Bayesian Techniques for Joint Sparse Signal Recovery: Theory and Algorithms

A Thesis

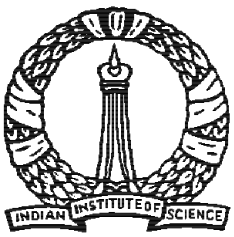
Submitted for the Degree of

Doctor of Philosophy

in the Faculty of Engineering

by

Saurabh Khanna



Electrical Communication Engineering
Indian Institute of Science, Bangalore
Bangalore – 560 012 (INDIA)

June 2018

Dedicated to

My parents

&

all contributing towards the collective human knowledge.

Acknowledgements

I am incredibly fortunate to have *Prof. Chandra R. Murthy* as my PhD supervisor. I am grateful to him for always patiently discussing my questions and musings, and encouraging me to work on my own research problems. I thank him for teaching me how to write good research papers. By setting an example for me, he has taught me the value of collaborative research and effective communication of research results. He works tirelessly to create an enabling environment for his students, which I find highly inspiring. I will be forever grateful to him for guiding and enabling my initial steps as a researcher. I am also thankful to all my teachers at IISc for their insightful lectures and talks that have helped shape my thought process.

Before becoming a full time student, my initial four years as a PhD student were shared between IISc and Texas Instruments (TI). I am extremely thankful to *Dr. Jaiganesh Balakrishnan* for his guidance and timely advices as a PhD co-advisor in TI. I am also grateful to *Hemanth* and *Manav* for doing all the groundwork which allowed me to pursue my PhD studies at IISc. I thank my fellow teammates *Sumeer, Zahir, Dheeraj and Balaji* for their unlimited patience and understanding. They made it possible for me to juggle my time between TI and IISc. At TI, I was inspired and helped by many colleagues, for which I will always be extremely grateful.

I thank my SPC labmates: *Mohit, Geethu, Parthajit, Abhay, Gana, Ranjitha, Bharath, Sanjeev, Venu, Sai, Sireesha, Lekshmi, Arun, Ashok, Akshay, Chethan, Monika, Ribhu, Ramu, Pradeep, Prabhasa, Vaibhav, Mohan, Bala, Mishfad, Sinchu, Vinuthna* and *Vinnu*, for the countless intellectual discussions and the fun times, which made my time in IISc such a wonderful experience. I am thankful for our Saturday group meetings where I got to learn about so many exciting ideas outside the confines of my usual research work. Also, thanks to my Next Gen Wireless labmates: *Jobin, Shilpa, Tirupathaiah, Rajat, Bala*

and *Sarvendranath* for sharing enjoyable discussions. Thanks to *Suma* for helping with all kinds of administrative tasks. Thanks to *Mohit* for our daily cuppa of tea and discussions, and for being such a great friend. This thesis has benefited from discussions with *Geethu, Mohit, Ranjitha, Lekshmi, Arun, Vinuthna, Akshay and Heng Qiao (from UCSD)*, for which I express my sincere thanks. Also, thanks to *Ashok* and *Monika*, who helped me in setting up the mobile USRP setup for conducting real-world experiments.

I thank my sister *Charu* for always lending a listening ear and encouraging me in the stressful times. None of this could have ever been possible without the unconditional love and support of my parents. They are my supportive angels who silently endure the uncertainties and burdens that come with being the parents of an aspiring researcher. I cannot thank them enough for selflessly sacrificing their time and aspirations so that I could pursue mine. Thanks!

Abstract

This thesis contributes new theoretical results, solution concepts, and algorithms concerning the Bayesian recovery of multiple joint sparse vectors from noisy and underdetermined linear measurements.

The thesis is written in two parts. The first part focuses on the recovery of nonzero support of multiple joint sparse vectors from their linear compressive measurements, an important canonical problem in multi-sensor signal processing. The support recovery performance of a well known Bayesian inference technique called Multiple Sparse Bayesian Learning (MSBL) is analyzed using tools from large deviation theory. New improved sufficient conditions are derived for perfect support recovery in MSBL with arbitrarily high probability. We show that the support error probability in MSBL decays exponentially fast with the number of joint sparse vectors and the rate of decay depends on the restricted eigenvalues and null space structure of the self Khatri-Rao product of the sensing matrix used to generate the measurements. New insights into MSBL's objective are developed which enhance our understanding of MSBL's ability to recover supports of size greater than the number of measurements available per joint sparse vector. These new insights are formalized into a novel covariance matching framework for sparsity pattern recovery.

Next, we characterize the restricted isometry property of a generic Khatri-Rao product matrix in terms of its restricted isometry constants (RICs). Upper bounds for the RICs of Khatri-Rao product matrices are of independent interest as they feature in the sample complexity analysis of several linear inverse problems of fundamental importance, including the above support recovery problem. We derive deterministic and probabilistic upper bounds for the RICs of Khatri-Rao product between two matrices. The newly obtained RIC bounds are then used to derive performance bounds for MSBL based support recovery.

Building upon the new insights about MSBL, a novel covariance matching based support

recovery algorithm is conceived. It uses a Rényi divergence objective which reverts to the MSBL's objective in a special case. We show that the Rényi divergence objective can be expressed as a difference of two submodular set functions, and hence it can be optimized via an iterative majorization-minimization procedure to generate the support estimate. The resulting algorithm is empirically shown to be several times faster than existing support recovery methods with comparable performance.

The second part of the thesis focuses on developing decentralized extensions of MSBL for in-network estimation of multiple joint sparse vectors from linear compressive measurements using a network of nodes. A common issue while implementing decentralized algorithms is the high cost associated with the exchange of information between the network nodes. To mitigate this problem, we examine two different approaches to reduce the amount of inter-node communication in the network. In the first decentralized extension of MSBL, the network nodes exchange information only via a small set of predesignated bridge nodes. For this bridge node based network topology, the MSBL optimization is then performed using decentralized Alternating Directions Method of Multipliers (ADMM). The convergence of decentralized ADMM in a bridge node based network topology for a generic consensus optimization is separately analyzed and a linear rate of convergence is established. Our second decentralized extension of MSBL reduces the communication complexity by adaptively censoring the information exchanged between the nodes of the network by exploiting the inherent sparse nature of the exchanged information. The performance of the proposed decentralized schemes is evaluated using both simulated as well as real-world data.

Glossary

AIC	:	Analog-to-Information-Converter
ADMM	:	Alternating Directions Methods of Multipliers
CB-DSBL	:	Consensus Based Distributed Sparse Bayesian Learning
DCS	:	Distributed Compressive Sensing
EM	:	Expectation Maximization
FB-DSBL	:	Fusion Based Distributed Sparse Bayesian Learning
FC	:	Fusion Center
i.i.d.	:	Independent and Identically Distributed
JSM	:	Joint Sparsity Model
JSSR	:	Joint Sparse Signal Recovery
KR	:	Khatri-Rao
KL	:	Kullback-Leibler
LASSO	:	Least Angle Absolute Shrinkage and Selection Operator
LHS	:	Left Hand Side
LLRT	:	Log Likelihood Ratio Test
MAP	:	Maximum a Posteriori
MIMO	:	Multiple-Input Multiple-Output
ML	:	Maximum Likelihood
MMSE	:	Minimum Mean Squared Error
MMV	:	Multiple Measurement Vector
MSBL	:	Multiple Sparse Bayesian Learning
NMSE	:	Normalized Mean Squared Error
NP-hard	:	Non-deterministic Polynomial-time hard
NSER	:	Normalized Support Error Rate
OMP	:	Orthogonal Matching Pursuit
PFA	:	Probability of False Alarm
RD-CMP	:	Rényi Divergence Based Covariance Matching Pursuit
RHS	:	Right Hand Side
RIC	:	Restricted Isometry Constant
RIP	:	Restricted Isometry Property
ROC	:	Receiver Operating Characteristics
SBL	:	Sparse Bayesian Learning
SMV	:	Single Measurement Vector
SNR	:	Signal-to-Noise Ratio
SOMP	:	Simultaneous Orthogonal Matching Pursuit
w.h.p.	:	With High Probability

Notation

Vectors and scalar random variables are denoted by boldface lowercase alphabets. Scalar variables are denoted by lowercase alphabets. Matrices are denoted by uppercase boldface alphabets and sets are denoted by uppercase script letters. The rest of the notation is listed in the table below:

Fields		
$\mathbb{R}$	:	Field of real numbers
$\mathbb{C}$	:	Field of complex numbers
$\mathbb{R}^n$	:	Set of all n-dimensional vectors
$\mathbb{R}_+^n$	:	Set of all real n-dimensional nonnegative vectors
S_+^n	:	Set of all $n \times n$ real positive semidefinite matrices
S_{++}^n	:	Set of all $n \times n$ real positive definite matrices

Sets		
$[n]$	:	$\{1, 2, \ldots, n\}$ for any positive integer n
$\lvert\mathcal{A}\rvert$	:	Cardinality of set $\mathcal{A}$
$\mathcal{A}^c$	:	Complement of set $\mathcal{A}$
$\mathcal{A} \cup \mathcal{B}$	:	Union of sets $\mathcal{A}$ and $\mathcal{B}$
$\mathcal{A} \cap \mathcal{B}$	:	Intersection of sets $\mathcal{A}$ and $\mathcal{B}$
$\mathcal{A} \backslash \mathcal{B}$	:	$\mathcal{A} \cup \mathcal{B}^c$, i.e., elements of $\mathcal{A}$ that are not in $\mathcal{B}$

For any vector $\mathbf{x} \in \mathbb{R}^n$		
$\bar{\mathbf{x}}$	:	Elementwise complex conjugate of $\mathbf{x}$
$\mathbf{x}(i)$	:	Element of $\mathbf{x}$ in i^{th} row

$\mathbf{x}_j$	:	Vector $\mathbf{x}$ at node/agent j
$\mathbf{x}_j^k$	:	Vector $\mathbf{x}$ at node/agent j in k^{th} iteration
$\text{supp}(\mathbf{x})$	:	Index set of nonzero rows in the vector $\mathbf{x}$
$\mathbf{x}_{\mathcal{S}}$	:	$\lvert\mathcal{S}\rvert \times 1$ sized vector retaining elements in $\mathbf{x}$ indexed by $\mathcal{S}$
$\lvert\lvert\mathbf{x}\rvert\rvert_0$	:	ℓ_0 norm of $\mathbf{x}$, evaluated as the number of nonzero elements in $\mathbf{x}$
$\lvert\lvert\mathbf{x}\rvert\rvert_1$	:	ℓ_1 norm of $\mathbf{x}$, evaluated as $\sum_{i=1}^{n} \lvert\mathbf{x}_i\rvert$
$\lvert\lvert\mathbf{x}\rvert\rvert_2$	:	ℓ_2 norm of $\mathbf{x}$, evaluated as $\sqrt{\sum_{i=1}^{n} \mathbf{x}_i^2}$
$\text{diag}(\mathbf{x})$	:	Matrix with $\mathbf{x}$ as its principle diagonal and rest of the entries set to zero
$\mathbf{x} \succeq a$	:	All entries of $\mathbf{x}$ are greater than or equal to scalar a

For any two real vectors x and y

$\mathbf{x} \circ \mathbf{y}$	:	Hadamard or Schur product between $\mathbf{x}$ and $\mathbf{y}$
$\mathbf{x} \otimes \mathbf{y}$	:	Kronecker product between $\mathbf{x}$ and $\mathbf{y}$
$\langle \mathbf{x}, \mathbf{y} \rangle$	:	Inner product of $\mathbf{x}$ and $\mathbf{y}$ evaluated as $\mathbf{x}^T\mathbf{y}$
$\mathbf{x} \leq \mathbf{y}$ or $\mathbf{y} \geq \mathbf{x}$	:	All entries in $\mathbf{y} - \mathbf{x}$ are nonnegative
$\mathbf{x} < \mathbf{y}$ or $\mathbf{y} > \mathbf{x}$	:	All entries in $\mathbf{y} - \mathbf{x}$ are strictly positive

Matrices

$\mathbf{A}^T$	:	Transpose of $\mathbf{A}$
$\mathbf{A}^H$	:	Complex conjugate transpose of $\mathbf{A}$
$\bar{\mathbf{A}}$	:	Elementwise complex conjugate of $\mathbf{A}$
$\mathbf{A}^{-1}$	:	Matrix inverse of $\mathbf{A}$
$\mathbf{A}^{\dagger}$	:	Generalized inverse of $\mathbf{A}$
$\mathbf{A}_{ij}$	:	Element of $\mathbf{A}$ in row i and column j
$\mathbf{A}_i$	:	i^{th} column of $\mathbf{A}$ unless specified otherwise
$\mathbf{A}_{\mathcal{S}}$	:	Submatrix comprising columns of $\mathbf{A}$ indexed by $\mathcal{S}$

$\text{tr}(\mathbf{A})$	:	Trace of matrix $\mathbf{A}$
$\|\mathbf{A}\|$ or $\det(\mathbf{A})$	:	Determinant of $\mathbf{A}$
$\text{rank}(\mathbf{A})$	:	Rank of matrix $\mathbf{A}$
$\text{spark}(\mathbf{A})$	:	Spark of matrix $\mathbf{A}$
$\text{krank}(\mathbf{A})$	:	Kruskal rank of matrix $\mathbf{A}$
$\text{supp}(\mathbf{A})$	:	Index set of nonzero rows in the matrix $\mathbf{A}$
$\text{Null}(\mathbf{A})$	:	Null space of $\mathbf{A}$
$\text{Col}(\mathbf{A})$	:	Subspace spanned by columns of $\mathbf{A}$
$\text{Row}(\mathbf{A})$	:	Subspace spanned by rows of $\mathbf{A}$
$\text{vec}(\mathbf{A})$	:	Vector obtained by columwise stacking of columns of $\mathbf{A}$
$\mathcal{R}(\mathbf{X})$	:	Index set of nonzero rows in the matrix $\mathbf{A}$
$\|\|\|\mathbf{A}\|\|\|_2$	:	Spectral norm of $\mathbf{A}$, evaluated as $\sup_{\mathbf{x}} \frac{\|\|\mathbf{Ax}\|\|_2}{\|\|\mathbf{x}\|\|_2}$
$\|\|\|\mathbf{A}\|\|\|_F$	:	Frobenius norm of $\mathbf{A}$, evaluated as the square-root of sum of squared elements in $\mathbf{A}$
$\|\|\|\mathbf{A}\|\|\|_1$	:	Maximum absolute column sum of $\mathbf{A}$
$\|\|\|\mathbf{A}\|\|\|_\infty$	:	Maximum absolute row sum of $\mathbf{A}$

For any two real matrices $\mathbf{A}$ and $\mathbf{B}$

$\mathbf{A} \circ \mathbf{B}$	:	Hadamard or Schur product between $\mathbf{A}$ and $\mathbf{B}$
$\mathbf{A} \otimes \mathbf{B}$	:	Kronecker product between $\mathbf{A}$ and $\mathbf{B}$
$\mathbf{A} \odot \mathbf{B}$	:	Columnwise Khatri-Rao product between $\mathbf{A}$ and $\mathbf{B}$
$\langle \mathbf{A}, \mathbf{B} \rangle$	:	Matrix inner product of $\mathbf{A}$ and $\mathbf{B}$ evaluated as $\text{tr}(\mathbf{A}^T\mathbf{B})$
$\mathbf{A} \leq \mathbf{B}$ or $\mathbf{B} \geq \mathbf{A}$	:	$\mathbf{B} - \mathbf{A}$ is positive semidefinite
$\mathbf{A} < \mathbf{B}$ or $\mathbf{B} > \mathbf{A}$	:	$\mathbf{B} - \mathbf{A}$ is positive definite

Probability

$\mathcal{N}(\mu, \sigma^2)$	:	Gaussian distribution with mean μ and variance σ^2
$\mathbb{E}[\cdot]$	:	Expectation operator
$\mathbb{E}[\mathbf{x}\|\mathbf{y}]$ or $\mathbb{E}_{\mathbf{y}}[\mathbf{x}]$	:	Conditional expectation of random variable $\mathbf{x}$

		conditioned on random variable $\mathbf{y}$
$\mathbb{P}(\mathcal{E})$	:	Probability of event $\mathcal{E}$
$p(\mathbf{x})$	:	Probability density function of random variable $\mathbf{x}$
$p(\mathbf{x}\|\mathbf{y})$	:	Conditional probability density function of random variable $\mathbf{x}$ given $\mathbf{y}$
$p(\mathbf{x};\tau)$	:	Probability density function of random variable $\mathbf{x}$ parameterized by deterministic variable τ

Calculus

∇f	:	First order derivative (gradient) of function f
$\nabla^2 f$	:	Second order derivative (Hessian) of function f
$\frac{\partial f}{\partial x}$	:	First order partial derivative of a multi-variate function f with respect to variable x
$\frac{\partial^2 f}{\partial x^2}$	:	Second order partial derivative of a multi-variate function f with respect to variable x

Constants

$\mathbf{I}_n$	:	$n \times n$ sized identity matrix
$\mathbf{0}_n$	:	$n \times 1$ sized all-zero vector
$\mathbf{1}_n$	:	$n \times 1$ sized all-ones vector

Miscellaneous

$\|x\|$	:	magnitude of scalar x

Contents

List of Figures

List of Tables

List of Publications

Journal Papers

1. **Saurabh Khanna** and Chandra R. Murthy, "Decentralized Joint-Sparse Signal Recovery: A Sparse Bayesian Learning Approach", *IEEE Transactions on Signal and Information Processing over Networks* , vol. 3, no. 1, pp. 29–45, Sept. 2016.

2. **Saurabh Khanna** and Chandra R. Murthy, "Communication Efficient Decentralized Sparse Bayesian Learning of Joint Sparse Signals", *IEEE Transactions on Signal and Information Processing over Networks*, vol. 3, no. 3, pp. 617–630, Nov. 2016.

3. **Saurabh Khanna** and Chandra R. Murthy, "On the Restricted Isometry of Columnwise Khatri-Rao Product", *IEEE Transactions on Signal Processing*, vol. 66, no. 5, Mar. 2018

4. **Saurabh Khanna** and Chandra R. Murthy, "On the Support Recovery of Jointly Sparse Gaussian Sources using Sparse Bayesian Learning", arXiv:1703.04930.

Conference Papers

1. **Saurabh Khanna** and Chandra R. Murthy, "Decentralized Bayesian learning of jointly sparse signals," *IEEE Global Communications Conference (GLOBECOM), 2014*, Austin, USA, Dec. 2014.

2. **Saurabh Khanna** and Chandra R. Murthy, "Rényi Divergence Based Covariance Matching Pursuit of Joint Sparse Support," *Proc. IEEE 18th Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Sapporo, Japan, Jul. 2017.

Chapter 1

General introduction

In which joint sparsity is everywhere...

1.1 Introduction

Rapid progress in sensing technology has led to the development of sensors that are capable of taking measurements at increasingly higher resolutions. The high dimensional sensor data thus generated poses several challenges concerning its storage and processing using a resource constrained hardware. Issues originating from large data dimensionality become even more apparent in the backdrop of multi-sensor signal processing, wherein data streams from multiple high-resolution sensors have to be processed simultaneously.

In the multi-sensor signal processing paradigm, signal processing and inference related tasks are peformed on data generated by multiple sensors. These sensors can be co-located, e.g., an antenna sensor array, or they can be physically separated in space as in a wireless sensor network. An important and widely used premise in multi-sensor signal processing is that *data generated by different sensors exhibits significant structure*: the data collected across the sensors could be highly correlated, or occupy a common

subspace, etc. Structure in the multi-sensor data could arise due to overlapping sensing regions, or due to the inherent spatiotemporal coupling of the sensed physical process. An interesting question thus arises: *how to efficiently acquire and store a set of high dimensional structured signals?* The answer to this question depends on the type of structure that exists among the signals. This thesis focuses on one type of structure, namely, that of joint sparsity, which is described next.

Latent Joint Sparse Structure in Multi-Sensor Signals

A recurring scenario in multi-sensor signal processing considers a network of sensors, where each sensor is tasked with learning the true signal subspace of a low dimensional physical process $P(\mathbf{x})$. The variable $\mathbf{x}$ here, for instance, can be a time or a spatial coordinate. The low-dimensional subspace associated with process P as perceived by different sensors[1] usually coincides with a common ground truth. This results in an interesting consequence: *the signals acquired by the sensors can be closely approximated by different linear combinations of the same elementary signals.* The elementary signals that model the mutual low dimensional structure of the sensor signals often come from an orthonormal or an overcomplete dictionary. Motivated by this viewpoint, the Simultaneous Sparse Approximation (SSA) multi-signal model was introduced by Tropp et al. in [5]. In the SSA model, depicted in Fig. 1.1, the signals $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_L \in \mathbb{R}^p$ from L distinct sensors are assumed to be generated according to the linear model:

$$\mathbf{s}_j = \mathbf{D}\mathbf{x}_j + \mathbf{e}_j, \ 1 \leq j \leq L. \tag{1.1}$$

[1] A linear sensing modality is assumed for all the sensors.

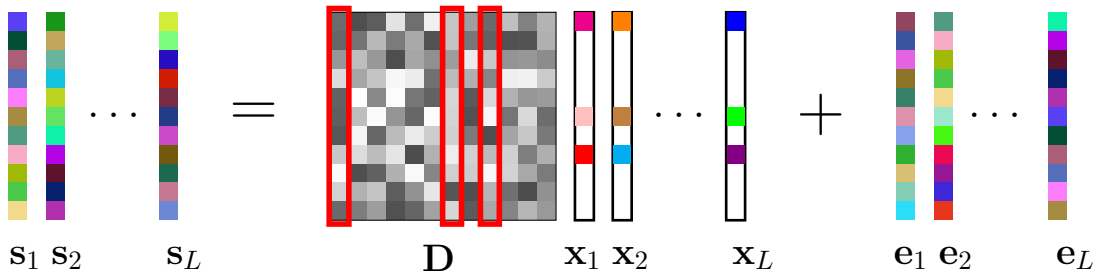


Figure 1.1: Simultaneous Sparse Approximation (SSA) model for multi-sensor signal ensemble $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_L$. Notice the common sparsity profile of the individual coefficient vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L$.

$\mathbf{D}$ here is a $p \times n$ sized dictionary matrix whose columns represent the elementary signals, a small fraction of whom can linearly approximate all of the sensor signals simultaneously. The vector $\mathbf{e}_j \in \mathbb{R}^p$ models the process noise due to any mismatch between the ground truth $\mathbf{s}_j$ and its linear approximation $\mathbf{D}\mathbf{x}_j$. Since the sensor signals $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_L$ are approximated by different linear combinations of the same $k(\ll p)$ columns of $\mathbf{D}$, as depicted in Fig. 1.1, the coefficient vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L \in \mathbb{R}^n$ contain exactly k nonzero entries belonging to the same set of rows. In other words, the coefficient vectors are jointly sparse.

Two or more sparse[2] vectors are said to be *jointly sparse*, if they share the same nonzero support, i.e., their nonzero coefficients belong to the same rows. The common index set of nonzero rows is referred to as the *support* of the jointly sparse vectors. A more realistic form of joint sparsity is the *approximate joint sparsity*, wherein the large coefficients of the vectors, which are few in number, belong to the same rows. Fig. 1.2 gives an example for each type of joint sparsity.

[2]A sparse vector has most of its coefficients equal to zero. A k-sparse vector has exactly k nonzero coefficients, and the rest of them are zero.

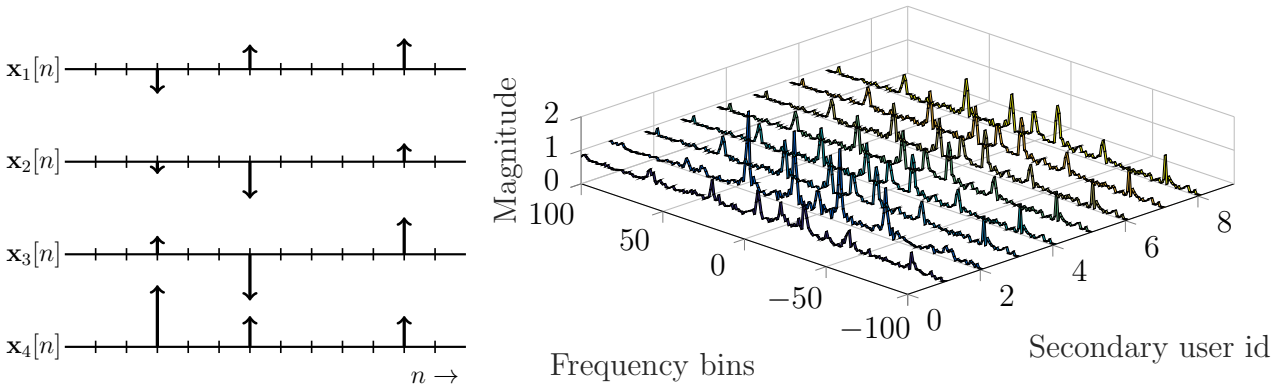


Figure 1.2: The left diagram shows an example of perfectly joint sparse signals. The right plot shows approximately joint sparse frequency domain representation of the signals received by 8 different secondary cell users in a cognitive radio application.

Compressive Acquisition of Multi-Sensor Signals

In many practical situations, the SSA model turns out to be a reasonably accurate model for structured multi-sensor signals. In that case, a simple compressive acquisition scheme for high-dimensional multi-sensor signals can be envisaged. The acquisition scheme enlists a linear encoder which inputs multi-sensor signal vectors $\{\mathbf{s}_j\}_{j=1}^{L} \in \mathbb{R}^p$, and outputs its low-dimensional linear projections $\{\mathbf{y}_j\}_{j=1}^{L} \in \mathbb{R}^m (m \ll p)$, as

$$\mathbf{y}_j = \mathbf{\Phi}\mathbf{s}_j, \quad 1 \leq j \leq L. \tag{1.2}$$

The encoder matrix $\mathbf{\Phi} \in \mathbb{R}^{m \times p}$ is fixed and does not change with the applied input. In this regard, the encoding process is non-adaptive, which is a desirable feature. As for decoding, the multi-sensor signal vectors $\{\mathbf{s}_j\}_{j=1}^{L}$ can be reconstructed from their low-dimensional projections $\{\mathbf{y}_j\}_{j=1}^{L}$ in the following two steps.

1. First, the coefficient vectors $\{\mathbf{x}_j\}_{j=1}^{L}$ in the SSA based representation of $\{\mathbf{s}_j\}_{j=1}^{L}$ are

recovered from $\{\mathbf{y}_j\}_{j=1}^{L}$. From (1.1), it follows that

$$\begin{aligned} \mathbf{y}_j &= \mathbf{\Phi}\left(\mathbf{D}\mathbf{x}_j + \mathbf{e}_j\right), \\ \text{or } \mathbf{y}_j &= \mathbf{A}\mathbf{x}_j + \mathbf{n}_j, \quad 1 \leq j \leq L. \end{aligned} \tag{1.3}$$

The matrix $\mathbf{A} = \mathbf{\Phi}\mathbf{D}$ is an $m \times n$ sized matrix which is typically known at the time of signal recovery. Vector $\mathbf{n}_j (= \mathbf{\Phi}\mathbf{e}_j)$ is the additive noise in the linear projections arising due to SSA model mismatches.[3] The first stage of decoding entails recovery of the joint sparse coefficient vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L$ from their underdetermined linear measurements $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_L$ generated according to (1.3).

2. Once the estimate of the coefficient vectors, i.e., $\{\hat{\mathbf{x}}_j\}_{j=1}^{L}$ are available, the unknown joint sparse signal vectors $\{\mathbf{s}_j\}_{j=1}^{L}$ are reconstructed as their linear approximations $\hat{\mathbf{s}}_j = \mathbf{D}\hat{\mathbf{x}}_j$.

Practical implementation of the above multi-sensor signal acquisition and reconstruction scheme entails solving the following canonical problem: *How to efficiently recover a joint sparse signal ensemble from its noisy and underdetermined linear measurements?* This comes under the purview of compressive sensing theory for multiple joint sparse vectors, which is the main focus of this thesis.

Compressive Sensing of a Joint Sparse Ensemble

According to the modern viewpoint in compression theory, a signal ensemble can be acquired or stored non-adaptively at a rate commensurate with its collective *information*

[3] In practice, the noise due to signal model mismatches is negligible in comparison to the measurement noise added while taking linear projections, and therfore it is ignored. For ease of analysis, going forward, we will assume that model mismatch noise is absent, i.e., $n_j \approx 0$ in (1.3).

content, which is typically several orders of magnitude smaller than the ambient signal dimension. This key idea underpins the widely popular theory of compressive sensing (CS) [6], which suggests that a high dimensional sparse vector can be sketched[4] to a small number of random linear projections. The sparsity of the signal allows one to reconstruct the original sparse vector from its linear projections in polynomial time.

The CS theory extends naturally to sketching of joint sparse signal ensembles, as first demonstrated in [7]. An ensemble of high dimensional joint sparse vectors in $\mathbb{R}^n$ can be embedded in a low dimensional subspace $\mathbb{R}^m (m \ll n)$, in the form of their linear projections onto an m-sized fixed set of vectors and subsequently reconstructed back when needed. Here, we could ask the following questions.

- *How to choose the fixed set of vectors used in obtaining the linear projections of the high dimensional jointly sparse vectors?*
- *What is the minimum number of linear projections that is necessary or sufficient for stable recovery of the original joint sparse signal ensemble?*
- *How to design polynomial time algorithms capable of recovering a joint sparse signal ensemble from its linear projections?*

In this thesis, we build on the compressive sensing theory for joint sparse signal ensembles, which deals with answering such fundamental questions.

[4]The term "sketching" is borrowed from the machine learning literature. A signal sketch has the same connotation as compressed, low dimensional representation of the signal. A sketch comprises multiple linear projections of the signal.

1.2 Multiple Measurement Vector Problem

Traditionally, the multiple measurement vector (MMV) problem refers to the estimation of a joint sparse signal ensemble from underdetermined linear measurements. The MMV problem and the term MMV itself were both first introduced in [7] by Cotter et al. In the MMV problem, the signal of interest is a matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L] \in \mathbb{R}^{n \times L}$ whose columns $\mathbf{x}_j$ are known to be joint 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. The goal is to recover $\mathbf{X}$ from its noisy, linear observations $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_L] \in \mathbb{R}^{m \times L}$, generated according to the linear model:

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{W}, \tag{1.4}$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}$ is a known measurement matrix and $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_L] \in \mathbb{R}^{m \times L}$ is additive measurement noise. For $L = 1$, the MMV problem reverts to its single measurement vector (SMV) form, which deals with the recovery of a single sparse vector from its linear compressive measurements. The columns of $\mathbf{Y}$ are generated by linearly projecting the individual columns of $\mathbf{X}$ onto the m rows of $\mathbf{A}$, i.e., $\mathbf{y}_j = \mathbf{A}\mathbf{x}_j + \mathbf{w}_j$ for $1 \leq j \leq L$. For $m < n$, the above linear system is underdetermined and therefore has infinitely many solutions for $\mathbf{X}$. However, if $\mathbf{A}$ satisfies certain isometry[5] properties, one can still obtain a unique row-sparse solution for $\mathbf{X}$.

In many real-world applications involving multi-sensor signal processing, the underlying core problem can be cast as the canonical MMV problem. Important application areas include magnetoencephalography [8], MIMO channel estimation [9, 10], event detection and

[5]Here isometry refers to the norm preserving behavior of a matrix.

target localization [11], radar signal processing [12], hyperspectral imaging [13], cooperative spectrum sensing [14, 15] distributed source coding [16], multi-modal recognition [17], anomaly detection [18], and one can imagine many other applications involving different types of multi-sensor processing.

1.3 Joint Sparse Support Recovery Problem

In many practical situations, the emphasis is primarily on correctly locating the nonzero rows of $\mathbf{X}$ in (1.4) rather than on recovering the entire $\mathbf{X}$. For instance, consider a cognitive radio network application, where the secondary cell users perform sub-Nyquist wideband spectrum sensing to estimate the spectrum usage of the primary cell users. This problem can be formulated as an MMV problem wherein finding the frequency bands occupied by the primary cell users is equivalent to locating the nonzero rows or the row-support of $\mathbf{X}$.

This gives rise to the *joint sparse support recovery* (JSSR) problem where the goal is to find the nonzero rows or the row support of $\mathbf{X}$, given $\mathbf{Y}$ in (1.4). Interestingly, unlike the nonzero coefficients in a k-sparse $\mathbf{X}$ which are recoverable only when $m \geq k$, the row-support of $\mathbf{X}$ can be uniquely identified even from $m < k$ measurements. In fact, in [16] it is shown that for i.i.d. Gaussian entries in $\mathbf{A}$ and $\mathbf{X}$, a non-iterative, correlation based algorithm called One Step Greedy Algorithm recovers the true support using only $m \geq 1$ measurement per signal with probability approaching one as $L \to \infty$. However, from a practitioner's viewpoint, one is more interested in non-asymptotic support recovery guarantees. In particular, we would like to examine *the necessary and sufficient conditions for exact recovery of a k-sparse row support of $\mathbf{X}$ in the noisy MMV problem, for a fixed k.*

1.4 The Evolution of MMV and JSSR Algorithms

In the following, we present a bird's eye view of the theory and algorithms developed for the canonical MMV and JSSR problems over the last decade.

1.4.1 Uniqueness under ℓ_0 norm

The earliest theoretical work focused on seeking guarantees for a unique joint sparse solution to the canonical ℓ_0 norm minimization problem:

$$L_0: \quad \min_{\mathbf{X}\in\mathbb{R}^{m\times L}} ||\mathbf{X}||_0 \quad \text{s.t. } \mathbf{AX}=\mathbf{Y}, \tag{1.5}$$

where $||\mathbf{X}||_0$ denotes the number of nonzero rows in $\mathbf{X}$. In [7,19], the authors showed that if $\text{rank}(\mathbf{Y}) \leq m$ and $k < \lceil (\text{spark}(\mathbf{A}) - 1 + \text{rank}(\mathbf{Y}))/2 \rceil$, then the L_0 problem admits a unique k-sparse solution. The sparsity bound depends on $\text{spark}(\mathbf{A})$, which is defined as the smallest integer p such that there exist p linearly dependent columns in $\mathbf{A}$. This result confirmed for the first time that the SMV bottleneck of $k < m/2$ for ℓ_0 norm based unique support recovery can be removed if multiple measurement vectors are used. Since $\text{spark}(\mathbf{A})$ can be as high as $m+1$ and $\text{rank}(\mathbf{Y})$ can also be as high as m, the above condition on k implies that supports of size $k < m$ are potentially uniquely recoverable. In the sequel, this limit on the maximum size of uniquely recoverable support will be referred to as the ℓ_0 *bound*. Since this early result, several joint sparse signal and support recovery algorithms have been proposed in the quest to meet the ℓ_0 bound.

1.4.2 Half way to the ℓ_0 bound

To circumvent the combinatorial hardness of the L_0 problem, [20] proposed using the mixed $\ell_{p,q}$ norm of $\mathbf{X}$ as a proxy for the ℓ_0 norm. Variants of the $\ell_{p,q}$ norm minimization problem with different combinations of p and q have been investigated independently in several works [7,19,21,22]. For $p \geq 1, q = 1$, it is shown in [19] that the $\ell_{p,q}$ norm minimization problem has a unique k-sparse solution, provided the measurement matrix $\mathbf{A}$ satisfies $\left|\left|\mathbf{A}_{\mathcal{S}}^{\dagger}\mathbf{a}_j\right|\right|_1 < 1$, 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 in Simultaneous Orthogonal Matching Pursuit (SOMP) [23], a greedy support reconstruction algorithm. In [24], the support recovery performance of various correlation based greedy and iterative hard-thresholding type algorithms is studied in the noiseless MMV setup. The authors obtain sufficient conditions for exact support recovery in terms of the asymmetric restricted isometry constants of the measurement matrix $\mathbf{A}$.

A limitation of the above algorithms is that their worst case performance[6] does not improve with rank($\mathbf{Y}$), and therefore they are called *rank-blind* [25]. However, their average case performances do improve with an increase in the number of MMVs [26]. These rank-blind methods fail to meet the ℓ_0-norm uniqueness criterion, i.e., $k < m$, and so far can guarantee a unique k-sparse solution only up to $k < m/2$.

1.4.3 Meeting the ℓ_0 bound

Inspired by the use of the popular MuSiC technique in sampling of multi-band signals [27] and in array signal processing, the authors in [25] propose two *rank aware* subspace pursuit

[6]For example, when all columns of the signal matrix $\mathbf{X}$ are the same.

algorithms: Rank Aware Orthogonal Matching Pursuit (RA-OMP) and Rank Aware Order Recursive Matching Pursuit (RA-ORMP), both capable of exact recovery of any k-sparse 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 (CS-MuSiC) [28] and subspace-augmented MuSiC (SA-MuSiC) [29] are capable of recovering any $k < \text{spark}(\mathbf{A}) - 1$ sized support as long as $k - \text{rank}(\mathbf{X})$ partial support can be estimated by another sparse signal recovery algorithm.

1.4.4 Beyond the ℓ_0 bound

A key insight was propounded in [30], that there often exists a latent structure in the MMV problem: the nonzero entries of $\mathbf{X}$ are uncorrelated. This signal structure is enforced by assuming that each column of $\mathbf{X}$ is i.i.d. $\mathcal{N}(0, \text{diag}(\boldsymbol{\gamma}))$, where $\boldsymbol{\gamma} \in \mathbb{R}_+^n$ is a non-negative vector of the variance parameters. Under this source model, identifying the nonzero rows of $\mathbf{X}$ is tantamount to estimating the support of $\boldsymbol{\gamma}$. In [30], the Co-LASSO algorithm was proposed for the recovery of $\boldsymbol{\gamma}$. Instead of working directly with linear observations $\mathbf{Y}$, Co-LASSO uses their covariance form, $\frac{1}{L}\mathbf{Y}\mathbf{Y}^T$, as input, and $\boldsymbol{\gamma}$ is recovered as a solution of the following non-negative ℓ_1 norm minimization problem:

$$\min_{\boldsymbol{\gamma} \in \mathbb{R}_+^n} ||\boldsymbol{\gamma}||_1 \quad \text{s.t. } (\mathbf{A} \odot \mathbf{A})\boldsymbol{\gamma} = \text{vec}\left(\frac{1}{L}\mathbf{Y}\mathbf{Y}^T - \sigma^2 \mathbf{I}_m\right). \tag{1.6}$$

In (1.6), the linear constraints are the vectorized form of the second order moment constraints, written as the covariance matching equation: $\frac{1}{L}\mathbf{Y}\mathbf{Y}^T = \mathbf{A}\text{diag}(\boldsymbol{\gamma})\mathbf{A}^T + \sigma^2\mathbf{I}_m$, where $\sigma^2\mathbf{I}_m$ denotes the noise covariance matrix. The psuedo-measurement matrix $\mathbf{A} \odot \mathbf{A}$ is the $m^2 \times n$ self Khatri-Rao product of $\mathbf{A}$ in (1.4) with itself. Since the constraints

in (1.6) comprise up to $(m^2+m)/2$ linearly independent equations[7], sparse $\boldsymbol{\gamma}$ with support size as high as $\mathcal{O}(m^2)$ is potentially recoverable. To uniquely recover the maximum level of sparsity, $k=(m^2+m)/2$, a necessary condition derived in [30, 31] dictates that the columnwise self Khatri-Rao product matrix $\mathbf{A}\odot\mathbf{A}$ has full Kruskal rank,[8] i.e., $\text{Krank}(\mathbf{A}\odot\mathbf{A})=(m^2+m)/2$.

Another highly popular MMV algorithm, MSBL [32], also imposes a shared Gaussian prior on the columns of $\mathbf{X}$, i.e.,

$$\mathbf{x}_j \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,\boldsymbol{\Gamma})\,;\quad \boldsymbol{\Gamma}=\text{diag}(\boldsymbol{\gamma}), \tag{1.7}$$

and hence it implicitly exploits the common support and the latent uncorrelatedness of the nonzero entries in $\mathbf{X}$. In MSBL, the hyperparameter vector $\boldsymbol{\gamma}$ is estimated by maximizing the type-II log-likelihood $\log p(\mathbf{Y};\boldsymbol{\gamma})$. Interestingly, similar to Co-LASSO, the support recovery performance of M-SBL is also closely tied to the properties of the psuedo measurement map, the self Khatri-Rao product, $\mathbf{A}\odot\mathbf{A}$ in (1.6). One of the key contributions of this thesis is to flesh out the explicit form of this relationship.

1.5 Bayesian Recovery of Joint Sparse Signals

A significant portion of this thesis is devoted towards developing new practical algorithms for solving the MMV and JSSR problems. We adopt the Bayesian inference approach to recover $\mathbf{X}$ and its sparse row-support from underdetermined, noisy linear measurements $\mathbf{Y}$ in (1.4). The signal of interest, $\mathbf{X}$, is treated as random and distributed according to

[7] $\mathbf{A}\odot\mathbf{A}$ has at most $(m^2+m)/2$ linearly independent rows.

[8] 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.

a columnwise joint sparsity inducing prior parameterized by hyperparameters Θ. Instead of generating a point estimate of $\mathbf{X}$, Bayesian inference outputs its posterior distribution, which is computed using the prior $p(\mathbf{X};\Theta^*)$, where Θ^* maximizes the Bayesian evidence $p(\mathbf{Y};\Theta)$ [33]. Two popular signal priors that have already been explored in the MMV literature are the Bernoulli-Gaussian prior (also known as the spike and slab prior) [34] and the Gaussian prior [32], as described in (1.7). The MMV algorithms that are based on these priors possess the *automatic relevance determination*[9] property [33], a hallmark of any Bayesian inference technique.

The choice of the Gaussian prior for $\mathbf{X}$ used in MSBL [32], and as depicted in (1.7), is particularly interesting from both theoretical and practical viewpoints. Since $\mathbf{X}$ is zero mean, the diagonal of the covariance matrix $\mathbf{\Gamma}$ directly represents the row-support of $\mathbf{X}$. Thus, finding the row-support of $\mathbf{X}$ from $\mathbf{Y}$ is tantamount to estimating the covariance matrix of $\mathbf{X}$ from its compressive measurements $\mathbf{Y}$. This covariance matrix estimation view of the support recovery problem proves to be very helpful, especially given the vast amount of literature available on the topic of covariance matrix estimation.

With a few modifications to the MSBL's Gaussian prior, one can easily model inter/intra column correlations within $\mathbf{X}$ as demonstrated in [35, 36]. The Gaussian prior is also conducive to Kalman filter based online implementations of Bayesian inference of $\mathbf{X}$ as shown in [9,37]. Perhaps the most convenient feature of the Gaussian prior is that it always results in an analytically friendly Gaussian likelihood of $\mathbf{Y}$ as long as the measurement process is linear. In light of the above, it therefore becomes necessary to further our theoretical understanding about Bayesian inference of $\mathbf{X}$ and its row-support under the

[9]Automatic Relevance Determination (ARD) here refers to the ability of an algorithm to correctly learn the number of nonzero rows in $\mathbf{X}$ directly from the observations $\mathbf{Y}$.

MSBL like Gaussian signal prior.

1.6 Distributed Estimation of Joint Sparse Signals

Recent years have witnessed an upward trend in the use of large sensor networks for performing important signal processing and inference tasks. This trend is primarily being fueled by the easy availability of low-cost sensors equipped with energy efficient communication radios and processors for local computations. Given the broad applicability of the MMV framework in the areas of signal processing, learning and inference, it is pertinent to develop efficient techniques for solving the MMV problem in a distributed fashion using a network of computing nodes.

A distributed version of the MMV problem deals with in-network estimation of multiple joint sparse vectors by a network of nodes where each network node is interested in estimating its local sparse vector from noisy, underdetermined, linear measurements. Since the local sparse vectors at the individual nodes share a common support, they can be jointly estimated from significantly fewer measurements compared to their independent reconstruction by their respective nodes.

In order to exploit the network-wide joint sparsity, the nodes can collaborate in a centralized or a decentralized fashion based on the underlying network topology. In a centralized approach, each node communicates its local observations to a central node or a fusion center which then recovers the unknown joint sparse vectors and transmits the reconstructed signals back to their respective nodes [7,32,38–40]. In contrast, decentralized implementation yields the same solution as the centralized approach, while processing the observations locally at each node and exchanging messages between one-hop neighbors in the network.

In addition to being inherently robust to node failures, decentralized implementations also tend to be more energy efficient when implemented over large networks.

The latter half of this thesis is devoted to developing new decentralized extensions of the MSBL algorithm that can be implemented over a network of computing nodes.

1.7 Thesis Outline and Contributions

This thesis has two parts. The first part discusses new theory, solution concepts and algorithms for solving the MMV problem, with the focus on the support recovery aspect of the problem. The second part of the thesis is devoted to developing new decentralized schemes for signal recovery in the MMV problem and addressing some of the common issues concerning their practical implementation.

In the following, we describe the main contributions of this thesis in a chapterwise fashion.

1.7.1 A New Perspective on Sparse Bayesian Learning

In chapter 2, we make an interesting connection between the MSBL algorithm and the Bregman matrix divergence minimization problem. The log-likelihood objective of MSBL is interpreted as a Bregman matrix divergence which reveals MSBL's relationship with the covariance matching approach, a popular technique in array signal processing. A formal framework for covariance matching based support recovery in the MMV problem is proposed and some preliminary observations are presented.

