbf{a}_i\|_2^2 \in [1 - \alpha, 1 + \alpha], \forall i \in [n]$.

(P2). *For* $K_{threshold} \triangleq 4K \left(\frac{\gamma_{max}}{\gamma_{min}} \left(\frac{1+\alpha}{1-\alpha}\right)^2\right)$, $\exists \beta > 0$ *such that* $\|(\mathbf{A} \odot \mathbf{A})\mathbf{v}\|_2^2 \geq \beta \|\mathbf{v}\|_2^2$ *for all k-sparse vectors* $\mathbf{v} \in \mathbb{R}^n$ *and* $1 \leq k \leq (K + K_{threshold})$.

Then, η *defined in* (23) *is lower bounded as*

$$
\eta \geq \frac{\gamma_{min}^2}{4 \left(\sigma^2 + \gamma_{max}\right)^2} \min \left(\frac{\beta}{\max\left(1, \delta_{(K + K_{threshold})}^2\right)}, \right.
$$
$$
\left. \frac{(1-\alpha)^2}{4m} \min_{\substack{\mathcal{S} \subseteq [n], \\ |\mathcal{S} \backslash \mathcal{S}^*| + |\mathcal{S}^* \backslash \mathcal{S}| > K_{threshold}}} \frac{|\mathcal{S} \cup \mathcal{S}^*|}{\max\left(1, \delta_{|\mathcal{S} \cup \mathcal{S}^*|}^2\right)} \right),
$$
$$
\tag{25}
$$

where $\delta_k \triangleq \max_{\mathcal{S} \subseteq [n]: |\mathcal{S}| \leq k} \left\| \mathbf{A}_{\mathcal{S}}^T \mathbf{A}_{\mathcal{S}} \right\|_2$ *for any $k \in [n]$.*

Proof. See Appendix H. $\qquad\square$

By substituting η in Proposition 4 with its lower bound in Proposition 6, one can upper bound κ_{cov} as follows.

Proposition 7. *For the same setting as Proposition 6,*

$$
\kappa_{cov} \leq \left[(K + 5)\left(\log m + \log K + \log \Delta_{\kappa_{cov}}\right) + 2 \log n\right]^+,
$$

where

$$
\Delta_{\kappa_{cov}} \triangleq \frac{24}{\gamma_{min}^2} \left(\frac{\gamma_{max}}{\gamma_{min}} - 1\right) \left(3 + \frac{2\gamma_{max}}{\sigma^2}\right) (\sigma^2 + \gamma_{max})^2
$$
$$
\times \max \left(\frac{\delta_{K + K_{threshold}}^2}{\beta}, \frac{9}{2} \frac{(1 + \alpha)^2}{(1 - \alpha)^2}\right)
$$

and $[\cdot]^+ = \max(\cdot, 0)$.

Proof. See Appendix I. $\qquad\square$

The above bounds for η and κ_{cov} are valid for any deterministic or randomly constructed sensing matrix $\mathbf{A}$. The validity of these bounds is contingent upon showing the existence of a strictly positive β that satisfies condition $P2$ of Proposition 6,

---

[4]Properties $P1$ and $P2$ ensure that $\mathrm{Null}(\mathbf{A} \odot \mathbf{A})$ does not contain any vectors of the form $\boldsymbol{\gamma} - \boldsymbol{\gamma}^*$ when $\|\boldsymbol{\gamma}\|_0 \geq K + K_{\mathrm{threshold}}$ and when $\|\boldsymbol{\gamma}\|_0 < K + K_{\mathrm{threshold}}$, respectively.

which tantamounts to showing that any submatrix of $\mathbf{A} \odot \mathbf{A}$ obtained by sampling its $K + K_{\mathrm{threshold}}$ columns is nonsingular. The quantity β is commonly referred to as the *restricted minimum singular value* of the self Khatri-Rao product $\mathbf{A} \odot \mathbf{A}$. To illustrate how the derived bounds for η and κ_{cov} scale with the MMV problem size, we consider two example scenarios in the following corollary wherein $\mathbf{A}$ is randomly constructed and the associated η is known to be strictly positive.

Corollary 3. *Let* $\mathbf{A}$ *be an* $m \times n$ *sized matrix with i.i.d.* $\mathcal{N}\left(0, \frac{1}{m}\right)$ *entries, and let* $m^2 \leq n$. *Then, we have*

(i) $\eta = \Omega\left(\frac{K}{n}\right)$ *for* $m = \Theta(K \log n)$,

(ii) $\eta = \Omega\left(\frac{\sqrt{K}}{n}\right)$ *for* $m = \Theta(\sqrt{K} \log n)$,

and $\kappa_{cov} = O\left(K \log K + K \log \log n + \log n\right)$ *in both cases, with probability exceeding* $1 - c_0 n^{-\Omega(1)}$.

Proof. See Appendix J. $\qquad\square$

V. EXACT SUPPORT RECOVERY USING SPARSE BAYESIAN LEARNING
- SUFFICIENT CONDITIONS

Define $\mathcal{M}_{m,n}^k(\alpha, \beta)$ to be the set of all $m \times n$ sized real valued sensing matrices $\mathbf{A}$ satisfying the following two properties:

1) $\|\mathbf{a}_i\|_2^2 \in [1 - \alpha, 1 + \alpha], \forall i \in [n]$, where $\mathbf{a}_i$ denotes the ith column of $\mathbf{A}$
2) $\|(\mathbf{A} \odot \mathbf{A})\mathbf{v}\|_2 \geq \beta \|\mathbf{v}\|_2^2$ for all k or less sparse vectors $\mathbf{v} \in \mathbb{R}^n$.

Equipped with the newly defined set $\mathcal{M}_{m,n}^k(\alpha, \beta)$ and the explicit bounds for η and κ_{cov} in Propositions 6 and 7, respectively, we now state the sufficient conditions for vanishing support error probability in cM-SBL.

Theorem 4. *Suppose* $\mathbf{X}$ *has row support* $\mathcal{S}^*$, $|\mathcal{S}^*| \leq K$, *and satisfies assumption* $A1$. *Let* $\hat{\boldsymbol{\gamma}}$ *denote a solution of the cM-SBL optimization in* (11). *Then, for any* $\delta \in (0, \frac{1}{2})$, $\mathrm{supp}\,(\hat{\boldsymbol{\gamma}}) = \mathcal{S}^*$ *with probability exceeding* $1 - 2\delta$, *provided the following two conditions are satisfied.*

C1. The sensing matrix $\mathbf{A} \in \mathcal{M}_{m,n}^{K_o}(\alpha, \beta)$, *where* $K_o = K\left(1 + 4\left(\frac{\gamma_{max}}{\gamma_{min}}\right)\left(\frac{1+\alpha}{1-\alpha}\right)^2\right)$ *for some* $\alpha \in (0, 1)$ *and* $\beta > 0$.

C2. The number of MMVs, L, *satisfies*

$$
L \geq \frac{8}{\eta} \max \left\{ \log\left(\frac{6enK}{\delta}\right), \kappa_{cov} \right\},
$$

where η *and* κ_{cov} *depend on* α *and* β *as described by the Propositions 6 and 7, respectively.*

Proof. Under condition **C1**, η as defined in (23) is rendered strictly positive due to Proposition 6. Further, condition **C2** ensures that the abstract MMV bound in Theorem 2 is satisfied. Therefore, it follows directly from Theorem 2 that $\mathbb{P}(\mathcal{E}_{\mathcal{S}^*}) \leq 2\delta$. $\qquad\square$

The following corollary of Theorem 4 states an additional condition besides **C1** and **C2** that guarantees support consistency of any solution of the M-SBL optimization.

Corollary 4 (Exact support recovery in M-SBL)**.** *For the same setting as Theorem 4, let* $\hat{\boldsymbol{\gamma}}$ *denote any global solution of the*

M-SBL optimization in (5). *If* $\hat{\gamma} \in \Theta_n$, *and the conditions* **C1** *and* **C2** *hold for any* $\delta \in (0, \frac{1}{2})$, *then* $\text{supp}(\hat{\gamma}) = \mathcal{S}^*$ *with probability exceeding* $1 - 2\delta$.

Proof. Since $\hat{\gamma}$ belongs to Θ_n and maximizes the M-SBL objective $\log p(\mathbf{Y}; \gamma)$, it follows that $\hat{\gamma}$ is also a solution to the constrained cM-SBL optimization in (11). Therefore, the statement of Corollary 4 follows from Theorem 4. $\square$

By Corollary 4, retrospective to the nonzero coefficients of the M-SBL solution $\hat{\gamma}$ lying inside $[\gamma_{\min}, \gamma_{\max}]$, the support error probability vanishes under conditions **C1** and **C2**. It is to be noted that any probablistic guarantees of support consistency of the M-SBL solution, $\hat{\gamma}$, for the case where $\hat{\gamma} \notin \Theta_n$ are yet to be established. We conjecture that any global solution of the M-SBL optimization belongs to the set Θ_n with overwhelming probability, where $\gamma_{\min}$ and $\gamma_{\max}$ are constants independent of the MMV problem dimensions. However, showing such a result remains an open problem for future research.

Compared to the existing M-SBL optimization-based support recovery guarantees in [29] and [30], which assume that $\hat{\gamma}$ and γ^* are K-sparse binary valued vectors, the sparsistency guarantees derived here are applicable with wider scope. Our guarantees are valid for any real-valued K-sparse γ^*, and for any M-SBL solution $\hat{\gamma}$ provided its nonzero coefficients are bounded between $\gamma_{\min}$ and $\gamma_{\max}$.

VI. DISCUSSION

We now consider a few special cases of the MMV problem and reinterpret the sufficient conditions specified in Theorem 4 which guarantee vanishing support error probability.

A. *The case of binary hyperparameters and known support size*

In [29] and [30], support recovery is treated as a multiple hypothesis testing problem by assuming that each column of $\mathbf{X}$ is i.i.d. $\mathcal{N}(0, \text{diag}(\mathbf{1}_{\mathcal{S}^*}))$, where $\mathcal{S}^*$ is the true K-sparse row-support of $\mathbf{X}$. In this case, finding the true support via type-II likelihood maximization as in (5) can be reformulated as a combinatorial search over all k-sparse vertices of the hypercube $\{0, 1\}^n$, as described below.

$$\hat{\gamma} = \underset{\gamma \in \{0,1\}^n, \|\gamma\|_0 \leq K}{\arg\max} \log p(\mathbf{Y}; \gamma). \qquad (26)$$

The binary valued hyperparameters can be accommodated as a special case of our source model by simply setting $\gamma_{\min} = \gamma_{\max} = 1$. For $\gamma_{\min} = \gamma_{\max}$, according to Proposition 4, the ϵ-net $\Theta^\epsilon(\mathcal{S})$ collapses to a single point for all $\mathcal{S} \in \mathcal{S}_n$, which ultimately amounts to $\kappa_{\text{cov}} = 0$. In [29], [30], the correct support has to be identified from $\binom{n}{K}$ candidate support hypothesis. Under this restrained setting, the lower bound for η in (25) simplifies to

$$\eta \geq \frac{\beta}{4(\sigma^2 + 1)^2 \delta_{2K}^2}, \qquad (27)$$

where β denotes the smallest restricted singular value of $\mathbf{A} \odot \mathbf{A}$ of order $2K$. By setting $\kappa_{\text{cov}} = 0$ and using (27) in Theorem 4, we can conclude that $\mathbb{P}(\mathcal{E}_{\mathcal{S}^*})$ is at most 2ϵ if

$$L \geq \frac{32(\sigma^2 + 1)^2 \delta_{2K}^2}{\beta} \log\left(\frac{6enK}{\epsilon}\right). \qquad (28)$$

For the special case where the sensing matrix $\mathbf{A}$ comprises i.i.d. $\mathcal{N}\left(0, \frac{1}{m}\right)$ entries and has $m = \Theta(K \log n)$ rows, we have $\delta_{2K} = O(1)$ from Corollary 1, and $\beta = \Omega(1)$ from [45, Theorem 3]. Therefore, it follows from (28) and Corollary 4 that $L \geq \Omega(\log n)$ suffices to ensure that any K or less sparse binary vector in $\mathbb{R}^n$ that maximizes the M-SBL objective also recovers the true row-support of $\mathbf{X}$ with overwhelming probability. In this regard, our MMV bound matches with the one derived in [29], $L \gg \frac{\log n}{\log \log n}$, up to an additional $\log \log n$ factor.

A more interesting case arises when $m = \Theta(\sqrt{K} \log n)$. From (60) and (63) in Appendix J, we observe that $\beta = \Theta(1)$ and $\delta_{2K} = O(\sqrt{K})$. Substituting this in (28), we conclude that $L = \Omega(K \log n)$ suffices to ensure that an M-SBL solution from the restrained set of K-sparse binary vectors recovers the true support with overwhelming probability.

B. *The case of continuous-valued hyperparameters and unknown support size*

The sparsistency of a continuous-valued solution of the M-SBL optimization has been investigated in [12] under the assumption that the nonzero amplitudes of γ^* are known a priori and their guarantees apply to the case where the search for γ is restricted to K or less sparse vectors in Θ_K. Our analysis dispenses with both of these restrictive assumptions.

When the sensing matrix $\mathbf{A}$ consists of i.i.d. $\mathcal{N}\left(0, \frac{1}{m}\right)$ entries, by substituting η and κ_{cov} in Theorem 4 with their respective lower and upper bounds from Corollary 3, we see that the M-SBL solution in Θ_n recovers the true support with vanishing support error probability provided $m = \Theta(K \log n)$ and $L = \Omega\left(\frac{n}{K} \log n + n \log K + n \log \log n\right)$. On the other hand, when $m = \Theta(\sqrt{K} \log n)$, $L = \Omega\left(\frac{n}{\sqrt{K}} \log n + n\sqrt{K} \log K + n\sqrt{K} \log \log n\right)$ is sufficient.