1.7.2 Restricted Isometry of Columnwise Khatri-Rao Product

In chapter 3, we study the restricted isometry property (RIP) of the Khatri-Rao product matrix form. The Khatri-Rao product (or columnwise Kronecker product) is an important

matrix type, that serves as a structured sensing matrix in many fundamental linear inverse problems. In context of the MMV problem, the restricted isometry constants (RICs) of the self Khatri-Rao product of the measurement matrix $\mathbf{A}$, denoted by $\mathbf{A} \odot \mathbf{A}$, feature in the support recovery performance analysis of MSBL and other covariance matching based support recovery techniques. Two types of upper bounds are derived for the k^{th} order RIC of a generic Khatri-Rao product matrix for different values of k. The first RIC bound is computed in terms of the individual RICs of the input matrices that participate in the Khatri-Rao product. The second RIC bound is probabilistic, and is specified in terms of the input matrix dimensions. A key result shown is that for a pair of $m \times n$ sized random matrices comprising independent and identically distributed subgaussian entries, their Khatri-Rao product satisfies k-RIP with arbitrarily high probability, provided $m \geq \mathcal{O}(k \log n)$. Our RIC bounds theoretically confirms that the Khatri-Rao product exhibits stronger restricted isometry compared to its constituent matrices for the same RIP order, thereby making a strong case for the use of covariance matching techniques for support recovery in the MMV problem. The proposed RIC bounds are of far-reaching consequence as they feature in the sample complexity analysis of several structured signal recovery and tensor decomposition problems.

1.7.3 New Support Recovery Guarantees in Sparse Bayesian Learning

In chapter 4, we investigate the support recovery performance of the MSBL algorithm in the MMV problem. We show that any global maximizer of MSBL's log-likelihood objective perfectly recovers the true common support of joint sparse Gaussian sources with arbitrarily high probability using only finitely many MMVs. In fact, the support

error probability decays exponentially fast with the number of MMVs, and the decay rate depends on the restricted eigenvalues and null space structure of the self Khatri-Rao product of the measurement matrix. Our analysis theoretically confirms that MSBL can recover k-sparse supports of joint sparse vectors from $m \geq \mathcal{O}(k \log n)$ measurements per vector. In the special case of noiseless measurements, we show that a single MMV suffices for perfect recovery of arbitrary k-sparse support in MSBL, provided any $k+1$ columns of the measurement matrix are linearly independent. Unlike existing support recovery guarantees for MSBL, our sufficient conditions are valid for the non-asymptotic settings, and do not require the orthogonality of the nonzero rows of the joint sparse signals.

1.7.4 Rényi Divergence Based Sparse Support Recovery

In chapter 5, we propose a novel joint sparse support recovery algorithm called Rényi Divergence based Covariance Matching Pursuit (RD-CMP). The proposed algorithm recovers the common support of the joint sparse signals as a "set" hyperparameter of a joint sparsity inducing Gaussian signal prior. The support hyperparameters are learned as a set-valued solution to a novel reverse information projection problem based on the α-Rényi information divergence. We show that the α-Rényi divergence objective is expressible as a difference of two submodular functions, and thus can be optimized via an iterative majorization-minimization procedure, with each iteration involving a greedy optimization step. Compared to existing covariance matching based joint sparse support recovery methods, RD-CMP is empirically shown to be orders of magnitude faster for signal dimensions in the 'big-data' regime.

1.7.5 Decentralized Joint Sparse Signal Recovey - An SBL Approach

In chapter 6, we propose a novel decentralized, iterative, sparse Bayesian learning algorithm named Consensus-Based Distributed Sparse Bayesian Learning (CB-DSBL) for in-network estimation of multiple joint sparse vectors by a network of nodes, using noisy and underdetermined linear measurements. It exploits the network-wide joint sparsity of the unknown sparse vectors to recover them from a significantly fewer number of local measurements compared to standalone sparse signal recovery schemes. In order to reduce the amount of internode communication and the associated overheads, the network nodes exchange messages with only a small set of bridge nodes. Under such a communication scheme, the convergence of the underlying bridge node-based alternating direction method of multiplier (ADMM) iterations used in our proposed algorithm is analyzed, and a linear convergence rate is shown. The convergence analysis results are used to optimize the convergence rate of the proposed decentralized MMV algorithm. Using Monte Carlo simulations as well as real-world-data-based experiments, we show that the proposed CB-DSBL algorithm has superior performance compared to existing decentralized algorithms.

1.7.6 Distributed Joint Sparse Signal Recovery Under Communication Constraints

Decentralized MMV algorithms suffer from high communication overheads due to frequent exchange of messages between network nodes. In chapter 7, we propose a decentralized extension of MSBL called Fusion-Based Distributed Sparse Bayesian Learning (FB-DSBL), in which the nodes collaborate by exchanging highly compressed messages to learn a

common joint sparsity inducing signal prior. The learnt signal prior is subsequently used by the nodes to compute maximum a posteriori probability (MAP) estimates of their local sparse vector. The size of the messages exchanged between nodes is substantially reduced by exchanging only those local signal prior parameters which are associated with the nonzero support detected via multiple composite log-likelihood ratio tests. The average message size is empirically shown to be proportional to the information rate of the unknown vectors. The proposed algorithm is interpreted as a degenerate case of a distributed consensus-based stochastic approximation algorithm for finding a fixed point of a function.

1.7.7 Conclusion and Future Work

The thesis is concluded in chapter 8, where we provide a brief summary of the main results in this thesis, and outline some new problems and directions for future investigations.

Chapter 2

Support Recovery Using Sparse Bayesian Learning - A Covariance Matching Viewpoint

In which MSBL objective is a Bregman matrix divergence & sparse supports are recovered by matching covariances...

The MSBL algorithm [32] was proposed by Wipf and Rao in 2007 as an extension of the Sparse Bayesian Learning (SBL) framework [41,42] to solve the multiple measurement vector (MMV) problem. In 2014, Balkan et al. [43] made an interesting observation about MSBL based support reconstruction in the noiseless MMV problem. They proved that the MSBL algorithm can recover supports of size larger than m, where m denotes the number of measurements per joint sparse vector in the MMV problem. Compared to the maximum size limit of m for uniquely recoverable supports in the conventional MMV algorithms, this was a significant improvement. In the present and next two chapters, we carry out an in-depth investigation seeking a better understanding of the superior support recovery performance of the MSBL algorithm.

In this chapter, we present a fresh perspective on support recovery using MSBL. At its core, MSBL is a Bayesian inference technique to recover the common support of multiple joint sparse signals from their underdetermined linear measurements. The support is recovered indirectly in the form of hyperparameters of a joint sparsity inducing Gaussian signal prior. We show that the maximum likelihood (ML) estimation of the hyperparameters in the MSBL algorithm can be interpreted as a Bregman matrix divergence minimization problem. This new interpretation of the MSBL algorithm unveils its connection with the covariance matching approach from array signal processing. Building further upon new insights about MSBL, we propose a general framework for sparse support recovery in the MMV setup. The new framework has the potential to spawn several new joint sparse support recovery algorithms.

2.1 Multiple Sparse Bayesian Learning (MSBL)

In this section, we review the Multiple Sparse Bayesian Learning (MSBL) algorithm [32], a type-II maximum likelihood (ML) procedure for estimation of joint sparse column vectors of $\mathbf{X}$ from linear compressive measurements in $\mathbf{Y}$. In MSBL, the columns of $\mathbf{X}$ are assumed to be i.i.d. $\mathcal{N}(0, \mathbf{\Gamma})$ distributed, i.e.,

$$p(\mathbf{X}; \boldsymbol{\gamma}) = \prod_{j=1}^{L} p(\mathbf{x}_j; \boldsymbol{\gamma}) = \prod_{j=1}^{L} \mathcal{N}(0, \mathbf{\Gamma}), \tag{2.1}$$

where $\mathbf{\Gamma} = \text{diag}(\boldsymbol{\gamma})$, and $\boldsymbol{\gamma} = [\boldsymbol{\gamma}(1), \boldsymbol{\gamma}(2), \ldots, \boldsymbol{\gamma}(n)]^T$ is an n length nonnegative vector of variance parameters. The parameter $\boldsymbol{\gamma}(i)$ models the common variance of the $\mathbf{x}_j(i)$ for $1 \leq j \leq L$, and the elements of $\boldsymbol{\gamma}$ are collectively called hyperparameters as they represent the parameters of the signal prior. Since the signal priors $p(\mathbf{x}_j; \boldsymbol{\gamma})$ for different j

are parameterized by a common $\boldsymbol{\gamma}$, if $\boldsymbol{\gamma}$ has a sparse support $\mathcal{S}$, then the MAP estimates of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L$ are also jointly sparse with the same support $\mathcal{S}$. The joint Gaussian prior in (2.1) promotes sparsity as it can be interpreted as a parameterized variational approximation of a sparsity inducing Student's t-distributed prior, as shown in [44]. In MSBL, the hyperparameters in $\boldsymbol{\gamma}$ are selected such that they maximize the Bayesian evidence $p(\mathbf{Y};\boldsymbol{\gamma})$. This is tantamount to finding a maximum likelihood (ML) estimate of $\boldsymbol{\gamma}$. Let $\hat{\boldsymbol{\gamma}}_{\mathrm{ML}}$ denote the ML estimate of $\boldsymbol{\gamma}$, i.e.,

$$\hat{\boldsymbol{\gamma}}_{\mathrm{ML}} = \underset{\boldsymbol{\gamma}\in\mathbb{R}_+^n}{\arg\max}\ \log p(\mathbf{Y};\boldsymbol{\gamma}). \tag{2.2}$$

Due to the linearity of the measurement model in (1.4), the Gaussian prior assumed for $\mathbf{x}_j$ induces Gaussian measurements, i.e., $p(\mathbf{y}_j;\boldsymbol{\gamma}) \sim \mathcal{N}(0, \sigma^2\mathbf{I}_m + \mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^T)$. For a fixed $\boldsymbol{\gamma}$, the measurement vectors $\mathbf{y}_j$ are mutually independent and it follows that

$$\begin{aligned} \log p(\mathbf{Y};\boldsymbol{\gamma}) &= \sum_{j=1}^{L} \log p(\mathbf{y}_j;\boldsymbol{\gamma}) \\ &\propto -L\log|\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}| - \mathrm{Tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1}\mathbf{Y}\mathbf{Y}^T\right), \end{aligned} \tag{2.3}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} = \sigma^2\mathbf{I}_m + \mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^T$. The log-likelihood $\log p(\mathbf{Y};\boldsymbol{\gamma})$ in (2.3) is a nonconvex function of $\boldsymbol{\gamma}$ and its global maximizer $\hat{\boldsymbol{\gamma}}_{\mathrm{ML}}$ cannot be obtained in closed form. However, its local maximizer can be found using fixed point iterations or the Expectation-Maximization (EM) procedure.

2.1.1 EM updates in MSBL

We now discuss the main steps of the EM algorithm for finding $\hat{\boldsymbol{\gamma}}_{\mathrm{ML}}$. Let $q_\theta(\mathbf{X})$ denote the variational approximation of true conditional density $p(\mathbf{X}|\mathbf{Y},\boldsymbol{\gamma})$ with variational parameter

set $\theta = \left(\{\tilde{\boldsymbol{\mu}}\}_{j=1}^{L}, \tilde{\boldsymbol{\Sigma}}\right)$. The variational parameters $\tilde{\boldsymbol{\mu}}_j$ and $\tilde{\boldsymbol{\Sigma}}$ denote the conditional mean and covariance of $\mathbf{x}_j$ given $\mathbf{y}_j$. Then, as shown in [45], the log likelihood in (2.3) admits the following decomposition.

$$\log p(\mathbf{Y};\boldsymbol{\gamma}) = \int q_\theta(\mathbf{X}) \log \frac{p(\mathbf{Y},\mathbf{X};\boldsymbol{\gamma})}{q_\theta(\mathbf{X})} \mathrm{d}\mathbf{X} \; + \; D\left(q_\theta(\mathbf{X}) \;||\; p(\mathbf{X}|\mathbf{Y};\boldsymbol{\gamma})\right) \tag{2.4}$$

where the term $D(q_\theta||p) = \int q_\theta(\mathbf{X}) \log \frac{q_\theta(\mathbf{X})}{p(\mathbf{X}|\mathbf{Y};\boldsymbol{\gamma})} \mathrm{d}\mathbf{X}$ is the *Kullback-Leibler* (KL) divergence between the probability densities $q_\theta(\mathbf{X})$ and $p(\mathbf{X}|\mathbf{Y};\boldsymbol{\gamma})$. Due to the non-negativity of $D(q_\theta||p)$ [46], the log likelihood is lower bounded by the first term in the RHS. In the *E-step*, we choose θ to make this variational lower bound tight by minimizing the KL divergence term.

$$\theta^{k+1} = \arg\min_{\theta} \; D(q_\theta(\mathbf{X}) \;||\; p(\mathbf{X}|\mathbf{Y},\boldsymbol{\gamma}^k)). \tag{2.5}$$

Here, k denotes the iteration index of EM algorithm. From the LMMSE theory [47], $p(\mathbf{x}_j|\mathbf{y}_j,\boldsymbol{\gamma}^k)$ is Gaussian with mean $\boldsymbol{\mu}_j^{k+1}$ and covariance $\boldsymbol{\Sigma}_j^{k+1}$ given by

$$\begin{aligned} \boldsymbol{\Sigma}^{k+1} &= \boldsymbol{\Gamma}^k - \boldsymbol{\Gamma}^k \mathbf{A}^T \left(\sigma^2 \mathbf{I}_m + \mathbf{A}\boldsymbol{\Gamma}^k \mathbf{A}^T\right)^{-1} \mathbf{A}\boldsymbol{\Gamma}^k \\ \boldsymbol{\mu}_j^{k+1} &= \sigma^{-2} \boldsymbol{\Sigma}^{k+1} \mathbf{A}^T \mathbf{y}_j. \end{aligned} \tag{2.6}$$

By choosing $\theta^{k+1} = \left(\left\{\boldsymbol{\mu}_j^{k+1}\right\}_{j=1}^{L}, \boldsymbol{\Sigma}^{k+1}\right)$ and $q_{\theta^{k+1}}(\mathbf{X}) \sim \prod_{j=1}^{L} \mathcal{N}(\mathbf{x}_j; \boldsymbol{\mu}_j^{k+1}, \boldsymbol{\Sigma}^{k+1})$, the KL divergence term in (2.5) can be driven to its minimum value of zero.

In the *M-step*, we choose $\boldsymbol{\gamma}$ which maximizes the tight variational lower bound obtained from the *E-step*:

$$\boldsymbol{\gamma}^{k+1} \;\; = \;\; \arg\max_{\boldsymbol{\gamma}} \int q_{\theta^{k+1}}(\mathbf{X}) \log \frac{p(\mathbf{Y},\mathbf{X};\boldsymbol{\gamma})}{q_{\theta^{k+1}}(\mathbf{X})} \mathrm{d}\mathbf{X}$$

$$= \arg\max_{\boldsymbol{\gamma}} \mathbb{E}_{\mathbf{X}\sim q_{\theta^{k+1}}} \left[\log p(\mathbf{Y}, \mathbf{X}; \boldsymbol{\gamma})\right]. \tag{2.7}$$

As shown in Appendix D.1, the optimization problem (2.7) can be recast as the following minimization problem.

$$\boldsymbol{\gamma}^{k+1} = \arg\min_{\boldsymbol{\gamma}\in\mathbb{R}^n_+} \sum_{j=1}^{L}\sum_{i=1}^{n} \left(\log \boldsymbol{\gamma}(i) + \frac{\boldsymbol{\Sigma}^k(i,i) + \boldsymbol{\mu}_j^k(i)^2}{\boldsymbol{\gamma}(i)}\right). \tag{2.8}$$

From the zero gradient optimality condition in (2.8), the M-step reduces to the following update rule:

$$\boldsymbol{\gamma}^{k+1}(i) = \frac{1}{L}\sum_{j=1}^{L} \left(\boldsymbol{\Sigma}^{k+1}(i,i) + \boldsymbol{\mu}_j^{k+1}(i)^2\right) \quad \text{for } 1 \le i \le n. \tag{2.9}$$

By repeatedly iterating between the E-step (2.6) and the M-step (2.9), the EM algorithm converges to either a local maximum or a saddle point of $\log p(\mathbf{Y}; \boldsymbol{\gamma})$ [48]. The MAP estimate of $\mathbf{x}_j$ is then obtained by substituting $\hat{\boldsymbol{\gamma}}_{\mathrm{ML}}$ in the expression for $\boldsymbol{\mu}_j$ in (2.6). It has been empirically observed that as the EM algorithm converges, the $\boldsymbol{\gamma}(i)$'s belonging to the inactive support tend to zero, resulting in sparse MAP estimates. In [32], it is empirically demonstrated that the EM procedure faithfully recovers the true row-support of $\mathbf{X}$ as $\mathrm{supp}(\hat{\boldsymbol{\gamma}}_{\mathrm{ML}})$, provided m and L are sufficiently large. The sufficient conditions for perfect support recovery in terms of m and L are derived in Chapter 4.

2.2 A New Interpretation of MSBL

We now present an interesting interpretation of MSBL's log-marginalized likelihood objective in (2.3) 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

$$\mathcal{D}_\varphi(\mathbf{X}, \mathbf{Y}) \triangleq \varphi(\mathbf{X}) - \varphi(\mathbf{Y}) - \langle \nabla\varphi(\mathbf{Y}), \mathbf{X} - \mathbf{Y} \rangle, \tag{2.10}$$

where $\varphi : S_{++}^n \to \mathbb{R}$ is a convex function with $\nabla\varphi(\mathbf{Y})$ as its first order derivative evaluated at point $\mathbf{Y}$. In (2.10), the matrix inner product $\langle \mathbf{X}, \mathbf{Y} \rangle$ is evaluated as $\operatorname{tr}\left(\mathbf{X}\mathbf{Y}^T\right)$. For the specific case of $\varphi(\cdot) = -\log|\cdot|$, a strongly convex function, we obtain the Bregman LogDet matrix divergence given by

$$\mathcal{D}_{\text{logdet}}(\mathbf{X}, \mathbf{Y}) = \operatorname{tr}\left(\mathbf{X}\mathbf{Y}^{-1}\right) - \log\left|\mathbf{X}\mathbf{Y}^{-1}\right| - n. \tag{2.11}$$

By termwise comparison of (2.3) and (2.11), we observe that the negative log likelihood $-\log p(\mathbf{Y}; \boldsymbol{\gamma})$ and $\mathcal{D}_{\text{logdet}}(\mathbf{R}_{\mathbf{yy}}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}})$ are the same up to a constant. In fact, it is shown in [49, 50] that every regular exponential family of probability distributions is associated with a unique Bregman divergence. For more details about the connection between the exponential family of distributions and Bregman divergences, the readers are referred to the excellent exposition in [50].

In the divergence term $\mathcal{D}_{\text{logdet}}(\mathbf{R}_{\mathbf{yy}}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}})$, the first argument $\mathbf{R}_{\mathbf{yy}} \triangleq \frac{1}{L}\mathbf{Y}\mathbf{Y}^T$ is the sample covariance matrix of the observations $\mathbf{Y}$ and the second argument $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} = \sigma^2\mathbf{I} + \mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^T$ is the parameterized covariance matrix of $\mathbf{Y}$. This connection between MSBL's log likelihood cost and the LogDet divergence reveals that by maximizing the MSBL cost, we seek a $\boldsymbol{\gamma}$ that minimizes the distance between $\mathbf{R}_{\mathbf{yy}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}$, with pointwise distances measured using the Bregman LogDet divergence. Thus, the MSBL algorithm, at its core, is essentially a *second order moment matching* or *covariance matching* procedure which selects $\boldsymbol{\gamma}$ such

that the associated covariance matrix $\mathbf{\Sigma}_\gamma$ is closest to the sample covariance matrix, in the Bregman LogDet divergence sense. In chapter 4, we theoretically show that if the second moment matching equations are too ill-conditioned, then MSBL fails to recover the true support of the joint sparse columns of $\mathbf{X}$.

The above interpretation of the MSBL cost as a Bregman matrix divergence beckons 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 MSBL cost and the Bregman (LogDet) matrix divergence to devise faster and more robust techniques for the type-II maximum likelihood procedure?

It is our strong opinion that evaluating the performance of other matrix divergences for covariance matching is a worthwhile exercise to pursue which can lead to the development of new, improved algorithms for the JSSR problem.

2.3 Covariance Matching Framework for Sparse Support Recovery

The Bregman matrix divergence minimization view of MSBL optimization can be formalized as the *covariance matching framework* for sparse support recovery in the canonical MMV problem. Some of the concepts that we present here are exploratory in nature and their in-depth investigation while merited is relegated to future work. We now describe the general framework as follows.

Just like in MSBL, the columns of $\mathbf{X}$ are assumed to be i.i.d. $\mathcal{N}(0, \mathbf{\Gamma})$ distributed, where $\mathbf{\Gamma} = \text{diag}(\boldsymbol{\gamma}), \boldsymbol{\gamma} \in \mathbb{R}^n_+$. As highlighted in [30], the diagonal covariance matrix $\mathbf{\Gamma}$ naturally captures the common support across the columns as well as the the uncorrelatedness of the nonzero coefficients within the individual columns of $\mathbf{X}$. Another convenient feature of the Gaussian prior for $\mathbf{X}$ is that it induces Gaussian measurements $\mathbf{y}_j \overset{i.i.d.}{\sim} \mathcal{N}(0, \mathbf{\Sigma}_{\boldsymbol{\gamma}})$, where $\mathbf{\Sigma}_{\boldsymbol{\gamma}} \triangleq \sigma^2 \mathbf{I}_m + \mathbf{A}\mathbf{\Gamma}\mathbf{A}^T$. In the sequel, we shall refer to $\mathbf{\Sigma}_{\boldsymbol{\gamma}}$ as the *parameterized MMV covariance matrix* as it depends on the hyperparameters $\boldsymbol{\gamma}$. Further, let $\hat{\mathbf{R}}_{\mathbf{yy}} \triangleq \frac{1}{L}\mathbf{Y}\mathbf{Y}^T$ denote the *empirical MMV covariance matrix* of the MMVs.

Consider a matrix function $d : \mathbb{S}^m_+ \times \mathbb{S}^m_+ \to \mathbb{R}_+$, which inputs two positive definite matrices and outputs a nonnegative real number that represents the degree of "nearness" or "distance" between the input matrices. Suitable candidates for d include matrix divergences, matrix norms, etc. In covariance matching framework, the row-support of $\mathbf{X}$ is recovered as $\text{supp}(\hat{\boldsymbol{\gamma}})$ where $\hat{\boldsymbol{\gamma}}$ is a solution of following constrained optimization.

$$\begin{aligned} \hat{\boldsymbol{\gamma}} \;\triangleq\; & \underset{\boldsymbol{\gamma} \in \mathbb{R}^n}{\arg\min} \; d\left(\hat{\mathbf{R}}_{\mathbf{yy}}, \underbrace{\sigma^2 \mathbf{I}_m + \mathbf{A}\mathbf{\Gamma}\mathbf{A}^T}_{\mathbf{\Sigma}_{\boldsymbol{\gamma}}}\right) \;+\; \lambda h(\boldsymbol{\gamma}) \\ & \text{subject to } \boldsymbol{\gamma} \succeq 0, \end{aligned} \tag{2.12}$$

where λ is a positive constant and $h : \mathbb{R}^n \to \mathbb{R}$ serves as a penalty function designed to regularize the solution space of $\boldsymbol{\gamma}$. In our case, h is a sparsity promoting function of $\boldsymbol{\gamma}$. Thus, in (2.12), we seek a sparse nonnegative vector $\boldsymbol{\gamma}$ such that the parameterized covariance $\mathbf{\Sigma}_{\boldsymbol{\gamma}}$ is a good approximation of the empirical covariance $\hat{\mathbf{R}}_{\mathbf{yy}}$. Finally, $\text{supp}(\hat{\boldsymbol{\gamma}})$ is declared as an estimate of the true row-support of $\mathbf{X}$.

Theorem 2.1. *If h is a concave and $n \geq \frac{m^2+m}{2}$, then $\hat{\boldsymbol{\gamma}}$ in (2.12) satisfies $||\hat{\boldsymbol{\gamma}}||_0 \leq \frac{m^2+m}{2}$.*

Proof. Suppose $\boldsymbol{\gamma}^*$ is a solution of (2.12). Then, $\boldsymbol{\gamma}^*$ is also a solution to the following

constrained optimization problem.

$$P_{d,h}: \quad \underset{\boldsymbol{\gamma}}{\text{maximize}} \; -h(\boldsymbol{\gamma}) \quad \text{such that } \mathbf{A}\boldsymbol{\Gamma}^*\mathbf{A}^T = \mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^T \text{ and } \boldsymbol{\gamma} \succeq 0. \tag{2.13}$$

To see this, first note that $\boldsymbol{\gamma}^*$ is $P_{d,h}$ feasible. Furthermore, for $\lambda > 0$, $h(\boldsymbol{\gamma}) \leq h(\boldsymbol{\gamma}^*)$ as $d(\hat{\mathbf{R}}_{\mathbf{yy}}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}) = d(\hat{\mathbf{R}}_{\mathbf{yy}}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}^*})$ for all feasible $\boldsymbol{\gamma}$ in $P_{d,h}$.

Note that (2.13) seeks maximization of a convex function $-h(\boldsymbol{\gamma})$ over a bounded convex polytope $\mathcal{A} = \left\{\boldsymbol{\gamma} \in \mathbb{R}^n : \boldsymbol{\gamma} \succeq 0, \mathbf{A}\boldsymbol{\Gamma}^*\mathbf{A}^T = \mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^T\right\}$. By [51, Chapter 7, Theorem 3], any solution of (2.13) is an extremum point of polytope $\mathcal{A}$, and thus a basic feasible solution of the vectorized linear constraints, $\text{vec}(\mathbf{A}\boldsymbol{\Gamma}^*\mathbf{A}^T) = (\mathbf{A} \odot \mathbf{A})\boldsymbol{\gamma}$. Since $\mathbf{A} \odot \mathbf{A}$ is an $m^2 \times n$ matrix, for $n \geq \frac{m^2+m}{2}$, it has at most $\frac{m^2+m}{2}$ linearly independent rows. Therefore, any basic feasible solution of (2.13) (including $\hat{\boldsymbol{\gamma}}$) must have at most $\frac{m^2+m}{2}$ nonzero elements. □

From Theorem 2.1, one can conclude that any concave penalty function h induces a sparse solution $\hat{\boldsymbol{\gamma}}$ in (2.13), assuming $m^2 < n$. The following theorem lays foward the sufficient conditions under which the support of $\hat{\boldsymbol{\gamma}}$ in (2.12) equals the true row-support of $\mathbf{X}$.

Theorem 2.2. *Suppose* $\mathbb{E}\left[\mathbf{x}_j\mathbf{x}_j^T\right] = \text{diag}(\boldsymbol{\gamma}^*)$ *and* $||\boldsymbol{\gamma}^*||_0 \leq k$. *Then, for* $L \to \infty$, *if* $||\hat{\boldsymbol{\gamma}}||_0 \leq k$, *then* $\hat{\boldsymbol{\gamma}} = \boldsymbol{\gamma}^*$, *provided Krank*$(\mathbf{A} \odot \mathbf{A}) \geq 2k$.

Proof. Let $\hat{\boldsymbol{\gamma}} \neq \boldsymbol{\gamma}^*$ be a k-sparse solution of (2.12). Then, $\hat{\boldsymbol{\gamma}}$ is also a solution of $P_{d,h}$, and therefore it must satisfy the feasibility condition: $\text{vec}(\hat{\mathbf{R}}_{\mathbf{yy}}) = (\mathbf{A} \odot \mathbf{A})\hat{\boldsymbol{\gamma}}$. Since the sample covariance $\hat{\mathbf{R}}_{\mathbf{yy}}$ is an asymptotically consistent estimate of the true covariance, it follows that as $L \to \infty$, $\hat{\mathbf{R}}_{\mathbf{yy}} = \mathbf{A}\text{diag}(\boldsymbol{\gamma}^*)\mathbf{A}^T$, or equivalently $\text{vec}(\hat{\mathbf{R}}_{\mathbf{yy}}) = (\mathbf{A} \odot \mathbf{A})\boldsymbol{\gamma}^*$. Therefore,

as $L \to \infty$, we have $(\mathbf{A} \odot \mathbf{A})\boldsymbol{\gamma}^* = (\mathbf{A} \odot \mathbf{A})\hat{\boldsymbol{\gamma}}$ or equivalently $(\mathbf{A} \odot \mathbf{A})(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) = 0$. Since both $\hat{\boldsymbol{\gamma}}$ and $\boldsymbol{\gamma}^*$ are at most k-sparse, $\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*$ is at most $2k$ sparse. This implies that there exists a $2k$ or less sparse vector in $\text{Null}(\mathbf{A} \odot \mathbf{A})$, which contradicts $\text{Krank}(\mathbf{A} \odot \mathbf{A}) \geq 2k$. Hence $\hat{\boldsymbol{\gamma}}$ must be equal to $\boldsymbol{\gamma}^*$. □

From Theorem 2.2, when the output of (2.12) is a k-sparse vector and $k \leq \lfloor \frac{\text{Krank}(\mathbf{A} \odot \mathbf{A})}{2} \rfloor$, the constrained optimization problem in (2.12) correctly recovers the true row-support of $\mathbf{X}$. In chapter 4, for the MSBL algorithm, we motivate how similar exact support recovery guarantees can be obtained for noisy measurements and finite L, albeit under a stronger condition based on a restricted null space property of $\mathbf{A} \odot \mathbf{A}$.

2.4 Examples of Covariance Matching Algorithms

The covariance matrices $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}$ and $\hat{\mathbf{R}}_{\mathbf{yy}}$ together with matrix nearness function d and penalty function h are the building blocks of the covariance matching framework. Different choices of d and h result in different support recovery algorithms.

2.4.1 The MSBL Algorithm

The MSBL optimization of log-likelihood function in (2.3) can be cast as the canonical covariance matching problem in (2.12) by choosing $d(\hat{\mathbf{R}}_{\mathbf{yy}}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}) = \text{tr}\left(\boldsymbol{\Sigma}^{-1}\hat{\mathbf{R}}_{\mathbf{yy}}\right) - m$, $h(\boldsymbol{\gamma}) = \log|\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}|$ and $\lambda = 1$. In chapter 4, we derive nonasymptotic probabilistic guarantees for perfect support recovery in MSBL.

2.4.2 Covariance Matching using Rényi Divergence

An interesting technique to recover hyperparameter vector $\boldsymbol{\gamma}$ is to solve the optimization:

$$\hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma} \in \mathbb{R}_+^n}{\arg\min}\ \mathcal{D}_\alpha\left(\tilde{p}, p_{\boldsymbol{\gamma}}\right) \tag{2.14}$$

where $\mathcal{D}_\alpha\left(\tilde{p}, p_{\boldsymbol{\gamma}}\right)$ is the α-Rényi divergence between the multivariate Gaussians $\tilde{p} \sim \mathcal{N}(0, \hat{\mathbf{R}}_{\mathbf{yy}})$ and $p_{\boldsymbol{\gamma}} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}})$. The objective function $\mathcal{D}_\alpha(\tilde{p}, p_{\boldsymbol{\gamma}})$ evaluates as

$$\begin{aligned} \mathcal{D}_\alpha(\tilde{p}, p_{\boldsymbol{\gamma}}) &= \frac{1}{2(1-\alpha)}\left(\log\left|(1-\alpha)\hat{\mathbf{R}}_{\mathbf{yy}} + \alpha\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\right| - \alpha\log|\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}| - (1-\alpha)\log\left|\hat{\mathbf{R}}_{\mathbf{yy}}\right|\right) \\ &\propto \log\left|(1-\alpha)\hat{\mathbf{R}}_{\mathbf{yy}} + \alpha\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\right| - \alpha\log|\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}| + \text{terms indep. of } \boldsymbol{\gamma}. \end{aligned} \tag{2.15}$$

We observe that by choosing $d(\hat{\mathbf{R}}_{\mathbf{yy}}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}) = \log\left|(1-\alpha)\hat{\mathbf{R}}_{\mathbf{yy}} + \alpha\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\right|$, $h(\boldsymbol{\gamma}) = -\log|\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}|$ and $\lambda = \alpha$, the Rènyi divergence minimization in (2.14) can be expressed as a canonical covariance matching problem for recovering $\boldsymbol{\gamma}$ and its sparse support. Based upon these observations, in chapter 5, we propose a novel support recovery algorithm called *Rènyi Divergence Covariance Matching Pursuit*.

2.4.3 Co-LASSO

Perhaps the most natural choices for functions d and h are $d(\hat{\mathbf{R}}_{\mathbf{yy}}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}) = |||\hat{\mathbf{R}}_{\mathbf{yy}} - \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}|||_F^2$ and $h(\boldsymbol{\gamma}) = ||\boldsymbol{\gamma}||_1$, which gives rise to the Co-LASSO problem.

$$\text{Co-LASSO: } \hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma} \in \mathbb{R}^{n]}}{\arg\min}\ |||\hat{\mathbf{R}}_{\mathbf{yy}} - \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}|||_F^2 + \lambda||\boldsymbol{\gamma}||_1. \tag{2.16}$$

Proposed in [30] by Pal and Vaidyanathan, the Co-LASSO algorithm is capable of recovering supports of size as high as $\mathcal{O}(m^2)$ from only m measurements per joint sparse

column of $\mathbf{X}$. The sample complexity of Co-LASSO for successful support recovery has been analyzed in [30].

2.5 Chapter Summary

In this chapter, we presented new insights about support recovery in the MSBL algorithm. We showed that the optimization of MSBL's log-likelihood objective can be interpreted as a Bregman matrix divergence minimization problem. This new interpretation of MSBL allows us to exploit the vast literature on Bregman divergence minimization in devising faster, more robust algorithms towards minimizing the MSBL cost function. We also proposed a general covariance matching based framework for support recovery using MMVs. Development of new support recovery algorithms using this framework along with their corresponding performance and complexity analysis could be an exciting research direction.

Chapter 3

Restricted Isometry of Columnwise Khatri-Rao Product

In which the Khatri-Rao product loves sparse vectors more...

3.1 Introduction

In the previous chapter, we discussed how the restricted isometry of the self Khatri-Rao product of the measurement matrix effects the performance of covariance matching based support recovery in the MMV problem. In general, the Khatri-Rao product of two matrices is an important matrix type which plays the role of a structured sensing matrix in several linear inverse problems of fundamental importance.

In this chapter[1], we analyze the restricted isometry property (RIP) of a generic columnwise Khatri-Rao product by seeking upper bounds for its k^{th} order restricted isometry

[1]Parts of this chapter have been published as S. Khanna and C. R. Murthy, "On the Restricted Isometry of Column-wise Khatri-Rao Product", *IEEE Transactions on Signal Processing*, vol. 66, no. 5, Mar. 2018. The discussion related to probabilistic RIC bounds for Khatri-Rao matrices is different and more up to date than in the published work.

constant (k-RIC) for different values of k. We derive two kinds of RIC bounds for a generic columnwise Khatri-Rao product matrix. The first RIC bound is a deterministic one and is computed in terms of the individual RICs of the input matrices participating in the Khatri-Rao product. The second RIC bound is probabilistic, and is specified in terms of the dimensions of the random input matrices. We show that the Khatri-Rao product of a pair of $m \times n$ sized random matrices comprising independent and identically distributed subgaussian entries satisfies k-RIP with arbitrarily high probability, provided m exceeds $\mathcal{O}(k \log n)$. We also theoretically confirm that the self Khatri-Rao product of a matrix exhibits a stronger restricted isometry property compared to the input matrix for the same RIP order.

Lately, in several machine learning problems, the necessary and sufficient conditions for successful signal recovery have been reported in terms of the RICs of a certain Khatri-Rao product matrix serving as a *pseudo* sensing matrix [52,53]. In light of this, the RIC bounds proposed in this chapter are quite timely, and pave the way towards obtaining order-wise tight sample complexity bounds for several fundamental learning problems.

3.2 Background

The Khatri-Rao product, denoted by the symbol $\odot$, is a columnwise Kronecker product, which was originally introduced by Khatri and Rao in [54]. For any two matrices $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2 \ldots, \mathbf{a}_p]$ and $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2 \ldots, \mathbf{b}_p]$ of sizes $m \times p$ and $n \times p$, respectively, the columnwise Khatri-Rao product $\mathbf{A} \odot \mathbf{B}$ is a matrix of dimension $mn \times p$ defined as

$$\mathbf{A} \odot \mathbf{B} = [\mathbf{a}_1 \otimes \mathbf{b}_1 \;\; \mathbf{a}_2 \otimes \mathbf{b}_2 \;\; \ldots \;\; \mathbf{a}_p \otimes \mathbf{b}_p], \tag{3.1}$$

where $\mathbf{a} \otimes \mathbf{b}$ denotes the Kronecker product [55] between vectors $\mathbf{a}$ and $\mathbf{b}$. That is, each column of $\mathbf{A} \odot \mathbf{B}$ is the Kronecker product between the respective columns of the two input matrices $\mathbf{A}$ and $\mathbf{B}$. In this chapter and elsewhere, we shall refer to the columnwise Khatri-Rao product as simply the Khatri-Rao product or the KR product. Since the Kronecker product $\mathbf{A} \otimes \mathbf{B}$ comprises all pairwise Kronecker product combinations of the columns of the input matrices, it can be shown that $\mathbf{A} \odot \mathbf{B} = (\mathbf{A} \otimes \mathbf{B})\mathbf{J}$, where $\mathbf{J}$ is a $p^2 \times p$ selection matrix with columns as a subset of the standard basis in $\mathbb{R}^{p^2}$ [56].

The Khatri-Rao product form is encountered in several linear inverse problems of fundamental importance. Recent examples include compressive sensing [57, 58], covariance matrix estimation [59, 60], direction of arrival estimation [52] and tensor decomposition [61]. In each of these examples, the KR product $\mathbf{A} \odot \mathbf{B}$, for certain $m \times n$ sized system matrices $\mathbf{A}$ and $\mathbf{B}$, plays the role of the sensing matrix used to generate linear measurements $\mathbf{y}$ of an unknown signal vector $\mathbf{x}$ according to

$$\mathbf{y} = (\mathbf{A} \odot \mathbf{B})\,\mathbf{x} + \mathbf{w}, \tag{3.2}$$

where $\mathbf{w}$ represents the additive measurement noise. It is now well established in the sparse signal recovery literature [6, 62, 63] that, if the signal of interest, $\mathbf{x}$, is a k-sparse[2] vector in $\mathbb{R}^n$, it can be stably recovered from its noisy underdetermined linear observations $\mathbf{y} \in \mathbb{R}^{m^2} (m^2 < n)$ in polynomial time provided that the sensing matrix (here, $\mathbf{A} \odot \mathbf{B}$) satisfies the restricted isometry property defined next.

[2] A vector is said to be k-sparse if at most k of its entries are nonzero.

Restricted Isometry Property

A matrix $\mathbf{\Phi} \in \mathbb{R}^{m \times n}$ is said to satisfy the *Restricted Isometry Property* (RIP) [64] of order k, if there exists a constant $\delta_k(\mathbf{\Phi}) \in (0, 1)$, such that for all k-sparse vectors $\mathbf{z} \in \mathbb{R}^n$,

$$(1 - \delta_k(\mathbf{\Phi}))||\mathbf{z}||_2^2 \leq ||\mathbf{\Phi z}||_2^2 \leq (1 + \delta_k(\mathbf{\Phi}))||\mathbf{z}||_2^2. \tag{3.3}$$

The smallest constant $\delta_k(\mathbf{\Phi})$ for which (3.3) holds for all k-sparse $\mathbf{z}$ is called the k^{th} order Restricted Isometry Constant or the k-RIC of $\mathbf{\Phi}$. Matrices with small k-RICs are good encoders for storing/sketching high dimensional vectors with k or fewer nonzero entries [65]. For example, $\delta_k(\mathbf{A} \odot \mathbf{B}) < 0.307$ is a sufficient condition for a unique k-sparse solution to (3.2) in the noiseless case, and its perfect recovery via the ℓ_1-norm minimization technique [66]. As pointed out earlier, in many structured signal recovery problems, the primary sensing matrix can be expressed as a columnwise Khatri-Rao product between two matrices. Thus, from a practitioner's viewpoint, it is pertinent to study the restricted isometry property of a columnwise Khatri-Rao product matrix.

Finding the exact kth order RIC of any matrix entails searching for the smallest and largest singular values among all possible k-column submatrices of the input matrix, which is, in general, an NP hard task [67]. Hence, we follow an alternative approach to analyzing the RIP of a KR product matrix by deriving tight upper bounds for its RICs.

3.2.1 Applications Involving Khatri-Rao Matrices

We briefly describe some practical examples where it is required to show the restricted isometry property of a KR product matrix.

Joint sparse support recovery from underdetermined linear measurements

Suppose $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L$ are unknown joint sparse signals in $\mathbb{R}^n$ with a common k-sized support denoted by an index set $\mathcal{S}$. A canonical problem in multi-sensor signal processing is concerned with the recovery of the common support $\mathcal{S}$ of the unknown signals from their noisy underdetermined linear measurements $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_L \in \mathbb{R}^m$ generated according to

$$\mathbf{y}_j = \mathbf{A}\mathbf{x}_j + \mathbf{w}_j, \quad 1 \leq j \leq L, \tag{3.4}$$

where $\mathbf{A} \in \mathbb{R}^{m \times n} (m < n)$ is a known measurement matrix, and $\mathbf{w}_j \in \mathbb{R}^n$ models the noise in the measurements. This problem arises in many practical applications such as MIMO channel estimation, cooperative wideband spectrum sensing in cognitive radio networks, target localization, and direction of arrival estimation. In [30], the support set $\mathcal{S}$ is recovered as the support of $\hat{\boldsymbol{\gamma}}$, the solution to the Co-LASSO problem:

$$\text{Co-LASSO: } \min_{\boldsymbol{\gamma} \succeq 0} \left|\left| \text{vec}(\hat{\mathbf{R}}_{\mathbf{yy}}) - (\mathbf{A} \odot \mathbf{A})\boldsymbol{\gamma} \right|\right|_2^2 + \lambda \left|\left|\boldsymbol{\gamma}\right|\right|_1, \tag{3.5}$$

where $\hat{\mathbf{R}}_{\mathbf{yy}} \triangleq \frac{1}{L} \sum_{j=1}^{L} \mathbf{y}_j \mathbf{y}_j^T$. From compressive sensing theory [62], the RIP of $\mathbf{A} \odot \mathbf{A}$ (also called the self Khatri-Rao product of $\mathbf{A}$) determines the stability of the sparse solution in the Co-LASSO problem. In MSBL [32], a different support recovery algorithm, $\mathbf{A} \odot \mathbf{A}$ satisfying $2k$-RIP guarantees unique k-sparse support recovery as $L \to \infty$ [53].

Sparse sampling of stationary graph signals

Let $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^T \in \mathbb{C}^n$ be a stochastic, zero mean and second-order stationary graph signal defined on n vertices of a graph $\mathcal{G}$. This implies that the graph signal $\mathbf{x}$ can be modeled as $\mathbf{x} = \mathbf{H}\mathbf{n}$, where $\mathbf{H}$ is any valid graph filter [68], and $n \sim \mathcal{N}(0, \mathbf{I}_n)$. The

covariance matrix $\mathbf{R_{xx}} = \mathbb{E}[\mathbf{x}\mathbf{x}^H]$ can then be expressed as

$$\begin{aligned} \mathbf{R_{xx}} &= \mathbf{H}\mathbb{E}[\mathbf{n}\mathbf{n}^H]\mathbf{H}^H = \mathbf{H}\mathbf{H}^H \\ &= \mathbf{U}\text{diag}(\mathbf{p})\mathbf{U}^H, \end{aligned} \tag{3.6}$$

where the nonnegative vector $\mathbf{p}$ refers to the graph power spectral density of the stationary graph signal $\mathbf{x}$ and the columns of $\mathbf{U}$ form Fourier like orthonormal basis for the graph signal.

As motivated in [68], in many applicaions, we are interested in reconstructing the sparse graph power spectral density $\mathbf{p}$ by observing a small subset of the graph vertices. Let $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_L$ denote the L independent obervations of the subsampled graph signal $\mathbf{x}$, i.e.,

$$\mathbf{y}_j = \mathbf{\Phi}\mathbf{x}_j, \quad 1 \leq j \leq L, \tag{3.7}$$

where $\mathbf{\Phi} \in \{0,1\}^{m \times n}$ is referred to as a binary subsampling matrix with $m(\ll n)$ rows, each row containing exactly one nonzero unity element. To recover the graph power spectral density $\mathbf{p}$ from the subsampled observations, we note that

$$\begin{aligned} \frac{1}{L}\sum_{j=1}^{L}\mathbf{y}_j\mathbf{y}_j^H &\approx \mathbf{\Phi}\mathbf{R_{xx}}\mathbf{\Phi}^T = \mathbf{\Phi}\mathbf{U}\text{diag}(\mathbf{p})\mathbf{U}^H\mathbf{\Phi}^T \\ \text{or, vec}\left(\frac{1}{L}\sum_{j=1}^{L}\mathbf{y}_j\mathbf{y}_j^H\right) &\approx (\mathbf{\Phi}\mathbf{U}^* \odot \mathbf{\Phi}\mathbf{U})\mathbf{p}, \end{aligned} \tag{3.8}$$

the superscript $*$ denoting conjugation without transpose. Here again, the recovery of a unique r-sparse solution for $\mathbf{p}$ in (3.8) can be guaranteed if the Khatri-Rao product $\mathbf{\Phi}\mathbf{U}^* \odot \mathbf{\Phi}\mathbf{U}$ satisfies the RIP of order $2r$.