By restricting the hyperparameter search in M-SBL to Θ_K, in [12], it is shown that the $L = \Omega\left(K \log^2\left(\frac{n}{K}\right) \log(nK)\right)$ is sufficient for exact support recovery when $m = \Omega(\sqrt{K} \log n)$. Comparing with our MMV bound, the same value of m is sufficient, but we pay an extra penalty of a roughly $\frac{n}{\sqrt{K}}$ factor in the MMV complexity, in order to circumvent the restrictive assumptions on the hyperparameter search domain in [12].

C. *The case when* $\mathbf{A} \odot \mathbf{A}$ *is full column rank*

The works in [22] and [46] have discussed random as well as deterministic constructions of the sensing matrix $\mathbf{A}$ for which the $m^2 \times n$ sized self Khatri-Rao product $\mathbf{A} \odot \mathbf{A}$ is full column rank provided $n \leq \frac{m^2 + m}{2}$. In this case, when the columns of $\mathbf{A}$ are approximately normalized, $\mathbf{A} \odot \mathbf{A}$ satisfies the Kth order NN-RNSP (Definition 1) by default for any $1 \leq K \leq n$, which in turn implies that η in (23) is always strictly positive. Hence, by Theorem 2, it follows that the support error probability decays exponentially with the number of MMVs even for $K = \Theta(m^2)$.

D. The case when $\mathbf{A} \odot \mathbf{A}$ is rank deficient

For $n > \frac{m^2+m}{2}$, the $m^2 \times n$ sized $\mathbf{A} \odot \mathbf{A}$ is rank-deficient. In [30] and [40], it is argued that this leads to parameter identifiability issues in M-SBL based support reconstruction. Specifically, the M-SBL objective can attain the same value for multiple distinct $\boldsymbol{\gamma}$, thereby fostering multiple global maxima with potentially different supports. However, both [30] and [40] do not take the non-negativity of the hyperparameters in $\boldsymbol{\gamma}$ into consideration. In Proposition 6, we have shown that as long as $\mathbf{A} \odot \mathbf{A}$ satisfies NN-RNSP of order K, η is always strictly positive and consequently K-sized supports can be recovered exactly with arbitrarily high probability using finitely many MMVs. Interestingly, the NN-RNSP condition can hold even when $\mathbf{A} \odot \mathbf{A}$ is column rank deficient, which allows cM-SBL to recover any $K = O(m^2)$-sized supports exactly even when $n > \frac{m^2+m}{2}$.

E. The case of noiseless measurements

If $K < \text{spark}(\mathbf{A}) - 1$, it can be shown that as the noise variance $\sigma^2 \to 0$, the error exponent $\mathcal{D}_{\frac{1}{2}}(p_{\boldsymbol{\gamma}}, p_{\boldsymbol{\gamma}^*})$ in (19) grows unbounded in the event of a support mismatch, $\text{supp}(\boldsymbol{\gamma}) \neq \text{supp}(\boldsymbol{\gamma}^*)$, culiminating in vanishing support error probability even when $L = 1$. This is formally proved below.

The $1/2$-Rényi divergence between two multivariate Gaussian densities $p_{\boldsymbol{\gamma}_i}(\mathbf{y}) \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}_i})$, $i = 1, 2$ is given by

$$
\begin{aligned}
\mathcal{D}_{\frac{1}{2}}\left(p_{\boldsymbol{\gamma}_1}, p_{\boldsymbol{\gamma}_2}\right) &= \log\left|\frac{\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_1} + \boldsymbol{\Sigma}_{\boldsymbol{\gamma}_2}}{2}\right| - \frac{1}{2}\log\left|\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_1}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_2}\right| \\
&= \log\left|\frac{\mathbf{H}^{1/2} + \mathbf{H}^{-1/2}}{2}\right|,
\end{aligned} \tag{29}
$$

where $\mathbf{H} \triangleq \boldsymbol{\Sigma}_{\boldsymbol{\gamma}_1}^{1/2}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_2}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_1}^{1/2}$ is referred to as the *discrimination matrix*. Since $\mathbf{H}$ is a normal matrix, it is unitarily diagonalizable. Let $\mathbf{H} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$, where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_m)$ with λ_i's being the strictly positive eigenvalues of $\mathbf{H}$ for any $\sigma^2 > 0$, and $\mathbf{U}$ being a unitary matrix with the eigenvectors of $\mathbf{H}$ as its columns. The $1/2$-Rényi divergence can be expressed in terms of λ_i as

$$
\begin{aligned}
\mathcal{D}_{\frac{1}{2}}\left(p_{\boldsymbol{\gamma}_1}, p_{\boldsymbol{\gamma}_2}\right) &= \sum_{i=1}^{m} \log\left(\left(\lambda_i^{1/2} + \lambda_i^{-1/2}\right)/2\right) \\
&\geq \log\left(\frac{1}{2}\left((\lambda_{\max}(\mathbf{H}))^{1/2} + (\lambda_{\max}(\mathbf{H}))^{-1/2}\right)\right).
\end{aligned} \tag{30}
$$

The above inequality is obtained by dropping all positive terms in the summation except the one term which corresponds to $\lambda_{\max}(\mathbf{H})$, the maximum eigenvalue of $\mathbf{H}$. Proposition 8 below relates $\lambda_{\max}(\mathbf{H})$ to the noise variance σ^2.

Proposition 8. *If $K < \text{spark}(\mathbf{A}) - 1$, then for any $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2 \in \mathbb{R}_+^n$ such that $\text{supp}(\boldsymbol{\gamma}_1) \neq \text{supp}(\boldsymbol{\gamma}_2)$ and $|\text{supp}(\boldsymbol{\gamma}_2)| \leq K$, the maximum eigenvalue of $\mathbf{H} \triangleq \boldsymbol{\Sigma}_{\boldsymbol{\gamma}_1}^{1/2}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_2}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_1}^{1/2}$ satisfies*

$$
\lambda_{max}(\mathbf{H}) \geq \frac{c_1}{\sigma^2}
$$

for some constant $c_1 > 0$ independent of σ^2.

Proof. See Appendix L. $\square$

According to Proposition 8, in the limit $\sigma^2 \to 0$, $\lambda_{\max}(\mathbf{H}) \to \infty$, and consequently, $\mathcal{D}_{\frac{1}{2}}(p_{\boldsymbol{\gamma}_1}, p_{\boldsymbol{\gamma}_2})$ grows unbounded (due to (30)) whenever $\text{supp}(\boldsymbol{\gamma}_1) \neq \text{supp}(\boldsymbol{\gamma}_2)$ and $K < \text{spark}(\mathbf{A}) - 1$. Based on this observation, we now state Theorem 5 which lays forward the sufficient conditions for exact support recovery in the noiseless case.

Theorem 5. *Consider the noiseless MMV problem, with observations $\mathbf{Y} = \mathbf{A}\mathbf{X}$ corresponding to an unknown $\mathbf{X}$ satisfying assumption A1. Suppose $\mathcal{S}^*$ is the true nonzero row support of $\mathbf{X}$ with $|\mathcal{S}^*| \leq K$. Further, let $\hat{\boldsymbol{\gamma}}$ be a solution of the MSBL optimization in (5), then $\text{supp}(\hat{\boldsymbol{\gamma}}) = \mathcal{S}^*$ almost surely, provided that $K < \text{spark}(\mathbf{A}) - 1$. This result holds even in the SMV case, i.e., when $L = 1$.*

Proof. Under assumption **A1**, there exists a $\boldsymbol{\gamma}^* \in \Theta_K$ such that every column in $\mathbf{X}$ is i.i.d. $\mathcal{N}(0, \text{diag}(\boldsymbol{\gamma}^*))$, and $\text{supp}(\boldsymbol{\gamma}^*) = \mathcal{S}^*$. Since $\hat{\boldsymbol{\gamma}}$ globally maximizes the MSBL objective $\mathcal{L}(\mathbf{Y}; \boldsymbol{\gamma})$, it follows that $\mathcal{L}(\mathbf{Y}; \hat{\boldsymbol{\gamma}}) \geq \mathcal{L}(\mathbf{Y}; \boldsymbol{\gamma}^*)$ if $\hat{\boldsymbol{\gamma}} \neq \boldsymbol{\gamma}^*$, i.e., the following chain of implications holds.

$$
\begin{aligned}
\{\text{supp}(\hat{\boldsymbol{\gamma}}) \neq \mathcal{S}^*\} &= \{\text{supp}(\hat{\boldsymbol{\gamma}}) \neq \text{supp}(\boldsymbol{\gamma}^*)\} \\
&\subseteq \{\hat{\boldsymbol{\gamma}} \neq \boldsymbol{\gamma}^*\} \\
&\subseteq \{\mathcal{L}(\mathbf{Y}; \hat{\boldsymbol{\gamma}}) \geq \mathcal{L}(\mathbf{Y}; \boldsymbol{\gamma}^*)\}.
\end{aligned}
$$

By applying Corollary 2, this further implies that

$$
\begin{aligned}
\mathbb{P}\left(\text{supp}(\hat{\boldsymbol{\gamma}}) \neq \mathcal{S}^*\right) &\leq \mathbb{P}\left(\mathcal{L}(\mathbf{Y}; \hat{\boldsymbol{\gamma}}) \geq \mathcal{L}(\mathbf{Y}; \boldsymbol{\gamma}^*)\right) \\
&\leq \exp\left(-\frac{L\mathcal{D}_{\frac{1}{2}}(p_{\hat{\boldsymbol{\gamma}}}, p_{\boldsymbol{\gamma}^*})}{4}\right).
\end{aligned}
$$

By using the lower bound in (30) for $\mathcal{D}_{\frac{1}{2}}(p_{\hat{\boldsymbol{\gamma}}}, p_{\boldsymbol{\gamma}^*})$, we have

$$
\mathbb{P}\left(\text{supp}(\hat{\boldsymbol{\gamma}}) \neq \mathcal{S}^*\right) \leq \left[\frac{1}{2}\left(\sqrt{\lambda_{\max}(\mathbf{H})} + \frac{1}{\sqrt{\lambda_{\max}(\mathbf{H})}}\right)\right]^{-\frac{L}{4}}, \tag{31}
$$

where $\mathbf{H} = \boldsymbol{\Sigma}_{\hat{\boldsymbol{\gamma}}}^{1/2}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}^*}^{-1}\boldsymbol{\Sigma}_{\hat{\boldsymbol{\gamma}}}^{1/2}$. Since $\boldsymbol{\gamma}^*$ is at most K-sparse, as long as $K < \text{spark}(\mathbf{A}) - 1$, by Proposition 8, $\sigma^2 \to 0$ results in $\lambda_{\max}(\mathbf{H}) \to \infty$ which in turn drives the RHS in (31) to zero for $L \geq 1$. $\square$

From Theorem 5, we conclude that, in the noiseless scenario ($\sigma^2 \to 0$) and for $\mathbf{X}$ satisfying assumption **A1**, MSBL requires only a single measurement vector ($L = 1$) to perfectly recover any $K < \text{spark}(\mathbf{A}) - 1$ sized support. If the sensing matrix $\mathbf{A}$ has full spark, i.e., $\text{spark}(\mathbf{A}) = m + 1$, MSBL can recover $m - 1$ or lesser sparse supports exactly from m noiseless measurements of a single sparse vector. It is noteworthy that $\mathbf{A}$ satisfies the full spark condition under very mild assumptions, e.g., the entries of $\mathbf{A}$ are drawn independently from a continuous probability distribution. This result is an improvement over the sufficient conditions for exact support recovery by MSBL in [23, Theorem 1]. Unlike in [23], we do not require the nonzero rows of $\mathbf{X}$ to be orthogonal. Also, our result improves over the $k \leq \frac{m}{2}$ condition shown in [29].

VII. CONCLUSIONS

We analyzed the sample complexity of error free recovery of the common nonzero support of multiple joint-sparse Gaussian

sources from their compressive measurements. We established the finite MMV, high-probability consistency of the nonzero support inferred from a constrained type-II ML estimate of the variance hyperparameters belonging to a correlation-aware Gaussian source prior. The nonzero coefficients of the type-II ML estimate are constrained to lie in a known interval $[\gamma_{\min}, \gamma_{\max}]$. Our support consistency guarantee also applies to any global solution of the M-SBL optimization when the nonzero coefficients of the solutions satisfy the same interval constraint.

We also showed that a single noiseless MMV suffices for perfect recovery of any K-sparse support, provided that $K < \mathrm{spark}(\mathbf{A}) - 1$. In case of noisy MMVs, we showed that any K-sized support can be recovered exactly with high probability using finitely many MMVs, provided $\mathbf{A} \odot \mathbf{A}$ admits a positive minimum restricted singular value of order $\Theta(K)$. We also presented an interesting interpretation of M-SBL's marginalized log-likelihood cost as a Bregman matrix divergence, which highlights that the M-SBL algorithm is, in principle, a covariance matching algorithm. There still remain the following open questions regarding M-SBL-based support recovery:

(i) What are necessary conditions for exact support recovery in terms of the number of required MMVs? Under the assumption of i.i.d. measurement matrices (unlike the common measurement matrix used in this work), such conditions have been derived in [47]; these have been extended to multiple support recovery in [48].

(ii) Is there a criterion under which all stationary points of the M-SBL objective also yield the correct support estimate?

(iii) How is the support recovery performance impacted by inter and intra vector correlations in the signals?

(iv) To derive the sufficient conditions for exact support recovery via cM-SBL, we assumed that $\gamma_{\max}$ and $\gamma_{\min}$ are roughly of the same order. This assumption needs to be certified, perhaps by analyzing the Karush–Kuhn–Tucker (KKT) conditions associated with M-SBL's nonnegative optimization.

Answering these questions would be interesting directions for future work.

APPENDIX

A. Proof of Proposition 1

By using the property: $\log |\mathbf{X}\mathbf{Y}| = \log |\mathbf{X}| + \log |\mathbf{Y}|$ for any positive definite matrices $\mathbf{X}$ and $\mathbf{Y}$, from (9), we have

$$
\begin{aligned}
&\mathcal{D}_{\frac{1}{2}}(p_1, p_2)\\
&= \log \left| \frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2} \right| - \frac{1}{2}\log|\boldsymbol{\Sigma}_1| - \frac{1}{2}\log|\boldsymbol{\Sigma}_2|\\
&= \log \left| \frac{\mathbf{I}_m + \boldsymbol{\Sigma}_1^{-1/2}\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1^{-1/2}}{2} \right| + \frac{1}{2}\log|\boldsymbol{\Sigma}_1| - \frac{1}{2}\log|\boldsymbol{\Sigma}_2|\\
&= \log \left| \frac{\mathbf{I}_m + \boldsymbol{\Sigma}_1^{-1/2}\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1^{-1/2}}{2} \right| + \frac{1}{2}\log \left| \boldsymbol{\Sigma}_1^{1/2}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1^{1/2} \right|\\
&= \log \left| \frac{\mathbf{I}_m + \mathbf{H}^{-1}}{2} \right| + \frac{1}{2}\log|\mathbf{H}|
\end{aligned}
$$

$$
= \log \left| \frac{\mathbf{H}^{1/2} + \mathbf{H}^{-1/2}}{2} \right|, \tag{32}
$$

where $\mathbf{H} \triangleq \boldsymbol{\Sigma}_1^{1/2}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1^{1/2}$. Denote the eigenvalue decomposition of $\mathbf{H}$ by $\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$, with the real and positive diagonal entries of $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \dots, \lambda_m)$ denoting the eigenvalues of $\mathbf{H}$, and the orthonormal columns of $\mathbf{U}$ denoting the eigenvectors of $\mathbf{H}$. By the non-negativity of the Bregman Log-Det divergence between positive definite matrices [42], for any $\mathbf{X} \in S_{++}^m$, it follows that $\log |\mathbf{X}| \geq \mathrm{tr}(\mathbf{I} - \mathbf{X}^{-1})$. Therefore, using (32), we have

$$
\begin{aligned}
\mathcal{D}_{\frac{1}{2}}(p_1, p_2) &\geq \mathrm{tr}\left(\mathbf{I}_m - \left(\frac{\mathbf{H}^{1/2} + \mathbf{H}^{-1/2}}{2} \right)^{-1} \right)\\
&= \sum_{i=1}^m \left(1 - \left(\frac{\lambda_i^{1/2} + \lambda_i^{-1/2}}{2} \right)^{-1} \right) \geq \sum_{i=1}^m \frac{(\lambda_i - 1)^2}{2(1 + \lambda_i)^2}\\
&= \frac{1}{2}\mathrm{tr}\left((\mathbf{H} - \mathbf{I}_m)(\mathbf{I}_m + \mathbf{H})^{-1}(\mathbf{H} - \mathbf{I}_m)(\mathbf{I}_m + \mathbf{H})^{-1} \right).
\end{aligned} \tag{33}
$$

Plugging back $\mathbf{H} = \boldsymbol{\Sigma}_1^{1/2}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1^{1/2}$ in (33), we obtain the desired lower bound for $\mathcal{D}_{\frac{1}{2}}(p_1, p_2)$ as shown below.

$$
\begin{aligned}
&D_{\frac{1}{2}}(p_1, p_2)\\
&\geq \frac{1}{2}\mathrm{tr}\Big((\boldsymbol{\Sigma}_1^{1/2}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1^{1/2} - \mathbf{I}_m)(\mathbf{I}_m + \mathbf{H})^{-1} \cdot\\
&\qquad\qquad (\boldsymbol{\Sigma}_1^{1/2}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1^{1/2} - \mathbf{I}_m)(\mathbf{I}_m + \mathbf{H})^{-1} \Big)\\
&= \frac{1}{2}\mathrm{tr}\Big(\boldsymbol{\Sigma}_1^{1/2}(\boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\Sigma}_1^{-1})\boldsymbol{\Sigma}_1^{1/2}(\mathbf{I}_m + \mathbf{H})^{-1} \cdot\\
&\qquad\qquad \boldsymbol{\Sigma}_1^{1/2}(\boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\Sigma}_1^{-1})\boldsymbol{\Sigma}_1^{1/2}(\mathbf{I}_m + \mathbf{H})^{-1} \Big)\\
&= \frac{1}{2}\mathrm{tr}\Big(\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1^{1/2}(\mathbf{I}_m + \mathbf{H})^{-1}\boldsymbol{\Sigma}_1^{1/2}\boldsymbol{\Sigma}_1^{-1} \cdot\\
&\qquad\qquad (\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1^{1/2}(\mathbf{I}_m + \mathbf{H})^{-1}\boldsymbol{\Sigma}_1^{1/2} \Big)\\
&= \frac{1}{2}\mathrm{tr}\Big[(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)\left(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1^{1/2}(\mathbf{I}_m + \mathbf{H})^{-1}\boldsymbol{\Sigma}_1^{-1/2} \right) \cdot\\
&\qquad\qquad (\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)\left(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1^{1/2}(\mathbf{I}_m + \mathbf{H})^{-1}\boldsymbol{\Sigma}_1^{-1/2} \right) \Big]\\
&= \frac{1}{2}\mathrm{tr}\Big[(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)\left(\boldsymbol{\Sigma}_2^{-1}(\mathbf{I}_m + \boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1})^{-1} \right) \cdot\\
&\qquad\qquad (\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)\left(\boldsymbol{\Sigma}_2^{-1}(\mathbf{I}_m + \boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1})^{-1} \right) \Big]\\
&= \frac{1}{2}\mathrm{tr}\Big[(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)(\boldsymbol{\Sigma}_2 + \boldsymbol{\Sigma}_1)^{-1}(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)(\boldsymbol{\Sigma}_2 + \boldsymbol{\Sigma}_1)^{-1} \Big].
\end{aligned}
$$

B. Proof of Corollary 1

For a fixed support $\mathcal{S}, |\mathcal{S}| \leq k$, from Proposition 3,

$$
\mathbb{P}\left(\|\mathbf{A}_{\mathcal{S}}\|_2 \geq \sqrt{m} + \sqrt{k} + \sqrt{6k\log n} \right) \leq 2e^{-3k\log n}.
$$

By taking the union bound over $\binom{n}{1} + \binom{n}{2} + \dots + \binom{n}{k} \leq \left(\frac{3en}{2} \right)^k$ submatrices of $\mathbf{A}$ containing k or fewer columns, we get

$$
\begin{aligned}
&\mathbb{P}\left(\bigcup_{\mathcal{S}\subset[n]:|\mathcal{S}|\leq k} \left\{ \|\mathbf{A}_{\mathcal{S}}\|_2 \geq \sqrt{m} + \sqrt{k} + \sqrt{6k\log n} \right\} \right)\\
&\qquad \leq \left(\frac{3ne}{2} \right)^k 2e^{-3k\log n} \leq 2e^{-3k\log n + k\log(3ne/2)} \leq \frac{2}{n^k},
\end{aligned}
$$

for $n > 5$.

C. Proof of Proposition 4

The following stepwise procedure shows how to construct a δ-net of $\Theta(\mathcal{S})$ (with respect to the Euclidean distance metric) which is entirely contained in $\Theta(\mathcal{S})$.

1) Consider an δ-blow up of $\Theta(\mathcal{S})$, denoted by

$$\Theta_{\uparrow\delta}(\mathcal{S}) \triangleq \{x : \exists x' \in \Theta(\mathcal{S}) \text{ such that } \|x - x'\|_2 \leq \delta\}.$$

2) Let $\Theta_{\uparrow\delta}^{\delta}(\mathcal{S})$ be a δ-net of $\Theta_{\uparrow\delta}(\mathcal{S})$. Some points in $\Theta_{\uparrow\delta}^{\delta}(\mathcal{S})$ may lie outside $\Theta(\mathcal{S})$.

3) Let $\mathcal{P}$ denote the set containing the projections of all points in $\Theta_{\uparrow\delta}^{\delta}(\mathcal{S}) \cap \Theta(\mathcal{S})^c$ onto the set $\Theta(\mathcal{S})$. By construction, $\mathcal{P} \subset \Theta(\mathcal{S})$, and $|\mathcal{P}| \leq |\Theta_{\uparrow\delta}^{\delta}(\mathcal{S}) \cap \Theta(\mathcal{S})^c|$.

4) Then, $\Theta^{\delta}(\mathcal{S}) \triangleq \left(\Theta_{\uparrow\delta}^{\delta}(\mathcal{S}) \cap \Theta(\mathcal{S})\right) \cup \mathcal{P}$ is a valid δ-net of $\Theta(\mathcal{S})$ which is entirely contained in $\Theta(\mathcal{S})$.

To prove the validity of the above δ-net construction, we need to show that for any $\gamma \in \Theta(\mathcal{S})$, there exists an element $\mathbf{a}$ in $\Theta^{\delta}(\mathcal{S})$ such that $\|\gamma - \mathbf{a}\|_2 \leq \delta$. Let γ be an arbitrary element in $\Theta(\mathcal{S})$. Then, γ also belongs to the larger set $\Theta_{\uparrow\delta}(\mathcal{S})$, and consequently, there exists $\gamma' \in \Theta_{\uparrow\delta}^{\delta}(\mathcal{S})$ such that $\|\gamma - \gamma'\|_2 \leq \delta$. Now, there are two cases. (i) $\gamma' \in \Theta(\mathcal{S})$, and (ii) $\gamma' \notin \Theta(\mathcal{S})$.

In case (i), $\gamma' \in \left(\Theta_{\uparrow\delta}^{\delta}(\mathcal{S}) \cap \Theta(\mathcal{S})\right)$, and hence also belongs to $\Theta^{\delta}(\mathcal{S})$. Further, $\|\gamma - \gamma'\|_2 \leq \delta$. Hence $\mathbf{a} = \gamma'$ will work.

In case (ii), $\gamma' \in \Theta_{\uparrow\delta}^{\delta}(\mathcal{S}) \cap \Theta(\mathcal{S})^c$. Let γ'' be the projection of γ' onto $\Theta(\mathcal{S})$, then γ'' must belong to $\mathcal{P}$, and hence must also belong to $\Theta^{\delta}(\mathcal{S})$. Note that since γ'' is the projection of γ' onto the *convex* set $\Theta(\mathcal{S})$, for any $\gamma \in \Theta(\mathcal{S})$, we have $\langle \gamma - \gamma'', \gamma' - \gamma''\rangle \leq 0$. Further, we have

$$\begin{aligned}
\delta &\geq \|\gamma - \gamma'\|_2^2 = \|(\gamma - \gamma'') + (\gamma'' - \gamma')\|_2^2 \\
&= \|\gamma - \gamma''\|_2^2 + \|\gamma'' - \gamma'\|_2^2 + 2\langle \gamma - \gamma'', \gamma'' - \gamma'\rangle \\
&\geq \|\gamma - \gamma''\|_2^2. \tag{34}
\end{aligned}$$

The last inequality is obtained by dropping the last two nonnegative terms in the RHS. From (34), $\mathbf{a} = \gamma''$ will work.

Since case (i) and (ii) together are exhaustive, $\Theta^{\delta}(\mathcal{S})$ in step-4 is a valid δ-net of $\Theta(\mathcal{S})$ which is entirely inside $\Theta(\mathcal{S})$.

Cardinality of $\Theta^{\delta}(\mathcal{S})$: The diameter of $\Theta(\mathcal{S})$ is $\sqrt{|\mathcal{S}|}(\gamma_{\max} - \gamma_{\min})$. Based on the construction in step-4, the cardinality of $\Theta^{\delta}(\mathcal{S})$ can be upper bounded as:

$$\begin{aligned}
|\Theta^{\delta}(\mathcal{S})| &\leq |\Theta_{\uparrow\delta}^{\delta}(\mathcal{S}) \cap \Theta(\mathcal{S})| + |\Theta_{\uparrow\delta}^{\delta}(\mathcal{S}) \cap \Theta(\mathcal{S})^c| \\
&= |\Theta_{\uparrow\delta}^{\delta}(\mathcal{S})| \\
&\leq \left|\delta\text{-net of } \ell_2\text{-ball of radius } \sqrt{|\mathcal{S}|}(\gamma_{\max} - \gamma_{\min})\right. \\
&\qquad \left. \text{in } \mathbb{R}^{|\mathcal{S}|}\right| \\
&\leq \max\left(1, \left(3\sqrt{|\mathcal{S}|}(\gamma_{\max} - \gamma_{\min})/\delta\right)^{|\mathcal{S}|}\right). \tag{35}
\end{aligned}$$

The last step is an extension of the volumetric arguments in [49] to show that the δ-covering number of a unit ball $\mathcal{B}_1(0)$ in $\mathbb{R}^k$ with respect to the standard Euclidean norm $\|\cdot\|_2$ satisfies $\mathcal{N}_{\text{cov}}^{\delta}(\mathcal{B}_1(0), \|\cdot\|_2) \leq (3/\delta)^k$. The max operation with unity covers the case when δ is larger than the diameter of $\Theta(\mathcal{S})$ and the case when $\gamma_{\max} = \gamma_{\min}$.

Now consider the modified net $\Theta^{\epsilon/C_{\mathcal{L},\mathcal{S}}}(\mathcal{S})$ obtained by setting $\delta = \frac{\epsilon}{C_{\mathcal{L},\mathcal{S}}}$ in steps 1-4, where $C_{\mathcal{L},\mathcal{S}}$ is the Lipschitz constant of $\mathcal{L}(\mathbf{Y}, \gamma)$ with respect to $\gamma \in \Theta(\mathcal{S})$. We claim that $\Theta^{\epsilon/C_{\mathcal{L},\mathcal{S}}}(\mathcal{S})$ is the desired set which simultaneously satisfies conditions (i) and (ii) stated in Proposition 4.