PARAFAC model for low-rank three-way arrays

Consider an $I \times J \times K$ tensor $\underline{\mathbf{X}}$ of rank r. We can express $\underline{\mathbf{X}}$ as the sum of r rank-one three way arrays as $\underline{\mathbf{X}} = \sum_{i=1}^{r} \mathbf{a}_i \circ \mathbf{b}_i \circ \mathbf{c}_i$, where $\mathbf{a}_i, \mathbf{b}_i, \mathbf{c}_i$ are loading vectors of dimension I, J, K, respectively, and $\circ$ denotes the vector outer product. The tensor $\underline{\mathbf{X}}$ itself can be arranged into a matrix as $\mathbf{X} = [\text{vec}(\mathbf{X}_1), \text{vec}(\mathbf{X}_2), \ldots, \text{vec}(\mathbf{X}_K)]$. In the parallel factor analysis (PARAFAC) model [69], the matrix $\mathbf{X}$ can be approximated as

$$\mathbf{X} \approx (\mathbf{A} \odot \mathbf{B})\mathbf{C}^T, \tag{3.9}$$

where $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$ are the loading matrices with columns as the loading vectors $\mathbf{a}_i$, $\mathbf{b}_i$ and $\mathbf{c}_i$, respectively. In many problems such as direction of arrival estimation using a 2D-antenna array, the loading matrix $\mathbf{C}$ turns out to be row-sparse matrix [70]. In such cases, the uniqueness of the PARAFAC model shown in (3.9) depends on the restricted isometry property of the Khatri-Rao product $\mathbf{A} \odot \mathbf{B}$.

3.2.2 Related work

Perhaps the most direct way to analyze the RICs of the KR product matrix is to use the eigenvalue interlacing theorem [71], which relates the singular values of any k-column submatrix of the KR product to the singular values of the Kronecker product. This is possible because any k columns of the KR product put together can be interpreted as a submatrix of the Kronecker product. However, barring the maximum and minimum singular values of the Kronecker product, there is no explicit characterization of its non-extremal singular values available yet that can be used to obtain tight bounds on the k-RIC of the KR product. Bounding the RICs of KR product using the extreme singular

values of the Kronecker product turns out to be too loose to be useful. In this context, it is noteworthy to mention that an upper bound for the k-RIC of the Kronecker product matrix has been derived in terms of the k-RICs of the input matrices in [57, 72]. However, the k-RIC of the KR product matrix has not been characterized yet.

Recently, [73, 74] gave probabilistic lower bounds for the minimum singular value of the columnwise KR product between two or more matrices. These bounds are limited to randomly constructed input matrices, and are polynomial in the matrix size. In [75], it is shown that for any two matrices $\mathbf{A}$ and $\mathbf{B}$, the Kruskal-rank[3] of $\mathbf{A} \odot \mathbf{B}$ has a lower bound in terms of K-rank($\mathbf{A}$) and K-rank($\mathbf{B}$). In fact, Krank($\mathbf{A} \odot \mathbf{B}$) is at least as high as $\max\left(\text{K-rank}(\mathbf{A}), \text{K-rank}(\mathbf{B})\right)$, thereby suggesting that $\mathbf{A} \odot \mathbf{B}$ satisfies a stronger restricted isometry property than both $\mathbf{A}$ and $\mathbf{B}$. The RIC bounds proposed in this chapter ratify this fact.

A closely related yet weaker notion of restricted isometry constant is the τ-robust K-rank, denoted by K-rank$_\tau$. For a given matrix $\mathbf{\Phi}$, the K-rank$_\tau(\mathbf{\Phi})$ is defined as the largest k for which every $n \times k$ submatrix of $\mathbf{\Phi}$ has its smallest singular value larger than $1/\tau$. In [73], it is shown that the τ-robust K-rank is super-additive, implying that the K-rank$_\tau$ of the Khatri-Rao product is strictly larger than individual K-rank$_\tau$ of the input matrices.

The ℓ_1-RIP of the Kronecker product $\mathbf{A} \otimes \mathbf{B}$ was analyzed in [60] with respect to vectorized d-distributed sparse matrices, when the input matrices $\mathbf{A}$ and $\mathbf{B}$ are the adjacency matrices of two independent uniformly random δ-left regular bipartite graphs. In the sequel, we instead analyze the RIP of the KR product $\mathbf{A} \odot \mathbf{B}$, which is equivalent to the RIP of Kronecker product $\mathbf{A} \otimes \mathbf{B}$ with respect to vectorized sparse diagonal matrices.

[3]The Kruskal rank of any matrix $\mathbf{A}$ is the largest integer r such that any r columns of $\mathbf{A}$ are linearly independent.

3.2.3 Contributions

In this chapter, we derive the following bounds on the k-RIC of the columnwise KR product of two $m \times n$ sized matrices $\mathbf{A}$ and $\mathbf{B}$.

1) A deterministic upper bound for the k-RIC of $\mathbf{A} \odot \mathbf{B}$ in terms of the k-RICs of the input matrices $\mathbf{A}$ and $\mathbf{B}$. The bound is valid for $k \leq m$, and for complex valued input matrices with unit ℓ_2-norm columns.

2a) A probabilistic upper bound for the k-RIC of $\mathbf{A} \odot \mathbf{B}$ in terms of k and the input matrix dimensions (m, n), for $\mathbf{A}, \mathbf{B}$ as real valued random matrices with i.i.d subgaussian elements. The probabilistic bound is polynomially tight with respect to the input matrix dimension n.

2b) A probabilistic upper bound for the k-RIC of the self KR product $\mathbf{A} \odot \mathbf{A}$ in terms of m, n, and k, for $\mathbf{A}$ as a real valued random matrix with i.i.d. subgaussian elements. The derivation of the RIC bound for the self KR product is more intricate as it involves showing sharp concentration inequalities for functions of *dependent* random variables.

A key idea used in our RIC analysis is the fact (stated formally as Proposition 3.1) that for any two matrices $\mathbf{A}$ and $\mathbf{B}$, the Gram matrix of their KR product $(\mathbf{A} \odot \mathbf{B})^H(\mathbf{A} \odot \mathbf{B})$ can be interpreted as the Hadamard product (element wise multiplication) between $\mathbf{A}^H\mathbf{A}$ and $\mathbf{B}^H\mathbf{B}$. The Hadamard product form turns out to be more analytically tractable than columnwise Kronecker product form of the KR matrix.

3.3 Deterministic k-RIC Bound

In this section, we present a deterministic upper bound on the k-RIC of a generic columnwise KR product $\mathbf{A} \odot \mathbf{B}$, for any two same-sized and complex valued matrices $\mathbf{A}$ and $\mathbf{B}$ with normalized columns. The bound is computed in terms of the k-RICs of $\mathbf{A}$ and $\mathbf{B}$.

Theorem 3.1. *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be* $m \times n$ *sized complex-valued matrices with unit* ℓ_2*-norm columns and satisfying the* k^{th} *order restricted isometry property with constants* $\delta_k^{\mathbf{A}}$ *and* $\delta_k^{\mathbf{B}}$*, respectively. Then, their columnwise Khatri-Rao product* $\mathbf{A} \odot \mathbf{B}$ *satisfies the restricted isometry property with k-RIC at most* δ^2*, where* $\delta \triangleq max\left(\delta_k^{\mathbf{A}}, \delta_k^{\mathbf{B}}\right)$*, i.e.,*

$$(1 - \delta^2)||\mathbf{z}||_2^2 \leq ||(\mathbf{A} \odot \mathbf{B})\mathbf{z}||_2^2 \leq (1 + \delta^2)||\mathbf{z}||_2^2 \tag{3.10}$$

holds for all k-sparse vectors $\mathbf{z} \in \mathbb{C}^n$*.*

Proof. The proof is given in Section 3.4. □

Remark 1: The RIC bound for $\mathbf{A} \odot \mathbf{B}$ in Theorem 3.1 is relevant only when $\delta_k(\mathbf{A})$ and $\delta_k(\mathbf{B})$ lie in $(0, 1)$, which is true only for $k \leq m$. In other words, the above k-RIC characterization for $\mathbf{A} \odot \mathbf{B}$ requires the input matrices $\mathbf{A}$ and $\mathbf{B}$ to be k-RIP compliant.

Remark 2: Since the input matrices $\mathbf{A}$ and $\mathbf{B}$ satisfy k-RIP with $\delta_k(\mathbf{A}), \delta_k(\mathbf{B}) \in (0, 1)$, it follows from Theorem 3.1 that $\delta_k(\mathbf{A} \odot \mathbf{B})$ is strictly smaller than $\max\left(\delta_k(\mathbf{A}), \delta_k(\mathbf{B})\right)$. If $\mathbf{B} = \mathbf{A}$, the special case of self Khatri-Rao product $\mathbf{A} \odot \mathbf{A}$ arises, for which

$$0 \ < \ \delta_k(\mathbf{A} \odot \mathbf{A}) < \delta_k^2(\mathbf{A}). \tag{3.11}$$

Above implies that the self Khatri-Rao product $\mathbf{A} \odot \mathbf{A}$ is a better restricted isometry compared to $\mathbf{A}$ itself. This observation is in alignment with the expanding Kruskal rank

and shrinking mutual coherence of the self Khatri-Rao product reported in [30]. In fact, for $k = 2$, the 2-RIC bound (3.11) exactly matches the mutual coherence bound shown in [30].

The k-RIC bound in Theorem 3.1 is useful only when $k \leq m$. For $k \in (m, m^2]$, using ([76], Theorem 1), one can show that $\delta_k(\mathbf{A} \odot \mathbf{B}) \leq \left(\sqrt{k} + 1\right) \delta_{\sqrt{k}}$, where $\delta_{\sqrt{k}} = \max\left(\delta_{\sqrt{k}}(\mathbf{A}), \delta_{\sqrt{k}}(\mathbf{B})\right)$. This bound, however, loses its tightness and quickly becomes unattractive for larger values of k. Finding a tighter k-RIC upper bound for the $k > m$ case remains an open problem.

To gauge the tightness of the proposed k-RIC bound for $\mathbf{A} \odot \mathbf{B}$, we present its simulation-based quantification for the case when the input matrices $\mathbf{A}$ and $\mathbf{B}$ are random Gaussian matrices with i.i.d. $\mathcal{N}(0, 1/m)$ entries. Fig. 3.1 plots $\delta_k(\mathbf{A})$, $\delta_k(\mathbf{B})$, $\delta_k(\mathbf{A} \odot \mathbf{B})$ and the upper bound $\overline{\delta_k(\mathbf{A} \odot \mathbf{B})} = \left(\max\left(\delta_k(\mathbf{A}), \delta_k(\mathbf{B})\right)\right)^2$ for a range of input matrix dimension m. The aspect ratio m/n of the input matrices is fixed to 0.5.[4] For computational tractability, we restrict our analysis to the cases $k = 2$ and 3. The RICs: $\delta_k(\mathbf{A})$, $\delta_k(\mathbf{B})$ and $\delta_k(\mathbf{A} \odot \mathbf{B})$ are computed by exhaustively searching for the worst conditioned submatrix comprising k columns of $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{A} \odot \mathbf{B}$, respectively. From Fig. 3.1, we observe that the proposed k-RIC upper bound becomes tighter as the input matrices grow in size. Interestingly, the experiments suggest that the KR product $\mathbf{A} \odot \mathbf{B}$ satisfies k-RIP in spite of the input matrices $\mathbf{A}$ and $\mathbf{B}$ failing to do so. A theoretical confirmation of this empirical observation remains an open problem.

[4]While the $m \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$ may represent highly underdetermined linear systems (when $m \ll n$), their $m^2 \times n$ sized Khatri-Rao product $\mathbf{A} \odot \mathbf{B}$ can become an overdetermined system. In fact, many covariance matching based sparse support recovery algorithms [30, 32, 77] exploit this fact to offer significantly better support reconstruction performance.

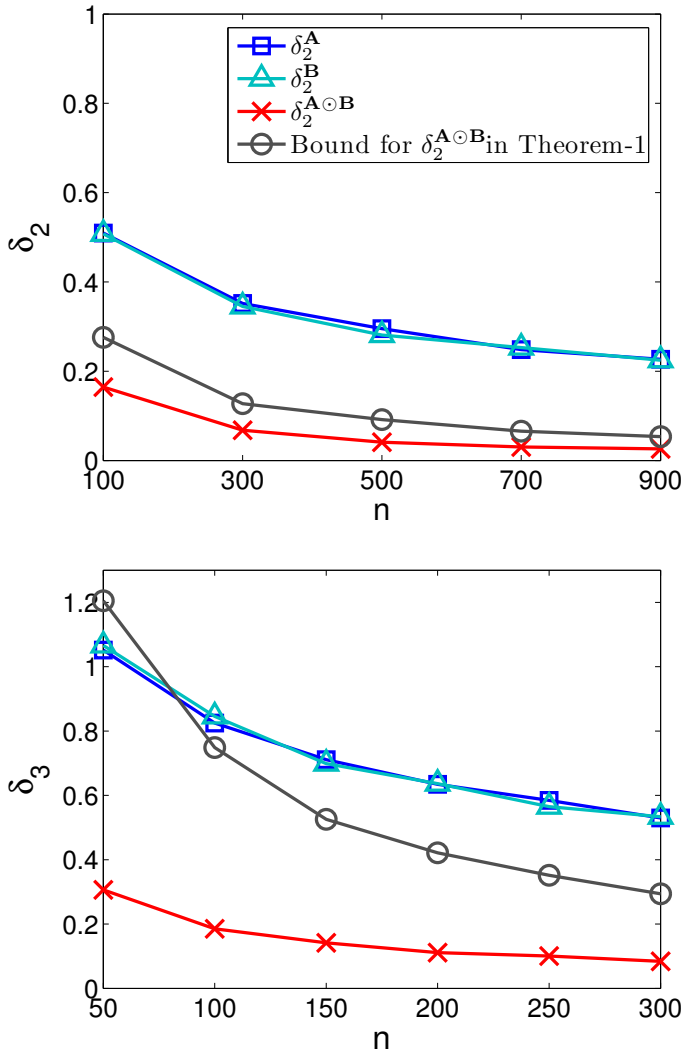


Figure 3.1: Variation of k-RICs of $\mathbf{A}$, $\mathbf{B}$, $\mathbf{A} \odot \mathbf{B}$ and the proposed upper bound with increasing input matrix dimensions. The top and the bottom plots are for $k = 2$ and 3, respectively. Each data point is averaged over 10 trials.

3.4 Proof of the Deterministic k-RIC Bound (Theorem 3.1)

We recommend a quick perusal of the preliminary mathematical concepts and results in Appendix A before proceeding with the derivation of the deterministic RIC bound in Theorem 3.1. The key idea used in bounding the k-RIC of the columnwise KR product $\mathbf{A} \odot \mathbf{B}$ is the observation that the Gram matrix of $\mathbf{A} \odot \mathbf{B}$ can be interpreted as a Hadamard product between the two correlation matrices $\mathbf{A}^H\mathbf{A}$ and $\mathbf{B}^H\mathbf{B}$, as mentioned in the following proposition.

Proposition 3.1 (Rao and Rao [78]). *For* $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{m\times n}$,

$$(\mathbf{A} \odot \mathbf{B})^H(\mathbf{A} \odot \mathbf{B}) = (\mathbf{A}^H\mathbf{A}) \circ (\mathbf{B}^H\mathbf{B}) \tag{3.12}$$

Proof. See [78, Proposition 6.4.2]. □

Then, by using the forward and reverse Kantorovich matrix inequalities presented in Appendix A, we obtain the proposed upper bound for k-RIC of $\mathbf{A} \odot \mathbf{B}$ in Theorem 3.1 as explained in the following arguments.

Without loss of generality, let $S \subset [n]$ be an arbitrary index set representing the nonzero support of $\mathbf{z}$ in (3.10), with $|S| \leq k$. Let $\mathbf{A}_S$ denote the $m\times|S|$ submatrix of $\mathbf{A}$, constituting $|S|$ columns of $\mathbf{A}$ indexed by the set S. Let $\mathbf{B}_S$ be constructed similarly. Since $\delta_k(\mathbf{A})$, $\delta_k(\mathbf{B}) < 1$, both $\mathbf{A}_S$, $\mathbf{B}_S$ have full column rank, and consequently the associated Gram matrices $\mathbf{A}_S^H\mathbf{A}_S$, $\mathbf{B}_S^H\mathbf{B}_S$ are Hermitian positive definite. Further, since $\mathbf{A}$ and $\mathbf{B}$ have unit norm columns, both $\mathbf{A}_S^H\mathbf{A}_S$ and $\mathbf{B}_S^H\mathbf{B}_S$ are correlation matrices with unit diagonal entries.

Using Proposition 3.1, we can write

$$(\mathbf{A}_S \odot \mathbf{B}_S)^H (\mathbf{A}_S \odot \mathbf{B}_S) = \mathbf{A}_S^H \mathbf{A}_S \circ \mathbf{B}_S^H \mathbf{B}_S. \tag{3.13}$$

Next, for $k \leq m$, by applying Lemma A.1 to the positive definite matrices $(\mathbf{A}_S^H \mathbf{A}_S)^{1/2}$ and $(\mathbf{B}_S^H \mathbf{B}_S)^{1/2}$, we get,

$$\begin{aligned} \mathbf{A}_S^H \mathbf{A}_S \circ \mathbf{B}_S^H \mathbf{B}_S &\leq \left((\mathbf{A}_S^H \mathbf{A}_S)^{\frac{1}{2}} \circ (\mathbf{B}_S^H \mathbf{B}_S)^{\frac{1}{2}} \right)^2 + \frac{1}{4}(M-m)^2 \mathbf{I}_k \\ &\leq \mathbf{I}_k + \frac{1}{4}(M-m)^2 \mathbf{I}_k, \end{aligned} \tag{3.14}$$

where the second inequality is a consequence of the unity bound on the spectral radius of the Hadamard product between correlation matrices, shown in Proposition A.5. In (3.14), M and m are upper and lower bounds for the maximum and minimum eigenvalues of $(\mathbf{A}_S^H \mathbf{A}_S)^{1/2} \otimes (\mathbf{B}_S^H \mathbf{B}_S)^{1/2}$, respectively. From the restricted isometry of $\mathbf{A}$ and $\mathbf{B}$, and by application of Proposition A.1, the minimum and maximum eigenvalues of $(\mathbf{A}_S^H \mathbf{A}_S)^{1/2} \otimes (\mathbf{B}_S^H \mathbf{B}_S)^{1/2}$ are lower and upper bounded by $\sqrt{(1-\delta_k^{\mathbf{A}})(1-\delta_k^{\mathbf{B}})}$ and $\sqrt{(1+\delta_k^{\mathbf{A}})(1+\delta_k^{\mathbf{B}})}$, respectively. By introducing $\delta \triangleq \max(\delta_k^{\mathbf{A}}, \delta_k^{\mathbf{B}})$, it is easy to check that the eigenvalues of $(\mathbf{A}_S^H \mathbf{A}_S)^{1/2} \otimes (\mathbf{B}_S^H \mathbf{B}_S)^{1/2}$ also lie inside the interval $[1-\delta, 1+\delta]$. Plugging $m = 1-\delta$ and $M = 1+\delta$ in (3.14), and by using (3.13), we get

$$(\mathbf{A}_S \odot \mathbf{B}_S)^H (\mathbf{A}_S \odot \mathbf{B}_S) \leq \left(1+\delta^2\right) \mathbf{I}_k. \tag{3.15}$$

Similarly, by applying Lemma A.2 to $\mathbf{A}_S^H \mathbf{A}_S$ and $\mathbf{B}_S^H \mathbf{B}_S$ with $m = 1-\delta$ and $M = 1+\delta$, we obtain

$$(\mathbf{A}_S^H \mathbf{A}_S)^{1/2} \circ (\mathbf{B}_S^H \mathbf{B}_S)^{1/2} \geq \left(\sqrt{1-\delta^2}\right) \mathbf{I}_k.$$

From Proposition A.3, we have $\mathbf{A}_S^H\mathbf{A}_S \circ \mathbf{B}_S^H\mathbf{B}_S \geq \left((\mathbf{A}_S^H\mathbf{A}_S)^{1/2} \circ (\mathbf{B}_S^H\mathbf{B}_S)^{1/2}\right)^2$. Therefore, we can write

$$\mathbf{A}_S^H\mathbf{A}_S \circ \mathbf{B}_S^H\mathbf{B}_S \geq \left(1-\delta^2\right)\mathbf{I}_k.$$

Further, using (3.13), we get

$$(\mathbf{A}_S \odot \mathbf{B}_S)^H(\mathbf{A}_S \odot \mathbf{B}_S) \geq \left(1-\delta^2\right)\mathbf{I}_k. \tag{3.16}$$

Finally, Theorem 3.1's statement follows from (3.15) and (3.16).

3.5 Probabilistic k-RIC Bound

The deterministic RIC bound for the columnwise KR product discussed in the Section 3.3 is useful only when the input matrices have unit norm columns. While using column normalized sensing matrices is standard practice in compressive sensing, we are often interested in the restricted isometry properties of Khatri-Rao product of randomly constructed matrices which adhere to the column normalization constraint only in the average sense. This concern is addressed to an extent by our second RIC bound for the columnwise KR product. This second RIC bound is probabilistic and is applicable to the KR product of random input matrices with i.i.d. subgaussian entries. Below, we define a subgaussian random variable and state some of its properties.

Definition 3.1. *(Subgaussian Random Variable): A zero mean random variable* $\mathbf{x}$ *is called subgaussian, if its tail probability is dominated by that of a Gaussian random variable. In other words, there exist constants* $C, K > 0$ *such that* $\mathbb{P}(|\mathbf{x}| \geq t) \leq Ce^{-t^2/K^2}$ *for* $t > 0$.

Gaussian, Bernoulli and all bounded random variables are subgaussian random variables.

For a subgaussian random variable, its p^{th} order moment grows no faster than $\mathcal{O}(p^{p/2})$ [79]. In other words, there exists $K_1 > 0$ such that

$$(\mathbb{E}\,|\mathbf{x}|^p)^{\frac{1}{p}} \leq K_1\sqrt{p}, \quad p \geq 1. \tag{3.17}$$

The minimum such K_1 is called the *subgaussian or* ψ_2 *norm* of the random variable $\mathbf{x}$, i.e.,

$$||\mathbf{x}||_{\psi_2} = \sup_{p\geq 1} p^{-1/2}\,(\mathbb{E}\,|\mathbf{x}|^p)^{\frac{1}{p}}\,. \tag{3.18}$$

Given a pair of random input matrices with i.i.d. subgaussian entries, Theorem 3.2 presents a new upper bound on the k-RIC of their columnwise KR product.

Theorem 3.2. *Suppose* $\mathbf{A}$ *and* $\mathbf{B}$ *are* $m \times n$ *random matrices with real i.i.d. subgaussian entries, such that* $\mathbb{E}\mathbf{A}_{ij} = 0$, $\mathbb{E}\mathbf{A}_{ij}^2 = 1$, *and* $||\mathbf{A}_{ij}||_{\psi_2} \leq \kappa$, *and similarly for* $\mathbf{B}$. *Then, the* k^{th} *order restricted isometry constant of* $\frac{\mathbf{A}}{\sqrt{m}} \odot \frac{\mathbf{B}}{\sqrt{m}}$, *denoted by* δ_k, *satisfies* $\delta_k \leq \delta$ *with probability at least* $1 - 10n^{-2(\gamma-1)}$ *for any* $\gamma > 1$, *provided that*

$$m \geq 4c\gamma\kappa^2\left(\frac{k\log n}{\delta}\right),$$

where c *a universal positive constant.*

Proof. The proof is provided in Appendix E.1. □

The normalization constant $\sqrt{m}$ used while computing the KR product $\frac{\mathbf{A}}{\sqrt{m}} \odot \frac{\mathbf{B}}{\sqrt{m}}$ ensures that the columns of the input matrices $\frac{\mathbf{A}}{\sqrt{m}}$, $\frac{\mathbf{B}}{\sqrt{m}}$ have unit average energy, i.e. $\mathbb{E}\,||\mathbf{a}_i/\sqrt{m}||_2^2 = \mathbb{E}\,||\mathbf{b}_i/\sqrt{m}||_2^2 = 1$ for $1 \leq i \leq n$. Column normalization is a key assumption towards correct modelling of the isotropic, norm-preserving nature of the effective sensing matrix $\frac{1}{m}(\mathbf{A} \odot \mathbf{B})$, an attribute found in most sensing matrices employed in

practice.

Theorem 3.2 implies that

$$\delta_k\left(\frac{\mathbf{A}}{\sqrt{m}} \odot \frac{\mathbf{B}}{\sqrt{m}}\right) \leq \mathcal{O}\left(\frac{k \log n}{m}\right) \tag{3.19}$$

with arbitrary high probability. Thus, the above k-RIC bound decreases as m increases, which is intuitively appealing. Interestingly, for fixed k and n, the above k-RIC upper bound for $\frac{\mathbf{A}}{\sqrt{m}} \odot \frac{\mathbf{B}}{\sqrt{m}}$ decays as $\mathcal{O}(\frac{1}{m})$. This is a significant improvement over the $\mathcal{O}(\frac{1}{\sqrt{m}})$ decay rate [62] already known for the individual k-RICs of the input subgaussian matrices $\frac{\mathbf{A}}{\sqrt{m}}$ and $\frac{\mathbf{B}}{\sqrt{m}}$. Thus, for any m, the Khatri-Rao product $\frac{\mathbf{A}}{\sqrt{m}} \odot \frac{\mathbf{B}}{\sqrt{m}}$ exhibits stronger restricted isometry property, with significantly smaller k-RICs compared to the k-RICs for the input matrices.

3.5.1 RIC bounds for the self Khatri-Rao product

In some applications, the effective sensing matrix can be expressed as the self-Khatri Rao product $\mathbf{X} \odot \mathbf{X}$ of a certain column normalized system matrix $\mathbf{X}$ with itself [30]. In Theorem 3.3 below, we present the k-RIC bound for the special case of self-Khatri-Rao product matrices.

Theorem 3.3. *Let $\mathbf{A}$ be an $m \times n$ random matrix with real i.i.d. subgaussian entries, such that $\mathbb{E}\mathbf{A}_{ij} = 0$, $\mathbb{E}\mathbf{A}_{ij}^2 = 1$, and $||\mathbf{A}_{ij}||_{\psi_2} \leq \kappa$. Then, the k^{th} order restricted isometry constant of the column normalized self Khatri-Rao product $\frac{\mathbf{A}}{\sqrt{m}} \odot \frac{\mathbf{A}}{\sqrt{m}}$ satisfies $\delta_k \leq \delta$ with probability at least $1 - 5n^{-2(\gamma-1)}$ for any $\gamma \geq 1$, provided*

$$m \geq 4c'\gamma\kappa^2\left(\frac{k \log n}{\delta}\right).$$

Here, $c' > 0$ is a universal constant.

Proof. See Appendix E.2. □

Theorem 3.3 implies that

$$\delta_k\left(\frac{\mathbf{A}}{\sqrt{m}} \odot \frac{\mathbf{A}}{\sqrt{m}}\right) \leq \mathcal{O}\left(\frac{k \log n}{m}\right) \tag{3.20}$$

with very high probability. The above k-RIC bound for the self Khatri-Rao product scales with m, n, and k in a similar fashion as the asymmetric Khatri-Rao product.

Remark 3: From Theorem 3.3, the $k-$ RIC of the columnwise Khatri-Rao product of two $m \times n$ sized subgaussian matrices becomes less than unity for $m = \mathcal{O}(k \log n)$. We suspect that the k-RIC of the Khatri-Rao product continues to be less than unity even for $m < k$. However, proving this is beyond the scope of the thesis.

3.6 Chapter Summary

In this chapter, we analyzed the restricted isometry property of the columnwise Khatri-Rao product matrix in terms of its restricted isometry constants. We gave two upper bounds for the k-RIC of a generic columnwise Khatri-Rao product matrix. The first k-RIC bound, a deterministic bound, is valid for the Khatri-Rao product of an arbitrary pair of complex valued input matrices of the same size with normalized columns. It is conveniently computed in terms of the k-RICs of the input matrices. We also gave a probabilistic RIC bound for the columnwise KR product of a pair of random matrices with i.i.d. subgaussian entries. The probabilistic RIC bound is one of the key components needed for computing tight sample complexity bounds for several machine learning algorithms.

The analysis of the RIP of Khatri-Rao product matrices in this chapter can be extended in multiple ways. The current RIC bounds can be extended to the Khatri-Rao product of three or more matrices. More importantly, in order to relate the RICs to the dimensions of the input matrices, we had to resort to the randomness in their entries. Removing this randomness aspect of our results could be an interesting direction for future work.

Chapter 4

Support Recovery Guarantees for Sparse Bayesian Learning

In which supports of size more than number of measurements are recoverable...

4.1 Introduction

In this chapter, we investigate the support recovery performance of the Multiple Sparse Bayesian Learning (MSBL) algorithm in the multiple measurement vector (MMV) framework. We derive new, improved sufficient conditions under which MSBL successfully recovers the correct nonzero support of a joint sparse Gaussian ensemble with vanishing support error probability. We show that the support error probability decays exponentially fast with the number of MMVs, and the decay exponent depends on the restricted eigenvalues and null space structure of self Khatri-Rao product of the measurement matrix. In the particular case of noiseless measurements, we show that a single MMV suffices for perfect recovery of the k-sparse support in MSBL, provided any $k+1$ columns of the

measurement matrix are linearly independent. Unlike the existing support recovery guarantees for MSBL, the new sufficient conditions derived in this chapter are non-asymptotic and cover a broader class of source signals than before.

4.2 Prior Work

MSBL's support recovery performance has been earlier investigated in [32] for $k < m$, and in [31, 43] for $k \geq m$. In these studies, it is assumed that the nonzero rows of $\mathbf{X}$ are mutually orthogonal. For finite L, the row-orthogonality condition is too restrictive for a deterministic $\mathbf{X}$ and almost never valid for a continuously distributed random source. Thus, the present support recovery guarantees for MSBL are in reality only applicable in the asymptotic sense when $L \to \infty$. Furthermore, the earlier analysis is restricted only to noiseless measurements. In contrast, the new sufficient conditions derived in this chapter are non-asymptotic, do not assume row orthogonality in the signal matrix $\mathbf{X}$, and account for the presence of measurement noise.

4.2.1 Information Theoretic Results on Support Recovery

There exist a handful of information theoretic results concerning the fundamental limits of sparse support recovery in the MMV problem. In [81], the 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. $\mathcal{N}(0, \mathrm{diag}(\mathbf{1}_{\mathcal{S}}))$, where $\mathcal{S}$ denotes the unknown support set. For $m = \Omega\left(k \log \frac{n}{k}\right)$, it is shown that $L \gg \frac{\log n}{\log \log n}$ suffices for vanishing support error probability. In this chapter, we extend this result to a more general signal prior on $\mathbf{X}$. Our

analysis indicates that reliable support recovery is possible even when m scales sublinearly in the support size k.

In [82], the support recovery problem is analyzed as a single-input-multi-output MAC communication problem. For number of nonzero rows fixed to k, $m = \frac{k \log n}{c(\mathbf{X})}$ 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. In [83], it has been shown that even fewer measurements $m = \Omega(\frac{k}{L} \log n)$ are sufficient when support size k grows sublinearly with n and each measurement vector is generated using a different sensing matrix.

Complementary to the existing results, our goal here is to analyze the support recovery in the MMV problem for a non-asymptotic setup. In particular, we seek sufficient conditions for vanishing support error probability in MSBL in terms of properties of the measurement matrix and the number of MMVs.

4.2.2 Contributions

The following are the main contributions of this chapter.

1. We derive new sufficient conditions under which a constrained version of MSBL with bounded hyperparameters perfectly recovers the true support of the joint sparse vectors with arbitrarily high probability in the noisy MMV problem.

2. We theoretically show that the support error probability decays exponentially with the number of MMVs. The error exponent is related to a restricted eigenvalue property and the null space structure of $\mathbf{A} \odot \mathbf{A}$, the self Khatri-Rao product of

the measurement matrix $\mathbf{A}$. Explicit upper bounds on the number of MMVs sufficient for vanishing support error probability in MSBL in both noisy and noiseless measurement cases are derived.

A key feature of our analysis is that our sufficient conditions are applicable to both random as well as deterministic constructions of $\mathbf{A}$.

4.3 Measurement and Source Models

4.3.1 Measurement Model

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 $\mathcal{S}^* \subseteq [n]$. We are interested in the recovery of the true support $\mathcal{S}^*$ from noisy underdetermined linear measurements $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_L$ generated as

$$\mathbf{y}_j = \mathbf{A}\mathbf{x}_j + \mathbf{w}_j, \quad 1 \leq j \leq L. \tag{4.1}$$

The measurement matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ is assumed to be a non-degenerate matrix, with $m \leq n$. By non-degeneracy of $\mathbf{A}$, it is implied that 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 assumed to be zero mean Gaussian distributed with diagonal covariance matrix $\sigma^2 \mathbf{I}_m$. The linear measurement model in (4.1) 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 support $\mathcal{S}^*$, $\mathbf{X}$ is a row sparse matrix, and its row-support $\mathcal{S}^*$ denoted by $\mathcal{R}(\mathbf{X})$ is equal to $\mathcal{S}^*$.

4.3.2 Signal Model

We make the following assumption about the unknown signal matrix $\mathbf{X}$.

Assumption 1: If the i^{th} 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 $\boldsymbol{\gamma}^*(i)$ belonging to the interval $[\gamma_{\min}, \gamma_{\max}]$.

Let K denote the maximum number of nonzero rows in $\mathbf{X}$, i.e., $|\mathcal{S}^*| \leq K$. Then, an immediate consequence of (**A1**) is that there exists a bounded, nonnegative, and at most K sparse vector, $\boldsymbol{\gamma}^* \in \mathbb{R}_+^n$, such that the columns $\mathbf{x}_j$ are i.i.d. $\mathcal{N}(0, \boldsymbol{\Gamma}^*)$ with $\boldsymbol{\Gamma}^* \triangleq \text{diag}(\boldsymbol{\gamma}^*)$. Note that under the assumption (**A1**), the nonzero elements in $\mathbf{X}$ are pairwise uncorrelated within and across its nonzero rows. Hence, $\mathbf{X}$ is compatible with the joint sparsity model JSM-2 proposed in [16].

Compared to [81] which also considers a Gaussian source model but with binary valued variance parameters, ours is a more general source model that allows for real valued and distinct row-variances. In fact, the signal model in [81] can be obtained as a special case by setting $\boldsymbol{\gamma}_{\min} = \boldsymbol{\gamma}_{\max} = 1$ in the assumption (**A1**).

4.4 Support Error Probability Analysis

We now proceed with the analysis of support recovery probability incurred by MSBL based support reconstruction. Under assumption (**A1**), we derive an MSBL specific Chernoff bound for the support error probability. We begin by introducing some of the frequently used notation in the table below.

Symbol	Description
K	Maximum number of nonzero rows in $\mathbf{X}$
$\mathcal{S}_k$	Collection of all support sets of k or lesser size, i.e., $\mathcal{S}_k = \{\mathcal{S} \subseteq [n], \lvert\mathcal{S}\rvert \leq k\}$
$\Theta(\mathcal{S})$	A bounded hyperparameter set associated with the support set $\mathcal{S}$, formally defined as $\Theta(\mathcal{S}) \triangleq \left\{\boldsymbol{\gamma} \in \mathbb{R}^n_+ : \text{supp}(\boldsymbol{\gamma}) = \mathcal{S}, \gamma_{\min} \preceq \boldsymbol{\gamma}_{\mathcal{S}} \preceq \gamma_{\max}\right\}$.
Θ	Collection of all nonnegative vectors in $\mathbb{R}^n_+$ whose nonzero elements belong to $[\gamma_{\min}, \gamma_{\max}]$. Also, $\Theta = \bigcup_{\mathcal{S} \in \mathcal{S}_n} \Theta(\mathcal{S})$.
$\mathcal{S}^*$	True row-support of $\mathbf{X}$.
$\boldsymbol{\gamma}^*$	Principal diagonal of the common covariance matrix $\boldsymbol{\Gamma}^*$ of the i.i.d. columns in $\mathbf{X}$. Also, $\text{supp}(\boldsymbol{\gamma}^*) = \mathcal{S}^*$.

By assumption **A1** on $\mathbf{X}$, we have $\boldsymbol{\gamma}^* \in \Theta$. Thus, in order to recover $\boldsymbol{\gamma}^*$ from $\mathbf{Y}$, we consider solving the following constrained version of the MSBL optimization problem in (2.2):

$$\text{cMSBL:} \quad \hat{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\gamma} \in \Theta} \mathcal{L}(\mathbf{Y}; \boldsymbol{\gamma}). \tag{4.2}$$

In the above, the objective $\mathcal{L}(\mathbf{Y}; \boldsymbol{\gamma}) \triangleq \log p(\mathbf{Y}; \boldsymbol{\gamma})$ is the same as the MSBL log-likelihood cost in (2.3). The row-support of $\mathbf{X}$ is then estimated as $\text{supp}(\hat{\boldsymbol{\gamma}})$, where $\hat{\boldsymbol{\gamma}}$ is a solution of (4.2). Let us consider the set of bad MMVs,

$$\mathcal{E}_{\mathcal{S}^*} \triangleq \left\{\mathbf{Y} \in \mathbb{R}^{m \times L} : \text{supp}\left(\hat{\boldsymbol{\gamma}}\right) \neq \mathcal{S}^*\right\}, \tag{4.3}$$

which result in erroneous estimation of the true support. In other words, $\mathcal{E}_{\mathcal{S}^*}$ is the collection of undesired MMVs for which MSBL's log-likelihood objective is globally maximized by $\boldsymbol{\gamma} \in \Theta(\mathcal{S})$, $\mathcal{S} \neq \mathcal{S}^*$. Therefore, one can rewrite $\mathcal{E}_{\mathcal{S}^*}$ as

$$\mathcal{E}_{\mathcal{S}^*} = \bigcup_{\mathcal{S} \in \mathcal{S}_n \setminus \{\mathcal{S}^*\}} \left\{\mathbf{Y} : \max_{\boldsymbol{\gamma} \in \Theta(\mathcal{S})} \mathcal{L}(\mathbf{Y}; \boldsymbol{\gamma}) \geq \max_{\boldsymbol{\gamma}' \in \Theta(\mathcal{S}^*)} \mathcal{L}(\mathbf{Y}; \boldsymbol{\gamma}')\right\}. \tag{4.4}$$

We are interested in identifying the conditions under which $\mathbb{P}(\mathcal{E}_{\mathcal{S}^*})$ can be made arbitrarily

small. Since $\max_{\boldsymbol{\gamma}' \in \Theta(\mathcal{S}^*)} \mathcal{L}(\mathbf{Y}; \boldsymbol{\gamma}') \geq \mathcal{L}(\mathbf{Y}; \boldsymbol{\gamma}^*)$, it follows that

$$\begin{aligned} \mathcal{E}_{\mathcal{S}^*} &\subseteq \bigcup_{\mathcal{S} \in \mathcal{S}_n \setminus \{\mathcal{S}^*\}} \left\{ \mathbf{Y} : \max_{\boldsymbol{\gamma} \in \Theta(\mathcal{S})} \mathcal{L}(\mathbf{Y}; \boldsymbol{\gamma}) \geq \mathcal{L}(\mathbf{Y}; \boldsymbol{\gamma}^*) \right\} \\ &= \bigcup_{\mathcal{S} \in \mathcal{S}_n \setminus \{\mathcal{S}^*\}} \bigcup_{\boldsymbol{\gamma} \in \Theta(\mathcal{S})} \{ \mathbf{Y} : \mathcal{L}(\mathbf{Y}; \boldsymbol{\gamma}) - \mathcal{L}(\mathbf{Y}; \boldsymbol{\gamma}^*) \geq 0 \}. \end{aligned} \tag{4.5}$$

The continuous union over infinitely many elements of $\Theta(\mathcal{S})$ in (4.5) 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 $\boldsymbol{\gamma}' \in \Theta^\epsilon(\mathcal{S})$ such that $|\mathcal{L}(\mathbf{Y}; \boldsymbol{\gamma}) - \mathcal{L}(\mathbf{Y}; \boldsymbol{\gamma}')| \leq \epsilon$. Proposition 4.1 gives an upper bound on the size of such an ϵ-net.

Proposition 4.1. *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. $|\Theta^\epsilon(\mathcal{S})| \leq \max\left\{1, \left(\frac{3C_{\mathcal{L},\mathcal{S}}(\gamma_{max} - \gamma_{min})\sqrt{|\mathcal{S}|}}{\epsilon}\right)^{|\mathcal{S}|}\right\}$, 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 F.1. □

From Proposition 4.1-(ii), we observe that the construction of $\Theta^\epsilon(\mathcal{S})$ depends on the Lipschitz continuity of the log-likelihood function $\mathcal{L}(\mathbf{Y}; \boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$ over the domain $\Theta(\mathcal{S})$. By virtue of the data-dependent nature of $\mathcal{L}(\mathbf{Y}; \boldsymbol{\gamma})$, its Lipschitz constant $\mathcal{C}_{\mathcal{L},\mathcal{S}}$ 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 $\mathcal{C}_{\mathcal{L},\mathcal{S}}$

is uniformly bounded solely in terms of the second order statistics of $\mathbf{Y}$. A possible choice of $\mathcal{G}$ could be

$$\mathcal{G} \triangleq \left\{ \mathbf{Y} \subset \mathbb{R}^{m \times L} : \left| \left| \left| \frac{1}{L} \mathbf{Y}\mathbf{Y}^T \right| \right| \right|_2 \leq 2\mathbb{E}\left[\mathbf{y}_1 \mathbf{y}_1^T \right] \right\}. \tag{4.6}$$

For $\mathbf{Y} \in \mathcal{G}$, $\mathcal{L}(\mathbf{Y}; \boldsymbol{\gamma})$ is uniformly Lipschitz continuous irrespective of $\mathbf{Y}$, and hence the ϵ-net $\Theta^{\epsilon}(\mathcal{S})$ can now be constructed entirely independent of $\mathbf{Y}$.

Since for arbitrary sets $\mathcal{A}$ and $\mathcal{B}$, $\mathcal{A} \subseteq (\mathcal{A} \cap \mathcal{B}) \cup \mathcal{B}^c$, the RHS in (4.5) relaxes as

$$\mathcal{E}_{\mathcal{S}^*} \subseteq \left\{ \bigcup_{\mathcal{S} \in \mathcal{S}_n \setminus \mathcal{S}^*} \bigcup_{\boldsymbol{\gamma} \in \Theta(\mathcal{S})} \{ \mathcal{L}(\mathbf{Y}; \boldsymbol{\gamma}) - \mathcal{L}(\mathbf{Y}; \boldsymbol{\gamma}^*) \geq 0 \} \cap \mathcal{G} \right\} \cup \ \mathcal{G}^c. \tag{4.7}$$

Let $\Theta^{\epsilon}(\mathcal{S})|_{\mathcal{G}}$ denote an ϵ-net of $\Theta(\mathcal{S})$ constructed under the assumption that $\mathbf{Y} \in \mathcal{G}$. Then, the continuous union over $\Theta(\mathcal{S})$ relaxes to a finite sized union over $\Theta^{\epsilon}(\mathcal{S})|_{\mathcal{G}}$ as shown below.

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

By applying the union bound, we obtain

$$\mathbb{P}\left(\mathcal{E}_{\mathcal{S}^*}\right) \ \leq \ \mathbb{P}\left(\mathcal{G}^c\right) \ + \sum_{\mathcal{S} \in \mathcal{S}_n \setminus \mathcal{S}^*} \sum_{\boldsymbol{\gamma} \in \Theta^{\epsilon}(\mathcal{S})|_{\mathcal{G}}} \mathbb{P}\left(\mathcal{L}(\mathbf{Y}; \boldsymbol{\gamma}) - \mathcal{L}(\mathbf{Y}; \boldsymbol{\gamma}^*) \geq -\epsilon \right) \tag{4.8}$$

From (4.8), $\mathbb{P}(\mathcal{E}_{\mathcal{S}^*})$ will be small if each of the constituent probability terms $\mathbb{P}\left(\mathcal{L}(\mathbf{Y}; \boldsymbol{\gamma}) - \mathcal{L}(\mathbf{Y}; \boldsymbol{\gamma}^*) \geq -\epsilon\right)$, $\boldsymbol{\gamma} \in \Theta^{\epsilon}(\mathcal{S})|_{\mathcal{G}}$ are sufficiently small so that their collective contribution remains small, and $\mathbb{P}(\mathcal{G}^c)$ is small. In Theorem 4.1, we show that each event within the summation in (4.8) is a large deviation event which occurs with an exponentially decaying probability.