To show condition (i), we observe that since $\Theta^{\epsilon/C_{\mathcal{L},\mathcal{S}}}(\mathcal{S})$ is an $(\epsilon/C_{\mathcal{L},\mathcal{S}})$-net of $\Theta(\mathcal{S})$ with respect to $\|\cdot\|_2$, for any $\gamma \in \Theta(\mathcal{S})$, there exists a $\gamma' \in \Theta^{\epsilon/C_{\mathcal{L},\mathcal{S}}}(\mathcal{S})$ such that $\|\gamma - \gamma'\|_2 \leq \epsilon/C_{\mathcal{L},\mathcal{S}}$. Since $\mathcal{L}(\mathbf{Y}, \gamma)$ is $C_{\mathcal{L},\mathcal{S}}$-Lipschitz in $\Theta(\mathcal{S})$, it follows that $|L(\mathbf{Y}, \gamma) - L(\mathbf{Y}, \gamma')| \leq C_{\mathcal{L},\mathcal{S}}\|\gamma - \gamma'\|_2 \leq \epsilon$.

Condition (ii) follows from (35) by setting $\delta = \epsilon/C_{\mathcal{L},\mathcal{S}}$.

D. Proof of Theorem 1

For continuous probability densities p_{γ} and p_{γ^*} defined on the observation space $\mathbb{R}^m$, for any $\epsilon > 0$, the tail probability of the random variable $\log(p_{\gamma}(\mathbf{Y})/p_{\gamma^*}(\mathbf{Y}))$ has a Chernoff upper bound with parameter $t > 0$ as shown below.

$$\begin{aligned}
\mathbb{P}\left(\log \frac{p_{\gamma}(\mathbf{Y})}{p_{\gamma^*}(\mathbf{Y})} \geq -\epsilon\right) &= \mathbb{P}\left(\sum_{j=1}^{L} \log \frac{p_{\gamma}(\mathbf{y}_j)}{p_{\gamma^*}(\mathbf{y}_j)} \geq -\epsilon\right) \\
&\leq \mathbb{E}_{p_{\gamma^*}}\left[\exp\left(t\sum_{j=1}^{L} \log \frac{p_{\gamma}(\mathbf{y}_j)}{p_{\gamma^*}(\mathbf{y}_j)}\right)\right] \exp(t\epsilon) \\
&= \left(\mathbb{E}_{p_{\gamma^*}}\left[\exp\left(t\log \frac{p_{\gamma}(\mathbf{y})}{p_{\gamma^*}(\mathbf{y})}\right)\right]\right)^L \exp(t\epsilon) \\
&= \left(\mathbb{E}_{p_{\gamma^*}}\left[\left(\frac{p_{\gamma}(\mathbf{y})}{p_{\gamma^*}(\mathbf{y})}\right)^t\right]\right)^L \exp(t\epsilon) \\
&= \left(\int_{\mathbf{y}} p_{\gamma}(\mathbf{y})^t p_{\gamma^*}(\mathbf{y})^{1-t} d\mathbf{y}\right)^L \exp(t\epsilon) \\
&= \exp\left(-L\left[t\left(-\frac{\epsilon}{L}\right) - (t-1)\mathcal{D}_t(p_{\gamma}, p_{\gamma^*})\right]\right). \tag{36}
\end{aligned}$$

In the above, the first and third steps follow from the independence of $\mathbf{y}_j$. The second step is the application of the Chernoff bound. The last step is obtained by using the definition of the Rényi divergence and rearranging the terms in the exponent.

By introducing the function $\psi(t) = (t-1)\mathcal{D}_t(p_{\gamma}, p_{\gamma^*})$, the Chernoff bound in (36) can be restated as

$$\mathbb{P}\left(\log \frac{p_{\gamma}(\mathbf{Y})}{p_{\gamma^*}(\mathbf{Y})} \geq -\epsilon\right) \leq \exp\left(-L\left[t\left(-\frac{\epsilon}{L}\right) - \psi(t)\right]\right). \tag{37}$$

For $t = \arg\sup_{t>0}\left(t\left(-\frac{\epsilon}{L}\right) - \psi(t)\right)$, the upper bound in (37) attains its tightest value $\exp\left(-L\psi^*\left(-\frac{\epsilon}{L}\right)\right)$, where ψ^* is the Legendre transform of ψ.

E. Proof of Theorem 2

Since $L \geq C\log\frac{2}{\delta}$, by Proposition 2, $\mathbb{P}(\mathcal{G}^c) \leq \delta$. Combined with $L \geq \frac{8\kappa_{\text{cov}}}{\eta}$, (22) can be rewritten as

$$\mathbb{P}(\mathcal{E}_{\mathcal{S}^*}) \leq \sum_{\mathcal{S} \in \mathcal{S}_n \setminus \mathcal{S}^*} \exp\left(-\frac{\eta L k_d^{\mathcal{S},\mathcal{S}^*}}{8}\right) + \delta. \tag{38}$$

The total number of support sets belonging to $\mathcal{S}_n \setminus \mathcal{S}^*$ which differ from the true support $\mathcal{S}^*$ in exactly k_d locations is

$\sum_{j=0}^{k_d} \binom{n-|\mathcal{S}^*|}{j}\binom{j+|\mathcal{S}^*|}{\min(k_d,j+|\mathcal{S}^*|)}$. Since $|\mathcal{S}^*| \leq K$, this summation can be further upper bounded by $(2nK)^{k_d}$. Thus, we can rewrite (38) as

$$
\begin{aligned}
\mathbb{P}(\mathcal{E}_{\mathcal{S}^*}) &\leq \delta + \sum_{k_d=1}^{n-|\mathcal{S}^*|} \sum_{\substack{\mathcal{S}\in\mathcal{S}_n\setminus\mathcal{S}^*,\\|(\mathcal{S}\setminus\mathcal{S}^*)\cup(\mathcal{S}^*\setminus\mathcal{S})|=k_d}} \exp\left(-\frac{\eta L k_d}{8}\right)\\
&\leq \delta + \sum_{k_d=1}^{n-|\mathcal{S}^*|} (2nK)^{k_d}\left(e^{-\frac{\eta L}{8}}\right)^{k_d}. \quad (39)
\end{aligned}
$$

Since $L \geq \frac{8}{\eta}\log\left(2nK\left(\frac{1+\delta}{\delta}\right)\right)$, $\mathbb{P}(\mathcal{E}_{\mathcal{S}^*})$ can be upper bounded by a geometric series as

$$
\mathbb{P}(\mathcal{E}_{\mathcal{S}^*}) \leq \delta + \sum_{k_d=1}^{\infty}\left(\frac{\delta}{1+\delta}\right)^{k_d} = \delta + \delta = 2\delta.
$$

F. Proof of Proposition 5

Let $\Delta\Gamma = \mathrm{diag}(\Delta\boldsymbol{\gamma})$, where $\Delta\boldsymbol{\gamma} \triangleq \boldsymbol{\gamma} - \boldsymbol{\gamma}^*$. Also, let $\boldsymbol{\Sigma}_{\gamma+\gamma^*} \triangleq \boldsymbol{\Sigma}_\gamma + \boldsymbol{\Sigma}_{\gamma^*}$. Then, using Proposition 1, $\mathcal{D}_{\frac{1}{2}}(p_\gamma, p_{\gamma^*})$ can be bounded as follows.

$$
\begin{aligned}
\mathcal{D}_{\frac{1}{2}}(p_\gamma, p_{\gamma^*}) &\geq \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_{\gamma+\gamma^*}^{-1}(\mathbf{A}\Delta\Gamma\mathbf{A}^T)\boldsymbol{\Sigma}_{\gamma+\gamma^*}^{-1}(\mathbf{A}\Delta\Gamma\mathbf{A}^T)\right)\\
&\geq \frac{\|\mathbf{A}\Delta\Gamma\mathbf{A}^T\|_F^2}{2\|\boldsymbol{\Sigma}_{\gamma+\gamma^*}\|_2^2} = \frac{\|(\mathbf{A}\odot\mathbf{A})\Delta\boldsymbol{\gamma}\|_2^2}{2\|\boldsymbol{\Sigma}_{\gamma+\gamma^*}\|_2^2}. \quad (40)
\end{aligned}
$$

In above, the second inequality is obtained by repeatedly applying the trace inequality: $\mathrm{tr}\left(\mathbf{A}^{-1}\mathbf{B}\right) \geq \mathrm{tr}(\mathbf{B})/\|\mathbf{A}\|_2$ for any positive definite $\mathbf{A}$ and positive semidefinite $\mathbf{B}$. The last step follows from the identity: $\mathrm{vec}(\mathbf{A}\Delta\Gamma\mathbf{A}^T) = (\mathbf{A}\odot\mathbf{A})\Delta\boldsymbol{\gamma}$.

Next, we derive an upper bound for the spectral norm of $\boldsymbol{\Sigma}_{\gamma+\gamma^*}$ as shown below.

$$
\begin{aligned}
\|\boldsymbol{\Sigma}_{\gamma+\gamma^*}\|_2 &= \left\|2\sigma^2\mathbf{I}_m + \mathbf{A}(\boldsymbol{\Gamma}+\boldsymbol{\Gamma}^*)\mathbf{A}^T\right\|_2\\
&\leq 2\sigma^2 + 2\gamma_{\max}\left\|\mathbf{A}_{\mathcal{S}\cup\mathcal{S}^*}^T\mathbf{A}_{\mathcal{S}\cup\mathcal{S}^*}\right\|_2. \quad (41)
\end{aligned}
$$

Finally, using (41) in (40), we obtain the desired lower bound for $\mathcal{D}_{\frac{1}{2}}(p_\gamma, p_{\gamma^*})$.

G. Proof of Theorem 3

Consider the unit norm m^2 length vector

$$
\mathbf{w} \triangleq \frac{\mathbf{1}_{\{1,m+2,2m+3,\ldots,m^2\}}}{\sqrt{m}}, \quad (42)
$$

where $\mathbf{1}_\mathcal{S}$ is a binary vector containing ones in the indices specified by the set $\mathcal{S}$ and zeros everywhere else. We call $\mathbf{w}$ the *Hadamard sampler*, as it samples the m rows of the Hadamard submatrix contained within $\mathbf{A}\odot\mathbf{A}$. Let $\mathbf{b} = (\mathbf{A}\odot\mathbf{A})^T\mathbf{w}$, then

$$
\mathbf{b}(i) = \frac{(\mathbf{a}_i\circ\mathbf{a}_i)^T\mathbf{1}_{[m]}}{\sqrt{m}} = \frac{\|\mathbf{a}_i\|_2^2}{\sqrt{m}} \quad \forall i\in[n], \quad (43)
$$

where $\mathbf{a}_i$ denotes the ith column of $\mathbf{A}$. Since we have assumed that $\|\mathbf{a}_i\|_2^2 \in [1-\alpha, 1+\alpha]$,

$$
\frac{1-\alpha}{\sqrt{m}}\mathbf{1}_{[n]} \preceq (\mathbf{A}\odot\mathbf{A})^T\mathbf{w} \preceq \frac{1+\alpha}{\sqrt{m}}\mathbf{1}_{[n]}. \quad (44)
$$

To ease the notation, let $\mathbf{X} = \mathbf{A}\odot\mathbf{A}$. Given projection matrices $\mathbf{w}\mathbf{w}^T$ and $\Pi = \mathbf{I}_{m^2} - \mathbf{w}\mathbf{w}^T$, one can write

$$\mathbf{v}^T\mathbf{X}^T\mathbf{X}\mathbf{v}$$

$$
\begin{aligned}
&= \mathbf{v}^T\mathbf{X}^T\mathbf{w}\mathbf{w}^T\mathbf{X}\mathbf{v} + \mathbf{v}^T\mathbf{X}^T\Pi\mathbf{X}\mathbf{v}\\
&\geq \mathbf{v}^T\mathbf{X}^T\mathbf{w}\mathbf{w}^T\mathbf{X}\mathbf{v}\\
&= \mathbf{v}_+^T\mathbf{X}^T\mathbf{w}\mathbf{w}^T\mathbf{X}\mathbf{v}_+ + \mathbf{v}_-^T\mathbf{X}^T\mathbf{w}\mathbf{w}^T\mathbf{X}\mathbf{v}_-\\
&\quad - 2\mathbf{v}_+^T\mathbf{X}^T\mathbf{w}\mathbf{w}^T\mathbf{X}\mathbf{v}_-\\
&\stackrel{(a)}{\geq} \frac{(1-\alpha)^2}{m}\left(\mathbf{v}_+^T\mathbf{1}_n\mathbf{1}_n^T\mathbf{v}_+\right) + \frac{(1-\alpha)^2}{m}\left(\mathbf{v}_-^T\mathbf{1}_n\mathbf{1}_n^T\mathbf{v}_-\right)\\
&\quad - 2\frac{(1+\alpha)^2}{m}\left(\mathbf{v}_+^T\mathbf{1}_n\mathbf{1}_n^T\mathbf{v}_-\right)\\
&= \frac{(1-\alpha)^2}{m}\left(\|\mathbf{v}_+\|_1^2 + \|\mathbf{v}_-\|_1^2\right) - 2\|\mathbf{v}_+\|_1\|\mathbf{v}_-\|_1\frac{(1+\alpha)^2}{m}\\
&= \frac{(1-\alpha)^2}{m}\left(\|\mathbf{v}_+\|_1^2 + \|\mathbf{v}_-\|_1^2\right)\\
&\quad \times \left[1 - \frac{2\|\mathbf{v}_+\|_1\|\mathbf{v}_-\|_1}{\|\mathbf{v}_+\|_1^2 + \|\mathbf{v}_-\|_1^2}\left(\frac{1+\alpha}{1-\alpha}\right)^2\right]. \quad (45)
\end{aligned}
$$

In above, step (a) follows from (44) and the nonnegativity of $\mathbf{v}_+$ and $\mathbf{v}_-$. We observe that for $\|\mathbf{v}_+\|_1 > 4\left(\frac{1+\alpha}{1-\alpha}\right)^2\|\mathbf{v}_-\|_1$, the ratio $\frac{2\|\mathbf{v}_+\|_1\|\mathbf{v}_-\|_1}{\|\mathbf{v}_+\|_1^2 + \|\mathbf{v}_-\|_1^2} \leq \frac{1}{2}\left(\frac{1-\alpha}{1+\alpha}\right)^2$, and therefore

$$
\mathbf{v}^T\mathbf{X}^T\mathbf{X}\mathbf{v} \geq \frac{(1-\alpha)^2}{2m}\left(\|\mathbf{v}_+\|_1^2 + \|\mathbf{v}_-\|_1^2\right).
$$

H. Proof of Proposition 6

Proof. Let us define $\Delta\boldsymbol{\gamma} = \boldsymbol{\gamma} - \boldsymbol{\gamma}^*$ which also splits as

$$
\Delta\boldsymbol{\gamma} = \Delta\boldsymbol{\gamma}_+ - \Delta\boldsymbol{\gamma}_-, \quad (46)
$$

where $\Delta\boldsymbol{\gamma}_+$ and $\Delta\boldsymbol{\gamma}_-$ are nonnegative vectors in $\mathbb{R}_+^n$ with non-overlapping supports and containing absolute values of positive and negative coefficients of $\Delta\boldsymbol{\gamma}$, respectively. Let $\mathcal{S}$ and $\mathcal{S}^*$ denote the nonzero supports of $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^*$, respectively. Suppose $\mathcal{S}$ and $\mathcal{S}^*$ differ in exactly $k_d^{\mathcal{S},\mathcal{S}^*}$ locations. By construction of $\Delta\boldsymbol{\gamma}_+$ and $\Delta\boldsymbol{\gamma}_-$, we have

$$
\|\Delta\boldsymbol{\gamma}\|_2^2 \geq k_d^{\mathcal{S},\mathcal{S}^*}\gamma_{\min}^2 \quad (47)
$$

$$
\left(|\mathcal{S}^*| - k_d^{\mathcal{S},\mathcal{S}^*}\right)_+ \gamma_{\min} \leq \|\Delta\boldsymbol{\gamma}_-\|_1 \leq |\mathcal{S}^*|\gamma_{\max} \quad (48)
$$

$$
\left(k_d^{\mathcal{S},\mathcal{S}^*} - |\mathcal{S}^*|\right)_+ \gamma_{\min} \leq \|\Delta\boldsymbol{\gamma}_+\|_1 \leq k_d^{\mathcal{S},\mathcal{S}^*}\gamma_{\max}
$$
$$
+ |\mathcal{S}^*|(\gamma_{\max} - \gamma_{\min}) \quad (49)
$$

We introduce $K_{\mathrm{threshold}} \triangleq \left(1 + 4\frac{\gamma_{\max}}{\gamma_{\min}}\left(\frac{1+\alpha}{1-\alpha}\right)^2\right)K$, and $\mathcal{B} \triangleq \left\{\mathcal{S}\in[n]: k_d^{\mathcal{S},\mathcal{S}^*} < K_{\mathrm{threshold}}\right\}$. Then, from (23), we have

$$
\eta = \min_{\mathcal{S}\subseteq[n]} \frac{\mathcal{D}_\mathcal{S}^*}{k_d^{\mathcal{S},\mathcal{S}^*}} = \min\left(\min_{\mathcal{S}\in\mathcal{B}} \frac{\mathcal{D}_\mathcal{S}^*}{k_d^{\mathcal{S},\mathcal{S}^*}}, \min_{\mathcal{S}\in\mathcal{B}^c} \frac{\mathcal{D}_\mathcal{S}^*}{k_d^{\mathcal{S},\mathcal{S}^*}}\right). \quad (50)
$$

Note that for $\mathrm{supp}(\boldsymbol{\gamma}) = \mathcal{S}$ and $\mathcal{S}\in\mathcal{B}$, from property $P2$, we have $\|(\mathbf{A}\odot\mathbf{A})\Delta\boldsymbol{\gamma}\|_2^2 \geq \beta\|\Delta\boldsymbol{\gamma}\|_2^2$. Using the lower bound on $\mathcal{D}_\mathcal{S}^*$ derived in Proposition 5, we can bound $\min_{\mathcal{S}\in\mathcal{B}} \frac{\mathcal{D}_\mathcal{S}^*}{k_d^{\mathcal{S},\mathcal{S}^*}}$ as follows.

$$
\min_{\mathcal{S}\in\mathcal{B}} \frac{\mathcal{D}_\mathcal{S}^*}{k_d^{\mathcal{S},\mathcal{S}^*}} = \min_{\substack{\boldsymbol{\gamma}\in\Theta(\mathcal{S}),\\\mathcal{S}\in\mathcal{B}}} \frac{\beta\|\Delta\boldsymbol{\gamma}\|_2^2}{4k_d^{\mathcal{S},\mathcal{S}^*}\left(\sigma^2 + \gamma_{\max}\sigma_{\max}^2(\mathbf{A}_{\mathcal{S}\cup\mathcal{S}^*})\right)^2}
$$

$$\geq \min_{\mathcal{S}\in\mathcal{B}} \frac{\beta k_d^{\mathcal{S},\mathcal{S}^*} \gamma_{\min}^2}{4k_d^{\mathcal{S},\mathcal{S}^*} \left(\sigma^2 + \gamma_{\max}\sigma_{\max}^2(\mathbf{A}_{\mathcal{S}\cup\mathcal{S}^*})\right)^2}$$

$$\geq \frac{\beta\gamma_{\min}^2}{4\left(\sigma^2+\gamma_{\max}\right)^2 \max(1,\delta_{(K+K_{\text{threshold}})}^2)}. \quad (51)$$

For the case where $\mathcal{S} \in \mathcal{B}^c$, i.e., $k_d^{\mathcal{S},\mathcal{S}^*} \geq K_{\text{threshold}}$, it follows from (48) and (49) that $\|\Delta\gamma_+\|_1 \geq 4\left(\frac{1+\alpha}{1-\alpha}\right)\|\Delta\gamma_-\|_1$. Therefore, we can invoke the restricted null space property of $\mathbf{A}\odot\mathbf{A}$ from Theorem 3 to bound $\min_{\mathcal{S}\in\mathcal{B}^c}\frac{\mathcal{D}_{\mathcal{S}}^*}{k_d^{\mathcal{S},\mathcal{S}^*}}$ as follows.

$$\min_{\mathcal{S}\in\mathcal{B}^c} \frac{\mathcal{D}_{\mathcal{S}}^*}{k_d^{\mathcal{S},\mathcal{S}^*}}$$

$$\geq \min_{\substack{\gamma\in\Theta(\mathcal{S}),\\\mathcal{S}\in\mathcal{B}^c}} \frac{(1-\alpha)^2\left(\|\Delta\gamma_+\|_1^2+\|\Delta\gamma_-\|_1^2\right)}{8mk_d^{\mathcal{S},\mathcal{S}^*}\left(\sigma^2+\gamma_{\max}\sigma_{\max}^2(\mathbf{A}_{\mathcal{S}\cup\mathcal{S}^*})\right)^2}$$

$$\geq \min_{\substack{\gamma\in\Theta(\mathcal{S}),\\\mathcal{S}\in\mathcal{B}^c}} \frac{(1-\alpha)^2\|\Delta\gamma_+\|_1^2}{8mk_d^{\mathcal{S},\mathcal{S}^*}(\sigma^2+\gamma_{\max})^2\left(\max\left(1,\sigma_{\max}^2(\mathbf{A}_{\mathcal{S}\cup\mathcal{S}^*})\right)\right)^2}$$

$$\geq \min_{\mathcal{S}\in\mathcal{B}^c} \frac{(1-\alpha)^2\gamma_{\min}^2\left(k_d^{\mathcal{S},\mathcal{S}^*}-|\mathcal{S}^*|\right)^2}{8mk_d^{\mathcal{S},\mathcal{S}^*}(\sigma^2+\gamma_{\max})^2\left(\max\left(1,\sigma_{\max}^2(\mathbf{A}_{\mathcal{S}\cup\mathcal{S}^*})\right)\right)^2}$$

$$\geq \frac{(1-\alpha)^2\gamma_{\min}^2}{8m\left(\sigma^2+\gamma_{\max}\right)^2} \min_{\mathcal{S}\in\mathcal{B}^c}\left(1-\frac{|\mathcal{S}^*|}{k_d^{\mathcal{S},\mathcal{S}^*}}\right)$$
$$\times \left(\min_{\mathcal{S}\in\mathcal{B}^c}\frac{k_d^{\mathcal{S},\mathcal{S}^*}-|\mathcal{S}^*|}{\left(\max\left(1,\sigma_{\max}^2(\mathbf{A}_{\mathcal{S}\cup\mathcal{S}^*})\right)\right)^2}\right)$$

$$\geq \frac{(1-\alpha)^2\gamma_{\min}^2}{8m\left(\sigma^2+\gamma_{\max}\right)^2}\left(1-\frac{K}{K_{\text{threshold}}}\right)$$
$$\times \left(\min_{\mathcal{S}\in\mathcal{B}^c}\frac{k_d^{\mathcal{S},\mathcal{S}^*}-|\mathcal{S}^*|}{\left(\max\left(1,\sigma_{\max}^2(\mathbf{A}_{\mathcal{S}\cup\mathcal{S}^*})\right)\right)^2}\right)$$

$$\geq \frac{(1-\alpha)^2\gamma_{\min}^2}{10m\left(\sigma^2+\gamma_{\max}\right)^2}\left(\min_{\mathcal{S}\in\mathcal{B}^c}\frac{k_d^{\mathcal{S},\mathcal{S}^*}-|\mathcal{S}^*|}{|\mathcal{S}\cup\mathcal{S}^*|}\right)$$
$$\times \left(\min_{\mathcal{S}\in\mathcal{B}^c}\frac{|\mathcal{S}\cup\mathcal{S}^*|}{\left(\max\left(1,\sigma_{\max}^2(\mathbf{A}_{\mathcal{S}\cup\mathcal{S}^*})\right)\right)^2}\right)$$

$$= \frac{(1-\alpha)^2\gamma_{\min}^2}{10m\left(\sigma^2+\gamma_{\max}\right)^2}\left(\min_{\mathcal{S}\in\mathcal{B}^c}\frac{k_d^{\mathcal{S},\mathcal{S}^*}-|\mathcal{S}^*|}{k_d^{\mathcal{S},\mathcal{S}^*}+|\mathcal{S}^*|}\right)$$
$$\times \left(\min_{\mathcal{S}\in\mathcal{B}^c}\frac{|\mathcal{S}\cup\mathcal{S}^*|}{\left(\max\left(1,\sigma_{\max}^2(\mathbf{A}_{\mathcal{S}\cup\mathcal{S}^*})\right)\right)^2}\right)$$

$$\geq \frac{(1-\alpha)^2\gamma_{\min}^2}{10m\left(\sigma^2+\gamma_{\max}\right)^2}\left(\frac{K_{\text{threshold}}-K}{K_{\text{threshold}}+K}\right)$$
$$\times \left(\min_{\mathcal{S}\in\mathcal{B}^c}\frac{|\mathcal{S}\cup\mathcal{S}^*|}{\left(\max\left(1,\sigma_{\max}^2(\mathbf{A}_{\mathcal{S}\cup\mathcal{S}^*})\right)\right)^2}\right)$$

$$\geq \frac{(1-\alpha)^2\gamma_{\min}^2}{16m\left(\sigma^2+\gamma_{\max}\right)^2}\left(\min_{\mathcal{S}\in\mathcal{B}^c}\frac{|\mathcal{S}\cup\mathcal{S}^*|}{\max(1,\delta_{|\mathcal{S}\cup\mathcal{S}^*|}^2)}\right). \quad (52)$$

In the above, the third to the last inequality follows by substituting $K_{\text{threshold}}$ with its lower bound $5K$. The penultimate inequality is obtained by noting that $|\mathcal{S}^*| \leq K$, and the ratio $\frac{k_d^{\mathcal{S},\mathcal{S}^*}-|\mathcal{S}^*|}{k_d^{\mathcal{S},\mathcal{S}^*}+|\mathcal{S}^*|}$ increases monotonically with $k_d^{\mathcal{S},\mathcal{S}^*}$ for $\mathcal{S} \in \mathcal{B}^c$. The final inequality is obtained by noting that the ratio $\frac{K_{\text{threshold}}-K}{K_{\text{threshold}}+K} \geq \frac{2}{3}$ for $\mathcal{S}\in\mathcal{B}^c$.

Substituting (51) and (52) in (50) and simplifying, we obtain the lower bound for η stated in the proposition. $\qquad\square$

I. Proof of Proposition 7

Proof. For any support $\mathcal{S} \subset [n]$, by setting $\epsilon = \frac{L\mathcal{D}_{\mathcal{S}}^*}{2}$ in Proposition 4, we have

$$|\Theta^\epsilon(\mathcal{S})|_{\mathcal{G}} \leq \max\left\{1, \left(\frac{6\sqrt{|\mathcal{S}|}(\gamma_{\max}-\gamma_{\min})C_{\mathcal{L},\mathcal{S}}}{L\mathcal{D}_{\mathcal{S}}^*}\right)^{|\mathcal{S}|}\right\}. \quad (53)$$

where $C_{\mathcal{L},\mathcal{S}}$ denotes the Lipschitz constant of $\mathcal{L}(\mathbf{Y};\gamma)$ with respect to γ over the bounded domain $\Theta^\epsilon(\mathcal{S})|_{\mathcal{G}}$. Proposition 9 characterizes the Lipschitz property of $\mathcal{L}(\mathbf{Y},\gamma)$.