Theorem 4.1. *Let $p_{\boldsymbol{\gamma}}(\mathbf{y})$ denote the marginal probability density of the columns of $\mathbf{Y}$ induced by the joint sparse columns of $\mathbf{X}$ drawn independently from $\mathcal{N}(0, \mathrm{diag}(\boldsymbol{\gamma}))$. 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.*

$$\mathbb{P}\left(\mathcal{L}(\mathbf{Y};\boldsymbol{\gamma}) - \mathcal{L}(\mathbf{Y};\boldsymbol{\gamma}^*) \geq -\epsilon\right) \;\leq\; \exp\left(-L\psi^*\left(-\frac{\epsilon}{L}\right)\right), \tag{4.9}$$

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

Proof. See Appendix F.2. □

Note that, when the measurement noise is Gaussian, the marginal density $p_{\boldsymbol{\gamma}}(\mathbf{y}_j)$ of the individual observations 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 term $\mathcal{D}_t(p_{\boldsymbol{\gamma}}, p_{\boldsymbol{\gamma}^*})$ in Theorem 4.1 is well defined. We now restate Theorem 4.1 as Corollary 4.1, which is the final form of the large deviation result for $\mathcal{L}(\mathbf{Y};\boldsymbol{\gamma})$ that will be used later for bounding $\mathbb{P}(\mathcal{E}_{\mathcal{S}^*})$.

Corollary 4.1. *For an arbitrary $\boldsymbol{\gamma} \in \mathbb{R}_+^n$, and the true variance parameters $\boldsymbol{\gamma}^*$, let the associated marginal densities $p_{\boldsymbol{\gamma}}$ and $p_{\boldsymbol{\gamma}^*}$ be as defined in Theorem 4.1, and suppose $\sigma^2 > 0$. Then, the log-likelihood $\mathcal{L}(\mathbf{Y};\boldsymbol{\gamma})$ satisfies the large deviation property*

$$\mathbb{P}\left(\mathcal{L}(\mathbf{Y};\boldsymbol{\gamma}) - \mathcal{L}(\mathbf{Y};\boldsymbol{\gamma}^*) \geq -\frac{L\mathcal{D}_{1/2}(p_{\boldsymbol{\gamma}}, p_{\boldsymbol{\gamma}^*})}{2}\right) \leq \exp\left(-\frac{L\mathcal{D}_{1/2}\left(p_{\boldsymbol{\gamma}}, p_{\boldsymbol{\gamma}^*}\right)}{4}\right). \tag{4.10}$$

Proof. The large deviation result is obtained by replacing $\psi^*\left(-\frac{\epsilon}{L}\right)$ in Theorem 4.1 by its lower bound $-\frac{t\epsilon}{L} - \psi(t)$, followed by setting $t = 1/2$ and $\epsilon = L\mathcal{D}_{1/2}(p_{\boldsymbol{\gamma}}, p_{\boldsymbol{\gamma}^*})/2$. □

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

Note that, in the above, we have used the sub-optimal choice $t = 1/2$ for the Chernoff parameter t, since its optimal value is not available in closed form. However, this sub-optimal selection of t is inconsequential as it figures only as a multiplicative constant in the final sample complexity. By using Corollary 4.1 in (4.8), we can bound $\mathbb{P}\left(\mathcal{E}_{\mathcal{S}^*}\right)$ as

$$\mathbb{P}(\mathcal{E}_{\mathcal{S}^*}) \leq \sum_{\mathcal{S}\in\mathcal{S}_n\setminus\mathcal{S}^*} \left|\Theta^{\epsilon}(\mathcal{S})|_{\mathcal{G}}\right| \exp\left(-\frac{L\mathcal{D}^*_{\mathcal{S}}}{4}\right) \; + \; \mathbb{P}\left(\mathcal{G}^c\right), \tag{4.11}$$

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

$$\mathcal{D}^*_{\mathcal{S}} \triangleq \inf_{\boldsymbol{\gamma}\in\Theta(\mathcal{S})} \mathcal{D}_{1/2}\left(p_{\boldsymbol{\gamma}}, p_{\boldsymbol{\gamma}^*}\right). \tag{4.12}$$

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

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_{\mathcal{S}^*}) &\leq \sum_{\mathcal{S}\in\mathcal{S}_n\setminus\mathcal{S}^*} \exp\left(-Lk_d^{\mathcal{S},\mathcal{S}^*}\left(\frac{\mathcal{D}^*_{\mathcal{S}}}{4k_d^{\mathcal{S},\mathcal{S}^*}} - \frac{\log\left|\Theta^{\epsilon}(\mathcal{S})|_{\mathcal{G}}\right|}{Lk_d^{\mathcal{S},\mathcal{S}^*}}\right)\right) + \mathbb{P}\left(\mathcal{G}^c\right) \\
&\leq \sum_{\mathcal{S}\in\mathcal{S}_n\setminus\mathcal{S}^*} \exp\left(-Lk_d^{\mathcal{S},\mathcal{S}^*}\left(\frac{\eta}{4} - \frac{\kappa_{\text{cov}}}{L}\right)\right) + \mathbb{P}\left(\mathcal{G}^c\right),
\end{aligned} \tag{4.13}$$

where

$$\eta \triangleq \min_{\mathcal{S}\subseteq[n]}\left(\frac{\mathcal{D}^*_{\mathcal{S}}}{k_d^{\mathcal{S},\mathcal{S}^*}}\right), \tag{4.14}$$

$$\text{and} \quad \kappa_{\text{cov}} \triangleq \max_{\mathcal{S}\subseteq[n]}\left(\frac{\log\left|\Theta^{L\mathcal{D}^*_{\mathcal{S}}}(\mathcal{S})|_{\mathcal{G}}\right|}{k_d^{\mathcal{S},\mathcal{S}^*}}\right). \tag{4.15}$$

Theorem 4.2. *Suppose $\mathcal{S}^*$ is the true row support of the unknown $\mathbf{X}$ satisfying assumption **A1**. Then, for any $\delta \in (0,1)$, $\mathbb{P}(\mathcal{E}_{\mathcal{S}^*}) \leq 2\delta$, if*

$$L \geq \max\left\{\frac{8}{\eta}\log\left(3enK\left(\frac{1+\delta}{\delta}\right)\right), \frac{8\kappa_{cov}}{\eta}, C\log\frac{2}{\delta}\right\}, \tag{4.16}$$

where η and κ_{cov} are as defined in (4.14) *and* (4.15)*, respectively, and $C > 0$ is a universal numerical constant.*

Proof. Since $L \geq C \log(2/\delta)$, by Proposition C.2, $\mathbb{P}(\mathcal{G}^c) \leq \delta$. Combined with $L \geq \frac{8\kappa_{\text{cov}}}{\eta}$, (4.13) can be rewritten as

$$\mathbb{P}(\mathcal{E}_{\mathcal{S}^*}) \leq \sum_{\mathcal{S} \in \mathcal{S}_n \backslash \mathcal{S}^*} \exp\left(-\frac{L k_d^{\mathcal{S},\mathcal{S}^*} \eta}{8}\right) + \delta. \tag{4.17}$$

The total number of support sets belonging to $\mathcal{S}_n \backslash \mathcal{S}^*$ which differ from the true support $\mathcal{S}^*$ in exactly k_d locations is $\sum_{j=0}^{k_d} \binom{n-|\mathcal{S}^*|}{j} \binom{j+|\mathcal{S}^*|}{\min(k_d, j+|\mathcal{S}^*|)}$, which can be further upper bounded by $(3enK)^{k_d}$. Thus, we can rewrite (4.17) as

$$\begin{aligned} \mathbb{P}(\mathcal{E}_{\mathcal{S}^*}) &\leq \delta + \sum_{k_d=1}^{n-|\mathcal{S}^*|} \sum_{\substack{\mathcal{S} \in \mathcal{S}_n \backslash \mathcal{S}^*, \\ |(\mathcal{S} \backslash \mathcal{S}^*) \cup (\mathcal{S}^* \backslash \mathcal{S})| = k_d}} \exp\left(-\frac{\eta L k_d}{8}\right) \\ &\leq \delta + \sum_{k_d=1}^{n-|\mathcal{S}^*|} (3enK)^{k_d} \left(e^{-\frac{\eta L}{8}}\right)^{k_d}. \end{aligned} \tag{4.18}$$

Since $L \geq \frac{8}{\eta} \log\left(3enK\left(\frac{1+\delta}{\delta}\right)\right)$, $\mathbb{P}(\mathcal{E}_{\mathcal{S}^*})$ can be bounded by an infinite geometric series as

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

This completes the proof. □

In Theorem 4.2, we finally have an abstract bound on the sufficient number of MMVs, L, which guarantees vanishing support error probability in cMSBL, for a fixed true support $\mathcal{S}^*$. However the MMV bound is meaningful only when η is strictly positive. We now proceed to deduce the conditions for

i. $\eta > 0$ (bounded away from zero by a strictly positive constant),

ii. η and κ_{cov} scaling favorably with the system dimensions.

4.4.1 Bounds for η and κ_{cov}

To understand how large η as defined in (4.14) can be, we first derive a lower bound for $\mathcal{D}^*_{\mathcal{S}}$. The following proposition provides a lower bound for $\mathcal{D}^*_{\mathcal{S}}$ for any $\mathcal{S} \subseteq [n]$.

Proposition 4.2. *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 $\boldsymbol{\gamma}, \boldsymbol{\gamma}^* \in \Theta$ 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*

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

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.4. □

From Proposition 4.2, 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 sparse vector), then $\eta = \min_{\mathcal{S}\in[n],\mathcal{S}\neq\mathcal{S}^*} \min_{\boldsymbol{\gamma}\in\Theta(\mathcal{S})} \mathcal{D}_{\frac{1}{2}}\left(p_{\boldsymbol{\gamma}}, p_{\boldsymbol{\gamma}^*}\right)/k_d^{\mathcal{S},\mathcal{S}^*}$ is always strictly positive. This condition can be formalized as a restricted null space property of $\mathbf{A} \odot \mathbf{A}$, defined next.

Definition 4.1. *A matrix is said to satisfy the non-negative restricted null space property (NN-RNSP) of order k if its null space does not contain any vectors that are expressible as a difference between k (or lesser) sparse nonnegative vector and an arbitrary nonnegative vector.*

From Proposition 4.2, it is evident that if the self Khatri-Rao product $\mathbf{A} \odot \mathbf{A}$ satisfies the NN-RNSP of order K, we have $\eta > 0$. We now show that if the ℓ_2 norm of the columns vectors in $\mathbf{A}$ are close to unity, then $\mathbf{A} \odot \mathbf{A}$ partially satisfies the NN-RNSP.

Theorem 4.3. *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 restricted null space property:*

$$||(\mathbf{A} \odot \mathbf{A})\mathbf{v}||_2^2 \geq \frac{(1-\alpha)^2}{2m} \left(||\mathbf{v}_+||_1^2 + ||\mathbf{v}_-||_1^2 \right) \tag{4.19}$$

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 F.3. □

An interesting consequence of Theorem 4.3 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 $||\boldsymbol{\gamma}||_0 \geq 4 \left(\frac{\gamma_{\max}}{\gamma_{\min}}\right) \left(\frac{1+\alpha}{1-\alpha}\right)^2 ||\boldsymbol{\gamma}^*||_0$. However, when the support sizes of $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^*$ are comparable, it is not as straightforward to ascertain the conditions under which $\boldsymbol{\gamma} - \boldsymbol{\gamma}^*$ does not lie in the null space of $\mathbf{A} \odot \mathbf{A}$. In Proposition 4.3, we state verifiable sufficient conditions that subsume the NN-RNSP condition for $\mathbf{A} \odot \mathbf{A}$, which in turn guarantees a strictly positive η.

Proposition 4.3. *Let the measurement matrix* $\mathbf{A}$ *satisfy the following two properties*[2]

[2]Condition 1 ensures that Null($\mathbf{A} \odot \mathbf{A}$) does not contain any vectors of the form $\boldsymbol{\gamma} - \boldsymbol{\gamma}^*$, where $||\boldsymbol{\gamma}||_0 \geq K + K_{\text{threshold}}$, whereas condition 2 ensures that NN-RNSP of order k holds for $||\boldsymbol{\gamma}||_0 < K + K_{\text{threshold}}$.

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

2. *For* $1 \leq k \leq (K + K_{threshold})$*, there exists a scalar* $\beta > 0$ *such that* $||(\mathbf{A} \odot \mathbf{A})\mathbf{v}||_2^2 \geq \beta ||\mathbf{v}||_2^2$*, for all* k*-sparse vectors* $\mathbf{v} \in \mathbb{R}^n$.

Then, η *as defined in* (4.14) *is bounded as*

$$\eta \geq \frac{\boldsymbol{\gamma}_{min}^2}{\left(\sigma^2 + 2\boldsymbol{\gamma}_{max}\right)^2} \min \left(\frac{\beta}{\delta^2_{(K+K_{threshold})}}, \frac{1}{4m\delta_n} \min_{\substack{\mathcal{S}\subseteq[n], \\ |\mathcal{S}\setminus\mathcal{S}^*|+|\mathcal{S}^*\setminus\mathcal{S}|>K_{threshold}}} \frac{|\mathcal{S}\cup\mathcal{S}^*|}{\delta_{|\mathcal{S}\cup\mathcal{S}^*|}} \right), \tag{4.20}$$

where $K_{threshold} \triangleq \left(1 + 4\frac{\boldsymbol{\gamma}_{max}}{\boldsymbol{\gamma}_{min}}\left(\frac{1+\alpha}{1-\alpha}\right)^2\right) K$ *and* $\delta_k \triangleq \max_{\mathcal{S}\subseteq[n]:|\mathcal{S}|\leq k} \left|\left|\left|\mathbf{A}_{\mathcal{S}}^T \mathbf{A}_{\mathcal{S}}\right|\right|\right|_2$.

Proof. See Appendix F.5. □

The lower bound for η in (4.20) can be used in Proposition 4.1 to obtain the following upper bound for κ_{cov}.

Proposition 4.4. *For the same setting as Proposition 4.3,*

$$\kappa_{cov} \leq (K+1) \log \left(\frac{6\sqrt{2}m\sqrt{K}(\boldsymbol{\gamma}_{max} - \boldsymbol{\gamma}_{min})\left(3\sigma^2 + 2\boldsymbol{\gamma}_{max}\right)\delta_K}{\boldsymbol{\gamma}_{min}\sigma^2\eta} \right), \tag{4.21}$$

where η *is bounded according to* (4.20).

Proof. See Appendix F.6. □

The above abstract bounds for η and κ_{cov} are applicable for both deterministic as well as random measurement matrix $\mathbf{A}$. To get clarity on how η and κ_{cov} scale with the problem dimensions, we now consider the scenario where $\mathbf{A}$ is randomly constructed. The following corollary states the simplified bounds for η and κ_{cov}, when $\mathbf{A}$ is drawn from a Gaussian ensemble.

Corollary 4.2. *Let* $\mathbf{A}$ *be an* $m \times n$ *sized matrix with independent and identically distributed* $\mathcal{N}\left(0, \frac{1}{\sqrt{m}}\right)$ *entries. Then, for* $m \geq \mathcal{O}(K \log n)$*, we have*

$$\begin{aligned} \eta &\geq \frac{c_1 \boldsymbol{\gamma}_{min}^2}{\left(\sigma^2 + 2\boldsymbol{\gamma}_{max}\right)^2} \left(\frac{K}{n}\right), \\ \text{and } \kappa_{cov} &\leq (K+1) \log \left(\frac{c_2 n \sqrt{K} \log n (\boldsymbol{\gamma}_{max} - \boldsymbol{\gamma}_{min}) \left(3\sigma^2 + 2\boldsymbol{\gamma}_{max}\right)^3}{\boldsymbol{\gamma}_{min}^3 \sigma^2} \right), \end{aligned}$$

with probability exceeding $1 - c_3 n^{-2}$*. Here* c_1, c_2 *are positive numerical constants and* $c_3 > 1$ *is a constant independent of the problem dimensions.*

Proof. See Appendix F.8. □

From Corollary 4.2, one can conclude that for $m \geq \mathcal{O}(K \log n)$, κ_{cov} overall scales as $\mathcal{O}(K \log n)$. It can be further shown that the simplified bounds for η and κ_{cov} derived for a Gaussian measurement matrix in Corollary 4.2 retain their orderwise relationship with respect to m, n and K, even for subgaussian measurement matrices.

Remark 1*:* While proving Corollary 4.2, it is required to show the existence of a strictly positive β that satisfies condition 2 of Proposition 4.3, which is tantamount to showing that any submatrix of $\mathbf{A} \odot \mathbf{A}$ obtained by sampling its K or fewer columns is nonsingular. This is indeed true for a subgaussian measurement matrix $\mathbf{A}$ with $m \geq \mathcal{O}(K \log n)$ rows, as a direct consequence of $\mathbf{A} \odot \mathbf{A}$ satisfying the K^{th} order restricted isometry property (Theorem 3.3). We conjecture that $\mathbf{A} \odot \mathbf{A}$ continues to satisfy the left-sided K-RIP (condition 2 in Proposition 4.3) even when m scales as $\mathcal{O}(\sqrt{K})$ for a fixed n. However, currently, a formal proof for such a result remains elusive.

4.5 Sufficient Conditions for Support Recovery

Equipped with explicit bounds for η and κ_{cov}, we can now state the sufficient conditions for vanishing support error probability in cMSBL as the following theorem.

Theorem 4.4. *Under assumption (A1), suppose* $\mathbf{X}$ *has row support* $\mathcal{S}^*$, $|\mathcal{S}^*| \leq K$. *Let* $\hat{\boldsymbol{\gamma}}$ *be the solution of cMSBL. Then,* $supp\,(\hat{\boldsymbol{\gamma}}) = \mathcal{S}^*$ *with probability exceeding* $1 - 2\delta$, *for any* $\delta \in (0, 1)$, *provided the following two conditions are satisfied.*

C1. *The measurement matrix* $\mathbf{A}$ *satisfies the conditions* 1 *and* 2 *in Proposition 4.3 with strictly positive parameters* α *and* β.

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} *are bounded according to* (4.20) *and* (4.21), *respectively.*

Proof. **C1** ensures that η is strictly positive and bounded as per (4.20). Further, **C2** ensures that the abstract MMV bound in (4.16) is satisfied. Then, by Theorem 4.2, it follows that $\mathbb{P}(\mathcal{E}_{\mathcal{S}^*}) \leq 2\delta$. □

In the following corollary, we state an extra condition besides **C1** and **C2** which extends the above cMSBL result to MSBL based support recovery.

Corollary 4.3. *For the same setting as Theorem 4.4, let* $\hat{\boldsymbol{\gamma}}$ *be the output of MSBL optimization in* (2.2). *If the conditions* ***C1*** *and* ***C2*** *hold for a given* $\delta \in (0, 1)$, *and* $\hat{\boldsymbol{\gamma}} \in \Theta$, *then* $supp(\hat{\boldsymbol{\gamma}}) = \mathcal{S}^*$ *with probability exceeding* $1 - 2\delta$.

Proof. Since $\hat{\boldsymbol{\gamma}}$ belongs to Θ and simultaneously maximizes the MSBL objective $\log p(\mathbf{Y};\boldsymbol{\gamma})$, it follows that $\hat{\boldsymbol{\gamma}}$ is also a solution to the cMSBL optimization in (4.2). Hence, the statement of Corollary 4.3 follows directly from Theorem 4.4 provided **C1** and **C2** are satisfied. □

According to Corollary 4.3, MSBL exhibits vanishing support error probability under conditions **C1** and **C2**, however only in a retrospective sense, i.e., when the MSBL output $\hat{\boldsymbol{\gamma}}$ belongs to the bounded domain Θ.

4.5.1 Average error probability

So far, we have analyzed the support error probability for the case where the true row support of $\mathbf{X}$ is fixed to $\mathcal{S}^*$. In reality, $\mathcal{R}(\mathbf{X})$ can assume any one of the $\binom{n}{1}+\binom{n}{2}+\cdots+\binom{n}{K}$ possible supports in the collection $\mathcal{S}_K$. The support error probability averaged over all possible supports can be evaluated as

$$P_{\text{avg}}^{\text{err}} = \sum_{\mathcal{S}^* \in \mathcal{S}_K} \mathbb{P}(\mathcal{R}(\mathbf{X}) = \mathcal{S}^*)\mathbb{P}\left(\mathcal{E}_{\mathcal{S}^*}\right), \tag{4.22}$$

where $\mathcal{E}_{\mathcal{S}^*}$ is as defined in (4.3). For example, if all supports in $\mathcal{S}_K$ are equiprobable, then $P_{\text{avg}}^{\text{err}} \leq 2\delta$ under **C1** and **C2** specified by Theorem 4.4.

4.6 Discussion

In this section, we interpret the sufficiency conditions **C1** and **C2** in Theorem 4.4 for vanishing support error probability in the context of various interesting cases.

4.6.1 The Case of Binary Hyperparameters

In [81], Tang and Nehorai formulated support recovery using MMVs as a multiple hypothesis testing problem by assuming that each column of $\mathbf{X}$ is i.i.d. $\mathcal{N}(0, \mathrm{diag}(\mathbf{1}_{\mathcal{S}^*}))$, where $\mathcal{S}^*$ is the k-sparse support set of $\mathbf{X}$. For this set parameterized signal prior, finding the true support via type-II likelihood maximization as in (2.2) is no longer a continuous variable optimization but rather a combinatorial search over all k-sparse vertices of the hypercube $\{0,1\}^n$, as described below.

$$\hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}\in\{0,1\}^n, \|\boldsymbol{\gamma}\|_0 = k}{\arg\max} \log p\left(\mathbf{Y}; \boldsymbol{\gamma}\right). \tag{4.23}$$

The binary valued, k-sparse hyperparameters can be accommodated as a special case of our signal model by setting $\boldsymbol{\gamma}_{\min} = \boldsymbol{\gamma}_{\max} = 1$. For $\boldsymbol{\gamma}_{\min} = \boldsymbol{\gamma}_{\max}$, according to Proposition 4.1, the ϵ-net $\Theta^{\epsilon}(\mathcal{S})$ collapses to a single point for all $\mathcal{S} \in \mathcal{S}_n$, which ultimately amounts to $\kappa_{\mathrm{cov}} = 1$. Further, in [81], the correct support has to be identified from $\binom{n}{K}$ possible candidate support hypothesis. For this restrained support recovery problem, the lower bound for η in (4.20) simplifies to

$$\eta \geq \frac{\beta}{\delta_{2K}^2}\left(\frac{1}{\sigma^2+2}\right), \tag{4.24}$$

and from Theorem 4.4, the support error probability $\mathbb{P}(\mathcal{E}_{\mathcal{S}^*})$ is guaranteed to at most 2ϵ, provided

$$L \geq 8\left(\sigma^2+2\right)\left(\frac{\delta_{2K}^2}{\beta}\right)\log\left(\frac{6enK}{\epsilon}\right), \tag{4.25}$$

for an arbitrary $\epsilon > 0$. For the measurement matrix $\mathbf{A}$ containing i.i.d $\mathcal{N}(0, 1/\sqrt{m})$ entries and $m = \mathcal{O}(K \log n)$ rows, from (F.24) and (F.27), both β and δ_{2K} behave as $\Theta(1)$.

Therefore, from (4.25), it can be concluded that $L = \mathcal{O}(\log(n/\epsilon))$ suffices to guarantee $\mathbb{P}(\mathcal{E}_{\mathcal{S}^*}) \leq 2\epsilon$. Compared to the sufficient condition, $L \gg \frac{\log n}{\log \log n}$ derived in [81], our MMV complexity bound contains an additional $\log \log n$ multiplicative factor.

4.6.2 The Case of Continuous Valued Hyperparameters

For the case of continuous valued hyperparameters, as the difference $\boldsymbol{\gamma}_{\max} - \boldsymbol{\gamma}_{\min}$ increases, the κ_{cov} term starts to dominant in condition **C2**, and that term, along with η, dictates the overall MMV complexity. Further, if the measurement matrix $\mathbf{A}$ contains i.i.d $\mathcal{N}\left(0, \frac{1}{\sqrt{m}}\right)$, it satisfies condition **C1** with constants α and β behaving as $\Theta(1)$ (as shown in (F.24) and (F.27)), with high probability, provided $m \geq \mathcal{O}\left(K \log n\right)$. Using the bounds for η and κ_{cov} from Corollary 4.2 and condition **C2**, we can conclude that

$$m \geq \mathcal{O}(K \log n) \text{ and } L \geq \mathcal{O}(n \log n) \tag{4.26}$$

together guarantee $\mathbb{P}(\mathcal{E}_{\mathcal{S}^*}) \leq 2\epsilon$. Further, if it can be shown that the output of MSBL or cMSBL is at most p-sparse, then the required MMV complexity relaxes to $L \geq \mathcal{O}(p \log n)$.

4.6.3 The case when $n = \mathbf{Rank}(\mathbf{A} \odot \mathbf{A})$

From [30, 73], there exist both random as well as deterministic constructions of $m \times n$ measurement matrix $\mathbf{A}$ for which $\mathbf{A} \odot \mathbf{A}$ is full column rank when $n \leq \frac{m^2+m}{2}$. If so, then $\mathbf{A} \odot \mathbf{A}$ satisfies NN-RNSP of order K (Definition 4.1) by default, for K up to $\frac{m^2+m}{2}$. This in turn implies that η in (4.14) is always strictly positive. Thus, from Theorem 4.2, it follows that $\mathbb{P}(\mathcal{E}_{\mathcal{S}^*})$ vanishes for any $\mathcal{S}^* \subseteq [n]$ for sufficiently large L.

4.6.4 The case when $n > \mathbf{Rank}(\mathbf{A} \odot \mathbf{A})$

For $n > \text{Rank}(\mathbf{A} \odot \mathbf{A})$ or $n > \frac{m^2+m}{2}$, $\mathbf{A} \odot \mathbf{A}$ becomes rank-deficient. In [31], it is argued that this leads to parameter identifiability issue in MSBL, i.e., the parameterized MSBL objective can assume a same value for different $\boldsymbol{\gamma}$. Thus, for $n > \frac{m^2+m}{2}$, the MSBL objective may foster multiple global maxima. However, in [31], the non-negativity of the hyperparameter $\boldsymbol{\gamma}$ was not taken into account. Proposition 4.3 shows that as long as $\mathbf{A} \odot \mathbf{A}$ satisfies NN-RNSP of order K (Definition 4.1), η is strictly positive and therefore it is possible to guarantee perfect K-sparse support recovery with arbitrarily high probability using only finitely many MMVs. Interestingly, the NN-RNSP condition can hold even when $\mathbf{A} \odot \mathbf{A}$ is column rank deficient. This allows cMSBL and MSBL to reconstruct K-sparse supports even when $n > \frac{m^2+m}{2}$. In fact, in the previous subsection, we have already showed that $m \geq \mathcal{O}(K \log n)$ suffices for consistent K-sparse support recovery using finitely many MMVs.[3]

4.6.5 Support Recovery from Noiseless Measurements

For $K < \text{spark}(\mathbf{A}) - 1$, it can be shown that as the noise variance $\sigma^2 \to 0$, the support error exponent $\mathcal{D}_{1/2}\left(p_{\boldsymbol{\gamma}}, p_{\boldsymbol{\gamma}^*}\right)$ in (4.10) grows in an unbounded fashion for all $\boldsymbol{\gamma} \in \Theta \backslash \Theta(\mathcal{S}^*)$. 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

$$\mathcal{D}_{1/2}\left(p_{\boldsymbol{\gamma}_1}, p_{\boldsymbol{\gamma}_2}\right) \;=\; \log \left| \frac{\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_1} + \boldsymbol{\Sigma}_{\boldsymbol{\gamma}_2}}{2} \right| - \frac{1}{2} \log \left| \boldsymbol{\Sigma}_{\boldsymbol{\gamma}_1} \boldsymbol{\Sigma}_{\boldsymbol{\gamma}_2} \right|$$

[3]Supported by the empirical performance of MSBL, we suspect that for subgaussian measurement matrix $\mathbf{A}$, $\mathbf{A} \odot \mathbf{A}$ satisfies NN-RNSP of order K even when $m = \mathcal{O}(\sqrt{K})$ for fixed n.

$$= \log \left| \frac{\mathbf{H}^{\frac{1}{2}} + \mathbf{H}^{-\frac{1}{2}}}{2} \right|. \tag{4.27}$$

where $\mathbf{H} \triangleq \mathbf{\Sigma}_{\boldsymbol{\gamma}_1}^{\frac{1}{2}} \mathbf{\Sigma}_{\boldsymbol{\gamma}_2}^{-1} \mathbf{\Sigma}_{\boldsymbol{\gamma}_1}^{\frac{1}{2}}$ is referred to as the *discrimination matrix*. Since $\mathbf{H}$ is a normal matrix, it is unitarily diagonalizable. Let $\mathbf{H} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_m)$ with λ_i's being the strictly positive eigenvalues of $\mathbf{H}$, 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

$$\begin{aligned} \mathcal{D}_{1/2}\left(p_{\boldsymbol{\gamma}_1}, p_{\boldsymbol{\gamma}_2}\right) &= \sum_{i=1}^{m} \log\left(\left(\lambda_i^{1/2} + \lambda_i^{-1/2}\right)/2\right) \\ &\geq \log\left(\frac{1}{2}\left(\left(\lambda_{\max}(\mathbf{H})\right)^{1/2} + \left(\lambda_{\max}(\mathbf{H})\right)^{-1/2}\right)\right). \end{aligned} \tag{4.28}$$

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 4.5 below relates $\lambda_{\max}(\mathbf{H})$ to the noise variance σ^2.

Proposition 4.5. *If $K <$ spark($\mathbf{A}$)-1, then for any K or lesser sparse $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2 \in \mathbb{R}_+^n$ such that supp($\boldsymbol{\gamma}_1$)$\backslash$supp($\boldsymbol{\gamma}_2$)$\neq \phi$, the maximum eigenvalue of $\mathbf{H} \triangleq \mathbf{\Sigma}_{\boldsymbol{\gamma}_1}^{\frac{1}{2}} \mathbf{\Sigma}_{\boldsymbol{\gamma}_2}^{-1} \mathbf{\Sigma}_{\boldsymbol{\gamma}_1}^{\frac{1}{2}}$ satisfies*

$$\lambda_{max}\left(\mathbf{H}\right) \geq \frac{c_1}{\sigma^2} \tag{4.29}$$

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

Proof. See Appendix F.9. □

According to Proposition 4.5, in the limit $\sigma^2 \to 0$, $\lambda_{\max}(\mathbf{H}) \to \infty$, and consequently, $\mathcal{D}_{1/2}\left(p_{\boldsymbol{\gamma}_1}, p_{\boldsymbol{\gamma}_2}\right)$ grows unbounded (due to (4.28)) whenever supp($\boldsymbol{\gamma}_1$) $\neq$ supp($\boldsymbol{\gamma}_2$) and $K <$ spark($\mathbf{A}$) $- 1$. Based on this observation, we now state Theorem 4.5 which lays

forward the sufficient conditions for exact support recovery in the noiseless case.

Theorem 4.5. *Consider the noiseless MMV problem, with observations* $\mathbf{Y} = \mathbf{AX}$ *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* K *or lesser sparse solution of the MSBL optimization problem in* (2.2)*, then* $supp(\hat{\boldsymbol{\gamma}}) = \mathcal{S}^*$ *almost surely, provided that* $K < spark(\mathbf{A}) - 1$*. This result holds even in the SMV case, i.e., when* $L = 1$*.*

Proof. Under assumption **A1**, there exists a $\boldsymbol{\gamma}^* \in \Theta$ such that every column in $\mathbf{X}$ is i.i.d. $\mathcal{N}(0, \mathrm{diag}(\boldsymbol{\gamma}^*))$, and $\mathrm{supp}(\boldsymbol{\gamma}^*) = \mathcal{S}^*$. 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 $\hat{\boldsymbol{\gamma}} \neq \boldsymbol{\gamma}^*$. Moreover, the following chain of implications holds.

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

By applying Corollary 4.1, this further implies that

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

By using the lower bound (4.28) for $\mathcal{D}_{1/2}(p_{\hat{\boldsymbol{\gamma}}}, p_{\boldsymbol{\gamma}^*})$, we have

$$\mathbb{P}(\mathrm{supp}(\hat{\boldsymbol{\gamma}}) \neq \mathcal{S}^*) \leq \left[\frac{1}{2}\left(\sqrt{\lambda_{\max}(\mathbf{H})} + \frac{1}{\sqrt{\lambda_{\max}(\mathbf{H})}}\right)\right]^{-\frac{L}{4}}, \tag{4.30}$$

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 4.5, $\sigma^2 \to 0$ results in $\lambda_{\max}(\mathbf{H}) \to \infty$ which in turn drives the RHS in (4.30) to zero for $L \geq 1$. □

From Theorem 4.5, we conclude that, *in the noiseless scenario and for* $\mathbf{X}$ *satisfying assumption* ***A1****, MSBL requires only a single measurement vector (*$L = 1$*) to perfectly recover any* $K <$ *spark*$(\mathbf{A}) - 1$ *sized support.* If the measurement matrix $\mathbf{A}$ contains independent entries drawn from a continuous probability distribution, then $\mathbf{A}$ has full spark, i.e. $\text{spark}(\mathbf{A}) = m + 1$ almost surely. When $\mathbf{A}$ has full spark, both SBL and MSBL can recover $m-1$ or lesser sparse supports exactly from m noiseless measurements per joint sparse vector. This result is in line with the sufficient conditions identified for successful support recovery by MSBL in [32, Theorem 1]. However, unlike in [32], the nonzero rows of $\mathbf{X}$ need not be orthogonal. Also, our result improves over the $k \leq m/2$ condition shown in [81].

4.6.6 Impact of Measurement Noise on Sufficient MMVs

For $\sigma^2 > 0$, the error exponent term $\mathcal{D}_{\mathcal{S}}^*$ in (4.11) is always bounded from above. This implies that unlike in the noiseless case, a single MMV is no longer sufficient, and multiple MMVs are needed to drive the error probability to zero.

A close inspection of the abstract MMV bound in (4.16) reveals that the noise variance influences the error probability in a twofold manner: (i) through η, and (ii) through the size of the ϵ-net or κ_{cov}. As σ^2 increases, η decreases polynomially (see (4.20)) while κ_{cov} increases at most logarithmically (see (4.21)). The overall effect is captured by condition **C2** in Theorem 4.4, which suggests that if the noise variance is very high relative to

$\gamma_{\max}$, it is sufficient to have a roughly quadratically larger number of MMVs to guarantee the desired probability of error. As the noise variance approaches zero, the MMV bound in **C2** loosens and is not informative.

4.7 Chapter Summary

In this chapter, we have derived new improved sufficient conditions for vanishing support error probability in the MMV problem for MSBL based support reconstruction. The sufficient conditions are specified in terms of the number of MMVs and properties of the measurement matrix. The new conditions cater to a wider, more useful class of Gaussian signals, and dispenses with the restrictive row-orthogonality condition on the signal matrix required in the previous results [32, 43].

Chapter 5

Rényi Divergence Based Sparse Support Recovery

In which Rényi divergence meets submodularity & covariance matching becomes faster...

5.1 Introduction

In this chapter[1], we propose a novel covariance matching based algorithm to recover the common nonzero support of multiple joint sparse signal vectors from their noisy linear compressive measurements. In the proposed algorithm, the true support of the unknown joint sparse vectors is recovered by minimizing a novel Rényi divergence cost function which is inspired from MSBL's type-II log likelihood objective. Interestingly, the Rényi divergence objective is expressible as a difference of two submodular set functions with the unknown support as a set variable. This allows us to efficiently optimize the Rényi

[1] This chapter is based on S. Khanna and C. R. Murthy, "Rényi Divergence Based Covariance Matching Pursuit of Joint Sparse Support", *Proc. IEEE 18th Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Sapporo, Japan, Jul. 2017

divergence objective via an iterative majorization-minimization (MM) procedure to recover the sparse signal support. The resulting support recovery scheme is called *Rényi Divergence based Covariance Matching Pursuit* or RD-CMP. Compared to existing covariance matching based support recovery methods, RD-CMP is empirically shown to have lower computational complexity as well as a lower rate of increase of complexity with the ambient signal dimension.

5.2 Issues with Existing Support Recovery Algorithms

Extensive effort have been spent in the recent years towards developing fast MMV solvers for the joint sparse signal/support recovery problem. Popular techniques include greedy strategies for support reconstruction [24, 38], regularization of the solution space using convex and non convex penalties [21,84], subspace methods based on the MuSiC criterion [28, 29], and Bayesian maximum a posterior probability (MAP) inference [32, 34]. Among these techniques, the greedy methods are typically computationally the fastest ones while the Bayesian MAP inference based methods have the state of the art performance.

A major limitation of most of the existing JSSR methods is that they implicitly assume the number of measurements acquired per joint sparse vector to be more than the size of the common support. In the previous chapters, we provided a preliminary justification of how it is possible to recover any k-sized support from fewer than k measurement per MMV using the covariance matching approach. Notable examples of covariance matching based MMV algorithms are MSBL [32] and Co-LASSO [30], both of which enforce a common Gaussian prior with a parameterized diagonal covariance matrix on the unknown signal vectors for the purpose of inducing a joint sparse solution. The diagonal nature of the

covariance matrix encapsulates the latent uncorrelatedness of the nonzero coefficients of the joint sparse vectors [30], while the diagonal entries themselves are representative of their common support. The parameterized covariance matrix can be learned directly from the measurements via the covariance matching technique which seeks to robustly minimize a certain distance between the empirical and the parameterized covariance matrices of the measurements. In MSBL, the covariance parameters are estimated via a type-II likelihood maximization procedure which can be interpreted as minimizing the Log-Det Bregman divergence between the empirical and the parameterized measurement covariance matrices. Likewise, Co-LASSO recovers the diagonal covariance parameters as the minimum ℓ_1 norm solution of the covariance matching constraints. Despite their superior support recovery performance compared to conventional MMV solvers, both MSBL and Co-LASSO suffer from high computational complexity, and hence are not suitable for large dimensional signals.

5.3 Mathematical Preliminaries

In this section, we present some preliminary concepts and results which will be used in the discussion later.

5.3.1 α-Rényi Divergence

Let $(\mathcal{X}, \mathcal{F})$ be a measurable space and P and Q be two probability measures on 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 by

$\mathcal{D}_\alpha\left(p||q\right)$ is defined as

$$\mathcal{D}_\alpha(p||q) \triangleq \frac{1}{(1-\alpha)} \log \int_{\mathcal{X}} p(x)^\alpha q(x)^{1-\alpha} \mu(dx). \tag{5.1}$$

For $\mathcal{D}_\alpha(p||q)$ finite, it is a nondecreasing function of α. For $\alpha \in [0,1)$, $\mathcal{D}_\alpha(p||q) < \mathcal{D}_{\mathrm{KL}}(p||q)$, with $\lim_{\alpha\to 1} \mathcal{D}_\alpha(p||q) = \mathcal{D}_{\mathrm{KL}}(p||q)$, [85], where $\mathcal{D}_{\mathrm{KL}}$ is the Kullback-Leibler divergence.

When $p = \mathcal{N}(0, \mathbf{\Sigma}_1)$ and $q = \mathcal{N}(0, \mathbf{\Sigma}_2)$, the α-Rényi divergence admits the following closed form expression [86],

$$\mathcal{D}_\alpha(p||q) = \frac{1}{2(1-\alpha)} \log \frac{|(1-\alpha)\mathbf{\Sigma}_1 + \alpha\mathbf{\Sigma}_2|}{|\mathbf{\Sigma}_1|^{1-\alpha} |\mathbf{\Sigma}_2|^{\alpha}}. \tag{5.2}$$

where $|\ .\ |$ denotes the determinant of the input matrix.

5.3.2 Submodular Set Functions and Related Properties

Let $f : 2^{\mathcal{V}} \to \mathbb{R}$ be a set function defined over the subsets of a ground set $\mathcal{V}$. Then f is *monotone* if and only if

$$\forall \mathcal{S} \subseteq \mathcal{T} \subseteq \mathcal{V}, \quad f(\mathcal{S}) \leq f(\mathcal{T}).$$

The set function $f : 2^{\mathcal{V}} \to \mathbb{R}$ is *submodular* if for every $\mathcal{A}, \mathcal{B} \subseteq \mathcal{V}$ satisfying $\mathcal{A} \subseteq \mathcal{B}$ and any $x \in \mathcal{V} \backslash \mathcal{B}$, it is true that

$$f(\mathcal{B} \cup \{x\}) - f(\mathcal{B}) \leq f(\mathcal{A} \cup \{x\}) - f(\mathcal{A}). \tag{5.3}$$

In other words, f satisfies the *law of diminishing returns* property. If f is submodular, then $-f$ is supermodular which implies that it satisfies the inequality (5.3) in the reverse direction. If (5.3) holds with equality, then f is called modular.

The following theorem by Nemhauser et al. discusses a recipe for approximate maximization of a submodular function subject to cardinality constraints.

Theorem 5.1. *Let f be a monotone, submodular and non-negative function on $\mathcal{V}$. For a fixed K, let*

$$\mathcal{S}^* \in \underset{\mathcal{S}\subseteq\mathcal{V},|\mathcal{S}|\leq K}{max} f(\mathcal{S}). \tag{5.4}$$

Then, an iterative greedy algorithm which starts with an empty solution set $\mathcal{S} = \phi$ and adds a new element $x \in \mathcal{V}\backslash\mathcal{S}$ to the current solution set $\mathcal{S}$ which results in maximal increment $f(\mathcal{S}\cup\{x\}) - f(\mathcal{S})$ in every iteration is a $(1-1/e)$-approximation algorithm for (5.4), i.e., after K iterations,

$$f(\mathcal{S}) \geq (1-1/e)\, f(\mathcal{S}^*), \quad \forall \mathcal{S} \subseteq \mathcal{V}, |\mathcal{S}| \leq K. \tag{5.5}$$

Proof. See [87]. □

Proposition 5.1. *For any positive definite matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, a generic $n\times p$ matrix $\mathbf{B}$ and constant $\beta > 0$, the set function $f(\mathcal{S}) = \log|\mathbf{A} + \beta\mathbf{B}_{\mathcal{S}}\mathbf{B}_{\mathcal{S}}^T|$ is monotone and submodular.*

Proof. Given in Appendix G.1. □

5.4 System and Source Models

5.4.1 System Model

In this chapter, we consider the joint sparse support recovery problem corresponding to the linear MMV measurement model: $\mathbf{Y} = \mathbf{AX} + \mathbf{W}$, with exactly the same definitions and assumptions as in section 4.3.1. Recall that $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L]$ is the $n \times L$ sized row-sparse matrix of interest here, and $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_L]$ and $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_L]$ are

the $m \times L$ sized observation and noise matrices, respectively. The goal here is to recover the true row support of $\mathbf{X}$ given the knowledge of $\mathbf{Y}$, $\mathbf{A}$ and σ^2.

5.4.2 Signal Prior Design

For the sake of algorithm development, we assume that the joint sparse columns of $\mathbf{X}$ are drawn independently from a common parameterized Gaussian distribution, i.e.,

$$p(\mathbf{x}_j) \sim \mathcal{N}(0, \gamma \mathbf{I}_{\mathcal{S}}), \quad 1 \leq j \leq L, \tag{5.6}$$

where $\mathbf{I}_{\mathcal{S}} = \text{diag}(\mathbf{1}_{\mathcal{S}})$, and $\mathbf{1}_{\mathcal{S}}$ is the binary support vector with ones at the index locations specified by set $\mathcal{S}$, and zeros elsewhere. Due to their common covariance matrix, all $\mathbf{x}_j$'s share the same support denoted by $\mathcal{S}$. The scalar variable γ models the common variance of the nonzero signal coefficients in $\mathbf{X}$. Thus, the support set $\mathcal{S}$ and the common variance γ together constitute the entire hyperparameters which parameterize the signal prior. The conditions for exact support recovery under this particular signal prior (and $\gamma = 1$) has been studied earlier in [81].

The Gaussian signal prior considered here is a special case of the MSBL signal prior [32]. Unlike in MSBL where each row of $\mathbf{X}$ is assigned a distinct variance parameter, here we assume that all nonzero elements in $\mathbf{X}$ share a common variance. In the following, we discuss how to learn the support set $\mathcal{S}$ directly from the observations $\mathbf{Y}$. The case of unknown variance hyperparameter γ is also discussed separately.