Proposition 9. *For $\mathcal{S} \in \mathcal{S}_n$, the log-likelihood $\mathcal{L}(\mathbf{Y};\gamma)$: $\Theta(\mathcal{S}) \to \mathbb{R}$ is Lipschitz continuous in γ as shown below.*

$$|\mathcal{L}(\mathbf{Y},\gamma_2) - \mathcal{L}(\mathbf{Y},\gamma_1)|$$
$$\leq \frac{mL}{\gamma_{min}}\left(1+\frac{\|\mathbf{R_{yy}}\|_2}{\sigma^2}\right)\|\gamma_2-\gamma_1\|_2,$$

for any $\gamma_1,\gamma_2\in\Theta(\mathcal{S})$. Here, $\mathbf{R_{yy}} \triangleq \frac{1}{L}\mathbf{YY}^T$.

Proof. See Appendix K. $\qquad\square$

By invoking the definitions of $K_{\text{threshold}}$ and set $\mathcal{B}$ from the proof of Proposition 6, we can rewrite κ_{cov} in (24) as

$$\kappa_{\text{cov}} = \max\left(\max_{\mathcal{S}\in\mathcal{B}}\frac{\log|\Theta^\epsilon(\mathcal{S})|_{\mathcal{G}}|}{k_d^{\mathcal{S},\mathcal{S}^*}}, \max_{\mathcal{S}\in\mathcal{B}^c}\frac{\log|\Theta^\epsilon(\mathcal{S})|_{\mathcal{G}}|}{k_d^{\mathcal{S},\mathcal{S}^*}}\right). \quad (54)$$

For the MMV set $\mathcal{G}$ as defined in (15) and $\mathbf{Y}\in\mathcal{G}$, we have $\|\mathbf{R_{yy}}\|_2 \leq 2\left(\sigma^2+\gamma_{\max}\delta_K\right)$. Then, for $\gamma_{\min}\neq\gamma_{\max}$ and by the Lipschitz continuity of $\mathcal{L}(\mathbf{Y};\gamma)$ as per Proposition 9, it follows that

$$\max_{\mathcal{S}\in\mathcal{B}}\frac{\log\left|\Theta^{\frac{L\mathcal{D}_{\mathcal{S}}^*}{2}}(\mathcal{S})|_{\mathcal{G}}\right|}{k_d^{\mathcal{S},\mathcal{S}^*}}$$

$$\leq \max_{\mathcal{S}\in\mathcal{B}}\frac{|\mathcal{S}|}{k_d^{\mathcal{S},\mathcal{S}^*}}\log\left(\frac{6\sqrt{|\mathcal{S}|}C_{\mathcal{L},\mathcal{S}}(\gamma_{\max}-\gamma_{\min})}{L\mathcal{D}_{\mathcal{S}}^*}\right)$$

$$\leq \max_{\mathcal{S}\in\mathcal{B}}\frac{|\mathcal{S}^*|+k_d^{\mathcal{S},\mathcal{S}^*}}{k_d^{\mathcal{S},\mathcal{S}^*}}$$
$$\times \log\left(\frac{6m\sqrt{|\mathcal{S}^*|+k_d^{\mathcal{S},\mathcal{S}^*}}(\gamma_{\max}-\gamma_{\min})(3+2\frac{\gamma_{\max}}{\sigma^2}\delta_K)}{\gamma_{\min}\mathcal{D}_{\mathcal{S}}^*}\right)$$

$$\leq (K+1)\log\left(m\sqrt{K+1}\Delta_1\right), \quad (55)$$

where $\Delta_1 = \frac{24(\gamma_{\max}-\gamma_{\min})\left(3+\frac{2\gamma_{\max}\delta_K}{\sigma^2}\right)(\sigma^2+\gamma_{\max})^2\max\left(1,\delta_{K+K_{\text{threshold}}}^2\right)}{\beta\gamma_{\min}^3}$. The last inequality is obtained by using (51) and noting that the RHS in the second inequality decreases monotonically with respect to $k_d^{\mathcal{S},\mathcal{S}^*}$ in the interval $[1, K_{\text{threshold}}]$.

The second max-term in (54) can be bounded as follows. By invoking Proposition 5, we first note that

$$\min_{\mathcal{S}\in\mathcal{B}^c}\mathcal{D}_{\mathcal{S}}^* \geq \min_{\substack{\gamma\in\Theta(\mathcal{S}),\\\mathcal{S}\in\mathcal{B}^c}}\frac{\|(\mathbf{A}\odot\mathbf{A})(\gamma-\gamma^*)\|_2^2}{4(\sigma^2+\gamma_{\max}\sigma_{\max}^2(\mathbf{A}_{\mathcal{S}\cup\mathcal{S}^*}))^2}$$

$$\geq \min_{\mathcal{S}\in\mathcal{B}^c} \frac{(1-\alpha)^2(k_d^{\mathcal{S},\mathcal{S}^*}-|\mathcal{S}^*|)^2\gamma_{\min}^2}{8m(\sigma^2+\gamma_{\max}\sigma_{\max}^2(\mathbf{A}_{\mathcal{S}\cup\mathcal{S}^*}))^2}$$

$$\geq \min_{\mathcal{S}\in\mathcal{B}^c} \left(\frac{(k_d^{\mathcal{S},\mathcal{S}^*}-|\mathcal{S}^*|)^2}{|\mathcal{S}\cup\mathcal{S}^*|^2}\right)\frac{(1-\alpha)^2\gamma_{\min}^2}{8m(1+\alpha)^2(\sigma^2+\gamma_{\max})^2}$$

$$= \min_{\mathcal{S}\in\mathcal{B}^c} \left(\frac{k_d^{\mathcal{S},\mathcal{S}^*}-|\mathcal{S}^*|}{k_d^{\mathcal{S},\mathcal{S}^*}+|\mathcal{S}^*|}\right)^2\left(\frac{1-\alpha}{1+\alpha}\right)^2\frac{\gamma_{\min}^2}{8m(\sigma^2+\gamma_{\max})^2}$$

$$\geq \frac{1}{18m}\left(\frac{1-\alpha}{1+\alpha}\right)^2\frac{\gamma_{\min}^2}{(\sigma^2+\gamma_{\max})^2}. \tag{56}$$

The last inequality follows from the fact that $|\mathcal{S}^*|\leq K$ and the ratio $\left(\frac{k_d^{\mathcal{S},\mathcal{S}^*}-|\mathcal{S}^*|}{k_d^{\mathcal{S},\mathcal{S}^*}+|\mathcal{S}^*|}\right)$ is nondecreasing in $|\mathcal{S}|$ for $\mathcal{S}\in\mathcal{B}^c$. Then, substituting the lower bound for $\mathcal{D}_\mathcal{S}^*$ from (56) in (53), and simplifying, we obtain

$$\max_{\mathcal{S}\in\mathcal{B}^c}\frac{\log\left|\Theta^{\frac{L\mathcal{D}_\mathcal{S}^*}{2}}(\mathcal{S})|_\mathcal{G}\right|}{k_d^{\mathcal{S},\mathcal{S}^*}}$$

$$\leq \max_{\mathcal{S}\in\mathcal{B}^c}\frac{|\mathcal{S}^*|+k_d^{\mathcal{S},\mathcal{S}^*}}{k_d^{\mathcal{S},\mathcal{S}^*}}\log\left(m^2\sqrt{|\mathcal{S}^*|+k_d^{\mathcal{S},\mathcal{S}^*}}\Delta_2\right)$$

$$\leq 4\log m + 2\log\Delta_2 + \frac{n}{n-K}\log n \tag{57}$$

where $\Delta_2 = \frac{108(\gamma_{\max}-\gamma_{\min})\left(3+2\frac{\gamma_{\max}\delta_K}{\sigma^2}\right)(\sigma^2+\gamma_{\max})^2(1+\alpha)^2}{(1-\alpha)^2\gamma_{\min}^3}$. The coefficient $\frac{n}{n-K}$ in (57) can be replaced by a suitable constant, say 2, even for the worst scaling $K=\Theta(n)$.

Finally, substituting the bounds from (55) and (57) in (54), we obtain the claimed upper bound for κ_{cov}. $\qquad\square$

J. Proof of Corollary 3

In order to bound η, we first derive probabilistic bounds for the parameters α, β and δ_K associated with the sensing matrix $\mathbf{A}$. We consider both cases: i. $m=\Theta(K\log n)$ and ii. $m=\Theta(\sqrt{K}\log n)$.

1) Lower bound for α: By invoking the Hanson-Wright concentration inequality [50], and taking the union bound over all columns of $\mathbf{A}$,

$$\mathbb{P}\left(\bigcup_{i\in[n]}\left\{\left|\|\mathbf{a}_i\|_2^2-1\right|\geq\alpha\right\}\right)\leq n\mathbb{P}\left(|\mathbf{a}_1^T\mathbf{I}_m\mathbf{a}_1-1|\geq\alpha\right)$$
$$\leq 2ne^{-cm\alpha^2}, \tag{58}$$

where c is a numerical constant. For both $m=\Theta(K\log n)$ and $m=\Theta(\sqrt{K}\log n)$, by setting $\alpha=\frac{1}{2}$ in (58), the squared ℓ_2-norm of columns of $\mathbf{A}$ lie inside the interval $\left[\frac{1}{2},\frac{3}{2}\right]$ with probability exceeding $1-n^{-\Theta(\sqrt{K})}$.

2) Lower bound for β: In Proposition 6, β refers to the smallest squared singular value among all submatrices obtained by sampling $K+K_{\text{threshold}}$ or fewer columns of $\mathbf{A}\odot\mathbf{A}$.

Note that the squared k^{th} order restricted minimum singular value of any matrix is lower bounded by $1-\delta_k$, δ_k being the kth order restricted isometry constant of the matrix. Thus, noting that $K_{\text{threshold}}=\Theta(K)$, and by invoking [51, Theorem 1] (a revised version of [52, Theorem 5]) with $\delta=\frac{1}{2}$, $b=2$, $B=\Theta(1)$, for both $m=\Theta(K\log n)$ and $m=\Theta(\sqrt{K}\log n)$, we

obtain the following tail probability for $\bar{\beta}$, the $(K+K_{\text{threshold}})^{\text{th}}$ order minimum restricted singular value of the centered-and-rescaled version of $\mathbf{A}\odot\mathbf{A}$ as defined in [52]:

$$\mathbb{P}\left(\bar{\beta}<\frac{1}{2}\right)\leq c_2 n^{-2}, \tag{59}$$

where c_2 is a numerical constant.

Further, by invoking [51, Theorem 4] with a large enough value of ξ, for both $m=\Theta(K\log n)$ and $m=\Theta(\sqrt{K}\log n)$, the squared restricted minimum singular values of the uncentered Khatri-Rao product $\mathbf{A}\odot\mathbf{A}$ and its centered-and-rescaled counterpart [52] can be guaranteed to differ by at most $\frac{1}{4}$ with probability exceeding $1-2n^{-\Theta(K)}$ and $1-2n^{-\Theta(\sqrt{K})}$, respectively. Therefore, we further have

$$\beta\geq\frac{1}{4} \tag{60}$$

with probability exceeding $1-\Theta(n^{-2})$.

3) Upper bounds for $\delta_{K+K_{\text{threshold}}}$ and δ_n: From Proposition 6, δ_k is defined as the maximum squared singular value among any k or fewer column submatrix of $\mathbf{A}$. Thus, by direct application of Corollary 1, for $k\leq n$,

$$\delta_k\leq\frac{\left(\sqrt{m}+\sqrt{k}+\sqrt{6k\log n}\right)^2}{m} \tag{61}$$

with probability exceeding $1-2n^{-k}$. From (61), for $m=\Theta(K\log n)$ we have

$$\delta_{K+K_{\text{threshold}}}=O(1),\text{ and }\delta_n=O\left(\frac{n}{K\log n}\right), \tag{62}$$

and $m=\Theta(\sqrt{K}\log n)$ guarantees that

$$\delta_{K+K_{\text{threshold}}}=O(\sqrt{K}),\text{ and }\delta_n=O\left(\frac{n}{\sqrt{K}\log n}\right), \tag{63}$$

with probability exceeding $1-n^{-\Theta(\sqrt{K})}$. In the above, the bounds for δ_n follow from Proposition 3.

Finally, for $m=\Theta(K\log n)$, by substituting the above bounds for α, β, δ_K and δ_n from (60) and (62) in Propositions 6 and 7, and simplifying, it can be verified that $\eta=\Omega\left(\frac{K}{n}\right)$ and $\kappa_{\text{cov}}\leq O(K\log K+K\log\log n+\log n)$, respectively, with probability exceeding $1-\Omega(n^{-2})$. Likewise, for the $m=\Theta(\sqrt{K}\log n)$ case, we have $\eta=\Omega\left(\frac{\sqrt{K}}{n}\right)$ and $\kappa_{\text{cov}}\leq O(K\log K+K\log\log n+\log n)$ with probability exceeding $1-c_3 n^{-\Omega(1)}$, where c_3 is a numerical constant.

K. Proof of Proposition 9

The log-likelihood $\mathcal{L}(\mathbf{Y};\gamma)$ can be expressed as the sum $f(\gamma)+g(\gamma)$ with $f(\gamma)=-L\log|\Sigma_\gamma|$ and $g(\gamma)=-L\text{tr}\left(\Sigma_\gamma^{-1}\mathbf{R}_{\mathbf{yy}}\right)$. Here, $\Sigma_\gamma=\sigma^2\mathbf{I}_m+\mathbf{A}\Gamma\mathbf{A}^T$.

First, we derive an upper bound for the Lipschitz constant of $f(\gamma)=-L\log|\Sigma_\gamma|$ for $\gamma\in\Theta(\mathcal{S})$. By the mean value theorem, any upper bound for $\|\nabla_\gamma f(\gamma)\|_2$ also serves as an upper bound for the Lipschitz constant of f. So, we derive an upper bound for $\|\nabla_\gamma f(\gamma)\|_2$. Note that $\left|\frac{\partial f(\gamma)}{\partial\gamma(i)}\right|=L(\mathbf{a}_i^T\Sigma_\gamma^{-1}\mathbf{a}_i)$ for

$i \in \mathcal{S}$, and 0 otherwise. Here, $\mathbf{a}_i$ denotes the ith column of $\mathbf{A}$. Then, $\|\nabla_{\boldsymbol{\gamma}} f(\boldsymbol{\gamma})\|_2$ can be upper bounded as shown below.

$$
\begin{aligned}
\|\nabla_{\boldsymbol{\gamma}} f(\boldsymbol{\gamma})\|_2 &\leq \|\nabla_{\boldsymbol{\gamma}} f(\boldsymbol{\gamma})\|_1 = L \sum_{i \in \mathcal{S}} \mathbf{a}_i^T \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \mathbf{a}_i \\
&= L \left(\text{tr} \left(\mathbf{A}_{\mathcal{S}}^T (\sigma^2 \mathbf{I}_m + \mathbf{A}_{\mathcal{S}} \boldsymbol{\Gamma}_{\mathcal{S}} \mathbf{A}_{\mathcal{S}}^T)^{-1} \mathbf{A}_{\mathcal{S}} \right) \right) \\
&= L \left(\text{tr} \left(\boldsymbol{\Gamma}_{\mathcal{S}}^{-1/2} \tilde{\mathbf{A}}_{\mathcal{S}}^T (\sigma^2 \mathbf{I}_m + \tilde{\mathbf{A}}_{\mathcal{S}} \tilde{\mathbf{A}}_{\mathcal{S}}^T)^{-1} \tilde{\mathbf{A}}_{\mathcal{S}} \boldsymbol{\Gamma}_{\mathcal{S}}^{-1/2} \right) \right) \\
&\overset{(a)}{=} L \left\| \boldsymbol{\Gamma}_{\mathcal{S}}^{-1} \right\|_2 \text{tr} \left(\tilde{\mathbf{A}}_{\mathcal{S}}^T (\sigma^2 \mathbf{I}_m + \tilde{\mathbf{A}}_{\mathcal{S}} \tilde{\mathbf{A}}_{\mathcal{S}}^T)^{-1} \tilde{\mathbf{A}}_{\mathcal{S}} \right) \\
&\overset{(b)}{\leq} (L \min(m, |\mathcal{S}|)) / \boldsymbol{\gamma}_{\min}.
\end{aligned} \tag{64}
$$

where $\tilde{\mathbf{A}}_{\mathcal{S}} = \mathbf{A}_{\mathcal{S}} \boldsymbol{\Gamma}_{\mathcal{S}}^{1/2}$. In the above, step (a) follows from the trace inequality $\text{tr}(\mathbf{AB}) \leq \|\|\mathbf{A}\|\|_2 \text{tr}(\mathbf{B})$ for any positive definite matrices $\mathbf{A}$ and $\mathbf{B}$. Step (b) follows from the observation that input argument of the trace operator has $\min(m, |\mathcal{S}|)$ nonzero eigenvalues, all of them less than unity.

We now shift focus to the second term $g(\boldsymbol{\gamma})$ of the loglikelihood. Note that $\left| \frac{\partial g(\boldsymbol{\gamma})}{\partial \gamma(i)} \right| = L(\mathbf{a}_i^T \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \mathbf{R}_Y \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \mathbf{a}_i)$ for $i \in \mathcal{S}$, and 0 otherwise. Then, $\|\nabla_{\boldsymbol{\gamma}} g(\boldsymbol{\gamma})\|_2$ can be upper bounded as

$$
\begin{aligned}
\|\nabla g(\boldsymbol{\gamma})\|_2 &\leq \|\nabla_{\boldsymbol{\gamma}} g(\boldsymbol{\gamma})\|_1 = L \sum_{i \in \mathcal{S}} \mathbf{a}_i^T \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \mathbf{R}_Y \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \mathbf{a}_i \\
&= L \left(\text{tr} \left(\mathbf{A}_{\mathcal{S}}^T \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \mathbf{R}_Y \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \mathbf{A}_{\mathcal{S}} \right) \right) \\
&\leq L \|\mathbf{R_Y}\|_2 \text{tr} \left(\mathbf{A}_{\mathcal{S}}^T \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \mathbf{A}_{\mathcal{S}} \right) \\
&= L \|\mathbf{R_Y}\|_2 \text{tr} \left(\boldsymbol{\Gamma}_{\mathcal{S}}^{-1/2} \tilde{\mathbf{A}}_{\mathcal{S}}^T \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \tilde{\mathbf{A}}_{\mathcal{S}} \boldsymbol{\Gamma}_{\mathcal{S}}^{-1/2} \right) \\
&\leq (L \|\mathbf{R_Y}\|_2 / \boldsymbol{\gamma}_{\min}) \text{tr} \left(\tilde{\mathbf{A}}_{\mathcal{S}}^T \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \tilde{\mathbf{A}}_{\mathcal{S}} \right) \\
&\leq \left(L \|\mathbf{R_Y}\|_2 \left\| \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \right\|_2 / \boldsymbol{\gamma}_{\min} \right) \text{tr} \left(\tilde{\mathbf{A}}_{\mathcal{S}}^T \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \tilde{\mathbf{A}}_{\mathcal{S}} \right) \\
&\overset{(a)}{\leq} \left(L \|\mathbf{R_Y}\|_2 \left\| \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \right\|_2 \min(m, |\mathcal{S}|) \right) / \boldsymbol{\gamma}_{\min} \\
&\overset{(b)}{\leq} (L \|\mathbf{R_Y}\|_2 \min(m, |\mathcal{S}|)) / \boldsymbol{\gamma}_{\min} \sigma^2.
\end{aligned} \tag{65}
$$

where $\tilde{\mathbf{A}}_{\mathcal{S}} \triangleq \mathbf{A}_{\mathcal{S}} \boldsymbol{\Gamma}_{\mathcal{S}}^{1/2}$. The inequality in $(65a)$ follows from $\left(\tilde{\mathbf{A}}_{\mathcal{S}}^T \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \tilde{\mathbf{A}}_{\mathcal{S}} \right)$ having $\min(m, |\mathcal{S}|)$ nonzero eigenvalues, all of them less than unity. The last inequality in $(65b)$ is due to $\left\| \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \right\|_2 \leq 1/\sigma^2$. Finally, the Lipschitz constant $C_{\mathcal{L}, \mathcal{S}}$ can be bounded as $C_{\mathcal{L}, \mathcal{S}} \leq \|\nabla_{\boldsymbol{\gamma}} f(\boldsymbol{\gamma})\|_2 + \|\nabla_{\boldsymbol{\gamma}} g(\boldsymbol{\gamma})\|_2$. Thus, by combining (64) and (65), and noting that $\min(m, |\mathcal{S}|) \leq m$, we obtain the desired result.

L. Proof of Proposition 8

We assume that $\text{supp}(\boldsymbol{\gamma}_1) \backslash \text{supp}(\boldsymbol{\gamma}_2) \neq \phi$. The proof also holds for the case where $\text{supp}(\boldsymbol{\gamma}_1) \subset \text{supp}(\boldsymbol{\gamma}_2)$ by swapping $\boldsymbol{\gamma}_1$ with $\boldsymbol{\gamma}_2$ and invoking the symmetry of $\mathcal{D}_{1/2}(p_{\boldsymbol{\gamma}_1}, p_{\boldsymbol{\gamma}_2})$ with respect to its input arguments.

Let μ^* be the largest eigenvalue of $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_1}^{\frac{1}{2}} \boldsymbol{\Sigma}_{\boldsymbol{\gamma}_2}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{\gamma}_1}^{\frac{1}{2}}$. Then,

$$
\begin{aligned}
\mu^* &\geq \frac{\text{tr} \left(\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_2}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{\gamma}_1} \right)}{m} \\
&= \frac{1}{m} \left[\sigma^2 \text{tr} \left(\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_2}^{-1} \right) + \text{tr} \left(\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_2}^{-1} \mathbf{A} \boldsymbol{\Gamma}_1 \mathbf{A}^T \right) \right] \\
&\geq \frac{1}{m} \text{tr} \left(\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_2}^{-1} \mathbf{A} \boldsymbol{\Gamma}_1 \mathbf{A}^T \right).
\end{aligned} \tag{66}
$$

Here, the second step is setting $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_1} = \sigma^2 \mathbf{I}_m + \mathbf{A} \boldsymbol{\Gamma}_1 \mathbf{A}^T$. The last inequality is obtained by dropping the strictly positive $\frac{\sigma^2}{m} \text{tr} \left(\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_2}^{-1} \right)$ term.

Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be the nonzero supports of $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$, respectively. Further, let the eigendecomposition of $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_2}$ be $\mathbf{U} \boldsymbol{\Lambda} \mathbf{U}^T$, where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_m)$, λ_i's are the eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_2}$ and $\mathbf{U}$ is a unitary matrix with columns as the eigenvectors of $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_2}$. Then, $\mathbf{U}$ can be partitioned as $[\mathbf{U}_2 \ \mathbf{U}_{2\perp}]$, where the columns of $\mathbf{U}_2$ and $\mathbf{U}_{2\perp}$ span the orthogonal complementary subspaces $\text{Col}(\mathbf{A}_{\mathcal{S}_2})$ and $\text{Col}(\mathbf{A}_{\mathcal{S}_2})^{\perp}$, respectively. Further, let $\boldsymbol{\Lambda}_2$ and $\boldsymbol{\Lambda}_{2\perp}$ be $|\mathcal{S}_2| \times |\mathcal{S}_2|$ and $((m - |\mathcal{S}_2|) \times (m - |\mathcal{S}_2|))$ sized diagonal matrices containing the eigenvalues in $\boldsymbol{\Lambda}$ corresponding to the eigenvectors in $\mathbf{U}_2$ and $\mathbf{U}_{2\perp}$, respectively. We observe that $\boldsymbol{\Lambda}_{2\perp} = \sigma^2 \mathbf{I}_{m-|\mathcal{S}_2|}$.

By setting $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_2}^{-1} = \mathbf{U}_2 \boldsymbol{\Lambda}_2^{-1} \mathbf{U}_2^T + \mathbf{U}_{2\perp} \boldsymbol{\Lambda}_{2\perp}^{-1} \mathbf{U}_{2\perp}^T$ in (66), we get

$$
\begin{aligned}
\mu^* &\geq \frac{1}{m} \big(\text{tr} \left(\mathbf{U}_2 \boldsymbol{\Lambda}_2^{-1} \mathbf{U}_2^T \mathbf{A} \boldsymbol{\Gamma}_1 \mathbf{A}^T \right) \\
&\qquad\qquad + \text{tr} \left(\mathbf{U}_{2\perp} \boldsymbol{\Lambda}_{2\perp}^{-1} \mathbf{U}_{2\perp}^T \mathbf{A} \boldsymbol{\Gamma}_1 \mathbf{A}^T \right) \big) \\
&\geq \frac{1}{m} \text{tr} \left(\boldsymbol{\Lambda}_{2\perp}^{-1} \mathbf{U}_{2\perp}^T \mathbf{A} \boldsymbol{\Gamma}_1 \mathbf{A}^T \mathbf{U}_{2\perp} \right),
\end{aligned}
$$

where the last inequality is due to nonnegativity of the first term. Since $\mathbf{U}_{2\perp}^T \mathbf{A}_{\mathcal{S}_2} = 0$ by construction of $\mathbf{U}_{2\perp}$,

$$
\begin{aligned}
\mu^* &\geq \frac{1}{m} \text{tr} \left(\boldsymbol{\Lambda}_{2\perp}^{-1} \mathbf{U}_{2\perp}^T \mathbf{A}_{\mathcal{S}_2^c} \boldsymbol{\Gamma}_{1, \mathcal{S}_2^c} \mathbf{A}_{\mathcal{S}_2^c}^T \mathbf{U}_{2\perp} \right) \\
&= \frac{1}{m\sigma^2} \sum_{i=1}^{m-|\mathcal{S}_2|} (\mathbf{u}_{2\perp, i})^T \mathbf{A}_{\mathcal{S}_2^c} \boldsymbol{\Gamma}_{1, \mathcal{S}_2^c} \mathbf{A}_{\mathcal{S}_2^c}^T \mathbf{u}_{2\perp, i} \\
&= \frac{1}{m\sigma^2} \sum_{i=1}^{m-|\mathcal{S}_2|} (\mathbf{u}_{2\perp, i})^T \mathbf{A}_{\mathcal{S}_1 \backslash \mathcal{S}_2} \boldsymbol{\Gamma}_{1, \mathcal{S}_1 \backslash \mathcal{S}_2} \mathbf{A}_{\mathcal{S}_1 \backslash \mathcal{S}_2}^T \mathbf{u}_{2\perp, i}.
\end{aligned} \tag{67}
$$

In the above, $\mathbf{u}_{2\perp, i}$ denotes the ith column of $\mathbf{U}_{2\perp}$. The last equality is obtained by observing that the nonzero elements of $\boldsymbol{\gamma}_{1, \mathcal{S}_2^c}$ are located in the index set $\mathcal{S}_1 \backslash \mathcal{S}_2$.