5.4.3 Type-II Maximum Likelihood Estimation of Support

Given the support set hyperparameter $\mathcal{S}$, the marginal probability density of the observations $\mathbf{Y}$ is given by

$$p(\mathbf{Y};\mathcal{S}) = \prod_{j=1}^{L} \mathcal{N}\left(\mathbf{y}_j; 0, \mathbf{\Sigma}(\mathcal{S})\right), \tag{5.7}$$

where $\mathbf{\Sigma}(\mathcal{S}) \triangleq \sigma^2\mathbf{I}_m + \gamma\mathbf{A}_{\mathcal{S}}\mathbf{A}_{\mathcal{S}}^T$ is the common measurement covariance matrix parameterized by set $\mathcal{S}$. In the Bayesian inference approach, the set hyperparameter $\mathcal{S}$ with maximal Bayesian evidence $p(\mathbf{Y};\mathcal{S})$ is selected to be the final estimate of the true support and is used for computing the maximum a posteriori (MAP) estimate of the unknown joint sparse columns of $\mathbf{X}$. Let $\hat{\mathcal{S}}_{\text{ML}}$ denote the set hyperparameter with the highest Bayesian evidence, i.e.,

$$\hat{\mathcal{S}}_{\text{ML}} = \underset{\mathcal{S}\subseteq[n]}{\arg\max}\ \log p(\mathbf{Y};\mathcal{S}). \tag{5.8}$$

The above hyperparameter selection technique is called the *type-II likelihood maximization* where rather than maximizing the joint likelihood $p(\mathbf{Y},\mathbf{X};\mathcal{S})$, we instead maximize the marginalized likelihood $p(\mathbf{Y};\mathcal{S})$. By using the expression (5.7) in (5.8) and dropping all terms independent of $\mathcal{S}$, we obtain

$$\hat{\mathcal{S}}_{\text{ML}} = \underset{\mathcal{S}\subseteq[n]}{\arg\min}\ \log\left|\mathbf{\Sigma}(\mathcal{S})\right| + \text{tr}\left(\mathbf{\Sigma}(\mathcal{S})^{-1}\hat{\mathbf{R}}_{\mathbf{yy}}\right), \tag{5.9}$$

where $\hat{\mathbf{R}}_{\mathbf{yy}} = \frac{1}{L}\mathbf{Y}\mathbf{Y}^T$ denotes the sample covariance matrix of the measurements. The above optimization is an NP-hard problem which requires a combinatorial search over all possible combinations of the support set $\mathcal{S}$ in order to find the set that maximizes the log-likelihood cost. In the following section, we present an alternative to the above log-likelihood cost which can be approximately optimized in polynomial time.

5.5 Rényi Divergence based Information Projection for Joint Sparse Support Recovery

Consider the *reverse information projection* problem:

$$\hat{\mathcal{S}} = \underset{\mathcal{S}\subseteq[n]}{\arg\min}\ \mathcal{D}_\alpha(\tilde{p}(\mathbf{Y}), p(\mathbf{Y};\mathcal{S})), \tag{5.10}$$

where $\hat{\mathcal{S}}$ corresponds to the reverse information projection of $\tilde{p}(\mathbf{Y}) = \prod_{j=1}^{L}\mathcal{N}(\mathbf{y}_j; 0, \hat{\mathbf{R}}_{\mathbf{yy}})$ onto the set of parameterized probability densities $p(\mathbf{Y};\mathcal{S}) = \prod_{j=1}^{L}\mathcal{N}(\mathbf{y}_j; 0, \sigma^2\mathbf{I}_m + \gamma\mathbf{A}_{\mathcal{S}}\mathbf{A}_{\mathcal{S}}^T)$. The above information projection is defined with respect to the α-Rényi divergence $D_\alpha(\tilde{p}, p)$ which is evaluated as per (5.1). The α-Rényi divergence based reverse-I projection problem generalizes the type-II likelihood maximization in the sense that as $\alpha \to 1$, the α-Rényi divergence converges to the Kullback-Leibler (KL) divergence [85], which is same as the negative of the type-II log-likelihood objective in (5.9).

By using the closed form (5.2) of the α-Rényi divergence between two zero mean multivariate Gaussian distributions, and dropping terms independent of $\mathcal{S}$, the reverse I-projection in (5.10) simplifies to

$$\hat{\mathcal{S}} = \underset{\mathcal{S}\subseteq[n]}{\arg\min}\ \log\left|(1-\alpha)\hat{\mathbf{R}}_{\mathbf{yy}} + \alpha\left(\sigma^2\mathbf{I}_m + \boldsymbol{\gamma}\mathbf{A}_{\mathcal{S}}\mathbf{A}_{\mathcal{S}}^T\right)\right| - \alpha\log\left|\sigma^2\mathbf{I}_m + \boldsymbol{\gamma}\mathbf{A}_{\mathcal{S}}\mathbf{A}_{\mathcal{S}}^T\right|. \tag{5.11}$$

Similar to the type-II ML optimization (5.9), the above reverse I-projection is also an NP hard problem and involves searching over all possible subsets of $[n]$. However, by using Proposition 5.1, the Rényi divergence objective in (5.11) can be interpreted as a difference of two submodular set functions with respect to the set variable $\mathcal{S}$. This interesting observation facilitates the abstraction of the reverse I-projection into the following canonical

set optimization problem:

$$\hat{\mathcal{S}} = \arg\min_{\mathcal{S}\subseteq[n]} f(\mathcal{S}) - g(\mathcal{S}), \tag{5.12}$$

where $f(\mathcal{S}) = \log\left|(1-\alpha)\hat{\mathbf{R}}_{\mathbf{yy}} + \alpha\left(\mathbf{\Sigma}(\mathcal{S})\right)\right|$ and $g(\mathcal{S}) = \alpha \log|\mathbf{\Sigma}(\mathcal{S})|$ are submodular set functions with respect to the set variable $\mathcal{S}$. Based on this decomposition, we now propose a polynomial time algorithm to find a support set $\hat{\mathcal{S}}$ which approximately minimizes the difference objective.

5.5.1 Minimizing Difference of Two Submodular Functions

In [88, 89], the Supermodular-Submodular (Sup-Sub) procedure is proposed as a polynomial time algorithm for approximate minimization of a set function expressible as a difference of two submodular set functions. Given a ground set $\mathcal{V}$, suppose $f : 2^{\mathcal{V}} \to \mathbb{R}$ and $g : 2^{\mathcal{V}} \to \mathbb{R}$ are two submodular set functions. Consider the following canonical problem:

$$\mathcal{S}^* = \arg\min_{\mathcal{S}\subseteq\mathcal{V}} f(\mathcal{S}) - g(\mathcal{S}), \tag{5.13}$$

where $\mathcal{S}$ is the optimization variable. In the Sup-Sub method, every iteration entails minimizing a supermodular upper bound $q(\mathcal{S})$ of the objective function in (5.13). If the upper bound $q(\mathcal{S})$ is tight at $\mathcal{S}_{k-1}$, the previous iterate value, then it follows that

$$f(\mathcal{S}_{k-1}) - g(\mathcal{S}_{k-1}) = q(\mathcal{S}_{k-1}) \geq q(\mathcal{S}_k) \geq f(\mathcal{S}_k) - g(\mathcal{S}_k), \tag{5.14}$$

where $\mathcal{S}_k$ is the minimizer of $q(\mathcal{S})$ at the k^{th} iteration. Thus, the objective decreases monotonically in every iteration, a hallmark of any majorization-minimization technique.

In the SupSub procedure, the upper bound for the difference objective in (5.13) is obtained by replacing the submodular $f(\mathcal{S})$ with its tight modular upper bound. For any

submodular f, there exists a tight modular upper bound $h^f_{\mathcal{X}}$ at point $\mathcal{X}$ [87, 88] which is defined as follows:

$$f(\mathcal{S}) \;\leq\; h^f_{\mathcal{X}}(\mathcal{S}) \;\triangleq\; f(\mathcal{X}) - \sum_{j\in\mathcal{X}\setminus\mathcal{S}} f(j|\mathcal{X}\setminus\{j\}) + \sum_{j\in\mathcal{S}\setminus\mathcal{X}} f(j|\phi), \tag{5.15}$$

where $f(\mathcal{A}|\mathcal{B})$ is evaluated as $f(\mathcal{A}\cup\mathcal{B}) - f(\mathcal{B})$. The above bound is tight at $\mathcal{X}$, i.e., $h^f_{\mathcal{X}}(\mathcal{X}) = f(\mathcal{X})$ and $h^f_{\mathcal{X}}(\mathcal{S}) \geq f(\mathcal{S})$ for any $\mathcal{S}\neq\mathcal{X}, \mathcal{S}\subseteq\mathcal{V}$.

Finally, in the k^{th} iteration of the Sup-Sub procedure, by invoking Theorem 5.1, the support set estimate $\mathcal{S}_k$ is updated as the approximate minimizer of the majorized supermodular objective $q(\mathcal{S}) = h^f_{\mathcal{S}_{k-1}}(\mathcal{S}) - g(\mathcal{S})$ obtained via greedy search.

5.5.2 RD-CMP Algorithm

Since the Rényi divergence cost in (5.11) is expressible as a difference of two submodular set functions, we can apply the Sup-Sub procedure just described to find the support set $\hat{\mathcal{S}}$ which minimizes it. We call this support set recovery scheme the *Rényi Divergence based Covariance Matching Pursuit (RD-CMP)*. The k^{th} iteration of RD-CMP comprises two steps:

1. *Majorization:* Construct a modular upper bound $h^f_{\mathcal{S}_{k-1}}(\mathcal{S})$ for the Rényi divergence objective's first term $f(\mathcal{S}) = \log\left|(1-\alpha)\hat{\mathbf{R}}_{\mathbf{yy}} + \alpha\,(\boldsymbol{\Sigma}(\mathcal{S}))\right|$ which satisfies $h^f_{\mathcal{S}_{k-1}}(\mathcal{S}) \geq f(\mathcal{S})$ for all $\mathcal{S}\subseteq[n]$, with equality at $\mathcal{S}_{k-1}$. Here, $\mathcal{S}_k$ denotes the estimated row support of $\mathbf{X}$ at the end of the k^{th} iteration of the RD-CMP algorithm.

2. *Minimization:* Update the support estimate according to:

$$\mathcal{S}_k = \underset{\mathcal{S}\subseteq[n]}{\arg\min}\; h^f_{\mathcal{S}_{k-1}}(\mathcal{S}) - \alpha\log\left|\sigma^2\mathbf{I}_m + \boldsymbol{\gamma}\mathbf{A}_{\mathcal{S}}\mathbf{A}_{\mathcal{S}}^T\right|.$$

By executing steps 1 and 2 repeatedly, the above iterations converge to a local minimizer of the Rényi divergence cost in (5.11). We now flesh out the majorization and minimization steps of RD-CMP in a greater detail.

Majorization Step

Let k denote the current iteration index. In the majorization step, the Rényi divergence objective in (5.11) is majorized by replacing the submodular component $f(\mathcal{S}) = \log\left|(1-\alpha)\hat{\mathbf{R}}_{\mathbf{yy}} + \alpha\left(\sigma^2\mathbf{I}_m + \boldsymbol{\gamma}\mathbf{A}_{\mathcal{S}}\mathbf{A}_{\mathcal{S}}^T\right)\right|$ with its modular upper bound which is tight at $\mathcal{S}_{k-1}$. Using (5.15), we construct a modular upper bound of f, which after dropping irrelevant constant terms has the following simplified form.

$$h^f_{\mathcal{S}_{k-1}}(\mathcal{S}) \triangleq \sum_{j\in\mathcal{S}_{k-1}\setminus\mathcal{S}} \log\left|1-\alpha\rho\mathbf{a}_j^T\mathbf{T}_{k-1}^{-1}\mathbf{a}_j\right| + \sum_{j\in\mathcal{S}\setminus\mathcal{S}_{k-1}} \log\left|1+\alpha\rho\mathbf{a}_j^T\mathbf{H}^{-1}\mathbf{a}_j\right|, \tag{5.16}$$

where $\rho = \frac{\gamma}{\sigma^2}$, $\mathbf{H} = \left(\alpha\mathbf{I}_m + (1-\alpha)\sigma^{-2}\hat{\mathbf{R}}_{\mathbf{yy}}\right)$ and $\mathbf{T}_k = \left(\mathbf{H} + \alpha\gamma\sigma^{-2}\mathbf{A}_{\mathcal{S}_k}\mathbf{A}_{\mathcal{S}_k}^T\right)$. The intermediate steps involved in the derivation of the final form of modular $h^f_{\mathcal{S}_{k-1}}$ in (5.16) are detailed in appendix G.2. By construction, for any $\mathcal{S}\subseteq[n]$, we have $h^f_{\mathcal{S}_{k-1}}(\mathcal{S}) \geq f(\mathcal{S})$ with equality at $\mathcal{S} = \mathcal{S}_{k-1}$.

Minimization Step

In this step, the support set estimate is updated to be the minimizer of the majorized Rényi divergence objective or equivalently the maximizer of the negative objective as shown below.

$$\mathcal{S}_k = \arg\max_{\mathcal{S}\subseteq[n]} g(\mathcal{S}) - h^f_{\mathcal{S}_{k-1}}(\mathcal{S}), \tag{5.17}$$

where $g(\mathcal{S}) = \alpha \log \left|\sigma^2 \mathbf{I}_m + \boldsymbol{\gamma} \mathbf{A}_{\mathcal{S}} \mathbf{A}_{\mathcal{S}}^T\right|$. Since the first and second terms of the objective in (5.17) are submodular and modular, respectively, the overall objective is submodular in $\mathcal{S}$. Consequently, its approximate maximizer can be constructed in a greedy manner by invoking Theorem 5.1.

The greedy construction of the solution of (5.17) begins with an empty set $\mathcal{S}_0$. At iteration t of the greedy procedure, a new index r^t is added to the support estimate $\mathcal{S}_{t-1}$ such that

$$r^t = \underset{r \in [n] \setminus \mathcal{S}_{t-1}}{\arg\max}\ g(\mathcal{S}_{t-1} \cup \{r\}) - h^f_{\mathcal{S}_{k-1}}(\mathcal{S}_{t-1} \cup \{r\}). \tag{5.18}$$

Here, k denotes the iteration index of the outer majorization-minimization loop of RD-CMP. The greedy procedure is terminated when the newly added column $\mathbf{a}_{r^t}$ of $\mathbf{A}$ ceases to increase the likelihood of observations $\mathbf{Y}$, i.e., $p(\mathbf{Y}; \mathcal{S}_{t-1} \bigcup \{\mathbf{a}_{r^t}\}) < p(\mathbf{Y}; \mathcal{S}_{t-1})$, which is formalized as

$$\log\left(1 + \gamma \mathbf{a}_{r^t}^T \mathbf{P}_{\mathcal{S}_{t-1}}^{-1} \mathbf{a}_{r^t}^T\right) > \frac{\mathbf{a}_{r^t}^T \mathbf{P}_{\mathcal{S}_{t-1}}^{-1} \mathbf{R}_{\mathbf{Y}} \mathbf{P}_{\mathcal{S}_{t-1}}^{-1} \mathbf{a}_{r^t}^T}{\gamma^{-1} + \mathbf{a}_{r^t}^T \mathbf{P}_{\mathcal{S}_{t-1}}^{-1} \mathbf{a}_{r^t}^T}, \tag{5.19}$$

where $\mathbf{P}_{\mathcal{S}_{t-1}} = (\sigma^2 \mathbf{I}_m + \gamma \mathbf{A}_{\mathcal{S}_{t-1}} \mathbf{A}_{\mathcal{S}_{t-1}}^T)$. The objective in (5.18) can be computed efficiently using rank one recursive updates of the log-determinant function, resulting in very low overall computational complexity of the proposed algorithm.

5.5.3 Unknown signal and noise variance

If the common variance $\boldsymbol{\gamma}$ of the nonzero elements in $\mathbf{X}$ and the noise variance σ^2 are unknown, they can be estimated as additional hyperparameters along with the support $\mathcal{S}$ directly from the observations. Starting with suitably chosen initial values $\boldsymbol{\gamma}_0$ and σ_0^2, the hyperparameters $\boldsymbol{\gamma}$, σ^2 are updated using the fixed point iterative updates below, which are derived from the zero gradient optimality conditions associated with their respective

likelihood maximization,

$$\gamma_k = \frac{\text{tr}\left(\mathbf{A}_{\mathcal{S}_k}\mathbf{Q}_{\mathcal{S}_k}(\mathbf{R_Y} - \sigma^2_{k-1}\mathbf{I}_m)\mathbf{Q}^T_{\mathcal{S}_k}\mathbf{A}^T_{\mathcal{S}_k}\right)}{\left|\left|\mathbf{A}^T_{\mathcal{S}_k}\mathbf{Q}_{\mathcal{S}_k}\mathbf{A}_{\mathcal{S}_k}\right|\right|^2_F}, \tag{5.20}$$

$$\sigma^2_k = \frac{\text{tr}\left(\mathbf{A}_{\mathcal{S}_k}\mathbf{Q}_{\mathcal{S}_k}(\mathbf{R_Y} - \boldsymbol{\gamma}_k\mathbf{A}_{\mathcal{S}_k}\mathbf{A}^T_{\mathcal{S}_k})\mathbf{Q}^T_{\mathcal{S}_k}\mathbf{A}^T_{\mathcal{S}_k}\right)}{||\mathbf{Q}_{\mathcal{S}_k}||^2_F} \tag{5.21}$$

where $\mathbf{Q}_{\mathcal{S}_k} = \mathbf{A}^T_{\mathcal{S}_k}\left(\sigma^2_{k-1}\mathbf{I}_m + \gamma_{k-1}\mathbf{A}_{\mathcal{S}_k}\mathbf{A}^T_{\mathcal{S}_k}\right)^{-1}$. Finally, Algorithm 1 summarizes the stepwise implementation of RD-CMP.

Algorithm 1 Rényi Divergence based Covariance Matching Pursuit (RD-CMP)

Input: $\mathbf{Y}$, $\mathbf{A}$, σ^2 and α

Notation: $f(\mathcal{S},\gamma) = \log\left|(1-\alpha)\mathbf{R_Y} + \alpha\left(\sigma^2\mathbf{I}_m + \gamma\mathbf{A}_{\mathcal{S}}\mathbf{A}^T_{\mathcal{S}}\right)\right|$,
$g(\mathcal{S},\gamma) = \alpha\log\left|\sigma^2\mathbf{I}_m + \gamma\mathbf{A}_{\mathcal{S}}\mathbf{A}^T_{\mathcal{S}}\right|$

Initialization: $k \leftarrow 0$, $\mathcal{S}_k = \{\phi\}$, $\gamma_k = \frac{\frac{1}{L}\text{tr}(\mathbf{Y}\mathbf{Y}^T) - m\sigma^2}{||\mathbf{A}||^2_F}$

1: **while** not converged (i.e., $\mathcal{S}_k \neq \mathcal{S}_{k-1}$) **do**
2: Compute the modular bound $h^f_{\mathcal{S}_{k-1}}$ according to (5.16).
3: Initialize greedy search to find the approximate maximizer of $g(\mathcal{X},\gamma^{k-1}) - h^f_{\mathcal{S}_{k-1}}(\mathcal{X})$: $t \leftarrow 0, \mathcal{X}_t = \{\phi\}$
4: **repeat**
5: $r^t = \underset{r\in[n]\setminus\mathcal{X}_{t-1}}{\arg\max}\; g(\mathcal{X}_{t-1}\cup\{r\},\gamma^{k-1}) - h^f_{\mathcal{S}_{k-1}}(\mathcal{X}_{t-1}\cup\{r\})$
6: $\mathcal{X}_t = \mathcal{X}_{t-1}\cup\{r^t\}$
7: $t \leftarrow t+1$.
8: **until** stopping criterion in (5.19) is met.
9: Update γ_k according to (5.20).
10: $k \leftarrow k+1$
11: **end while**

Output: $\hat{\mathcal{S}} = \mathcal{S}_k$.

The performance of RD-CMP is sensitive to the choice of the parameter α which defines the order of the Rényi divergence used in the information projection step. While a theoretical analysis of α's effect on the support recovery performance is warranted, we conjecture that α controls how much importance is assigned to the observations $\mathbf{Y}$ while learning the support hyperparameters. At lower SNRs, it is preferable to choose a higher value of α and vice versa. In our experiments, $\alpha = \sigma^2$ was found to perform uniformly

well over a wide range of SNR and signal dimensions, and it is also the value used in all of our experiments.

5.5.4 Fast Recursive RD-CMP Updates

By exploiting the rank-one update property of the log det function, one can compute the submodular objective $g(\mathcal{S}_{t-1}\cup\{r\})-h^f_{\mathcal{S}_{k-1}}(\mathcal{S}_{t-1}\cup\{r\})$ in (5.18) in a fast recursive manner as discussed below.

We note that the first objective term $g(\mathcal{S}_{t-1}\cup\{r\})$ decomposes as

$$\begin{aligned}
&g(\mathcal{S}_{t-1}\cup\{r\})\\
&=\alpha\log\left|\sigma^2\mathbf{I}_m+\gamma\mathbf{A}_{\mathcal{S}_{t-1}\cup\{r\}}\mathbf{A}^T_{\mathcal{S}_{t-1}\cup\{r\}}\right|\\
&=\alpha\log\left|\sigma^2\mathbf{I}_m+\gamma\mathbf{A}_{\mathcal{S}_{t-1}}\mathbf{A}^T_{\mathcal{S}_{t-1}}+\gamma\mathbf{a}_r\mathbf{a}_r^T\right|\\
&=\alpha\log\left|\sigma^2\mathbf{I}_m+\gamma\mathbf{A}_{\mathcal{S}_{t-1}}\mathbf{A}^T_{\mathcal{S}_{t-1}}\right|+\alpha\log\left|\mathbf{I}_m+\gamma\mathbf{a}_r\mathbf{a}_r^T(\sigma^2\mathbf{I}_m+\gamma\mathbf{A}_{\mathcal{S}_{t-1}}\mathbf{A}^T_{\mathcal{S}_{t-1}})^{-1}\right|\\
&=\alpha\log\left|\sigma^2\mathbf{I}_m+\gamma\mathbf{A}_{\mathcal{S}_{t-1}}\mathbf{A}^T_{\mathcal{S}_{t-1}}\right|+\alpha\log\left|1+\gamma\mathbf{a}_r^T(\sigma^2\mathbf{I}_m+\gamma\mathbf{A}_{\mathcal{S}_{t-1}}\mathbf{A}^T_{\mathcal{S}_{t-1}})^{-1}\mathbf{a}_r\right|. \qquad (5.22)
\end{aligned}$$

Since the first term in (5.22) does not depend on r, only the Rayleigh terms $R_t(r)\triangleq\mathbf{a}_r^T(\sigma^2\mathbf{I}_m+\gamma\mathbf{A}_{\mathcal{S}_{t-1}}\mathbf{A}^T_{\mathcal{S}_{t-1}})^{-1}\mathbf{a}_r$ have to be compuated for all $r\in[n]$. The Rayleigh terms $R_t(r)$ can be computed directly from $R_{t-1}(r)$ using the following recursive updates.

$$\text{Initialization:}\quad \mathbf{Z}_0=\mathbf{I}_m \text{ and } R_0(r)=||\mathbf{a}_r||_2^2\ \forall r\in[n] \qquad (5.23)$$

$$\mathbf{w}_{t-1} = \mathbf{Z}^{-1}_{t-1}\mathbf{a}_{r^{t-1}} \qquad (5.24)$$

$$\mathbf{b}_{t-1} = \mathbf{A}^T\mathbf{w}_{t-1} \qquad (5.25)$$

$$\mathbf{Z}_t = \mathbf{Z}_{t-1}-\frac{\mathbf{w}_{t-1}\mathbf{w}^T_{t-1}}{\frac{\sigma^2}{\gamma}+\mathbf{b}_{t-1}(r^{t-1})} \qquad (5.26)$$

$$R_t(r) \;=\; R_{t-1}(r) - \frac{\mathbf{b}_{t-1}(r)^2}{\frac{\sigma^2}{\gamma} + \mathbf{b}_{t-1}(r^{t-1})} \quad \forall r \in [n] \tag{5.27}$$

where r^{t-1} is the new index added to the support set estimate (Step 5 in Algorithm 1) in the $(t-1)^{\text{th}}$ iteration of the greedy search executed every RD-CMP iteration.

Likewise, from (5.15), we observe that the modular upper bound $h^f_{\mathcal{S}_{k-1}}(\mathcal{S}_{t-1} \cup \{r\})$ can be rewritten as

$$\begin{aligned} h^f_{\mathcal{S}_{k-1}}(\mathcal{S}_{t-1} \cup \{r\}) \;&=\; h^f_{\mathcal{S}_{k-1}}(\mathcal{S}_{t-1}) - \mathbb{I}_{\{r \in \mathcal{S}_{k-1}\}} \left[f(\mathcal{S}_{k-1}) - f(\mathcal{S}_{k-1} \setminus \{r\}) \right] \\ &\qquad + \mathbb{I}_{\{r \notin \mathcal{S}_{k-1}\}} \left[f(\{r\}) - f(\phi) \right] \\ &=\; h^f_{\mathcal{S}_{k-1}}(\mathcal{S}_{t-1}) + \mathbb{I}_{\{r \in \mathcal{S}_{k-1}\}} \log\left(1 - \alpha\rho \mathbf{a}_r^T \mathbf{T}_{k-1}^{-1} \mathbf{a}_r\right) \\ &\qquad + \mathbb{I}_{\{r \notin \mathcal{S}_{k-1}\}} \log\left(1 + \alpha\rho \mathbf{a}_r^T \mathbf{H}^{-1} \mathbf{a}_r\right). \end{aligned} \tag{5.28}$$

From (5.28), the outer loop of RD-CMP (indexed by k) requires computation of only the incremental terms. The quadratic terms $\mathbf{a}_r^T \mathbf{T}_{k-1}^{-1} \mathbf{a}_r$ and $\mathbf{a}_r^T \mathbf{H}^{-1} \mathbf{a}_r$ in (5.28) can be evaluated with $\mathcal{O}(m^3 + nm^2)$ computational complexity.

Table 5.1 provides a step-wise breakdown of the computational complexity incurred in a single iteration of RD-CMP. Interestingly, the per-iteration computational complexity of RD-CMP does not depend on L, which can be beneficial in situations where the number of MMVs is very large.

5.5.5 Convergence

The RD-CMP algorithm is a majorization-minimization procedure designed to minimize the Rényi divergence objective in (5.11). In the minimization step, the greedy search based optimization of the submodular objective in (5.17) is however only in the approximate

Table 5.1: Per Iteration Computational Complexity of RD-CMP

Steps in RD-CMP iteration	**Computational complexity**
Computation of $g(\mathcal{S})$	$\mathcal{O}(nm + m^2)$
Computation of $h^f_{\mathcal{S}_{k-1}}(\mathcal{S})$	$\mathcal{O}(nm^2 + m^3 + n)$
Greedy search in (5.17)	$\mathcal{O}(n)$
Stopping criterion in (5.19)	$\mathcal{O}(m^2)$
Update $\boldsymbol{\gamma}$ acc. to (5.20)	$\mathcal{O}(nm^2)$

sense. This makes it difficult to analyze the convergence of RD-CMP algorithm. However, in our experiments, the RD-CMP algorithm always converges within 10-20 iterations of its outer loop.

5.5.6 Possible Extensions

An interesting variation of the proposed RD-CMP algorithm is to use an MSBL like vector prior prescribed in (2.1) for support inference. In this case, RD-CMP reverts to the following multivariate optimization problem,

$$\hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}\in\mathbb{R}^n_+}{\arg\min}\ \log\left|(1-\alpha)\hat{\mathbf{R}}_{\mathbf{yy}} + \alpha\left(\sigma^2\mathbf{I}_m + \mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^T\right)\right| - \alpha\log\left|\sigma^2\mathbf{I}_m + \mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^T\right|, \qquad (5.29)$$

where $\boldsymbol{\Gamma} = \text{diag}(\boldsymbol{\gamma})$. Since, for any positive definite matrix $\mathbf{B}$, the function $-\log|\mathbf{B} + \mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^T|$ is a strongly convex with respect to $\boldsymbol{\gamma}$, the objective in (5.29) can be viewed as a difference of two convex functions. This allows the objective to be minimized via DC (Difference of Convex functions) programming [90]. Once $\hat{\boldsymbol{\gamma}}$ is obtained, $\text{supp}(\hat{\boldsymbol{\gamma}})$ can be declared as an estimate of the true support.

5.5.7 Distributed Implemetation

One can also envisage a distributed or parallel implementation of the RD-CMP algorithm using a network of computing nodes, where each node has access to a single measurement vector. In the distributed setup, each nodes execute an independent instantiation of the RD-CMP algorithm using a rank one sample covariance computed from the locally observed measurement vector. At the end of every RD-CMP's outer loop iteration, the nodes in the network can exchange their local support estimates with the neighboring nodes based on the underlying network topology. Each node can then refine its local support estimate by fusing the support estimates from the neighboring nodes using voting based support fusion and refinement rules discussed in [91] and [92].

5.6 Numerical Experiments

In this section, we present the simulation results comparing RD-CMP with existing covariance matching based MMV solvers, namely MSBL [32], Co-LASSO [30] and SParse Iterative Covariance-based Estimation (SPICE) [93] in terms of their support recovery performance and computational complexity. Non-COMET algorithms are not considered here as they fail completely when the support size k is larger than m. In each experiment, the nonzero entries of $\mathbf{X}$ and the elements of $\mathbf{A}$ are independently drawn according to a zero mean Gaussian distribution with variance 1 and $1/m$, respectively. In MSBL and Co-LASSO, the final support estimate is declared to be the set of indices whose variance exceeds the average of estimated signal and noise variances.

Fig. 5.1 plots the phase transition of support recovery performance of RD-CMP at an SNR of 10 dB. We observe that RD-CMP, due to its covariance matching roots, is able to

recover k-sparse supports from fewer than k measurements. Furthermore, the quadratic nature of the phase transition boundary suggests that RD-CMP can recover of size as high as $\mathcal{O}(m^2)$.

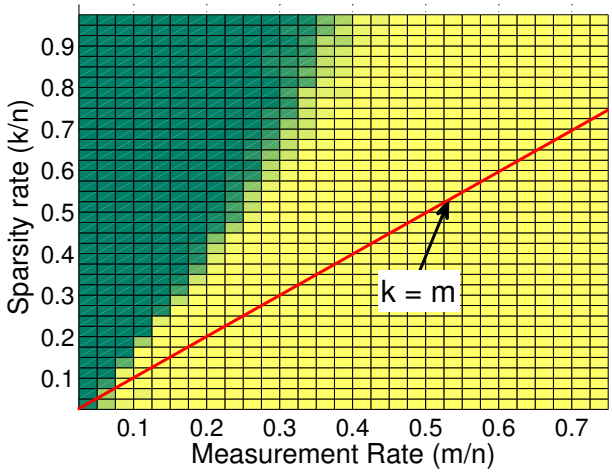


Figure 5.1: Support recovery phase transition plot of RD-CMP. The yellow region marks the measurement rate (m/n) and the sparsity rate (k/n) combinations for which RD-CMP successfully recovers the true support of $\mathbf{X}$ up to 99% accuracy. The green region marks the $(m/n, k/n)$ tuples for which support recovery accuracy is below 99%. Simulation parameters: $n = 200$, $L = 200$, SNR $= 10$ dB and number of trials $= 100$.

Fig. 5.2 compares the support detection and false alarm rates of different MMV algorithms. We note that SPICE and MSBL offers the most balanced performance in terms of tradeoff between support detection and false alarm rates. However, they are also the most computationally intensive algorithms. RD-CMP has lower false alarm rates than Co-LASSO and M-SBL. It also has a better detection rate compared to Co-LASSO. In Fig. 5.3, we compare average run-times of the MMV algorithms for a wide range of signal dimensions spanning $n = 10^2$ to 10^6. We were unable to run Co-LASSO and SPICE beyond $n = 10^4$ due to their prohibitively high memory requirements. Co-LASSO requires storing an $m^2 \times n$ sized $\mathbf{A} \odot \mathbf{A}$ Khatri-Rao matrix in the memory which renders it impractical for large signal dimensions. To mimic a typical sparse setting, the support

size K scales as $\lceil 50 \log_{10} n \rceil$ and the number of effective MMV measurements scales as $mL = \lceil 50K \log_{10} n \rceil$. We observe that the average runtimes for RD-CMP are the lowest among the MMV algorithms considered here, and scales favorably with increasing signal dimension n.

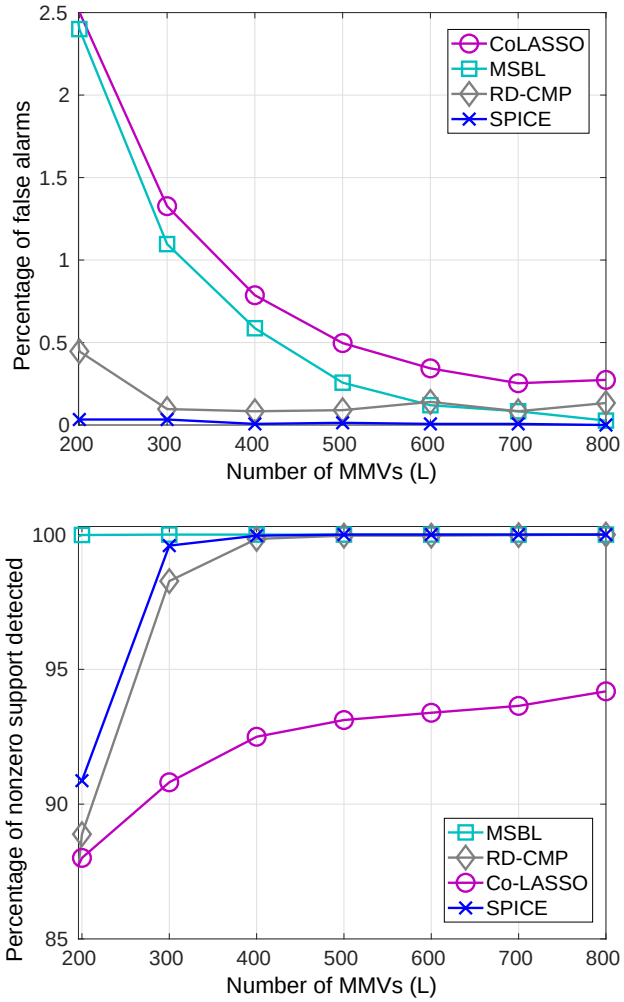


Figure 5.2: Average false alarm (top) and detection rates (bottom) for the recovered support versus the number of MMVs. Other parameters: $n = 500$, K (no. of nonzero rows in $\mathbf{X}$) $= 200$, $m = 100$, SNR $= 10$ dB, and #trials $= 100$.

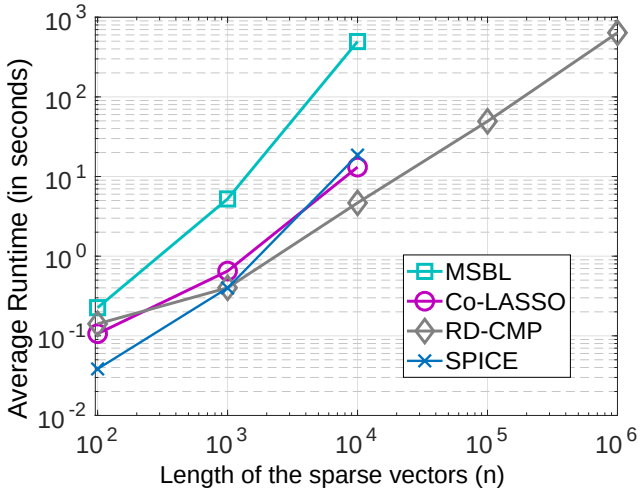


Figure 5.3: Average runtime of the MMV algorithms versus signal dimension n. Other simulation parameters: $K = \lceil 50 \log_{10} n \rceil$, $m = \lceil \frac{3K}{4} \rceil$, $L = \lceil \frac{50K}{m} \log_{10} n \rceil$, and SNR $= 10$ dB.

5.7 Chapter Summary

In this chapter, we proposed a novel covariance matching based MMV solver called RD-CMP which recovers the true support of unknown joint sparse vectors from their compressive measurements by minimizing a novel Rényi divergence objective which serves as an approximation to the marginalized log-likelihood of the measurements. The proposed algorithm is capable of recovering k-sparse support from fewer than k measurements. Based on simulation results, we can conclude that the proposed RD-CMP algorithm is significantly faster compared to existing covariance matching based MMV algorithms such as MSBL and Co-LASSO, making it a better candidate for handling large dimensional signals.

Chapter 6

Decentralized Joint Sparse Signal Recovery - An SBL Approach

In which joint sparse signal recovery gets decentralized...

6.1 Introduction

In this chapter[1], we solve the multiple measurement vector problem in a distributed setting. The distributed MMV problem deals with the estimation of multiple joint sparse vectors from their noisy underdetermined linear measurements using a network of computing nodes. Each vector in the joint sparse ensemble is estimated by a distinct node in the network with access to its linear measurements. Additionally, the nodes are allowed to collaborate among themselves in order to exploit the network-wide joint sparsity, and

[1]This chapter is based on S. Khanna and C. R. Murthy, "Decentralized Joint-Sparse Signal Recovery: A Sparse Bayesian Learning Approach", *IEEE Transactions on Signal and Information Processing over Networks* , vol. 3, no. 1, pp. 29–45, Sept. 2016.

jointly estimate their local sparse vectors using fewer measurements and with better accuracy. In [38], it has been shown that as the network grows in size, a k-sparse ensemble can be recovered exactly from as few as k noiseless measurements per node. Any such substantial reduction in the number of required measurements is an attractive proposition, especially for applications where the measurement acquisition costs are of concern.

In this chapter, we assume that the joint sparse vectors follow the Type-2 Joint Sparsity Model (JSM-2) [38], i.e., the nonzero coefficients within and across the vectors are uncorrelated. Distributed estimation of JSM-2 signals has many real-world applications such as MIMO channel estimation [9, 10, 94], cooperative spectrum sensing [14, 95], decentralized event detection [84] and acoustic source localization [96].

To further motivate the joint sparsity signal structure in a distributed setup, consider the problem of detection/classification of randomly occurring events in a field by multiple sensor nodes. Each sensor node j, $1 \leq j \leq L$, employs a dictionary $\Psi_j = [\psi_j^1; \psi_j^2 \dots \psi_j^n]$, whose each column ψ_j^i is the signature corresponding to the i^{th} event, one out of the n events which can potentially occur. In many cases, due to the inability to accurately model the sensing process, the signature vectors ψ_j^i are simply chosen to be the past recordings of j^{th} sensor corresponding to standalone occurrence of the i^{th} event, averaged across multiple experiments [97]. This procedure can result in distinct dictionaries at the individual nodes. For any k ($\ll n$) events occurring simultaneously, a noisy sensor recording might belong to multiple subspaces, each spanned by different subsets of columns of the local dictionary. In such a scenario, enforcing joint sparsity across the sensor nodes can resolve the ambiguity in selecting the correct subset of columns at each sensor node.

Distributed algorithms for JSM-2 signal recovery come in two flavors - centralized and

decentralized. In the centralized approach, each node transmits its local measurements to a fusion center (FC) which runs a joint-sparse signal recovery algorithm. The FC then transmits the reconstructed sparse signal estimates back to their respective nodes. In contrast, in a decentralized approach, the goal is to obtain the same centralized solution at all nodes by allowing each node to exchange information with its single-hop neighbors in addition to processing its local measurements. Besides being inherently robust to node failures, decentralized schemes also tend to be more energy efficient as the inter-node communication is restricted to relatively short ranges covering only single communication links. Here, we focus on the decentralized approach for solving the sparse signal recovery problem under the JSM-2 signal model.

We propose a decentralized extension of the Multiple Sparse Bayesian Learning (MSBL) algorithm [32] for estimation of multiple joint sparse vectors from their noisy and underdetermined linear measurements. The proposed algorithm, called *Consensus Based Distributed Sparse Bayesian Learning* or CB-DSBL, exploits the network-wide joint sparsity of the unknown sparse vectors to recover them using a significantly fewer number of local measurements compared to standalone sparse signal recovery schemes. To reduce the amount of inter-node communication and the associated overheads, the nodes are permitted to exchange messages with only a small set of predesignated bridge nodes. Using Monte Carlo simulations as well as real-world data based experiments, we demonstrate the superior signal reconstruction performance of CB-DSBL compared to the existing decentralized algorithms.

6.2 Background on Decentralized Joint Sparse Signal Recovery

6.2.1 Literature Survey

In this section, we review the existing decentralized algorithms for JSM-2 signal recovery.

Early work in this area focused on the development of decentralized extensions of existing centralized MMV algorithms. DCOMP [92] and DCSP [98] belong to this category, they are the decentralized versions of greedy MMV algorithms SOMP [38] and SSP [98], respectively. Both these algorithms have very low computational complexity but they do not perform as well as convex relaxation and Bayesian recovery methods. Moreover, both DCOMP and DCSP assume a priori knowledge of the size of the nonzero support set, which could be unknown or hard to estimate. In [99], three distributed greedy pursuit algorithms: DiOMP, DiSP and DiFROGS have been proposed, and an extensive simulation based performance comparison study has been conducted for a wide range of network connection densities. Among these three algorithms, DiOMP turns out to be the best all-round performer. However, DCOMP is more accurate than DiOMP in terms of support detection, as demonstrated in [92].

DR-LASSO [100] is an iterative decentralized algorithm which uses alternating minimization to optimize a convex regularized objective which includes a mixed ℓ_1-ℓ_2 norm based joint sparsity inducing penalty. In [95], the same ℓ_1-ℓ_2 norm penalty is used in a cooperative spectrum sensing setup to estimate the sparse spectral occupancy patterns perceived by multiple cognitive radios via independent channels. The decentralized re-weighted ℓ_1 norm minimization algorithm or DRL-1 [84] employs a sum-log-sum penalty to induce a joint-sparse solution. The non-convex objective of DRL-1 is better at promoting sparsity

than the ℓ_1 norm penalty [101]. In DRL-1, the non-convex objective is minimized by replacing it with a surrogate convex function made up of weighted ℓ_1 norm terms, and the weights are updated in each iteration. In both DR-LASSO and DRL-1, the sparsity of the solution is controlled by a regularization parameter λ which biases the joint sparsity inducing penalty term in the objective. The correct amount of regularization or the optimal value of λ is typically chosen via cross-validation, which is not practical unless additional training data is available.

In Bayesian approach, the amount of regularization is tuned automatically by the procedure of selecting an appropriate member prior from a parameterized family of joint sparsity inducing priors. The selected prior is the one with the maximal Bayesian evidence [33, 102]. The learned signal prior is subsequently used to compute the maximum a posteriori probability (MAP) estimate of the sparse vectors. DCS-AMP [103] is one such Bayesian algorithm which uses approximate message passing to efficiently learn the parameters of a joint sparsity inducing Bernoulli-Gaussian family of priors. DCS-FBMP [104] also uses a Bernoulli-Gaussian signal prior, and constructs the sparse support incrementally by using a greedy approach to maximize the log-posterior probability of the support parameters. However, it is designed to work only with star and ring topology networks. Turbo Bayesian Compressive Sensing (Turbo-BCS) [105], another decentralized algorithm, adopts a more relaxed zero mean Gaussian signal prior, with the variance hyperparameters themselves distributed according to an exponential distribution. The relaxed signal prior improves the MSE performance without compromising on the sparsity of the solution. Turbo-BCS, however, involves direct exchange of signal estimates between the nodes, which renders it unsuitable for applications where it is necessary to preserve the privacy

Table 6.1: Comparison of decentralized joint-sparse signal recovery algorithms

Decentralized algorithm	**Per node, per iteration computational complexity**	**Per node, per iteration communication complexity**	**Privacy of local signal estimates**	**Tunable parameters (if any)**	**Assumes a priori knowledge of sparsity level**
DCSP [107]	$\mathcal{O}(mn + \zeta n + k \log n + m^2)$	$\mathcal{O}(\zeta n + k \log n)$	Yes	None	Yes
DCOMP [92]	$\mathcal{O}(n\zeta + L)$	$\mathcal{O}(\zeta n + L)$	Yes	None	Yes
DRL-1 [84]	$\mathcal{O}((n^2 + m^3 + nm^2)r_{\max} + \zeta n)$	$\mathcal{O}(\zeta n)$	Yes	Yes	No
DR-LASSO [100]	$\mathcal{O}(n^2 m T_1 + \zeta n T_2)$	$\mathcal{O}(\zeta n T_2)$	Yes	Yes	No
Turbo-BCS [105]	$\mathcal{O}(n^3 + nL + nk^2 + k^3 + mk)$	$\mathcal{O}(kL)$	No	None	No
DCS-AMP [103]	$\mathcal{O}(mn + \zeta n + c_1 n)$	$\mathcal{O}(\zeta n)$	Yes	Yes	No
CB-DSBL (proposed)	$\mathcal{O}(n^2 + m^3 + nm^2 + \zeta n r_{\max})$	$\mathcal{O}(\zeta n r_{\max})$	Yes	None	No

1. n, m, k and L stand for the dimension of unknown sparse vector, number of local measurements per node, number of nonzero coefficients in the true support and network size, respectively.
2. ζ is the maximum number of communication links activated per node, per communication round.
3. $r_{\max}$ is the number of inner loop ADMM iterations executed per CB-DSBL iteration.
4. In DRL-1, $r_{\max}$ is the number of inner loop ADMM iterations used to obtain an inexact solution to the weighted ℓ_1 norm based subproblem.
5. T_1 and T_2 denote the number of iterations of the two different inner loop iterations executed per DR-LASSO iteration.

of the local signals.

For a more up-to-date review of existing decentralized JSM-2 algorithms, the readers are referred to an excellent survey article [106] by Wimalajeewa et al.