We now prove that if $K < \text{spark}(\mathbf{A}) - 1$, then there exists at least one strictly positive term in the above summation. Let us assume the contrary, i.e., let each term in the summation in (67) be equal to zero. This implies that the columns of $\mathbf{U}_{2\perp}$ belong to $\text{Null}(\boldsymbol{\Gamma}_{1, \mathcal{S}_1 \backslash \mathcal{S}_2}^{1/2} \mathbf{A}_{\mathcal{S}_1 \backslash \mathcal{S}_2}^T)$, which means that they also belong to $\text{Null}(\mathbf{A}_{\mathcal{S}_1 \backslash \mathcal{S}_2} \boldsymbol{\Gamma}_{1, \mathcal{S}_1 \backslash \mathcal{S}_2} \mathbf{A}_{\mathcal{S}_1 \backslash \mathcal{S}_2}^T)$. Since, for a symmetric matrix, the row and column spaces are equal and orthogonal to the null space of the matrix, it follows that $\text{Col}(\mathbf{A}_{\mathcal{S}_1 \backslash \mathcal{S}_2} \boldsymbol{\Gamma}_{1, \mathcal{S}_1 \backslash \mathcal{S}_2} \mathbf{A}_{\mathcal{S}_1 \backslash \mathcal{S}_2}^T)$ (same as $\text{Col}(\mathbf{A}_{\mathcal{S}_1 \backslash \mathcal{S}_2} \boldsymbol{\Gamma}_{1, \mathcal{S}_1 \backslash \mathcal{S}_2}^{1/2})$) is spanned by the columns of $\mathbf{U}_2$, or equivalently by columns of $\mathbf{A}_{\mathcal{S}_2}$. Thus, every column in $\mathbf{A}_{\mathcal{S}_1 \backslash \mathcal{S}_2}$ can be expressed as a linear combination of columns in $\mathbf{A}_{\mathcal{S}_2}$. Since $|\mathcal{S}_2| \leq K$, this contradicts our initial assumption that $K + 1 < \text{spark}(\mathbf{A})$. Therefore, we conclude that there is at least one strictly positive term in the summation in (67), and consequently there exists a constant $c_1 > 0$ such that $\mu^* \geq c_1/\sigma^2$.

REFERENCES

[1] M. Masood, L. Afify, and T. Al-Naffouri, "Efficient coordinated recovery of sparse channels in massive MIMO," *IEEE Trans. Signal Process.*, vol. 63, no. 1, pp. 104–118, 2015.

[2] Y. Barbotin, A. Hormati, S. Rangan, and M. Vetterli, "Estimation of sparse MIMO channels with common support," *IEEE Trans. Commun.*, vol. 60, no. 12, pp. 3705–3716, 2012.

[3] R. Prasad, C. Murthy, and B. Rao, "Joint channel estimation and data detection in MIMO-OFDM systems: A sparse Bayesian learning approach," *IEEE Trans. Signal Process.*, vol. 63, no. 20, pp. 5369–5382, Oct. 2015.

[4] D. Baron, M. F. Duarte, M. B. Wakin, S. Sarvotham, and R. G. Baraniuk, "Distributed compressive sensing," *CoRR*, vol. abs/0901.3403, 2009. [Online]. Available: http://arxiv.org/abs/0901.3403

[5] H. Zhu, G. Giannakis, and A. Cano, "Distributed in-network channel decoding," *IEEE Trans. Signal Process.*, vol. 57, no. 10, pp. 3970–3983, Oct. 2009.

[6] S. Ji, D. Dunson, and L. Carin, "Multitask compressive sensing," *IEEE Trans. Signal Process.*, vol. 57, no. 1, pp. 92–106, Jan. 2009.

[7] T. Erseghe, "A distributed and maximum-likelihood sensor network localization algorithm based upon a nonconvex problem formulation," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 1, no. 4, pp. 247–258, Dec. 2015.

[8] H. Qiao and P. Pal, "Guaranteed localization of more sources than sensors with finite snapshots in multiple measurement vector models using difference co-arrays," *IEEE Trans. Signal Process.*, vol. 67, no. 22, pp. 5715–5729, Nov 2019.

[9] Z. Fanzi, C. Li, and Z. Tian, "Distributed compressive spectrum sensing in cooperative multihop cognitive networks," *IEEE J. Sel. Top. Signal Process.*, vol. 5, no. 1, pp. 37–48, Feb. 2011.

[10] Y. Wang, A. Pandharipande, Y. L. Polo, and G. Leus, "Distributed compressive wide-band spectrum sensing," in *Information Theory and Applications Workshop*, Feb. 2009, pp. 178–183.

[11] J. Bazerque and G. Giannakis, "Distributed spectrum sensing for cognitive radio networks by exploiting sparsity," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1847–1862, Mar. 2010.

[12] A. Fengler, S. Haghighatshoar, P. Jung, and G. Caire, "Non-Bayesian activity detection, large-scale fading coefficient estimation, and unsourced random access with a massive mimo receiver," *ArXiv*, vol. abs/1910.11266, 2019.

[13] S. Cotter, B. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Trans. Signal Process.*, vol. 53, no. 7, pp. 2477–2488, Jul. 2005.

[14] J. Chen and X. Huo, "Theoretical results on sparse representations of multiple-measurement vectors," *IEEE Trans. Signal Process.*, vol. 54, no. 12, pp. 4634–4643, Dec. 2006.

[15] J. Tropp, A. Gilbert, and M. Strauss, "Simultaneous sparse approximation via greedy pursuit," in *Proc. ICASSP*, vol. 5, 2005, pp. 721–724.

[16] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. Birkhäuser Basel, 2013.

[17] B. D. Rao, K. Engan, and S. Cotter, "Diversity measure minimization based method for computing sparse solutions to linear inverse problems with multiple measurement vectors," in *Proc. ICASSP*, vol. 2, May 2004, pp. 369–372.

[18] J. D. Blanchard, M. Cermak, D. Hanle, and Y. Jing, "Greedy algorithms for joint sparse recovery," *IEEE Trans. Signal Process.*, vol. 62, no. 7, pp. 1694–1704, Apr. 2014.

[19] K. Lee, Y. Bresler, and M. Junge, "Subspace methods for joint sparse recovery," *IEEE Trans. Inf. Theory*, vol. 58, no. 6, pp. 3613–3641, Jun. 2012.

[20] J. M. Kim, O. K. Lee, and J. C. Ye, "Compressive music: Revisiting the link between compressive sensing and array signal processing," *IEEE Trans. Inf. Theory*, vol. 58, no. 1, pp. 278–301, Jan. 2012.

[21] M. Mishali and Y. C. Eldar, "Reduce and boost: Recovering arbitrary sets of jointly sparse vectors," *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 4692–4702, Oct. 2008.

[22] P. Pal and P. P. Vaidyanathan, "Pushing the limits of sparse support recovery using correlation information," *IEEE Trans. Signal Process.*, vol. 63, no. 3, pp. 711–726, Feb. 2015.

[23] D. Wipf and B. Rao, "An empirical Bayesian strategy for solving the simultaneous sparse approximation problem," *IEEE Trans. Signal Process.*, vol. 55, no. 7, pp. 3704–3716, Jul. 2007.

[24] O. Balkan, K. Kreutz-Delgado, and S. Makeig, "Localization of more sources than sensors via jointly-sparse Bayesian learning," *IEEE Signal Process. Lett.*, vol. 21, no. 2, pp. 131–134, Feb. 2014.

[25] S. Khanna and C. R. Murthy, "Rényi divergence based covariance matching pursuit of joint sparse support," in *Sig. Process. Advances in Wireless Commun. (SPAWC)*, Jul. 2017, pp. 1 – 5.

[26] S. Khanna and C. R. Murthy, "Sparse recovery from multiple measurement vectors using exponentiated gradient updates," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1485–1489, Oct 2018.

[27] L. Ramesh and C. R. Murthy, "Sparse support recovery via covariance estimation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 6633–6637.

[28] C. C. Yee and Y. F. Atchade, "On the sparse bayesian learning of linear models," *Communications in Statistics - Theory and Methods*, vol. 46, no. 15, pp. 7672–7691, 2017. [Online]. Available: https://doi.org/10.1080/03610926.2016.1158837

[29] G. Tang and A. Nehorai, "Performance analysis for sparse support recovery," *IEEE Trans. Inf. Theory*, vol. 56, no. 3, pp. 1383–1399, Mar. 2010.

[30] A. Koochakzadeh, H. Qiao, and P. Pal, "On fundamental limits of joint sparse support recovery using certain correlation priors," *IEEE Transactions on Signal Processing*, vol. 66, no. 17, pp. 4612–4625, Sep. 2018.

[31] E. van den Berg and M. P. Friedlander, "Theoretical and empirical results for recovery from multiple measurements," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2516–2527, May 2010.

[32] Y. C. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 5302–5316, Nov. 2009.

[33] M. E. Davies and Y. C. Eldar, "Rank awareness in joint sparse recovery," *IEEE Trans. Inf. Theory*, vol. 58, no. 2, pp. 1135–1146, Feb. 2012.

[34] Y. Jin and B. D. Rao, "Support recovery of sparse signals in the presence of multiple measurement vectors," *IEEE Trans. Inf. Theory*, vol. 59, no. 5, pp. 3139–3157, May 2013.

[35] S. Park, N. Y. Yu, and H. N. Lee, "An information-theoretic study for joint sparsity pattern recovery with different sensing matrices," *IEEE Trans. Inf. Theory*, vol. PP, no. 99, pp. 1–1, 2017.

[36] L. Ramesh, C. R. Murthy, and H. Tyagi, "Sample-measurement tradeoff in support recovery under a subgaussian prior," in *2019 IEEE International Symposium on Information Theory (ISIT)*, 2019, pp. 2709–2713.

[37] J. Zhu and D. Baron, "Performance limits with additive error metrics in noisy multimeasurement vector problems," *IEEE Trans. Signal Process.*, vol. 66, no. 20, pp. 5338–5348, 2018.

[38] Junan Zhu and D. Baron, "Performance regions in compressed sensing from noisy measurements," in *2013 47th Annual Conference on Information Sciences and Systems (CISS)*, 2013, pp. 1–6.

[39] J. Zhu, D. Baron, and F. Krzakala, "Performance limits for noisy multimeasurement vector problems," *IEEE Trans. Signal Process.*, vol. 65, no. 9, pp. 2444–2454, 2017.

[40] P. Pal and P. P. Vaidyanathan, "Parameter identifiability in sparse Bayesian learning," in *Proc. ICASSP*, May 2014, pp. 1851–1855.

[41] Z. Chen and W. Yu, "Phase transition analysis for covariance based massive random access with massive mimo," in *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, 2019, pp. 36–40.

[42] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *J. Mach. Learn. Res.*, vol. 6, pp. 1705–1749, Dec. 2005.

[43] M. Gil, F. Alajaji, and T. Linder, "Rényi divergence measures for commonly used univariate continuous distributions," *Information Sciences*, vol. 249, pp. 124 – 131, Nov. 2013.

[44] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *CoRR*, vol. abs/1011.3027, 2011. [Online]. Available: http://arxiv.org/abs/1011.3027

[45] S. Khanna and C. R. Murthy, "Corrections to "On the restricted isometry of the columnwise Khatri–Rao product"," *IEEE Trans. Signal Process.*, vol. 67, no. 9, pp. 2387–2388, May 2019.

[46] A. Bhaskara, M. Charikar, A. Moitra, and A. Vijayaraghavan, "Smoothed analysis of tensor decompositions," in *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, 2014, pp. 594–603.

[47] L. Ramesh, C. R. Murthy, and H. Tyagi, "Sample-measurement tradeoff in support recovery under a subgaussian prior," *IEEE Trans. Inf. Theory*, vol. 67, no. 12, pp. 8140–8153, 2021.

[48] ——, "Multiple support recovery using very few measurements per sample," *IEEE Trans. Signal Process.*, vol. 70, pp. 2193–2206, 2022.

[49] R. Vershynin, "On the role of sparsity in compressed sensing and random matrix theory," in *Proc. CAMSAP*, Dec. 2009, pp. 189–192.

[50] M. Rudelson and R. Vershynin, "Hanson-wright inequality and subgaussian concentration," *Electron. Commun. Probab.*, vol. 18, p. 9 pp., 2013.

[51] S. Khanna and C. R. Murthy, "Detailed proof of Theorem 5 in "On the Restricted Isometry Property of centered self Khatri-Rao products"," *Supplementary report*, 2021.

[52] A. Fengler and P. Jung, "On the restricted isometry property of centered self Khatri-Rao products," *CoRR*, vol. arXiv:1905.09245, 2019.

Saurabh Khanna (S'14–M'18) received the B.Tech. degree in Electrical Engineering from the Indian Institute of Technology Kanpur, India, in 2007. He received the Ph.D. degree in Electrical Communication Engineering from Indian Institute of Science, Bangalore, India, in 2018. From 2007 to 2016, he worked as a design engineer with Texas Instruments, Bangalore, India. From 2018 to 2020, he was a postdoctoral research fellow in the Department of Electrical and Computer Engineering at the National University of Singapore. Since Sept. 2020, he is working as an Assistant Professor in the Department of Electronics and Communication Engineering at the Indian Institute of Technology Roorkee, India.

His research interests are in the areas of compressive sensing and sparsity-assisted inference, inverse problems concerning structured datasets, and signal processing for communications.

Chandra R. Murthy (S'03–M'06–SM'11) received the B.Tech. degree in Electrical Engineering from the Indian Institute of Technology Madras, India, in 1998, the M.S. and Ph.D. degrees in Electrical and Computer Engineering from Purdue University and the University of California, San Diego, USA, in 2000 and 2006, respectively. In Sept. 2007, he joined the Department of Electrical Communication Engineering at the Indian Institute of Science, Bangalore, India, where he is currently working as a Professor.

His research interests are in the areas of energy harvesting communications, multiuser MIMO systems, and sparse signal recovery techniques applied to wireless communications. His paper won the best paper award in the Communications Track at NCC 2014 and a paper co-authored with his students won the student best paper award at the IEEE ICASSP 2018 and IEEE ISIT 2021. He has over 75 journal papers and 100 conference papers to his credit. He was an associate editor for the IEEE SIGNAL PROCESSING LETTERS during 2012-16, the IEEE TRANSACTIONS ON COMMUNICATIONS during 2017-22, and the SADHANA JOURNAL during 2016-18. He was an elected member of the IEEE SPCOM Technical Committee for the years 2014-19. He is currently serving as an area editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and as an associate editor for the IEEE TRANSACTIONS ON INFORMATION THEORY.