6.2.2 Contributions

Our main contributions in this chapter are as follows:

1. We propose a novel decentralized, iterative, Bayesian joint-sparse signal recovery algorithm called *Consensus Based Distributed Sparse Bayesian Learning* or CB-DSBL. Our proposed algorithm works by establishing network wide consensus with respect to the estimated parameters of a joint sparsity inducing signal prior. The learnt signal prior is subsequently used by the individual nodes to obtain MAP estimates of their local sparse signal vectors.

2. The proposed algorithm employs the Alternating Direction Method of Multipliers (ADMM) to solve a series of iteration dependent consensus optimization problems

which require the nodes to exchange messages with each other. To reduce the associated communication overheads, we adopt a bandwidth efficient inter-node communication scheme. This scheme entails the nodes exchanging messages with a predesignated subset of its single-hop neighbors identified as *bridge nodes*, as motivated in [1]. In this connection, we analytically establish the relationship between the selected set of bridge nodes and the convergence rate of the ADMM iterations. For the bridge-node based inter-node communication scheme, we show a linear rate of convergence for the ADMM iterations when applied to a generic consensus optimization problem. The analysis is useful in obtaining a closed form expression for the tunable parameter of our proposed joint sparse signal recovery algorithm, ensuring its fast convergence.

3. We empirically demonstrate the superior MSE and support recovery performance of CB-DSBL compared to existing decentralized algorithms. The experimental results are presented for both synthetic and real world data. In the latter case, we illustrate the performance of the different algorithms for the application of cooperative wideband spectrum sensing in cognitive radios.

In Table 6.1, we compare the existing decentralized joint-sparse signal recovery schemes with respect to their per iteration computational and communication complexity, privacy of local estimates, presence/absence of tunable parameters and dependence on prior knowledge of the sparsity level. As highlighted in the comparison in Table 6.1, CB-DSBL belongs to a handful of decentralized algorithms for joint-sparse signal recovery which do not require a priori knowledge of the sparsity level, rely only on single-hop communication,

and do not involve direct exchange of local signal estimates between network nodes. Besides this, unlike loopy Belief Propagation (BP) or Approximate Message Passing (AMP) based Bayesian algorithms, CB-DSBL does not suffer from any convergence issues even when the local measurement matrix at each node is dense or not randomly constructed.

6.3 System Model

We consider L computing nodes connected as a network described by a bi-directional graph $\mathcal{G} = (\mathcal{J}, \mathcal{A})$, where $\mathcal{J} = \{1, 2, \ldots, L\}$ is the set of vertices in $\mathcal{G}$, each vertex representing a distinct node in the network. Set $\mathcal{A}$ contains the edges in $\mathcal{G}$, each edge representing a single-hop error-free communication link between a distinct pair of nodes. Each node is interested in estimating an unknown k-sparse vector $\mathbf{x}_j \in \mathbb{R}^n$ from m locally acquired noisy linear measurements $\mathbf{y}_j \in \mathbb{R}^m$. The local measurement vector $\mathbf{y}_j$ at node j is generated according to the linear model

$$\mathbf{y}_j = \mathbf{\Phi}_j \mathbf{x}_j + \mathbf{w}_j, \quad 1 \leq j \leq L \tag{6.1}$$

where, $\mathbf{\Phi}_j \in \mathbb{R}^{m \times n}$ is a full row rank sensing matrix and $\mathbf{w}_j \in \mathbb{R}^m$ is the measurement noise modeled as zero mean Gaussian distributed with covariance matrix $\sigma_j^2 \mathbf{I}_m$. The sparse vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L$ at different nodes follow the JSM-2 signal model [16]. This implies that all $\mathbf{x}_j$ share a common support, represented by the index set $\mathcal{S}$. From the JSM-2 model, it also follows that the nonzero coefficients of the sparse vectors are independent within and across the vectors.

The goal is to recover the sparse vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L$ at the respective nodes using decentralized processing. In addition to processing the locally available data $\left\{\mathbf{y}_j, \mathbf{\Phi}_j, \sigma_j^2\right\}$

at j^{th} node, each node must collaborate with its single-hop neighboring nodes to exploit the network wide joint sparsity of the sparse vectors. Ideally, the decentralized algorithm should be able to generate the centralized solution at each node, as if each node has access to the global information, i.e., $\left\{\mathbf{y}_j, \mathbf{\Phi}_j, \sigma_j^2\right\}_{j\in\mathcal{J}}$.

6.4 Centralized Bayesian Learning of JSM-2 Signals

In this section, we briefly recall the centralized MSBL algorithm [32] for JSM-2 signal recovery and extend it to support distinct measurement matrices $\mathbf{\Phi}_j$ and noise variances σ_j^2 at each node. The centralized algorithm runs at an FC, which assumes complete knowledge of network wide information, $\left\{\mathbf{y}_j, \mathbf{\Phi}_j, \sigma_j^2\right\}_{j=1}^{L}$. For ease of notation, we introduce two variables $\mathbf{X} \triangleq \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L\}$ and $\mathbf{Y} \triangleq \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_L\}$ to be used in the sequel. Similar to MSBL, each of the sparse vectors $\mathbf{x}_j, j \in \mathcal{J}$ is assumed to be distributed according to a parameterized signal prior $p(\mathbf{x}_j; \boldsymbol{\gamma})$ shown below.

$$p(\mathbf{x}_j; \boldsymbol{\gamma}) = \prod_{i=1}^{n} p\left(\mathbf{x}_j(i); \boldsymbol{\gamma}\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\boldsymbol{\gamma}(i)}} \exp\left(-\frac{\mathbf{x}_j(i)^2}{2\boldsymbol{\gamma}(i)}\right).$$

Further, the joint signal prior $p(\mathbf{X}; \boldsymbol{\gamma})$ is assumed to be

$$p(\mathbf{X}; \boldsymbol{\gamma}) = \prod_{j\in\mathcal{J}} p(\mathbf{x}_j; \boldsymbol{\gamma}). \tag{6.2}$$

In the above, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}(0), \boldsymbol{\gamma}(1), \ldots, \boldsymbol{\gamma}(n))^T$ is an n dimensional hyperparameter vector, whose i^{th} entry, $\boldsymbol{\gamma}(i)$, models the common variance of $\mathbf{x}_j(i)$ for $1 \leq j \leq L$. Since the signal priors $p(\mathbf{x}_j; \boldsymbol{\gamma})$ are parameterized by a common $\boldsymbol{\gamma}$, if $\boldsymbol{\gamma}$ has a sparse support $\mathcal{S}$, then the MAP estimates of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L$ will also be jointly sparse with common support $\mathcal{S}$. Just like in MSBL, the hyperparameter vector $\boldsymbol{\gamma}$ is estimated as the the maximizer of the total

Bayesian evidence or the network-wide log-likelihood $\log p(\mathbf{Y};\boldsymbol{\gamma})$.

Let $\hat{\boldsymbol{\gamma}}_{\mathrm{ML}}$ denote the maximum likelihood (ML) estimate of hyperparameters of the joint source prior:

$$\hat{\boldsymbol{\gamma}}_{\mathrm{ML}} = \underset{\boldsymbol{\gamma}\succeq 0}{\arg\max}\ \log p(\mathbf{Y};\boldsymbol{\gamma}) \tag{6.3}$$

where $p(\mathbf{Y};\boldsymbol{\gamma})$ is a type-2 likelihood function obtained by marginalizing the joint density $p(\mathbf{Y},\mathbf{X};\boldsymbol{\gamma})$ with respect to the unknown vectors in $\mathbf{X}$ as shown below.

$$p(\mathbf{Y};\boldsymbol{\gamma}) = \prod_{j=1}^{L}\int p(\mathbf{y}_j|\mathbf{x}_j)p(\mathbf{x}_j;\boldsymbol{\gamma})\mathrm{d}\mathbf{x}_j = \prod_{j=1}^{L}\mathcal{N}\left(0, \boldsymbol{\Phi}_j\boldsymbol{\Gamma}\boldsymbol{\Phi}_j^T + \sigma_j^2\mathbf{I}_m\right). \tag{6.4}$$

Here $\boldsymbol{\Gamma} = \mathrm{diag}(\boldsymbol{\gamma})$. We note that $\hat{\boldsymbol{\gamma}}_{\mathrm{ML}}$ cannot be derived in closed form by directly maximizing the likelihood in (6.4) with respect to $\boldsymbol{\gamma}$. Hence, as suggested in the SBL framework [41], we use the expectation maximization (EM) procedure to maximize $\log p(\mathbf{Y};\boldsymbol{\gamma})$ by treating $\mathbf{X}$ as hidden variables. Each EM iteration comprises the following two steps:

$$\begin{aligned}
&\text{E step:} \quad \mathcal{Q}(\boldsymbol{\gamma}|\boldsymbol{\gamma}^k) = \mathbb{E}_{\mathbf{X}|\mathbf{Y},\boldsymbol{\gamma}^k}\left[\log p(\mathbf{Y},\mathbf{X};\boldsymbol{\gamma})\right] \\
&\text{M step:} \quad \boldsymbol{\gamma}^{k+1} = \underset{\boldsymbol{\gamma}\in\mathbb{R}_+^n}{\arg\max}\,\mathcal{Q}(\boldsymbol{\gamma}|\boldsymbol{\gamma}^k)
\end{aligned} \tag{6.5}$$

where k denotes the iteration index of EM algorithm. From the LMMSE theory [47], we know that the conditional density $p(\mathbf{x}_j|\mathbf{y}_j,\boldsymbol{\gamma}^k)$ used to compute $\mathcal{Q}(\boldsymbol{\gamma}|\boldsymbol{\gamma}^k)$ is Gaussian distributed with mean $\boldsymbol{\mu}_j^{k+1}$ and covariance $\boldsymbol{\Sigma}_j^{k+1}$ given by

$$\begin{aligned}
\boldsymbol{\Sigma}_j^{k+1} &= \boldsymbol{\Gamma}^k - \boldsymbol{\Gamma}^k\boldsymbol{\Phi}_j^T\left(\sigma_j^2\mathbf{I}_m + \boldsymbol{\Phi}_j\boldsymbol{\Gamma}^k\boldsymbol{\Phi}_j^T\right)^{-1}\boldsymbol{\Phi}_j\boldsymbol{\Gamma}^k \\
\boldsymbol{\mu}_j^{k+1} &= \sigma_j^{-2}\boldsymbol{\Sigma}_j^{k+1}\boldsymbol{\Phi}_j^T\mathbf{y}_j.
\end{aligned} \tag{6.6}$$

As shown in Appendix H.1, the M-step optimization in (6.5) can be recast as the following

minimization problem.

$$\boldsymbol{\gamma}^{k+1} = \arg\min_{\boldsymbol{\gamma}\in\mathbb{R}_+^n} \sum_{j\in\mathcal{J}} \sum_{i=1}^{n} \left(\log \boldsymbol{\gamma}(i) + \frac{\boldsymbol{\Sigma}_j^k(i,i) + \boldsymbol{\mu}_j^k(i)^2}{\boldsymbol{\gamma}(i)} \right). \tag{6.7}$$

From the zero gradient optimality condition in (6.7), the M-step reduces to the following update rule:

$$\boldsymbol{\gamma}^{k+1}(i) = \frac{1}{L} \sum_{j\in\mathcal{J}} \left(\boldsymbol{\Sigma}_j^{k+1}(i,i) + \boldsymbol{\mu}_j^{k+1}(i)^2 \right) \quad \text{for } 1 \le i \le n. \tag{6.8}$$

By repeatedly iterating between the E-step (6.6) and the M-step (6.8), $\boldsymbol{\gamma}^k$ converges to either a local maximum or a saddle point of $\log p(\mathbf{Y};\boldsymbol{\gamma})$ [48]. Once $\hat{\boldsymbol{\gamma}}_{\mathrm{ML}}$ is obtained, the MAP estimate of $\mathbf{x}_j$ is evaluated by substituting $\hat{\boldsymbol{\gamma}}_{\mathrm{ML}}$ in the expression for $\boldsymbol{\mu}_j$ in (6.6). It has been observed empirically that as the EM algorithm converges, the $\boldsymbol{\gamma}(i)$'s belonging to the inactive support tend to zero, hence resulting in a sparse ML estimate of $\boldsymbol{\gamma}$.

6.5 Decentralized Bayesian Learning of JSM-2 Signals

6.5.1 Algorithm Development

In this section, we develop a decentralized version of the centralized algorithm discussed in the previous section. For notational convenience, we introduce an n length vector $\mathbf{a}_j^k = \left(a_{j,1}^k, a_{j,2}^k, \ldots, a_{j,n}^k\right)^T$ maintained at node j, where $a_{j,i}^k = \boldsymbol{\Sigma}_j^k(i,i) + \boldsymbol{\mu}_j^k(i)^2$ with $\boldsymbol{\Sigma}_j^k$ and $\boldsymbol{\mu}_j^k$ as defined in (6.6).

We observe that the solution of the M-step optimization in (6.8) can be interpreted as computing an average of the L vectors $\left\{\mathbf{a}_j^{k+1}\right\}_{j=1}^{L}$. The same solution can also be obtained

by solving a different minimization problem

$$\boldsymbol{\gamma}^{k+1} = \underset{\boldsymbol{\gamma} \in \mathbb{R}^n_+}{\arg\min} \sum_{j \in \mathcal{J}} \|\boldsymbol{\gamma} - \mathbf{a}_j^{k+1}\|_2^2. \tag{6.9}$$

Unlike the non-convex M-step objective function in (6.7), the surrogate objective function in (6.9) is convex in $\boldsymbol{\gamma}$ and therefore can be minimized in a distributed manner using powerful convex optimization techniques. An alternate form of (6.9) amenable to distributed optimization is given by

$$\begin{aligned} &\min_{\boldsymbol{\gamma}_j \in \mathbb{R}^n_+,\, j \in \mathcal{J}} \sum_{j \in \mathcal{J}} \|\boldsymbol{\gamma}_j - \mathbf{a}_j^{k+1}\|_2^2 \\ &\text{subject to } \boldsymbol{\gamma}_j = \boldsymbol{\gamma}_{j'} \quad \forall \;\; j \in \mathcal{J},\; j' \in \mathcal{N}_j \end{aligned} \tag{6.10}$$

where $\mathcal{N}_j$ denotes the set of single-hop neighbors of node j. The equality constraints in (6.10) ensure its equivalence to the unconstrained optimization in (6.9). Here, the number of equality constraints is equal to $|\mathcal{A}|$, the total number of single-hop links in the network. In a naive decentralized implementation of (6.10), the number of messages exchanged between the nodes grow linearly with the number of consensus constraints. By restricting the nodes to exchange information only through a small set of pre-designated nodes called *bridge nodes*, the number of consensus constraints can be drastically reduced while preserving the equivalence of (6.9) and (6.10). Let $\mathcal{B} \subseteq \mathcal{J}$ denote the set of all bridge nodes in the network and $\mathcal{B}_j \subseteq \mathcal{B}$ denote the set of bridge nodes which are single-hop neighbors of node j, then (6.10) can be rewritten as

$$\begin{aligned} &\underset{\boldsymbol{\gamma}_j \in \mathbb{R}^n_+,\, j \in \mathcal{J}}{\text{minimize}} \sum_{j \in \mathcal{J}} \|\boldsymbol{\gamma}_j - \mathbf{a}_j^{k+1}\|_2^2 \\ &\text{subject to } \boldsymbol{\gamma}_j = \boldsymbol{\gamma}_b \quad \forall \;\; j \in \mathcal{J},\; b \in \mathcal{B}_j. \end{aligned} \tag{6.11}$$

The auxiliary variables $\boldsymbol{\gamma}_b$, called *bridge parameters*, are used to establish consensus among $\boldsymbol{\gamma}_j$. Each bridge parameter $\boldsymbol{\gamma}_b$ is a non negative n length vector maintained by the bridge node b. Fig. 6.1 illustrates the selection of bridge nodes according to Lemma 6.1, in a sample network. As motivated in [1], [108], using bridge nodes to impose network wide consensus allows us to trade off between the communication cost and robustness of the distributed optimization algorithm.[2] The following Lemma provides sufficient conditions on the choice of the bridge node set $\mathcal{B}$ under which (6.9) and (6.11) are equivalent. The proof for the Lemma can be found in [1].

Lemma 6.1. *For a connected graph $\mathcal{G}$, if the bridge node set $\mathcal{B} \subseteq \mathcal{J}$ satisfies the following conditions*

1. *Each node j must be connected to at least one bridge node in $\mathcal{B}$, i.e., $\mathcal{B}_j \neq \phi$ for any $j \in \mathcal{J}$, and,*

2. *If two nodes j_1 and j_2 are single-hop neighbors, then $\mathcal{B}_{j_1} \bigcap \mathcal{B}_{j_2} \neq \phi$ for any $j_1, j_2 \in \mathcal{J}$,*

then, in the solution to (6.11), $\boldsymbol{\gamma}_j$'s are equal for all $j \in \mathcal{J}$.

We employ the *Alternating Directions Method of Multipliers* (ADMM) algorithm [109] to solve the convex optimization problem in (6.11). ADMM is the state of the art dual ascent algorithm for solving constrained convex optimization problems, offering a linear convergence rate and a natural extension to a decentralized implementation. We start by

[2]In an alternate embodiment of the proposed algorithm, the message exchanges could be restricted to occur only through the (trustworthy) bridge nodes, thereby avoiding direct communication between the nodes. In this case, the role of the bridge nodes could be to enforce consensus in $\boldsymbol{\gamma}$ across the nodes, and these nodes need not directly participate in signal reconstruction.

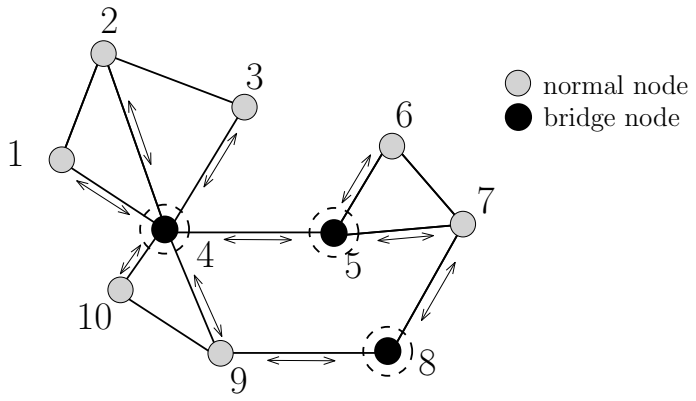


Figure 6.1: Selection of bridge nodes in a sample network consisting of 10 nodes. In the proposed scheme, only those edges that have at least one of the vertices as a bridge node are used for communication. The remaining edges are not used for communication. For example, node 9 communicates only with bridge nodes 4 and 8.

constructing an augmented Lagrangian, L_ρ, given by

$$\begin{aligned} L_\rho(\boldsymbol{\gamma}_\mathcal{J}, \boldsymbol{\gamma}_\mathcal{B}, \boldsymbol{\lambda}) &\triangleq \sum_{j\in\mathcal{J}} \|\boldsymbol{\gamma}_j - \mathbf{a}_j^{k+1}\|_2^2 \\ &+ \sum_{j\in\mathcal{J}}\sum_{b\in\mathcal{B}_j} (\boldsymbol{\lambda}_j^b)^T(\boldsymbol{\gamma}_j - \boldsymbol{\gamma}_b) + \frac{\rho}{2}\sum_{j\in\mathcal{J}}\sum_{b\in\mathcal{B}_j}\|\boldsymbol{\gamma}_j - \boldsymbol{\gamma}_b\|_2^2, \end{aligned} \tag{6.12}$$

where $\boldsymbol{\lambda}_j^b$ denotes the $n \times 1$ sized Lagrange multiplier vector corresponding to the equality constraint $\boldsymbol{\gamma}_j = \boldsymbol{\gamma}_b$ and ρ is a positive scalar which biases the quadratic consensus penalty term. For lighter notation, we define concatenated vectors $\boldsymbol{\gamma}_\mathcal{J} = \{\boldsymbol{\gamma}_1^T, \boldsymbol{\gamma}_2^T, \ldots, \boldsymbol{\gamma}_L^T\}^T$ and $\boldsymbol{\gamma}_\mathcal{B} = \{\boldsymbol{\gamma}_{b_1}^T, \ldots, \boldsymbol{\gamma}_{b_{|\mathcal{B}|}}^T\}^T$. We also define the $nN_C \times 1$ concatenated Lagrange multiplier vector $\boldsymbol{\lambda}$, where N_C is the number of equality constraints in (6.11). The solution to (6.11) is then obtained by executing the following ADMM iterations until convergence:

$$\boldsymbol{\gamma}_\mathcal{J}^{r+1} = \arg\min_{\boldsymbol{\gamma}_\mathcal{J}} L_\rho(\boldsymbol{\gamma}_\mathcal{J}, \boldsymbol{\gamma}_\mathcal{B}^r, \boldsymbol{\lambda}^r) \tag{6.13}$$

$$\boldsymbol{\gamma}_\mathcal{B}^{r+1} = \arg\min_{\boldsymbol{\gamma}_\mathcal{B}} L_\rho(\boldsymbol{\gamma}_\mathcal{J}^{r+1}, \boldsymbol{\gamma}_\mathcal{B}, \boldsymbol{\lambda}^r) \tag{6.14}$$

$$(\boldsymbol{\lambda}_j^b)^{r+1} = (\boldsymbol{\lambda}_j^b)^r + \rho(\boldsymbol{\gamma}_j^{r+1} - \boldsymbol{\gamma}_b^{r+1}) \tag{6.15}$$

$\forall j \in \mathcal{J}, b \in \mathcal{B}_j$. Here, r denotes the ADMM iteration index. In (6.13-6.14), the primal variables, $\boldsymbol{\gamma}_{\mathcal{J}}$ and $\boldsymbol{\gamma}_{\mathcal{B}}$, are updated in a Gauss-Seidel fashion by minimizing the augmented Lagrangian, L_ρ, evaluated at the previous estimate of the dual variable $\boldsymbol{\lambda}$. By adding an extra quadratic penalty term to the original Lagrangian, the objective in (6.14) is no longer affine in $\boldsymbol{\gamma}_{\mathcal{B}}$ and hence has a bounded minimizer. The dual variable $\boldsymbol{\lambda}$ is updated via a gradient-ascent step (6.15) with a step-size equal to the ADMM parameter ρ. This particular choice of step-size ensures the dual feasibility of the iterates $\{\boldsymbol{\gamma}_{\mathcal{J}}^{r+1}, \boldsymbol{\gamma}_{\mathcal{B}}^{r+1}, \boldsymbol{\lambda}^{r+1}\}$ for all r [109]. Since the augmented Lagrangian L_ρ is strictly convex with respect to $\boldsymbol{\gamma}_{\mathcal{J}}$ and $\boldsymbol{\gamma}_{\mathcal{B}}$ individually, the zero gradient optimality conditions for (6.13) and (6.14) translate into simple update equations for $\boldsymbol{\gamma}_j$ and $\boldsymbol{\gamma}_b$:

$$\begin{aligned} \boldsymbol{\gamma}_j^{r+1} &= \frac{2\mathbf{a}_j^{k+1} + \sum_{b\in\mathcal{B}_j}\left(\rho\boldsymbol{\gamma}_b^r - (\boldsymbol{\lambda}_j^b)^r\right)}{2+\rho|\mathcal{B}_j|} \quad \forall\, j \in \mathcal{J} \\ \boldsymbol{\gamma}_b^{r+1} &= \frac{\sum_{j\in\mathcal{N}_b}(\rho\boldsymbol{\gamma}_j^{r+1} + (\boldsymbol{\lambda}_j^b)^r)}{\rho|\mathcal{N}_b|} \quad \forall\, b \in \mathcal{B}. \end{aligned} \tag{6.16}$$

Here $\mathcal{N}_b$ denotes the set of nodes connected to bridge node b. As shown in Appendix H.2, by eliminating the Lagrange multiplier terms from (6.15) and (6.16), the update rule for $\boldsymbol{\gamma}_b$ can be further simplified to

$$\boldsymbol{\gamma}_b^{r+1} = \frac{1}{|\mathcal{N}_b|}\sum_{j\in\mathcal{N}_b}\boldsymbol{\gamma}_j^{r+1} \quad \forall\, b \in \mathcal{B}. \tag{6.17}$$

In section 6.5.6, we compare the bridge node based ADMM discussed above with other decentralized optimization techniques in the literature. We show empirically that the bridge node based ADMM scheme is able to flexibly trade off between communication complexity, robustness to node failures, speed of convergence, and signal reconstruction performance. Moreover, the ADMM iterations (6.13)-(6.15) can be adapted to handle time

Algorithm 2 Consensus Based Distributed Sparse Bayesian Learning (CB-DSBL)

Input: $k \leftarrow 0$
$\boldsymbol{\gamma}_j^k \leftarrow 10^{-3}\mathbf{1}_{n\times 1} \quad \forall j \in \mathcal{J}$
$\boldsymbol{\gamma}_b^k, (\boldsymbol{\lambda}_j^b)^k \leftarrow 0 \quad \forall j \in \mathcal{J},\ b \in \mathcal{B}_j$
while $(k < k_{\max})\ \&\ \left(\Delta\boldsymbol{\gamma}_{\mathcal{J}} > \epsilon\right)$ **do**
E step: Each node s_j, $j \in \mathcal{J}$, updates $\mathbf{a}_j^k$ according to (6.6).
M step: $r \leftarrow 0$, $\boldsymbol{\gamma}_{\mathcal{J}}^r \leftarrow \boldsymbol{\gamma}_{\mathcal{J}}^k$, $\boldsymbol{\gamma}_{\mathcal{B}}^r \leftarrow \boldsymbol{\gamma}_{\mathcal{B}}^k$, $(\boldsymbol{\lambda})^r \leftarrow (\boldsymbol{\lambda})^k$
while $r < r_{\max}$ **do**
1. All nodes $s_{j\in\mathcal{J}}$ update their local estimate of hyperparameters $\boldsymbol{\gamma}_j^r$ according to (6.16).
2. All nodes $s_{j\in\mathcal{J}}$ transmit the updated $\boldsymbol{\gamma}_j^{r+1}$ estimate to connected bridge nodes $s_{b\in\mathcal{B}_j}$.
3. Each bridge node $s_{b\in\mathcal{B}}$ updates its bridge variable $\boldsymbol{\gamma}_b^r$ according to (6.17).
4. All bridge nodes $s_{b\in\mathcal{B}}$ transmit updated bridge hyperparameters $\boldsymbol{\gamma}_b^{r+1}$ to nodes in their neighborhood $\mathcal{N}_b$.
5. All nodes $s_{j\in\mathcal{J}}$ update their Lagrange multipliers $(\boldsymbol{\lambda}_j^b)^r, b \in \mathcal{B}_j$ according to (6.15).
6. $r \leftarrow r + 1$
end while
$\boldsymbol{\gamma}_{\mathcal{J}}^k \leftarrow \boldsymbol{\gamma}_{\mathcal{J}}^r$, $\boldsymbol{\gamma}_{\mathcal{B}}^k \leftarrow \boldsymbol{\gamma}_{\mathcal{B}}^r$, $(\boldsymbol{\lambda})^k \leftarrow (\boldsymbol{\lambda})^r$
$k \leftarrow k + 1$
$\Delta\boldsymbol{\gamma}_{\mathcal{J}} \leftarrow ||\boldsymbol{\gamma}_{\mathcal{J}}^k - \boldsymbol{\gamma}_{\mathcal{J}}^{k-1}||_2$
end while
Output: $\boldsymbol{\gamma}_{\mathcal{J}}^k$.

varying, asynchronous networks, as suggested in [110]. During asynchronous operation, the dynamic assignment of bridge nodes can be avoided by treating all nodes as bridge nodes.

6.5.2 Consensus Based Distributed Sparse Bayesian Learning

We now propose the CB-DSBL algorithm. Essentially, it is a decentralized EM algorithm for finding the ML estimate of the hyperparameters $\boldsymbol{\gamma}$. The algorithm comprises two nested loops. In the outer loop, each node performs the E-step (6.6) in a standalone manner. In the inner loop, ADMM iterations are performed to solve the M-step optimization in a decentralized manner. Upon convergence of the outer loop, each node $j \in \mathcal{J}$ has the same ML estimate of $\boldsymbol{\gamma}$, which is then used to obtain a MAP estimate of the local sparse vector $\mathbf{x}_j$, similar to the centralized algorithm. The steps of the CB-DSBL algorithm are detailed in Algorithm 1.

Each ADMM iteration in the M-step of the CB-DSBL algorithm involves two rounds of communication (Steps 2 and 4) between the nodes. In the first communication round, each node $j \in J$ transmits $\boldsymbol{\gamma}_j \in \mathbb{R}^n$ to its $|\mathcal{B}_j|$ single-hop neighbors. In the second communication round, each bridge node $b \in \mathcal{B}$ transmits $\boldsymbol{\gamma}_b \in \mathbb{R}^n$ to its $|\mathcal{N}_b|$ single-hop neighbors. Thus, in each M-step, $2n \sum_{j \in \mathcal{J}} |\mathcal{B}_j|$ real numbers are exchanged between the nodes and their respective bridge nodes. A pragmatic way to select the bridge node set $\mathcal{B}$ is to sort the nodes in decreasing order of their nodal degrees and retain the least number of top most $|\mathcal{B}|$ nodes satisfying the conditions in Lemma 6.1. In section 6.6, we show empirically that this method of selecting bridge nodes is able to significantly reduce the overall communication complexity of the algorithm.

The inter-node communication can be further optimized by executing only a finite number of ADMM iterations per M-step.[3] In a practical embodiment of the algorithm, running a single ADMM iteration per M-step is sufficient for the CB-DSBL to converge. As shown in Fig. 6.2, beyond two or three ADMM iterations per M-step, there is only a marginal improvement in the quality of solution as well the convergence speed. Fig. 6.3 shows that even with a single ADMM iteration per M-step, CB-DSBL typically converges quite rapidly to the centralized solution.

It is also noteworthy to mention that, in CB-DSBL, the nodes are allowed to exchange only their local estimates of the common hyperparameter $\boldsymbol{\gamma}$. Thus, the proposed algorithm is well suited for applications which require the nodes to keep their local signal estimates private.

[3]Even further reduction in internode communication is possible in subsequent rounds of ADMM ($r \geq 2$). Each node j needs to exchange only incremental changes in $\boldsymbol{\gamma}_j^r$, as the initial value $\boldsymbol{\gamma}_j^0$ is already available at the neighboring nodes from the first round of communication.

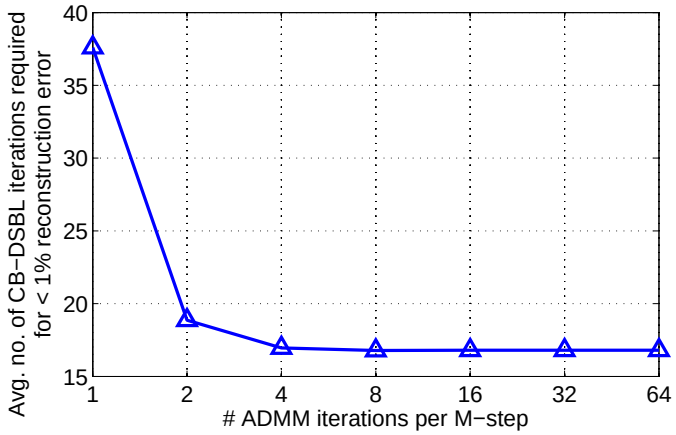


Figure 6.2: This plot illustrates the sensitivity of CB-DSBL's outer loop iterations to the number of ADMM iterations executed per M-step in the inner loop of the algorithm. Each point in the curve represents the average number of overall CB-DSBL iterations needed to achieve less than 1% signal reconstruction error for a given number of ADMM iterations executed in the inner loop. Simulation parameters used: $n = 100$, $m = 10$, $L = 10$, 5% sparsity, SNR = 30 dB and #trials = 100.

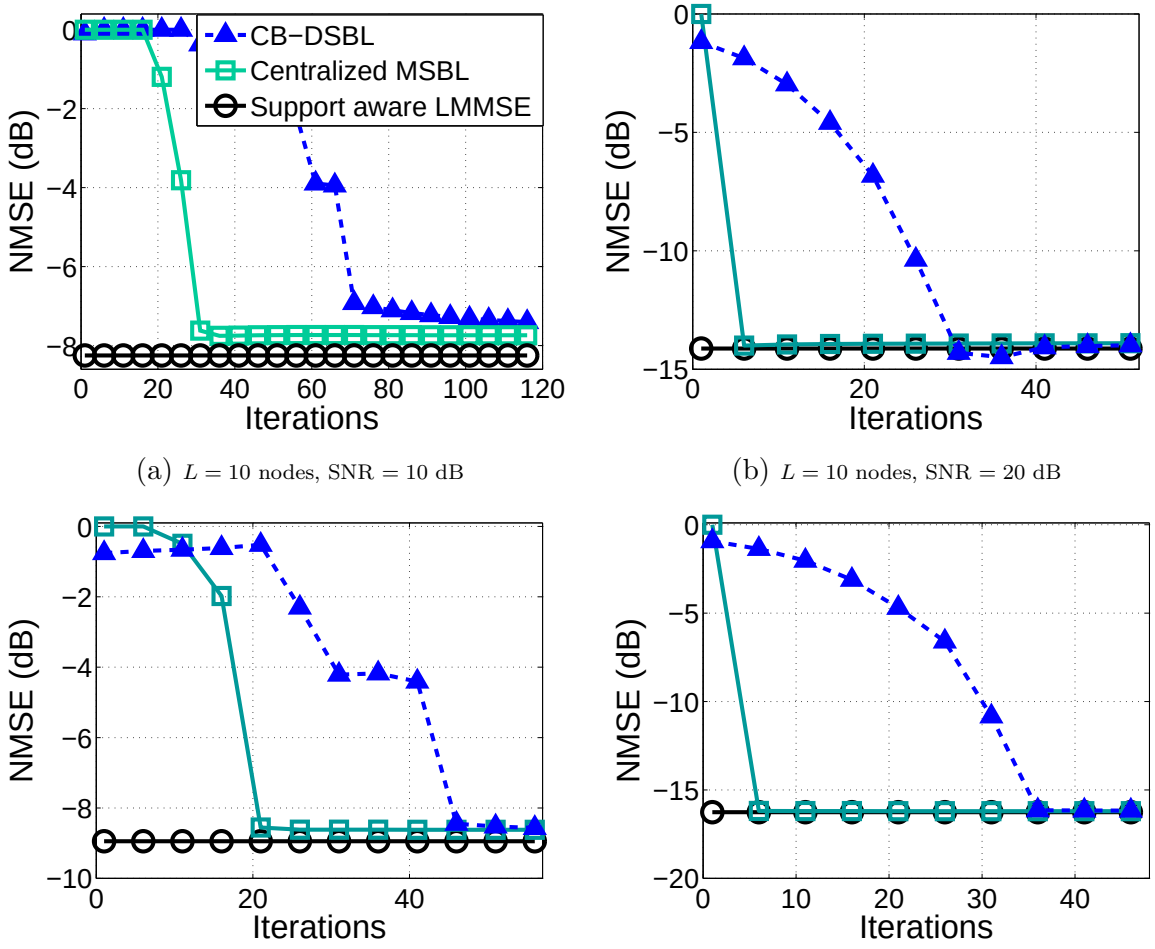
6.5.3 Convergence of ADMM Iterations

In this section, we analyze the convergence of the ADMM iterations (6.15), (6.16) and (6.17) derived for the M-step optimization in CB-DSBL. By doing so, we aim to highlight the effects of the bridge node set $\mathcal{B}$ and the augmented Lagrangian parameter ρ on the convergence of the ADMM iterations.

ADMM has been a popular choice for solving both convex [1, 3, 84, 109, 111] and more recently non-convex [112] optimization problems as well, in a distributed setup. In its classical form, ADMM solves the following constrained optimization problem:

$$\min_{\mathbf{x},\mathbf{z}} f(\mathbf{x}) + g(\mathbf{z})$$
$$\text{subject to } \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} = \mathbf{c}, \tag{6.18}$$

where $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{z} \in \mathbb{R}^m$ are the primal variables. The matrices $\mathbf{A}, \mathbf{B}$ and the vector $\mathbf{c}$



(a) $L = 10$ nodes, SNR $= 10$ dB

(b) $L = 10$ nodes, SNR $= 20$ dB

(c) $L = 20$ nodes, SNR $= 10$ dB

(d) $L = 20$ nodes, SNR $= 20$ dB

Figure 6.3: Convergence of decentralized CB-DSBL to centralized MSBL solution for different network sizes and SNRs. The CB-DSBL variant used here executes a single ADMM iteration per EM iteration. Other simulation parameters: $n = 50$, $m = 10$ and 10% sparsity.

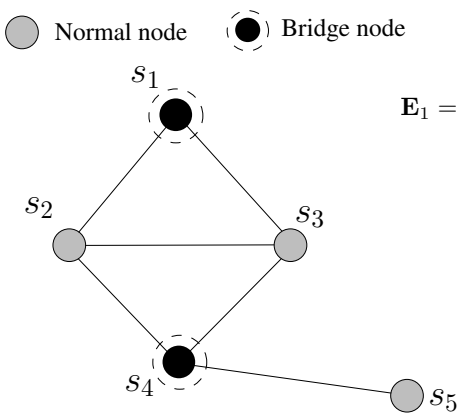
appearing in the linear equality constraint are of appropriate dimensions. The functions $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^m \to \mathbb{R}$ are convex with respect to $\mathbf{x}$ and $\mathbf{z}$, respectively. In [113], the authors have shown linear convergence rate for the classical ADMM iterations under the assumptions of strict convexity and Lipschitz gradient on one of f or g, along with full row rank assumptions for the matrix $\mathbf{A}$. However, in the ADMM formulation of a decentralized consensus optimization problem, the coefficient matrix $\mathbf{A}$ is seldom of full row

rank. In [114], the full row rank condition of $\mathbf{A}$ was relaxed and linear rate of convergence was established for decentralized ADMM iterations for a generic convex optimization with linear consensus constraints similar to (6.10). In [115], the convergence of ADMM for solving an average consensus problem has been analyzed for both noiseless and noisy communication links. In both [114] and [115], the secondary variables indicated by the entries of $\mathbf{z}$ have a one to one correspondence with the communication links between the network nodes. However, such a bijection is not valid for the bridge variables used in our work for enforcing consensus between the primal variables. Due to this, the convergence results of [114, 115] are not applicable to our case. In the sequel, we present the analysis of the convergence of decentralized ADMM iterations for the bridge node based internode communication scheme.

We start by defining block matrices $\mathbf{E}_1 = \mathbf{C}_1 \otimes \mathbf{I}_n$ and $\mathbf{E}_2 = \mathbf{C}_2 \otimes \mathbf{I}_n$ of sizes $nN_C \times nL$ and $nN_C \times n|\mathcal{B}|$, respectively. The rows of $\mathbf{C}_1$ and $\mathbf{C}_2$ encode the N_C equality constraints in (6.11) such that if the i^{th} equality constraint is $\boldsymbol{\gamma}_j = \boldsymbol{\gamma}_{b_k}$, $b_k \in \mathcal{B}$, then $\mathbf{C}_1(i,j) = 1$ and $\mathbf{C}_2(i,k) = -1$; with the rest of the entries in the i^{th} row equal to zero. It can be shown that the minimum and maximum number of bridge nodes connected to any node in the network is the same as the minimum and maximum eigenvalues of $\mathbf{E}_1^T\mathbf{E}_1$, denoted by $\sigma^2_{\min}$ and $\sigma^2_{\max}$, respectively. Fig. 6.4 illustrates the construction of the block matrices $\mathbf{E}_1$ and $\mathbf{E}_2$ for an example network consisting of 5 nodes. Using the newly defined terms, the optimization problem in (6.11) can be rewritten compactly as

$$\min_{\boldsymbol{\gamma}_{\mathcal{J}}, \boldsymbol{\gamma}_{\mathcal{B}}} f(\boldsymbol{\gamma}_{\mathcal{J}}) \quad \text{s.t. } \mathbf{E}_1\boldsymbol{\gamma}_{\mathcal{J}} + \mathbf{E}_2\boldsymbol{\gamma}_{\mathcal{B}} = 0 \tag{6.19}$$

where $f : \mathbb{R}^{nL} \to \mathbb{R}$ denotes the objective function in (6.11), which depends only on $\boldsymbol{\gamma}_{\mathcal{J}}$.



$$\mathbf{E}_1 = \begin{bmatrix} \mathbf{I}_n & 0 & 0 & 0 & 0 \\ 0 & \mathbf{I}_n & 0 & 0 & 0 \\ 0 & \mathbf{I}_n & 0 & 0 & 0 \\ 0 & 0 & \mathbf{I}_n & 0 & 0 \\ 0 & 0 & \mathbf{I}_n & 0 & 0 \\ 0 & 0 & 0 & \mathbf{I}_n & 0 \\ 0 & 0 & 0 & 0 & \mathbf{I}_n \end{bmatrix} \quad \mathbf{E}_2 = \begin{bmatrix} -\mathbf{I}_n & 0 \\ -\mathbf{I}_n & 0 \\ 0 & -\mathbf{I}_n \\ -\mathbf{I}_n & 0 \\ 0 & -\mathbf{I}_n \\ 0 & -\mathbf{I}_n \\ 0 & -\mathbf{I}_n \end{bmatrix}$$

$$\mathbf{E}_1^T\mathbf{E}_1 = \begin{bmatrix} \mathbf{I}_n & 0 & 0 & 0 & 0 \\ 0 & 2\mathbf{I}_n & 0 & 0 & 0 \\ 0 & 0 & 2\mathbf{I}_n & 0 & 0 \\ 0 & 0 & 0 & \mathbf{I}_n & 0 \\ 0 & 0 & 0 & 0 & \mathbf{I}_n \end{bmatrix}$$

Figure 6.4: Construction of block matrices $\mathbf{E}_1$ and $\mathbf{E}_2$ for a sample 5 node network. The matrices $\mathbf{E}_1$ and $\mathbf{E}_2$ are together used to enforce the linear consensus constraints in (6.10), as shown in (6.19). Note the correspondence between the diagonal coefficients of $\mathbf{E}_1^T\mathbf{E}_1$ and the number of bridge node connections per node.

The augmented Lagrangian L_ρ corresponding to (6.19) can also be rewritten compactly as

$$\begin{aligned} L_\rho(\boldsymbol{\gamma}_\mathcal{J}, \boldsymbol{\gamma}_\mathcal{B}, \boldsymbol{\lambda}) &= f(\boldsymbol{\gamma}_\mathcal{J}) + \boldsymbol{\lambda}^T(\mathbf{E}_1\boldsymbol{\gamma}_\mathcal{J} + \mathbf{E}_2\boldsymbol{\gamma}_\mathcal{B}) \\ &+ \frac{\rho}{2}(\mathbf{E}_1\boldsymbol{\gamma}_\mathcal{J} + \mathbf{E}_2\boldsymbol{\gamma}_\mathcal{B})^T(\mathbf{E}_1\boldsymbol{\gamma}_\mathcal{J} + \mathbf{E}_2\boldsymbol{\gamma}_\mathcal{B}). \end{aligned} \tag{6.20}$$

By construction, the block matrix $\mathbf{E}_1$ has full column rank, as all its columns are mutually disjoint in support. However, $\mathbf{E}_1$ can be row rank deficient due to repeated rows caused by a node being connected to multiple bridge nodes, which is often the case. Theorem 6.1 below summarizes the convergence of the ADMM iterations (6.15), (6.16) and (6.17) to their fixed point. The result in Theorem 6.1 holds for any f that is strongly convex with strong convexity constant m_f, and has a Lipschitz continuous gradient with Lipschitz constant M_f.

Theorem 6.1. *Let* $\{\boldsymbol{\gamma}_\mathcal{J}^*, \boldsymbol{\gamma}_\mathcal{B}^*\}$ *and* $\boldsymbol{\lambda}^*$ *denote the unique primal and dual optimal solutions of (6.19), and vector* $\mathbf{u}$ *be constructed as* $\mathbf{u} = [(\mathbf{E}_2\boldsymbol{\gamma}_\mathcal{B})^T \ \ \boldsymbol{\lambda}^T]^T$ *(similarly for* $\mathbf{u}^r, \mathbf{u}^*$*). Then,*

it holds that

1. *The sequence* $\mathbf{u}^r$ *is Q-linearly*[4] *convergent to* $\mathbf{u}^*$*, i.e.,*

$$\|\mathbf{u}^{r+1} - \mathbf{u}^*\|_{\mathbf{G}} \leq \frac{1}{1+\delta}\|\mathbf{u}^r - \mathbf{u}^*\|_{\mathbf{G}} \tag{6.21}$$

where δ *is evaluated as*

$$\delta = \max_{\mu,\nu\geq 1}\left\{ min\left(\frac{2m_f}{\frac{\nu M_f^2}{\rho(\nu-1)\sigma_{min}^2} + \mu\rho\sigma_{max}^2}, \frac{\sigma_{min}^2}{\nu\sigma_{max}^2}, \frac{\mu-1}{\mu} \right) \right\}. \tag{6.22}$$

2. *The primal sequence* $\boldsymbol{\gamma}_{\mathcal{J}}^r$ *is R-linearly*[5] *convergent to* $\boldsymbol{\gamma}_{\mathcal{J}}^*$*, i.e.,*

$$\|\boldsymbol{\gamma}_{\mathcal{J}}^{r+1} - \boldsymbol{\gamma}_{\mathcal{J}}*\|_2 \leq \frac{1}{2m_f}\|\mathbf{u}^r - \mathbf{u}^*\|_{\mathbf{G}} \tag{6.23}$$

where $\|\cdot\|_{\mathbf{G}}$ *is the weighted norm with respect to the diagonal matrix* $\mathbf{G} = \text{diag}(\rho I_{n|\mathcal{B}|}, \rho^{-1} I_{N_C})$.

Proof. See Appendices H.3 and H.4. □

According to Theorem 6.1, the primal optimality gap $||\boldsymbol{\gamma}_{\mathcal{J}}^r - \boldsymbol{\gamma}_{\mathcal{J}}^*||_2$ decays R-linearly with each ADMM iteration. Moreover, since $\boldsymbol{\gamma}_{\mathcal{J}}^*$ is primal feasible, there is consensus among $\boldsymbol{\gamma}_j, j \in J$ upon convergence, implying that each node effectively minimizes the centralized M-step cost function in (6.7).

[4] A sequence $x_k : \mathcal{Z}_+ \to \mathbb{R}$ is said to be *Q-linearly* convergent to L, if there exists a $\mu \in (0,1)$ such that $\lim_{k\to\infty} \frac{|x_{k+1}-L|}{|x_k-L|} = \mu$ [114].

[5] A sequence $x_k : \mathcal{Z}_+ \to \mathbb{R}$ is said to be *R-linearly* convergent to L, if there exists a Q-linearly convergent sequence y_k which converges to zero such that $\lim_{k\to\infty} |x_k - L| \leq y_k$.

6.5.4 Selection of the Augmented Lagrangian Parameter ρ

From (6.21) and (6.23) in Theorem 6.1, we observe that to optimize the decay of the primal optimality gap between $\boldsymbol{\gamma}_{\mathcal{J}}^{r}$ and $\boldsymbol{\gamma}_{\mathcal{J}}^{*}$ in each ADMM iteration, the augmented Lagrangian parameter ρ has to be chosen such that it maximizes δ in (6.22). Theorem 6.2 reveals the optimal value of ρ and the corresponding value of δ.

Theorem 6.2. *The optimal value of augmented Lagrangian parameter ρ which uniquely maximizes the δ as defined in* (6.22) *is given by*

$$\rho_{opt} = \frac{M_f}{\sigma_{max}\sigma_{min}} \left[\frac{\sqrt{(\kappa-1)^2 + 4\kappa\kappa_f^2} + (\kappa-1)}{\sqrt{(\kappa-1)^2 + 4\kappa\kappa_f^2} - (\kappa-1)} \right]^{\frac{1}{2}}. \tag{6.24}$$

The corresponding maximal value of δ is given by

$$\delta_{opt} = \frac{2}{\left(\kappa + 1 + \sqrt{(\kappa-1)^2 + 4\kappa\kappa_f^2}\right)} \tag{6.25}$$

where $\kappa_f = \dfrac{M_f}{m_f}$ represents the condition number of the objective function in (6.19) and $\kappa = \dfrac{\sigma_{max}^2}{\sigma_{min}^2}$ is the ratio of the maximum and minimum eigenvalues of $\mathbf{E}_1^T\mathbf{E}_1$.

Proof. See Appendix H.5. □

From (6.25), we observe that the convergence rate of the ADMM iteration in the M-step of the CB-DSBL algorithm depends upon two factors: κ and κ_f. κ close to its minimum value of unity results in faster convergence of the ADMM iterations. Since the ratio $\kappa = \sigma_{\max}^2 \backslash \sigma_{\min}^2$ is also equal to the ratio of maximum and minimum number of bridge nodes per node in the network, a rule of thumb for bridge node selection would be to ensure that each node is connected to more or less the same number of bridge nodes.

The convergence rate also depends upon κ_f, the parameter that determines how well conditioned the function f is. For the case where f is the objective function in (6.11), it is easy to check that $m_f = M_f = 2$ and $\kappa_f = 1$. Thus, specific to CB-DSBL, the optimal ADMM parameter ρ is given by $\rho_{\text{opt}} = \frac{2}{\sigma^2_{\min}}$ and the corresponding $\delta_{\text{opt}} = \frac{1}{\kappa+1}$. For a given network connectivity graph $\mathcal{G}$, this ρ_{opt} can be computed off-line and programmed in each node. As shown in Fig. 6.5, the average MSE and mean number of iterations vary widely with ρ, an inappropriate choice of ρ resulting in slow convergence and poor reconstruction performance. Also, the ρ_{opt} computed in (6.24) is very close to the ρ that results in both the fastest convergence as well as the lowest average MSE.

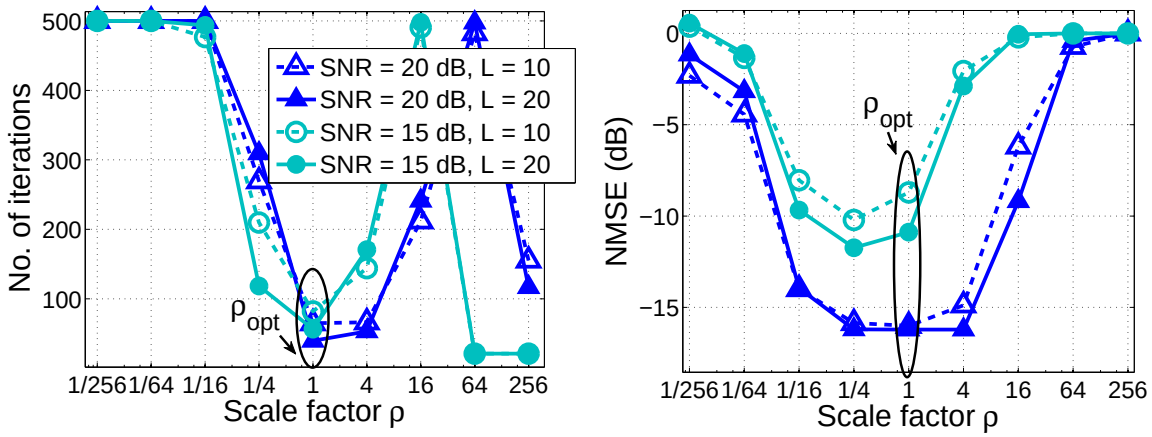


Figure 6.5: Left and right plots show the sensitivity of the number of iterations required for convergence and NMSE, respectively with respect to the ADMM parameter ρ. The scale factor $\rho = 1$ corresponds to ρ_{opt} in (6.24).

6.5.5 Computational Complexity of CB-DSBL

In this section, we discuss the computational complexity of the steps involved in a single iteration of the CB-DSBL algorithm. The local E-step requires $\mathcal{O}(n^2 + nm^2 + m^3)$ elementary operations at each node. The M-step is executed as multiple (say, $r_{\max}$) ADMM

iterations. A single ADMM iteration involves updating of the local hyperparameter estimate $\boldsymbol{\gamma}_j$ and Lagrange multipliers, which takes $\mathcal{O}(\zeta n)$ computations per node, ζ being the highest number of bridge nodes assigned per node in the network. Further, each bridge node $b \in \mathcal{B}$ has to perform an additional $\mathcal{O}(\zeta n)$ computations to update the local bridge parameters $\boldsymbol{\gamma}_b$ in every ADMM iteration. Thus, the overall computational complexity of a single CB-DSBL algorithm at each node is $\mathcal{O}(n^2 + nm^2 + m^3 + \zeta n r_{\max})$. As desired, the computational complexity does not scale with L, i.e., the total number of nodes in the network.

6.5.6 Other CB-DSBL Variations

There are several alternatives to the aforementioned bridge node based ADMM technique that could potentially be used to solve the M-step optimization in (6.10). In this section, we present empirical results comparing the performance and communication complexity of four different variations of the proposed CB-DSBL algorithm based on (i) bridge node based ADMM [1] (ii) Distributed ADMM (D-ADMM) [3] (iii) Consensus averaging Method of Multipliers (CA-MoM) [2], and (iv) EXact firsT ordeR Algorithm (EXTRA) [4]. The first three ADMM implementations differ in their use of auxiliary variables in enforcing the consensus constraints in (6.10). Each of these decentralized algorithms is endowed with at least $\mathcal{O}(\frac{1}{k})$ convergence rate, where k stands for the iteration count. Besides these four, there are proximal gradient based methods [116,117] relying on Nesterov-type acceleration techniques which also offer linear convergence rates. However, these algorithms require the objective function to be bounded and involve multiple communication rounds per

iteration, which is of major concern in our setup. As shown in Fig. 6.6, the proposed CB-DSBL variant relying on the bridge node based ADMM scheme is the most communication efficient one.

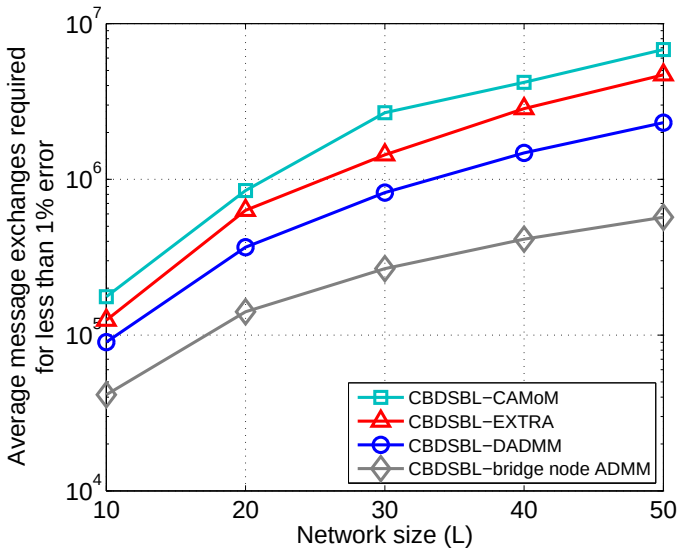


Figure 6.6: Comparison of the communication complexity of CB-DSBL variants based on 'bridge node' ADMM [1], CA-MoM [2], D-ADMM [3] and EXTRA [4] algorithms. The plot shows the average number of messages exchanged between nodes in order to achieve less than 1% signal reconstruction error (-20 dB NMSE). Other simulation parameters: $n = 50$, $m = 10$, 10% sparsity, SNR $= 30$ dB, # trials $= 500$.

6.5.7 Implementation Issues

The CB-DSBL algorithm can be seen as a decentralized EM algorithm to find the ML estimate of the hyperparameters $\boldsymbol{\gamma}$ of a sparsity inducing prior. CB-DSBL, not surprisingly, also inherits the property of the EM algorithm of converging to a local maximum of the ML cost function $\log p(\mathbf{Y};\boldsymbol{\gamma})$. However, getting trapped in a local maximum is not an issue, as it has been shown in [41] that all local maxima of the $\log p(\mathbf{Y};\boldsymbol{\gamma})$ are at most m-sparse and hence qualify as reasonably good solutions to our original sparse model estimation problem. In our work, we initialize the EM algorithm with $\boldsymbol{\gamma}$ whose all entries are close

to zero.

In practice, hard thresholding of $\boldsymbol{\gamma}$ is required to identify the nonzero support set. In this work, we remove all coefficients from the active support set for which $\boldsymbol{\gamma}(i), 1 \leq i \leq n$ is below the local noise variance. It must be noted that if the local noise variance at each node is unknown, it can also be estimated along with $\boldsymbol{\gamma}$ within the EM framework, as discussed in [32].

Another common issue is that of the wide variation in the energy of the nonzero entries of $\mathbf{x}_j$ across the network. Specifically, in distributed event classification by sensors of different types [97], each sensor node may employ its own distinct sensing modality and hence may perceive a different SNR. In such cases, a preconditioning step which normalizes the local response vector to unit energy is recommended for fast convergence of the CB-DSBL algorithm. The final signal estimates can be re-adjusted to undo the pre-conditioning.

6.6 Simulations

In this section, we present simulation results to examine the performance and complexity aspects of the proposed CB-DSBL algorithm when compared with existing decentralized algorithms: DRL-1 [84], DCOMP [92] and DCSP [107]. The centralized MSBL [32] is also included in the study as a performance benchmark for CB-DSBL. Since DRL-1 [84] has been shown to have superior performance compared to the mixed ℓ_1-ℓ_2 norm penalty based algorithms, we skip ADM-MMV [39] from our comparisons. The CB-DSBL variant considered here executes two ADMM iterations in the inner loop for every EM iteration in the outer loop. The value of the augmented Lagrangian parameter, ρ, is chosen according to (6.24). For each experiment, the set $\mathcal{B}$ of bridge nodes is selected as described in section

6.5.2. The local measurement matrices $\mathbf{\Phi}_j$ are chosen to be normalized Gaussian random matrices. The nonzero signal coefficients are sampled independently from the Rademacher distribution, unless mentioned otherwise. For each trial, the connections between the nodes are assumed according to a randomly generated Erdös-Renyi graph with a node connection probability of 0.8. In the final step of MSBL and CB-DSBL algorithms, the active support is identified by element-wise thresholding the local hyperparameter vector $\boldsymbol{\gamma}_j$ at node j using the threshold $4\sigma_j^2$, where σ_j^2 denotes the local measurement noise variance.

6.6.1 Performance versus SNR

In the first set of experiments, we compare the normalized mean squared error (NMSE) and the normalized support error rate (NSER) of different algorithms for a range of SNRs. The support-aware LMMSE estimator sets the MSE performance benchmark for all the support agnostic algorithms considered here. The NMSE and NSER error metrics are defined as

$$\begin{aligned} \text{NMSE} &= \frac{1}{L}\sum_{j=1}^{L}\frac{||\mathbf{x}_j - \hat{\mathbf{x}}_j||_2^2}{||\mathbf{x}_j||_2^2} \\ \text{NSER} &= \frac{1}{L}\sum_{j=1}^{L}\frac{|\mathcal{S}\backslash\hat{\mathcal{S}}_j| + |\hat{\mathcal{S}}_j\backslash\mathcal{S}|}{|\mathcal{S}|} \end{aligned}$$

where $\mathcal{S}$ is the true common support and $\hat{\mathcal{S}}_j$ is the support estimated at node j. The network size is fixed to $L = 10$ nodes. As seen in Fig. 6.7, CB-DSBL matches the performance of centralized MSBL in all cases. For higher SNR ($\geq$ 15 dB), it can be seen that both MSBL and proposed CB-DSBL are MSE optimal. CB-DSBL also outperforms DRL-1 and DCOMP in terms of both MSE and support recovery. This is attributed to the fact that the Gaussian prior used in CB-DSBL with its alternate interpretation as a parameterized

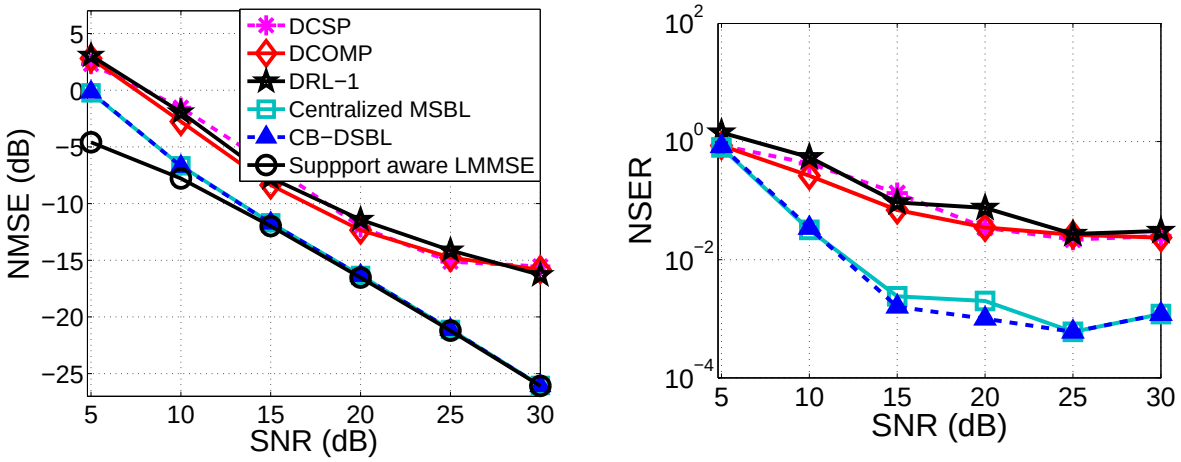
variational approximation of the sparsity inducing Student's t-distribution is more capable of inducing sparsity in comparison to the sum-log-sum penalty used in DRL-1. The poor performance of DCOMP is primarily due to its sequential approach towards support recovery which prevents any corrections to be applied to the support estimate at each step of the algorithm. Contrary to [107], DCSP fails to perform better than DCOMP. This is because DCSP works only when the number of measurements exceeds $2k$, where k is the size of the nonzero support.

6.6.2 Tradeoff between Measurement Rate and Network Size

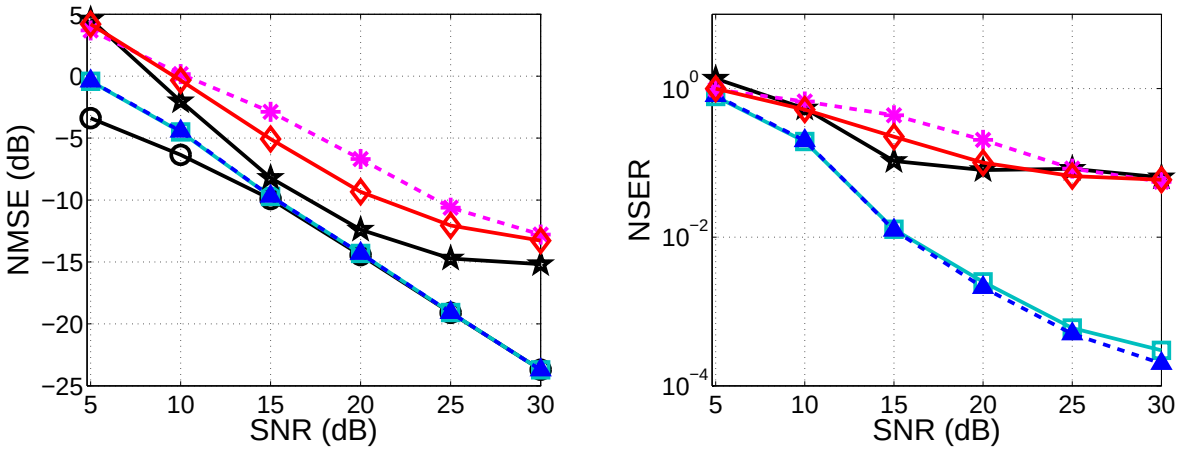
In the second set of experiments, we characterize the NMSE phase transition of the different algorithms in the L-(m/n) plane to identify the minimum measurement rate (m/n) needed to ensure less than 1% signal reconstruction error (or, NMSE ≤ -20 dB), for different network sizes (L), and a fixed sparsity rate $(k/n = 0.1)$. As shown in Fig. 6.8, for the same network size, CB-DSBL is able to successfully recover the unknown signals at a much lower measurement rate compared to DRL-1, DCOMP and DCSP. This plot brings out the significant benefit of using collaboration between nodes and taking advantage of the JSM-2 model in reducing the number of measurements required per node for successful signal recovery.

6.6.3 Performance versus Measurement Rate (m/n)

In the third set of experiments, we compare the algorithms with respect to their ability to recover the exact support for different undersampling ratios. As seen in Fig. 6.9, for a similar network size, CB-DSBL is able to exploit the joint sparsity structure better than DCOMP, DCSP and DRL-1, and can correctly recover the support from significantly fewer



(a) Nonzero coefficients drawn from Rademacher distribution



(b) Nonzero coefficients drawn from Gaussian distribution

Figure 6.7: Left and right figures in the above plot the NMSE and NSER, respectively for different SNRs. Other simulation parameters: $L = 10$ nodes, $n = 50$, $m = 10$ and 10% sparsity.

number of measurements per node. Once again, CB-DSBL has identical support recovery performance as the centralized MSBL, which was one of our design goals.

6.6.4 Robustness to Random Node Failures

Here, we demonstrate empirically that increasing the number of bridge nodes in the CB-DSBL algorithm makes it more robust to random node failures. As shown in Fig. 6.10, by

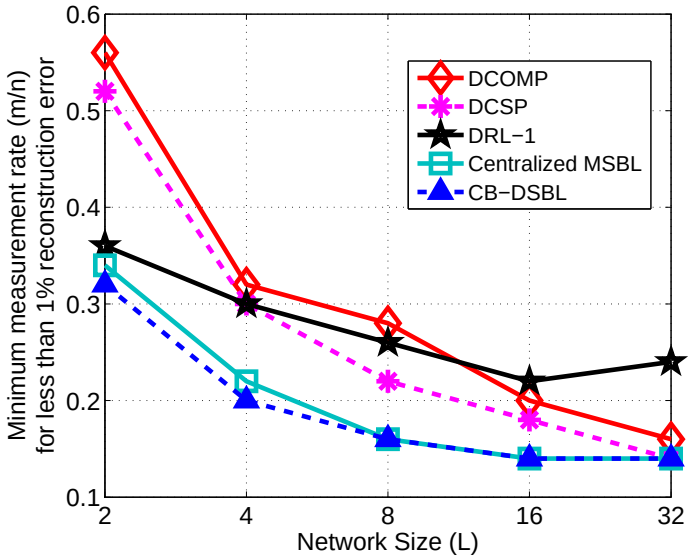


Figure 6.8: NMSE phase transition plots of different algorithms illustrating the dependence of minimum measurement rate required to guarantee less than 1% signal reconstruction error on the network size, for signal sparsity rate fixed at 10%. Other simulation parameters: $n = 50$ and SNR $= 30$ dB.

gradually increasing the density of bridge nodes in the network, the CB-DSBL algorithm is able to tolerate higher rates of node failures without compromising on signal reconstruction performance. More interestingly, only a relatively small fraction of nodes need to be bridge nodes ($< 10\%$ of the total network size) to ensure that CB-DSBL operates robustly in the face of random node failures.

6.6.5 Communication Complexity

Lastly, we compare the decentralized algorithms with respect to the total number of messages exchanged between the nodes during the estimation of the unknown vectors. As seen in Fig. 6.11, the greedy algorithms DCSP and DCOMP are the most communication efficient algorithms as these algorithms have the lowest per iteration communication complexity (see Table 6.1) and run for very few iterations. CB-DSBL and DRL-1 on the other

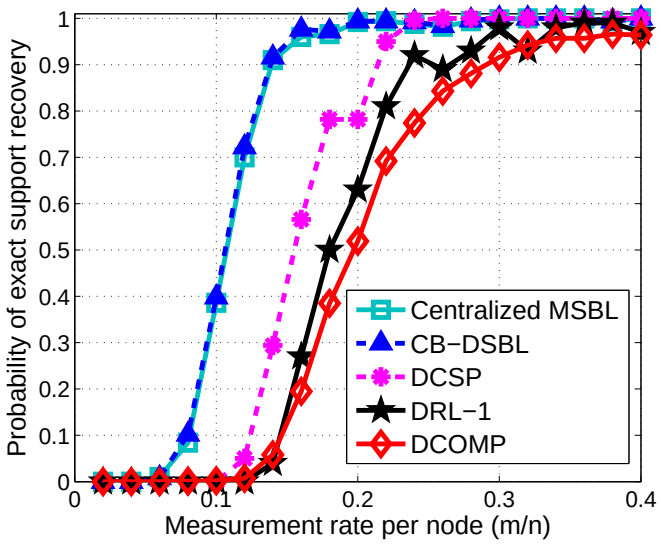


Figure 6.9: Exact support recovery probability versus measurement rate. Simulation parameters: $n = 50$, 10% sparsity, SNR = 15 dB and $L = 10$ nodes.

hand have higher communication complexity, with the proposed scheme requiring fewer overall message exchanges compared to DRL-1.

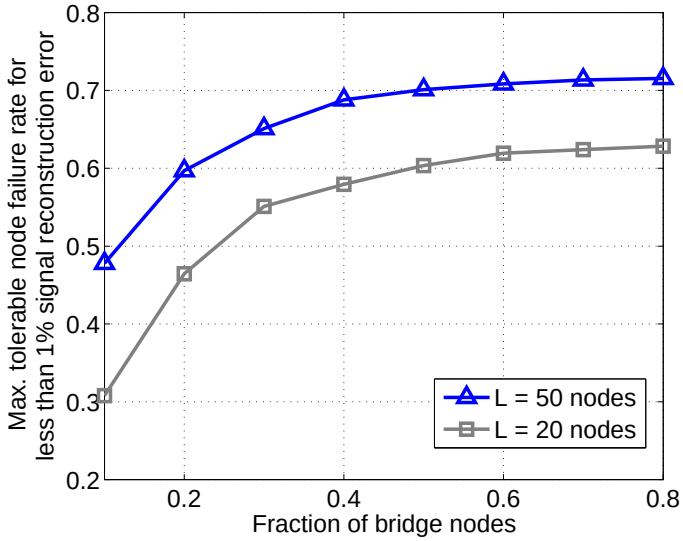


Figure 6.10: Plot illustrating the trade off between the density of bridge nodes and the robustness of the proposed CB-DSBL algorithm to random node failures. For a given fraction of bridge nodes (no. of bridge nodes / L), each point on the curve represents the average node failure rate that can be tolerated by CB-DSBL while still achieving less than 1% signal reconstruction error.

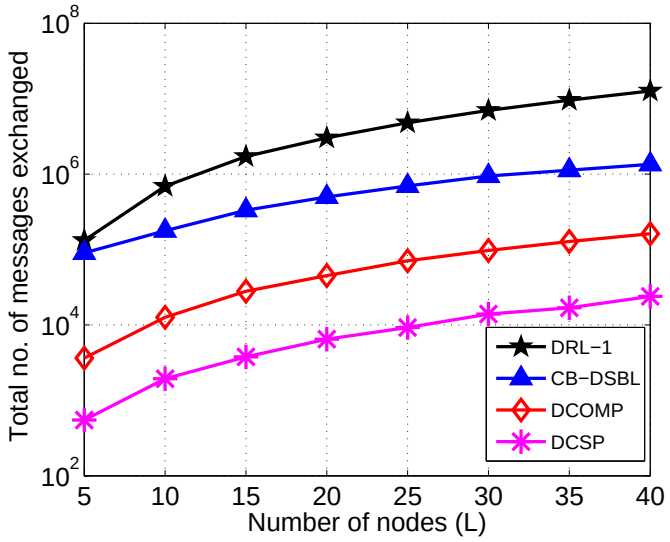


Figure 6.11: Average total number of messages exchanged between the nodes for different network sizes. Each message comprises a single real number. Simulation parameters: $n = 50$, $m = 10$, 10% sparsity, SNR = 20 dB.

6.7 Distributed Wideband Spectrum Sensing - A Real World Example

To find out how CB-DSBL and other competing distributed JSM-2 algorithms perform in a real world setup, we consider the wideband spectrum sensing problem described below. In wideband spectrum sensing, the goal is to efficiently estimate the occupancy of radio spectrum spanning a wide range of frequencies. Spectrum sensing is a crucial component in the implementation of cognitive radio (CR) networks. In a CR network, the secondary users perform spectrum sensing in order exploit the spectral opportunities which arise due to sparse utilization of the radio spectrum by the primary/licensed users [14, 15, 118–120].

6.7.1 Compressive Wideband Spectrum Sensing System Model

Consider a CR network comprising P primary users and L secondary users. The total available radio spectrum is partitioned into B frequency sub-bands of equal size. The secondary users sense the spectrum to determine the spectrum holes, i.e., the sub-bands that are not occupied by primary users. Most spectrum sensing techniques consist of two main steps [119, 120]. The first step is to obtain a frequency domain representation of the received wideband signal. This is followed by multiple sub-band level energy detection tests in order to estimate the wideband spectral occupancy of the primary users. Given below is the frequency domain representation of the sampled baseband signal received by the j^{th} secondary user:

$$\mathbf{r}_j^f = \sum_{i=1}^{P} \mathbf{D}_{p,j}^f \mathbf{s}_p^f + \mathbf{w}_j^f \tag{6.26}$$

where $\mathbf{D}_{p,j}^f = \text{diag}\left(\mathbf{h}_{p,j}^f\right)$ is an $n \times n$ diagonal channel gain matrix representing the wireless channel between primary user p and secondary user j. The diagonal matrix $\mathbf{D}_{p,j}^f$ alludes to

the frequency selective nature of channel across the sub-bands, but a flat fading behavior within sub-bands. The $n \times 1$ sized complex vectors: $\mathbf{s}_p^f$, $\mathbf{r}_j^f$ and $\mathbf{w}_j^f$ denote the frequency domain versions of the signal transmitted by the p^{th} primary user, and the signal and noise received at the j^{th} secondary user, respectively. Since the primary users collectively transmit in very few of the B sub-bands, each $\mathbf{s}_p^f, 1 \leq p \leq P$ can be modeled as an approximately sparse vector with most of its coefficients close to zero, and with a few large coefficients coinciding with the sub-band(s) occupied by the primary user p. Consequently, the vector $\mathbf{D}_{p,j}^f \mathbf{s}_p^f$ is also approximately sparse, and so is the summation $\sum_{i=1}^{P} \mathbf{D}_{p,j}^f \mathbf{s}_p^f$ in (6.26).

Since the received signal component $\mathbf{x}_j^f = \sum_{i=1}^{P} \mathbf{D}_{p,j}^f \mathbf{s}_p^f$ in (6.26) is a wideband signal, acquiring it at Nyquist or higher rate can be prohibitive. Often, due to bandwidth and sampling rate constraints, the secondary users resort to sliding-window, narrow-band processing, i.e., covering a small number of sub-bands at a time, in order to determine the band spectrum occupancy [121]. In a compressive sensing based approach [15], each secondary user implements an *Analog-to-Information-Converter* (AIC), which directly outputs low rate compressive measurements of the received wideband signal. Each secondary user acquires m compressive measurement samples in the form of an $m \times 1$ vector $\mathbf{y}_j^t$ as shown below,

$$\mathbf{y}_j^t = \mathbf{\Phi}_j \mathbf{r}_j^t, \qquad 1 \leq j \leq L, \tag{6.27}$$

where $\mathbf{r}_j^t$ is the discrete-time representation of the received signal sampled at Nyquist rate, f_S. $\mathbf{\Phi}_j$ is the $m \times n$ sized compressive sampling matrix. Since $m \ll n$, the effective sampling rate $(m/n)f_S$ is significantly lower than the conventional sampling rate f_S. Combining

(6.26) and (6.27), we can write,

$$\mathbf{y}_j^t \;=\; \mathbf{\Phi}_j\mathbf{F}^H\left(\sum_{i=1}^{P}\mathbf{D}_{p,j}^f\mathbf{s}_p^f\right)+\mathbf{\Phi}_j\mathbf{F}^H\mathbf{w}_j^f \;=\; \mathbf{\Phi}_j\mathbf{F}^H\mathbf{x}_j^f+\tilde{\mathbf{w}}_j^t \tag{6.28}$$

where $\mathbf{F}$ is the DFT matrix of order n, $\tilde{\mathbf{w}}_j^t$ is the effective measurement noise vector of length m, and $\mathbf{x}_j^f = \sum_{i=1}^{P}\mathbf{D}_{p,j}^f\mathbf{s}_p^f$ is an $n \times 1$ vector denoting the frequency domain representation of the received signal collectively transmitted by all primary users. Since the secondary users perceive a common spectral occupancy pattern associated with the primary users, the frequency domain vectors $\mathbf{x}_1^f, \mathbf{x}_2^f, \ldots, \mathbf{x}_L^f$ exhibit joint sparsity, and therefore can be recovered efficiently by a JSM-2 signal recovery algorithm.

6.7.2 Experimental Setup

For collecting experimental data, *Universal Software Radio Peripheral* (USRP) units (model N-210) were used to realize the primary and the secondary users. Due to the limited bandwidth of the USRP hardware, a scaled down version of the wideband spectrum sensing problem is considered, where the total available frequency band of 1 MHz, centered at 1.1 GHz, is divided into 128 non-overlapping, 7.8125 KHz wide sub-bands. A single USRP unit was configured to mimic five primary users transmitting QPSK modulated RF signal in sub-bands $\{(-59,-49),(-21),(-7),(2,3,4,5,6,7),(21)\}$ (see Fig. 6.12). A separate USRP unit was configured to collect $n = 256$ samples of down-converted baseband signal, sampled at twice the Nyquist rate. A Random Modulator Pre-Integrator (RMPI) based AIC [122] was simulated in Matlab by taking $\mathbf{\Phi}_j$ to be a column normalized Bernoulli ($p = 0.5$) random matrix with ± 1 entries, to generate $m = 32$ compressive

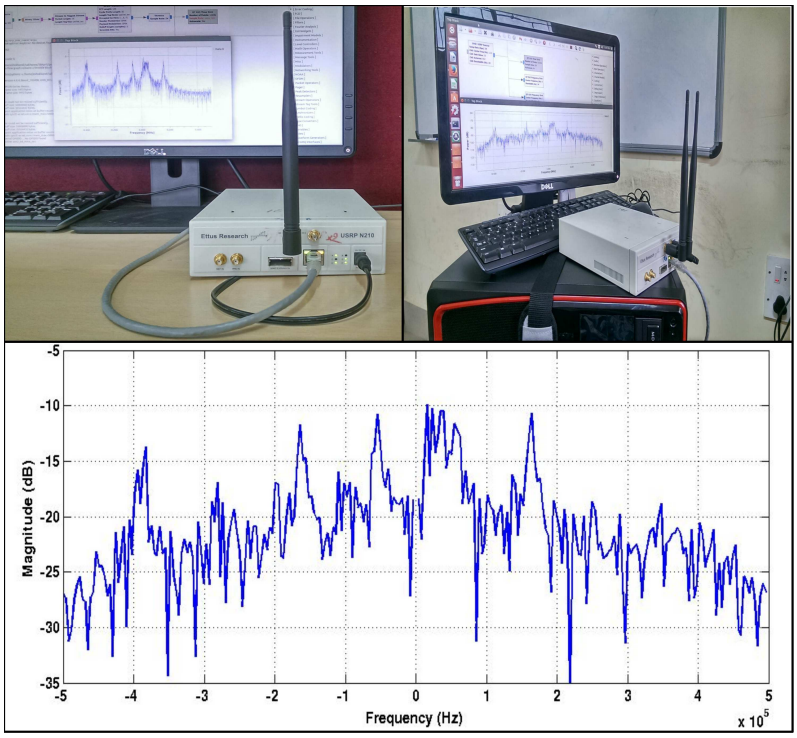


Figure 6.12: Top left: a USRP unit configured as a wideband transmitter, transmitting in 11 out of 128 frequency sub-bands. Top right: a mobile USRP station configured to capture the entire wideband signal at Nyquist sampling rate. Bottom: frequency spectrum of the down-converted baseband signal received by one of the secondary users. The five peaks correspond to the five active primary users.

measurements according to (6.27), which were then fed to the recovery algorithms. Multiple recordings were taken at 10 distinct spatial locations, one for each secondary user, in order to capture the effect of both temporal and spatial variations of the wireless channel. The SNR recorded at the CR nodes varied from -2.4dB to 7.8dB. The SNR here is defined as the ratio of total power in signal sub-bands to that in noise sub-bands. All performance metrics are averaged over 100 independent trials.

6.7.3 Performance Analysis

We compare the Receiver Operating Characteristics (ROC) of various decentralized JSM-2 recovery algorithms. For MSBL, CB-DSBL, and DRL-1, the ROC plots are obtained by

sweeping the threshold value used to identify the occupied frequency bins. For DCOMP and DCSP, the ROC plots are obtained by sweeping the sparsity size k input to these greedy algorithms. To account for spectral leakage, we adopt a pragmatic approach for computing the detection and false alarm probabilities. If a frequency bin is declared as active, it qualifies as a successful detection provided that the detected bin or one of its immediate left or right bins coincides with one of the signal sub-bands. Otherwise, a false alarm is declared. Fig. 6.13 compares the ROCs of different recovery schemes for measurement compression rate of 12.5%. Due to the hard-sparse nature of the output of DCOMP and DCSP algorithms, their false alarm rates never attain the maximal value of one, as reflected in the figure. In spectrum sensing, the goal is to achieve high detection probability while maintaining a low false alarm rate. In this regard, both centralized M-SBL and the proposed CB-DSBL outperform DCOMP, DCSP and DRL-1. In our experiments, both M-SBL and CB-DSBL are able to identify occupied spectrum bins and spectrum holes with 90% and 80% accuracy, respectively. In comparison, for DCSP and DRL-1, in order to achieve the same accuracy of detecting 90% of the occupied bins, these algorithms miss 35% of the unoccupied bins. Also, when fed with the correct sparsity level (k), the greedy algorithms DCSP and DCOMP have poor detection performance (less than 60% accuracy). In contrast, owing to their automatic relevance determination property [102], both CB-DSBL and M-SBL are able to correctly infer the sparsity level of the occupied spectrum directly from the measurements, resulting in superior performance.

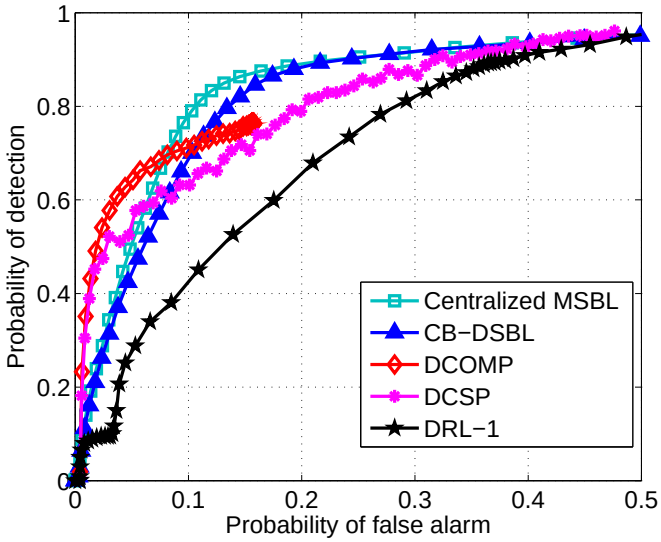


Figure 6.13: ROC performance for 12.5% measurement compression ratio, $L = 10$, and SNR ranging from -2.4dB to 7.8dB across the secondary users.

6.8 Chapter Summary

In this chapter, we proposed a novel iterative Bayesian algorithm called CB-DSBL for decentralized estimation of joint-sparse signals by multiple nodes in a network. The CB-DSBL algorithm is an ADMM based decentralized EM procedure to efficiently learn the parameters of a joint sparsity inducing signal prior which is shared by all the nodes, and is subsequently used in the MAP estimation of the local signals. From the simulation results, we can conclude that CB-DSBL outperforms existing decentralized algorithms: DRL- 1, DCOMP and DCSP, in terms of both signal reconstruction error as well as support recovery.

We also established the R-linear convergence of the underlying decentralized ADMM iterations. For the bridge node based network topology considered in this chapter, our ADMM convergence results are of independent interest and are applicable to any consensus driven optimization of a strongly convex and separable objective function.

Chapter 7

Decentralized Joint Sparse Signal Recovery under Communication Constraints

In which sharing becomes intelligent and decentralization becomes efficient...

7.1 Introduction

In this chapter[1], we revisit the topic of decentralized joint sparse signal recovery from the communication complexity standpoint. Decentralized algorithms often incur a high communication cost associated with the exchange of messages between the network nodes. Most existing decentralized algorithms for joint sparse signal recovery involve multiple exchanges of $\mathcal{O}(n)$ sized messages between the network nodes, where n is the dimension of the unknown sparse vectors. Since n is typically large, the $\mathcal{O}(n)$ scaling of message size

[1] This chapter is based on S. Khanna and C. R. Murthy, "Communication Efficient Decentralized Sparse Bayesian Learning of Joint Sparse Signals", *IEEE Transactions on Signal and Information Processing over Networks*, vol. 3, no. 3, pp. 617–630, Nov. 2016.

poses a severe bottleneck concerning the communication cost of in-network signal recovery. In practice, each network node needs to operate within a limited budget for time and power resources dedicated towards inter-node communications. Thus, any sizable reduction in the amount of information exchanged between the nodes is highly desirable as it has a direct positive impact on the operating lifetime of the network. Herein lies the challenge of devising schemes for reconstruction of sparse signals locally at their parent nodes, while exploiting their joint sparsity to improve the accuracy of their joint reconstruction, by using minimal communication between the nodes. Further, in many security-related applications, it is also vital to ensure that the messages exchanged between the nodes cannot be misused by a malicious node or an eavesdropper to reconstruct the signal vector of another node in the network. Thus, it is preferable to consider decentralized schemes which maintain the privacy of local measurements and local signal estimates.

To address the above challenges, in this chapter we propose an iterative, decentralized Bayesian algorithm called *Fusion Based Distributed Sparse Bayesian Learning* (FB-DSBL) in which the network nodes collaborate by exchanging highly compressed messages to learn a common joint sparsity inducing signal prior from local observations. The learnt signal prior is subsequently used by each node to compute the maximum a posteriori probability (MAP) estimate of its respective sparse vector. Since the inter-node communication can be expensive, the size of the messages exchanged between nodes is reduced substantially by exchanging only those local signal prior parameters which are associated with the nonzero support detected via multiple composite log-likelihood ratio tests. The average message size is empirically shown to be proportional to the information rate of the unknown vectors. The proposed Sparse Bayesian Learning (SBL) based distributed algorithm allows

the nodes to exploit the underlying joint sparsity of their local signal vectors. In turn, this enables the nodes to recover sparse vectors with a significantly lower number of measurements compared to the standalone SBL algorithm. Using Monte Carlo simulations, we demonstrate that the proposed FB-DSBL has superior MSE and support recovery performance compared to the existing decentralized algorithms that are of similar or higher communication complexity.

7.2 Background

7.2.1 Literature Survey

In Chapter 6, we surveyed the existing decentralized algorithms that have been proposed in the literature for joint sparse signal recovery. Except for DCSP [98] with its $\mathcal{O}(s \log n)$ sized message, where s denotes the size of the nonzero support of unknown joint sparse signals, all other distributed algorithms involve an exchange of $\mathcal{O}(n)$ sized messages per iteration between the nodes. In this chapter, we aim to fill a gap in the existing literature for a decentralized algorithm which is well endowed in terms of both signal recovery performance as well as low communication complexity.

7.2.2 Contributions

Our main contributions in this chapter are as follows:

1. We propose a novel, decentralized, iterative algorithm for in-network estimation of joint sparse vectors. We call our algorithm *Fusion based Decentralized Sparse Bayesian Learning* (FB-DSBL) to emphasize that the algorithm fuses the support estimates across the nodes, is decentralized and is based on the sparse Bayesian

learning algorithm. In the proposed algorithm, each node computes the MAP estimate of its local sparse vector, using a network-wide joint parameterized prior. The parameterized prior is itself learned on-the-fly using highly compressed messages exchanged over the network. The combined effect of exchanging compressed messages and using a joint parameterized prior is accurate reconstruction of the sparse vectors at the individual nodes using far fewer measurements compared to their independent reconstruction at every node.

2. In order to reduce the communication complexity, we propose a scheme to reduce the size of the messages exchanged between the nodes. Each node shares with its single-hop neighbors, only those components of the joint parameterized prior which are associated with the active locations in the local instantaneous estimate of the true support. These active locations are identified via multiple log-likelihood ratio tests. We also propose a scheme for refinement of the local estimates of the joint prior by using the compressed messages received from the single-hop neighbors. We show empirically that FB-DSBL requires the exchange of only $\mathcal{O}(s \log n)$ sized messages between the nodes, where s is the size of the nonzero support.

3. We show that FB-DSBL can be analyzed under the stochastic approximation framework by interpreting its iterations as degenerate distributed Robbins-Monro iterations for finding a fixed point of a certain function. Based upon this interpretation, we also propose FB-DSBL†, a stochastic approximation inspired generalization of the FB-DSBL algorithm. Through simulations, we show that its performance is marginally better, but its convergence is slower, and the communication cost is higher compared to the FB-DSBL algorithm.

Unlike DCSP and DCOMP, FB-DSBL does not require the actual sparsity level to be known apriori, nor does it require the exchange of the raw measurements or sparse vector estimates between the nodes. Thus, it is suitable for applications where data privacy needs to be protected.

7.3 System Model

The system model remains unchanged from section (6.3). A network consisting of L computing nodes is considered where the nodes are enumerated by the index set $\mathcal{J} = \{1, 2, \ldots, L\}$. The network connectivity is represented by an undirected graph $\mathcal{G}$ whose vertices have a one-to-one correspondence with the network nodes. An edge between the i^{th} and j^{th} vertices of $\mathcal{G}$ represents an error free communication link between node i and node j. Each node $j \in \mathcal{J}$ communicates only with its single-hop neighboring nodes, belonging to an index set denoted by $\mathcal{N}_j$.

Each sensor node j is interested in estimating an unknown s-sparse vector $\mathbf{x}_j \in \mathbb{R}^n$ from its $m(\ll n)$ noisy linear measurements $\mathbf{y}_j \in \mathbb{R}^m$, generated according to the local measurement model:

$$\mathbf{y}_j = \mathbf{\Phi}_j \mathbf{x}_j + \mathbf{w}_j, \quad j \in \mathcal{J}. \tag{7.1}$$

Here, $\mathbf{\Phi}_j$ is an $m \times n$ measurement matrix which is assumed to be known locally at node j. The measurement noise vector w_j is assumed to be zero mean Gaussian with known covariance matrix $\sigma_j^2 \mathbf{I}_m$. The sparse vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L$ are assumed to be *joint sparse* i.e., their nonzero coefficients belong to the same but unknown index set $\mathcal{S} \subset [n]$, with $|\mathcal{S}| = s$. In accordance with JSM-2 [38], the nonzero coefficients are uncorrelated within and across the vectors.

The goal here is to estimate the local sparse vector $\mathbf{x}_j$ at node j, for each $j \in \mathcal{J}$, in a decentralized fashion. Decentralized processing here implies that each node is capable of processing its local measurements and is allowed to exchange messages with only its single neighbors. In particular, we seek to devise algorithms that (a) exploit the joint sparsity structure of the vectors to be recovered; and (b) require minimal communication overhead between the nodes.

7.3.1 Sparse Bayesian Learning - A Recap

We now briefly review the Sparse Bayesian Learning (SBL) framework [41] which lies at the core of our proposed decentralized algorithm. The SBL algorithm is suitable when each node j estimates its local unknown sparse vector $\mathbf{x}_j$ in a standalone fashion, using only local measurements $\mathbf{y}_j$ generated according to (7.1). In SBL, a MAP estimate of the unknown sparse vector $\mathbf{x}_j$ is sought by imposing a fictitious parameterized signal prior on $\mathbf{x}_j$ as shown below.

$$p(\mathbf{x}_j;\boldsymbol{\gamma}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\boldsymbol{\gamma}(i)}} \exp\left(-\frac{|\mathbf{x}_j(i)|^2}{2\boldsymbol{\gamma}(i)}\right). \tag{7.2}$$

The hyperparameter $\boldsymbol{\gamma}(i)$ models the variance of $\mathbf{x}_j(i)$. The unknown hyperparameter vector $\boldsymbol{\gamma} \in \mathbb{R}_+^n$ is chosen to maximize the Bayesian evidence $p(\mathbf{y}_j;\boldsymbol{\gamma})$, which is equivalent to *maximum likelihood* (ML) estimation of $\boldsymbol{\gamma}$. The ML estimate of $\boldsymbol{\gamma}$ is obtained via the iterative Expectation-Maximization (EM) procedure, comprising the following two steps:

$$\begin{aligned} &\text{E-step:} \quad Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^k) = \mathbb{E}_{\mathbf{x}_j|\mathbf{y}_j,\boldsymbol{\gamma}^k}[\log p(\mathbf{y}_j, \mathbf{x}_j;\boldsymbol{\gamma})] \\ &\text{M-step:} \quad \boldsymbol{\gamma}^{k+1} = \underset{\boldsymbol{\gamma}\in\mathbb{R}_+^n}{\arg\max}\, Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^k), \end{aligned} \tag{7.3}$$

where k denotes the EM iteration index. The expectation in the E-step is evaluated using the a posteriori probability density $p(\mathbf{x}_j|\mathbf{y}_j, \boldsymbol{\gamma}^k)$, with mean $\boldsymbol{\mu}_j^k$ and covariance $\boldsymbol{\Sigma}_j^k$. According to the LMMSE theory [47], $\boldsymbol{\mu}_j^k$ and $\boldsymbol{\Sigma}_j^k$ are given by

$$\begin{aligned} \boldsymbol{\Sigma}_j^k &= \left(\left(\boldsymbol{\Gamma}^k\right)^{-1} + \frac{\boldsymbol{\Phi}_j^T \boldsymbol{\Phi}_j}{\sigma_j^2} \right)^{-1} \\ \boldsymbol{\mu}_j^k &= \sigma_j^{-2} \boldsymbol{\Sigma}_j^k \boldsymbol{\Phi}_j^T \mathbf{y}_j, \end{aligned} \tag{7.4}$$

where $\boldsymbol{\Gamma} = \mathrm{diag}(\boldsymbol{\gamma})$. The M-step optimization yields the following update rule for $\boldsymbol{\gamma}$:

$$\boldsymbol{\gamma}^{k+1} = \left(\boldsymbol{\mu}_j^k\right)^2 + \mathrm{diag}\left(\boldsymbol{\Sigma}_j^k\right). \tag{7.5}$$

By repeatedly iterating between (7.4) and (7.5), the EM algorithm converges to a local maximum ($\hat{\boldsymbol{\gamma}}_{\mathrm{ML}}$) of the ML cost $\log p(\mathbf{y}_j; \boldsymbol{\gamma})$. Once $\hat{\boldsymbol{\gamma}}_{\mathrm{ML}}$ is found, the MAP estimate of the local sparse vector $\mathbf{x}_j$ is given by the posterior mean $\boldsymbol{\mu}_j$ in (7.4) evaluated at $\hat{\boldsymbol{\gamma}}_{\mathrm{ML}}$. In [41], it has been shown that all local maxima of $\log p(\mathbf{y}_j; \boldsymbol{\gamma})$ are at most m-sparse and hence, the SBL algorithm promotes a sparse[2] MAP estimate of $\mathbf{x}_j$.

SBL can be easily extended to handle the estimation of multiple joint sparse vectors, resulting in its multiple measurement vector or MMV variant called MSBL [32]. As discussed in great detail in previous chapters, MSBL is a centralized algorithm which enforces joint sparsity of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L$ by assuming each vector $\mathbf{x}_j$ to be drawn independently from a common $\mathcal{N}(0, \boldsymbol{\Gamma})$ signal prior, where $\boldsymbol{\Gamma} = \mathrm{diag}(\boldsymbol{\gamma})$. Thus, the joint distribution of

[2]In practice, to obtain a hard-sparse estimate, estimated coefficients with variance at least one order of magnitude lower than noise variance are forced to zero and excluded from the active support.

$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L]$ is given by

$$p(\mathbf{X}; \boldsymbol{\gamma}) = \prod_{j=1}^{L} p(\mathbf{x}_j; \boldsymbol{\gamma}) = \prod_{j=1}^{L} \frac{1}{(2\pi)^{\frac{n}{2}} |\boldsymbol{\Gamma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\mathbf{x}_j^T \boldsymbol{\Gamma}^{-1} \mathbf{x}_j\right). \tag{7.6}$$

Once again, the ML estimate of $\boldsymbol{\gamma}$ is sought by maximizing the joint log-likelihood $\log p(\mathbf{Y}; \boldsymbol{\gamma}) = \sum_{j=1}^{L} \log p(\mathbf{y}_j; \boldsymbol{\gamma})$, where $p(\mathbf{y}_j; \boldsymbol{\gamma}) = \mathcal{N}(0, \sigma_j^2 \mathbf{I}_m + \boldsymbol{\Phi}_j \boldsymbol{\Gamma} \boldsymbol{\Phi}_j^T)$. As shown in [32], an iterative EM procedure can be derived to obtain the ML estimate of $\boldsymbol{\gamma}$ in the MMV setup using the following M-step update rule:

$$\boldsymbol{\gamma}^{k+1} = \frac{1}{L} \sum_{j=1}^{L} \left(\left(\mu_j^k\right)^2 + \operatorname{diag}\left(\boldsymbol{\Sigma}_j^k\right) \right). \tag{7.7}$$

Here, μ_j^k and $\boldsymbol{\Sigma}_j^k$ are evaluated according to (7.4). A naïve centralized implementation of MSBL would be a fusion center (FC) based approach, in which each node j communicates its $\mathcal{O}(m)$ sized local measurement vector $\mathbf{y}_j$ to the FC. The FC recovers all the joint sparse vectors and transmits the recovered vectors back to their respective nodes. In the case of distinct measurement matrices at the nodes, each node incurs an additional communication cost of sending its local $m \times n$ sized measurement matrix to the FC. The communication cost associated with the exchange of measurement matrices between the nodes and FC can become prohibitively high, especially if the measurement matrices vary with time. To keep the communication costs low, we seek a decentralized implementation of the MSBL algorithm. In previous chapter, we proposed a decentralized algorithm called CB-DSBL, for in-network estimation of the common hyperparameters $\boldsymbol{\gamma}$. Although closely matching MSBL in performance, its each iteration entailed an exchange of $n \times 1$ sized vectors between the single-hop neighboring nodes. In the following section, we propose a new decentralized extension of the SBL algorithm called FB-DSBL, which can estimate the common $\boldsymbol{\gamma}$ with

significantly reduced communication complexity.

7.4 FB-DSBL Algorithm

In this section, we propose the FB-DSBL algorithm, in which each network node learns the same joint sparsity inducing signal prior as the centralized MSBL algorithm, but in a decentralized fashion. Each node $j(\in \mathcal{J})$ maintains a separate SBL type signal prior $p(\mathbf{x}_j) \sim \mathcal{N}(0, \text{diag}(\boldsymbol{\gamma}_j))$ for its local sparse vector $\mathbf{x}_j$ of interest. Here, $\boldsymbol{\gamma}_j$ denotes the $n \times 1$ hyperparameters of the local SBL prior at node j. We introduce two quantities: *hard support estimate* and *soft support estimate*, which will be used in the sequel. The *hard support estimate* $\mathbf{b}_j \in \{0,1\}^n$, is an estimate of true binary support $\mathcal{S}$ of the sparse vector at node j. Its computation is discussed in Section 7.4.1. For a given $\mathbf{b}_j$ at node j, we define the *soft support estimate* $\mathbf{g}_j \triangleq \boldsymbol{\gamma}_j \circ \mathbf{b}_j$, where $\circ$ is the element-wise multiplication operator. To take advantage of the common support of the unknown signal vectors, each node j shares its soft support estimate $\mathbf{g}_j$ with its single-hop neighbors after every local SBL iteration. Unlike $\boldsymbol{\gamma}_j$, its censored copy $\mathbf{g}_j$ is highly sparse and hence, node j can share it with its one-hop neighbors at a substantially lower communication cost. The soft support estimates gathered from the neighboring nodes are subsequently used by node j to refine its local estimate of $\boldsymbol{\gamma}_j$ obtained from a standalone EM step.

We now outline the main steps involved in a single iteration of FB-DSBL:

Step-1 Each node j updates its local hyperparameters $\boldsymbol{\gamma}_j$ according to the local SBL update rule given by (7.4) and (7.5).

Step-2 Each node j generates a local hard support estimate $\mathbf{b}_j$. Node j broadcasts the soft support estimate $\mathbf{g}_j = \boldsymbol{\gamma}_j \circ \mathbf{b}_j$ to its single-hop neighbors in $\mathcal{N}_j$.

Step-3 Upon receiving $\mathbf{g}_l$ from all of its neighboring nodes l in $\mathcal{N}_j$, each node j fuses the hard support estimates $\mathbf{b}_l = supp(\mathbf{g}_l)$ to generate an improved binary estimate of the true support $\mathcal{S}$ denoted by $\mathbf{b}_j^{\text{fused}}$.

Step-4 Each node j updates its local hyperparameters $\boldsymbol{\gamma}_j$ to assimilate the available extrinsic information $\{\mathbf{g}_l \mid l \in \mathcal{N}_j\}$ conditioned on $\mathbf{b}_j^{\text{fused}}$.

Step-5 Repeat steps 1-4 until convergence.

Step 1 is the standard EM (SBL) iteration executed locally by each node in a standalone fashion using its local observations and its past local estimate of the hyperparameters. The remaining steps update the hyperparameters at each node based on the coarse information about the common support received from the neighboring nodes, for use in the next iteration of SBL. Since the network is connected, the neighborhoods $\mathcal{N}_j, j \in \mathcal{J}$ are partially overlapping, which allows FB-DSBL to exploit the joint sparsity across the entire network. The details of the computations involved in steps 2, 3 and 4 are fleshed out in the subsections 7.4.1, 7.4.2 and 7.4.3, respectively.

7.4.1 Computation of Local Hard Support Estimates

Since, under the JSM-2 model, a common $\boldsymbol{\gamma}$ is sought across the nodes, it is desirable that, at the end of the iterations, the estimates of $\boldsymbol{\gamma}_j(i)'$s across the nodes are equal if $i \in \mathcal{S}$, and 0 otherwise. Hence, we essentially want to exchange the values of the entries of $\boldsymbol{\gamma}_j^k$ that are likely to correspond to the true common support $\mathcal{S}$, which is a real valued vector of length s, where s is the sparsity of the vectors to be recovered. If the set of indices that contain the true support $\mathcal{S}$ with high probability, along with their corresponding hyperparameter values, is exchanged across the nodes, that would suffice for estimating/updating the local

estimate of $\boldsymbol{\gamma}$. Hence, the first step is to estimate the true support $\mathcal{S}$, which we denote by $\mathbf{b}_j$, at node j. We do this by setting up a hypothesis test for each of the indices, as described in the sequel.

Recall that $\mathbf{b}_j$ is an n length binary vector representing an estimate of the support at node j, and is computed locally at the node based on its measurements. The i^{th} coefficient of $\mathbf{b}_j$ is detected through a composite hypothesis test, by testing hypothesis $\mathcal{H}_0$, that the coefficient is zero, against hypothesis $\mathcal{H}_1$, that the coefficient is nonzero. Such an approach was first used in [124], where the authors proposed index-wise log-likelihood ratio tests (LLRTs) to prune the hyperparameters corresponding to the inactive coefficients for faster convergence of the EM iterations in the SBL algorithm in a centralized setup. Unlike in [124], our goal here is to use the LLRTs to identify the set of hyperparameters corresponding to the most likely support which will be exchanged with other nodes. For an index $i \in [n]$, we setup the following hypothesis test:

$$\begin{aligned} \mathcal{H}_0 : &\quad \mathbf{x}_j(i) = 0 \\ \mathcal{H}_1 : &\quad \mathbf{x}_j(i) \neq 0 \end{aligned} \tag{7.8}$$

or equivalently,

$$\begin{aligned} \mathcal{H}_0 : &\quad \boldsymbol{\gamma}_j(i) = 0 \\ \mathcal{H}_1 : &\quad \boldsymbol{\gamma}_j(i) > 0. \end{aligned} \tag{7.9}$$

To keep the notation light, the specificity of the hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ to node j and coefficient i is not highlighted, however, it is implicitly assumed. Unlike (7.8), the hypothesis test in (7.9) is a one sided test, and hence, a uniformly most powerful (UMP) test may

exist. For the i^{th} index, we decide in favor of $\mathcal{H}_1$, if

$$\log \frac{p(\mathbf{y}_j; \mathcal{H}_1)}{p(\mathbf{y}_j; \mathcal{H}_0)} \geq \theta \tag{7.10}$$

where θ is the detection threshold. Note that, the likelihood $p(\mathbf{y}_j; \boldsymbol{\gamma})$ equals $\mathcal{N}(0, \sigma_j^2 \mathbf{I}_m + \boldsymbol{\Phi}_j \tilde{\boldsymbol{\Gamma}}_{j,i}^k \boldsymbol{\Phi}_j^T + \boldsymbol{\gamma}_j(i) \boldsymbol{\Phi}_{j,i} \boldsymbol{\Phi}_{j,i}^T)$ under $\mathcal{H}_1$ and $\mathcal{N}(0, \sigma_j^2 \mathbf{I}_m + \boldsymbol{\Phi}_j \tilde{\boldsymbol{\Gamma}}_{j,i}^k \boldsymbol{\Phi}_j^T)$ under $\mathcal{H}_0$. Here, k denotes the current iteration index and, $\tilde{\boldsymbol{\Gamma}}_{j,i}^k = \text{diag}\big(\boldsymbol{\gamma}_j^k(1), \ldots, \boldsymbol{\gamma}_j^k(i-1), 0, \boldsymbol{\gamma}_j^k(i+1), \ldots, \boldsymbol{\gamma}_j^k(n)\big)$, with the $(i,i)^{\text{th}}$ diagonal entry set equal to zero. By substituting the likelihood functions and simplifying the LLR, it is shown in Appendix I.1 that the above hypothesis test decides in favor of $\mathcal{H}_1$ for index i if

$$\mathcal{T}_j^i(\mathbf{y}_j) = \left\{ \boldsymbol{\Phi}_{j,i}^T \left(\sigma_j^2 \mathbf{I}_m + \boldsymbol{\Phi}_j \tilde{\boldsymbol{\Gamma}}_{j,i}^k \boldsymbol{\Phi}_j^T \right)^{-1} \mathbf{y}_j \right\}^2 \geq \theta', \tag{7.11}$$

and decides in favor of $\mathcal{H}_0$ otherwise. Here, θ' denotes the detection threshold. Since the detection test metric $\mathcal{T}_j^i$ is independent of the parameter under test, i.e., $\boldsymbol{\gamma}_j(i)$, a UMP test for $\boldsymbol{\gamma}_j(i)$ exists. We normalize the detection test metric $\mathcal{T}_j^i$ to a standard chi-squared distributed random variable with a single degree of freedom under $\mathcal{H}_0$, resulting in the following Neyman-Pearson (NP) test:

$$\text{Decide } \mathcal{H}_1 \text{ for index } i \text{ if } \quad \bar{\mathcal{T}}_j^i = \frac{\left\{ \boldsymbol{\Phi}_{j,i}^T \left(\sigma_j^2 \mathbf{I}_m + \boldsymbol{\Phi}_j \tilde{\boldsymbol{\Gamma}}_j^k \boldsymbol{\Phi}_j^T \right)^{-1} \mathbf{y}_j \right\}^2}{\boldsymbol{\Phi}_{j,i}^T \left(\sigma_j^2 \mathbf{I}_m + \boldsymbol{\Phi}_j \tilde{\boldsymbol{\Gamma}}_{j,i}^k \boldsymbol{\Phi}_j^T \right)^{-1} \boldsymbol{\Phi}_{j,i}} \geq \bar{\theta}. \tag{7.12}$$

For a desired probability of false alarm (PFA) $\alpha \in [0,1]$, the normalized threshold $\bar{\theta}$ is computed offline as

$$\bar{\theta} = \left(\mathcal{Q}^{-1} \left(\frac{\alpha}{2} \right) \right)^2, \tag{7.13}$$

where $\mathcal{Q}(\cdot)$ is the standard Q-function. In the proposed scheme, α is an algorithm parameter which is common across all the network nodes.

In the k^{th} iteration of the proposed algorithm, node j generates its hard support estimate $\mathbf{b}_j^k$ by performing the NP test (7.12) for each index $i = 1$ to n. Subsequently, node j also computes the corresponding soft support estimate $\mathbf{g}_j^k = \boldsymbol{\gamma}_j^k \circ \mathbf{b}_j^k$, and broadcasts it to its single-hop neighboring nodes in $\mathcal{N}_j$.

Remark: In [125], the active coefficients are identified by performing component-wise maximization of marginalized likelihood across individual hyperparameters. In this scheme, node j declares i^{th} index as active if the ratio $q_{j,i}^2/s_{j,i}$ is greater than one, where $q_{j,i} = \boldsymbol{\Phi}_{j,i}^T \left(\sigma_j^2 \mathbf{I}_m + \boldsymbol{\Phi}_j \tilde{\boldsymbol{\Gamma}}_j^k \boldsymbol{\Phi}_j^T\right)^{-1} \mathbf{y}_j$ and $s_{j,i} = \boldsymbol{\Phi}_{j,i}^T \left(\sigma_j^2 \mathbf{I}_m + \boldsymbol{\Phi}_j \tilde{\boldsymbol{\Gamma}}_{j,i}^k \boldsymbol{\Phi}_j^T\right)^{-1} \boldsymbol{\Phi}_{j,i}$. The ratio $q_{j,i}^2/s_{j,i}$ is shown to be a proxy for SNR of the i^{th} component, thereby suggesting that the active indices must have associated SNR greater than 0 dB. In [126], this rule is generalized to $q_{j,i}^2/s_{j,i} > \eta$, where η is the predefined SNR of the i^{th} component. It is interesting to note that the ratio $q_{j,i}^2/s_{j,i}$ is the same as the chi-squared LLRT statistic (7.12) derived for the proposed FB-DSBL algorithm. In fact, if η is set equal to the NP threshold $\bar{\theta}$, as defined in (7.13), we obtain the LLRT based criterion for active support detection. Hence, the selection of the PFA parameter α in FB-DSBL offers a principled mechanism to control the sparsity of the hard support estimates, and consequently, the size of the messages exchanged between the network nodes.

7.4.2 Fusion of Hard Support Estimates from Local Neighborhood

In this subsection, we discuss how node j combines the soft support estimates $\left\{\mathbf{g}_l^k \mid l \in \mathcal{N}_j\right\}$ received from its single-hop neighbors to obtain a more accurate estimate of the true

binary support $\mathcal{S}$. In each iteration of the proposed algorithm, node j computes a fused binary support estimate called $\mathbf{b}_j^{\text{fused}}$ by applying an element-wise *majority rule* to the locally available binary support estimates $\mathbf{b}_l = supp(\mathbf{g}_l)$, where $l \in \mathcal{N}_j \cup \{j\}$. For index $1 \leq i \leq n$,

$$\mathbf{b}_j^{\text{fused},k}(i) \triangleq \begin{cases} 1 & \text{if } |\mathcal{A}_i^{j,k}| > \left\lceil \dfrac{|\mathcal{N}_j|}{2} \right\rceil \\ 0 & \text{otherwise} \end{cases} \tag{7.14}$$

where $\mathcal{A}_i^{j,k} \triangleq \left\{ l \in \mathcal{N}_j \cup \{j\} : \mathbf{b}_l^k(i) = 1, \mathbf{b}_l^k = supp(\mathbf{g}_l^k) \right\}$. The above fused estimate of $\mathcal{S}$ is subsequently used in section 7.4.3 to further refine the local hyperparameter estimate $\boldsymbol{\gamma}_j^k$ obtained from the local SBL update.

In the proposed algorithm, element-wise majority rule has been used to fuse the binary support estimates from the neighboring nodes. The primary reason for this choice is the lack of knowledge of the probability of detection P_{D} associated with the binary support estimates of the one-hop neighbors. Since the direct exchange of local measurements and measurement matrices between network nodes is not allowed, the optimal "*K-out-of-N*" fusion rule is not implementable. Further motivation for choice of majority fusion rule comes from [127] where it has been shown that the optimal fusion rule for the binary decisions from non-identical sensors has a similar structure as the majority rule.

7.4.3 Updating Local Hyperparameters using Extrinsic Information

Now, we present a scheme to update the hyperparameters estimate $\boldsymbol{\gamma}_j^k$ at node j by assimilating the available extrinsic information. The extrinsic information at node j refers to $\left\{ \mathbf{g}_l^k \mid l \in \mathcal{N}_j \right\}$, i.e., the collection of soft support estimates gathered from the neighboring nodes. From [32,128], the optimal scheme to update the local hyperparameter estimate $\boldsymbol{\gamma}_j^k$

at node j is to replace it with the average of the current hyperparameter estimates from all network nodes. However, the optimal scheme cannot be implemented as each node has access to only censored copies of the current hyperparameter estimates of its single-hop neighbors.

The problem of parameter estimation from censored measurements in a distributed setup has been studied in the literature [129, 130]. One of the ways to circumvent the non-availability of hyperparameters associated with indices under the $\mathcal{H}_0$ hypothesis is to replace the missing hyperparameters with their respective ML estimates. We observe that the probability of detection P_{D} for a one sided LLRT (similar to (7.12)) for i^{th} index is functionally dependent on $\boldsymbol{\gamma}_j(i)$. Then, under the assumption that the expression for $P_{\mathrm{D}}(\boldsymbol{\gamma}(i))$ is the same for all the neighboring nodes, the missing $\boldsymbol{\gamma}(i)$ can be chosen such that the associated P_{D} maximizes the likelihood of locally available binary decisions from the neighboring nodes regarding the i^{th} index. However, this scheme is not suitable for the current situation where P_{D} of a neighboring node also depends on its local measurement matrix and local measurement noise power, which are not available globally. Moreover, for a practical network topology, the number of available binary decisions can be insufficient for robust estimation of the missing parameters.

We now propose a suboptimal but pragmatic rule to update the current local hyperparameter estimate $\boldsymbol{\gamma}_j^k(i)$, $i \in [n]$. The proposed update rule is designed to approximate the MSBL update rule (7.7). Conditioned on the majority vote $\mathbf{b}_j^{\mathrm{fused},k}(i)$ computed according to (7.14), we propose separate update rules in the following two cases.

Case I: $\mathbf{b}_j^{\mathbf{fused},k}(i) = 0$ If the majority vote $\mathbf{b}_j^{\mathrm{fused},k}(i)$ suggests $\mathcal{H}_0$ at the i^{th} location in the k^{th} iteration, $\boldsymbol{\gamma}_j^k(i)$ at node j is set equal to the average of the estimated hyperparameter

of all nodes in $\mathcal{N}_j$, with censored hyperparameters replaced with zero. The updated local hyperparameter $\bar{\boldsymbol{\gamma}}_j^k(i)$ is given by:

$$\bar{\boldsymbol{\gamma}}_j^k(i) = \frac{\boldsymbol{\gamma}_j^k(i) + \sum_{l \in \mathcal{N}_j} \mathbf{g}_l^k(i)}{1 + |\mathcal{N}_j|}. \tag{7.15}$$

For index $i \notin \mathcal{S}$, since $\boldsymbol{\gamma}_j^k(i)$ finally converges to zero for all $j \in \mathcal{J}$, replacing the missing hyperparameters with zero turns out be a good approximation.

Case II: $\mathbf{b}_j^{\mathbf{fused},k}(i) = 1$ If the majority vote $\mathbf{b}_j^{\text{fused},k}(i)$ suggests $\mathcal{H}_1$ at the i^{th} location in the k^{th} iteration, we propose to set the hyperparameter $\boldsymbol{\gamma}_j^k(i)$ at node j to equal the average of its own hyperparameter $\boldsymbol{\gamma}_j^k(i)$ and the hyperparameters received from only those neighboring nodes that are in agreement with the majority vote. The updated local hyperparameter $\bar{\boldsymbol{\gamma}}_j^k(i)$ is given by:

$$\bar{\boldsymbol{\gamma}}_j^k(i) = \frac{\boldsymbol{\gamma}_j^k(i) + \sum_{l \in \mathcal{N}_j} \mathbf{g}_l^k(i)}{1 + \sum_{l \in \mathcal{N}_j} \mathbf{b}_l(i)}. \tag{7.16}$$

The selective averaging in (7.16) can be seen as an unbiased approximation of the hyperparameter update (7.7) used in the centralized MSBL algorithm by allowing only the neighboring nodes that are in agreement with the majority vote, $\mathbf{b}_j^{\text{fused},k}(i)$, to contribute to the average. Note that, for a fully connected network, when α equals one, the proposed FB-DSBL algorithm is tantamount to executing the M-step update (7.7) via a decentralized local averaging algorithm. Finally, we summarize the steps involved in proposed FB-DSBL algorithm as Algorithm 1. In Table 7.1, we provide a stepwise breakdown of the per iteration computational and communication complexity of the algorithm.

In the above exposition, a majority vote based selective averaging procedure has been

Algorithm 3 FB-DSBL: Fusion Based Decentralized Sparse Bayesian Learning

Input: $\{\mathbf{y}_j, \mathbf{\Phi}_j, \sigma_j\}_{j=1}^{L}$, and α **Initializations:** $k \leftarrow 1 \quad \boldsymbol{\gamma}_j^0 \leftarrow 10^{-3} \cdot \mathbf{1}_{n\times 1} \quad \forall\, j \in \mathcal{J} \quad \Delta = 2\epsilon$

while $(k < k_{\max})$ and $(\Delta > \epsilon)$ **do**

1a. ***Local E step:*** Each node j updates its posterior mean $\boldsymbol{\mu}_j^{k-1}$ and variance $\boldsymbol{\Sigma}_j^{k-1}$ according to (7.4).

1b. ***Local M step:*** Each node j updates it local hyperparameter vector: $\boldsymbol{\gamma}_j^k = \text{diag}(\boldsymbol{\Sigma}_j^{k-1}) + (\boldsymbol{\mu}_j^{k-1})^2$.

2a. Each node j generates hard support estimate $\mathbf{b}_j$ by performing index-wise LLRTs as shown in (7.12).

2b. Each node j computes the soft support estimate $\mathbf{g}_j^k = \boldsymbol{\gamma}_j^k \circ \mathbf{b}_j^k$ and broadcasts it to all the nodes in $\mathcal{N}_j$.

3. Each node j computes $\mathbf{b}_j^{\text{fused}}$ by fusing $\left\{\mathbf{b}_l \triangleq supp(\mathbf{g}_l)\right\}_{l\in\mathcal{N}_j}$ using the majority rule (7.14).

4. Each node j assimilates available extrinsic information by updating $\boldsymbol{\gamma}_j^k$ to $\bar{\boldsymbol{\gamma}}_j^k$ according to: For $1 \le i \le n$,

If $\mathbf{b}_j^{\text{fused}}(i) = 0$: $\bar{\boldsymbol{\gamma}}_j^k(i) = \frac{\boldsymbol{\gamma}_j^k(i) + \sum_{l\in\mathcal{N}_j} \mathbf{g}_l^k(i)}{1+|\mathcal{N}_j|}$

If $\mathbf{b}_j^{\text{fused}}(i) = 1$: $\bar{\boldsymbol{\gamma}}_j^k(i) = \frac{\boldsymbol{\gamma}_j^k(i) + \sum_{l\in\mathcal{N}_j} \mathbf{g}_l^k(i)}{1+\sum_{l\in\mathcal{N}_j} \mathbf{b}_l(i)}$

5. $\boldsymbol{\gamma}_j^k \leftarrow \bar{\boldsymbol{\gamma}}_j^k$, $\Delta \leftarrow \min_{j\in\mathcal{J}} \frac{||\boldsymbol{\gamma}_j^k - \boldsymbol{\gamma}_j^{k-1}||_2}{||\boldsymbol{\gamma}_j^{k-1}||_2}$ and $k \leftarrow k+1$.

end while

Output: For $1 \le j \le L$, $\hat{\mathbf{x}}_{j,\text{MAP}} \leftarrow \mu_j^k$

proposed which combines both the available censored soft samples and the associated hard decisions to estimate the index wise local average of the unknown variance parameter $\boldsymbol{\gamma}(i)$. The proposed updates can be understood as a solution to a theoretical formalism which we will now discuss independently in a simpler setup.

Consider a toy problem in which the goal is to find the average of L scalar random variables x_i, generated according to the model $x_i = x_{\text{true}} + w_i, 1 \le i \le L$, where x_{true} is an unknown nonnegative parameter and w_i is a zero mean measurement noise with an unknown probability distribution. Without any knowledge about the statistics of the measurement noise, a reasonable estimator of x_{true} is the sample average $\bar{x}$ which can be

computed by solving the optimization: $\bar{x} = \arg\min_x \frac{1}{L}\sum_{i=1}^{L}(x - x_i)^2$. Say, in addition to the soft samples x_i, we also have access to side information $\{b_i\}_{i\in[L]}$, where each b_i is an independently generated hard decision variable which takes values 0 or 1 to indicate wheather x_{true} is zero or nonzero, respectively. The hard decisions b_i could be in error, and as motivated from many practical setups similar to ours, the associated type-1 and type-2 error probabilities are unknown. A natural formalism which incorporates this extra side information for improved estimation of x_{true} is to solve the regularized optimization problem:

$$\hat{x} = \arg\min_x \frac{1}{L}\sum_{i=1}^{L}(x - x_i)^2 + \lambda x^2 \tag{7.17}$$

where a quadratic penalty x^2 has been introduced to attract the solution towards zero, with the strength of attraction governed by a regularization parameter $\lambda \geq 0$. Clearly, the value of λ must reflect the available side information $\{b_i\}_{i\in[L]}$. Intuitively, λ should increase with the number of hard decisions indicating x_{true} to be zero.

It is easy to check that (7.17) has a closed form solution: $\hat{x} = \bar{x}\left(\frac{1}{1+\lambda}\right)$. The multiplicative factor $\frac{1}{1+\lambda}$ causes shrinkage of the original unregularized solution $\bar{x}$, thereby accounting for the side information which may be suggesting x_{true} to be zero. To ensure that the shrinkage is proportional to the number of zero hard decisions, we suggest choosing λ as the ratio: $\lambda = \frac{Z}{NZ}$, where Z and $NZ = (L - Z)$ denote the number of hard decisions b_i equal to zero and nonzero, respectively. Note that this choice of λ allows the penalty term to completely disappear when all hard decisions suggest x_{true} to be nonzero, resulting in $\hat{x} = \bar{x}$. Next, we complicate the toy problem further by censoring those soft samples x_i, for which the associated b_i is zero. Due to the censoring of the soft samples, the solution $\hat{x}$ of the regularized optimization (7.17) can no longer be computed, as $\hat{x}$ depends

on $\bar{x}$ whose evaluation requires the uncensored values of all soft samples. To proceed, we evaluate $\hat{x} = \bar{x}\left(\frac{1}{1+\lambda}\right)$ with $\bar{x}$ replaced with its unbiased proxy, $\tilde{x} = \frac{1}{NZ}\sum_{i:b_i\neq 0} x_i$ computed only using uncensored soft samples. Thus, we obtain the regularized estmate $\hat{x} = \left(\frac{1}{NZ}\sum_{i:b_i\neq 0} x_i\right)\left(\frac{NZ}{Z+NZ}\right) = \frac{1}{L}\sum_{i:b_i\neq 0} x_i$, which interestingly is also the FB-DSBL update (7.15) with x_i's and b_i's representing the per index soft and hard support estimates, respectively.

To account for potential errors in the hard decisions b_i's, the regularized solution (7.17) is accepted only if $Z > NZ$, i.e., when the majority of the hard decisions indicate that x_{true} is zero, otherwise the unregularized but unbiased estimate $\tilde{x} = \frac{1}{NZ}\sum_{i:b_i\neq 0} x_i$ is accepted to be the solution. This also provides a theoretical underpinning for the other FB-DSBL update (7.16). In the absence of of knowledge of the type-1 and type-2 error probabilities, the majority rule based selection between the regularized and unregularized solutions turns out to be surprisingly effective in practice, as illustrated via simulations in Section 7.6.

Table 7.1: Computational & communication complexity analysis of a single iteration of FB-DSBL

Steps in FB-DSBL iteration	**Computational complexity per node**	**Communication complexity per node**
Local E-step (7.4)	$\mathcal{O}(n^2m + m^3)$	0
Local M-step (7.5)	$\mathcal{O}(n)$	0
Compute $\mathbf{b}_j, \mathbf{g}_j$ (7.12)	$\mathcal{O}(nm^3)$	0
Compute $\mathbf{b}_j^{\text{fused}}$ (7.14)	$\mathcal{O}(nL)$	$\mathcal{O}(sL\log n)$
Compute $\bar{\boldsymbol{\gamma}}_j$ (7.15, 7.16)	$\mathcal{O}(nL)$	0

Table 7.2: Comparison in terms of communication cost

Decentralized algorithm	**Total number of message exchanges per iteration**
FB-DSBL (proposed)	$\mathcal{O}(sL^2 \log n)$
DCSP [98]	$\mathcal{O}(sL^2 \log n)$
DCOMP [92]	$\mathcal{O}(nL^2)$
CB-DSBL [128]	$\mathcal{O}(nL^2)$
DRL-1 [84]	$\mathcal{O}(nL^2)$

7.4.4 Inter-node Communication

As discussed in Section 7.4.1, in every FB-DSBL iteration, each node j broadcasts its local soft support estimate $\mathbf{g}_j$ to its single-hop neighbors in $\mathcal{N}_j$. Although $\mathbf{g}_j$ is an n length vector, in practice, it is found to be a highly sparse vector i.e., most of its entries are equal to zero. From Fig. 7.1, it can be seen that for a fixed sparsity rate (s/n), the fraction of nonzero coefficients in the soft support estimate remains roughly constant with increasing signal dimension n, which is desirable. Further, from Fig. 7.2, it can be seen that the sparsity of the soft support estimates grows linearly with the sparsity rate (s/n) of the unknown vectors for fixed signal dimension n. Figures 7.1 and 7.2 together imply that by encoding the locations and magnitudes of only the nonzero entries of $\mathbf{g}_j$, the size of the messages exchanged amongst the network nodes can be restricted to $\mathcal{O}(s \log n)$. The additional $\log n$ bits are needed to encode the locations of the individual nonzero coefficients in the soft support estimates. Fig. 7.3 shows one such example of a frame structure for encoding $\mathbf{g}_j$ using $\mathcal{O}(s \log n)$ information bits. Compared to $\mathcal{O}(n)$ sized messages exchanged in existing decentralized joint sparse signal recovery algorithms [84,92,128], the proposed algorithm requires significantly lower communication bandwidth

for the recovery of the jointly sparse vectors at the individual nodes. [ht]

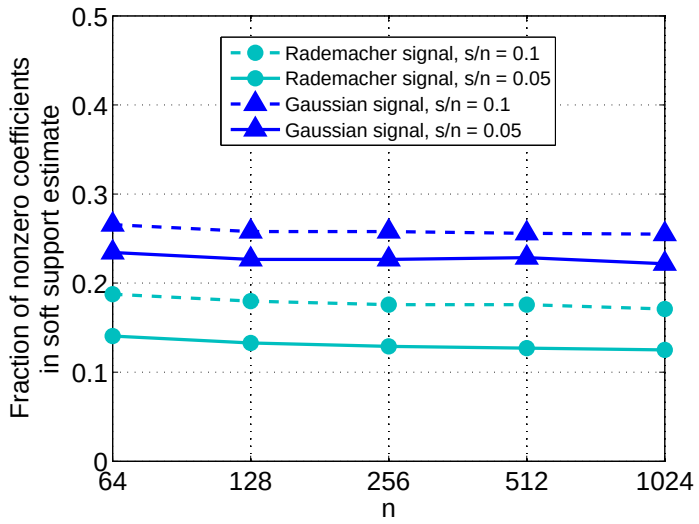


Figure 7.1: Fraction of nonzero entries in the soft support estimate $\mathbf{g}_j$ plotted against increasing value of n. The fraction of nonzero entries in $\mathbf{g}_j$ displayed here is averaged across nodes and iterations. The sparse soft support estimates are exchanged between the nodes in each FB-DSBL iteration. The flat curves indicate that the sparsity of the soft support estimates does not change with increasing signal dimension n. Here, SNR = 20 dB, the network size $L = 10$ nodes and the number of measurements at each node is $m = s \log n/s$.

7.5 A Stochastic Approximation View of FB-DSBL

We now present an interesting interpretation of the proposed FB-DSBL algorithm as a degenerate case of a stochastic approximation based distributed algorithm for maximum likelihood estimation of the unknown model parameters $\boldsymbol{\gamma}$. We show that the FB-DSBL updates are a special case of a distributed Robbins-Monro type stochastic approximation updates [131, 132], when the PFA parameter α is set to one. For $\alpha = 1$, the threshold $\bar{\theta}$ used in the index wise LLRT (7.12) is zero. This in turn causes the hard support estimates $\mathbf{b}_j$'s to be evaluated as $\mathbf{1}_n$, by each node. Thus, the nodes exchange their local hyperparameters estimates as is, without any censoring. In this case, the k^{th} iteration of

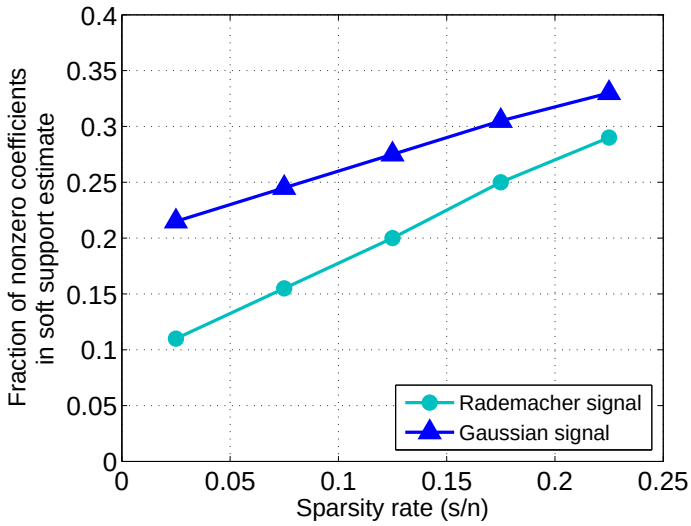


Figure 7.2: Fraction of nonzero entries in the soft support estimate $\mathbf{g}_j$ plotted against increasing value of sparsity rate (s/n) for fixed n equal to 200. The fraction of nonzero entries in $\mathbf{g}_j$ displayed here is averaged across nodes and iterations. Here, SNR = 20 dB, the network size $L = 10$ nodes and the number of measurements at each node is $m = s \log n/s$.

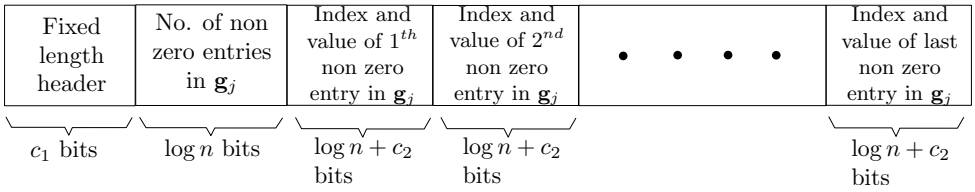


Figure 7.3: $\mathcal{O}(s \log n)$ sized example data packet format encoding the local soft support estimate $\mathbf{g}_j$, which is broadcast by node j. The scalar constant c_2 controls the quantization noise of nonzero entries of $\mathbf{g}_j$.

FB-DSBL at node j comprises the following two steps:

$$\text{Combined EM step:} \quad \tilde{\boldsymbol{\gamma}}_j^k = G_j(\boldsymbol{\gamma}_j^{k-1}, \mathbf{y}_j), \tag{7.18}$$

$$\text{Consensus step:} \quad \boldsymbol{\gamma}_j^k = \frac{1}{|\mathcal{N}_j|} \sum_{l \in \mathcal{N}_j} \tilde{\boldsymbol{\gamma}}_l^k. \tag{7.19}$$

The function $G_j : \mathbb{R}_+^n \times \mathbb{R}^m \to \mathbb{R}_+^n$ replaces the combination of the local E and M-steps given by (7.4) and (7.5), respectively, and is evaluated as

$$G_j(\boldsymbol{\gamma}, \mathbf{y}) = \operatorname{diag}\left[\left(\boldsymbol{\Gamma}^{-1} + \frac{\boldsymbol{\Phi}_j^T\boldsymbol{\Phi}_j}{\sigma_j^2}\right)^{-1} + \frac{1}{\sigma^4}\left(\boldsymbol{\Gamma}^{-1} + \frac{\boldsymbol{\Phi}_j^T\boldsymbol{\Phi}_j}{\sigma^2}\right)^{-1}\boldsymbol{\Phi}_j^T\mathbf{y}\mathbf{y}^T\boldsymbol{\Phi}_j\left(\boldsymbol{\Gamma}^{-1} + \frac{\boldsymbol{\Phi}_j^T\boldsymbol{\Phi}_j}{\sigma_j^2}\right)^{-1}\right]$$

where $\boldsymbol{\Gamma} = \operatorname{diag}(\boldsymbol{\gamma})$. Let us define the function $G : \mathbb{R}_+^{nL} \times \mathbb{R}^{mL} \to \mathbb{R}_+^{nL}$ as

$$G(\boldsymbol{\gamma}_\ominus, \mathbf{y}_\ominus) = \left[G_1(\boldsymbol{\gamma}_1, \mathbf{y}_1), G_2(\boldsymbol{\gamma}_2, \mathbf{y}_2), \ldots, G_L(\boldsymbol{\gamma}_L, \mathbf{y}_L)\right]^T, \tag{7.20}$$

where $\boldsymbol{\gamma}_\ominus = \left(\boldsymbol{\gamma}_1^T, \boldsymbol{\gamma}_2^T, \ldots, \boldsymbol{\gamma}_L^T\right)^T$ and $\mathbf{y}_\ominus = \left(\mathbf{y}_1^T, \mathbf{y}_2^T, \ldots, \mathbf{y}_L^T\right)^T$ are $nL \times 1$ and $mL \times 1$ sized concatenated vectors representing the local hyperparameter estimates and the local observations, respectively. With these new definitions, we can rewrite the network wide FB-DSBL iterations in a compact vector form,

$$\boldsymbol{\gamma}_\ominus^k = (\mathbf{W} \otimes \mathbf{I}_n) G\left(\boldsymbol{\gamma}_\ominus^{k-1}, \mathbf{y}_\ominus\right), \tag{7.21}$$

where $\otimes$ denotes the Kronecker product, and $\mathbf{W}$ is an $L \times L$ matrix with entry $\mathbf{W}_{jl} = \frac{1}{|\mathcal{N}_j|+1}$ when $l \in \mathcal{N}_j \cup \{j\}$ and 0 otherwise. The weight matrix $\mathbf{W}$ defined in this manner encodes the local averaging based consensus step described in (7.19). Other consensus strategies can be realized in this framework by appropriately choosing the entries of $\mathbf{W}$. For example, by modeling $\mathbf{W}$ as an appropriate random matrix, the vector representation in (7.21) can support the $\alpha \neq 1$ case. In the ensuing arguments, we elucidate the structural similarities between the FB-DSBL vector iterations in (7.21) and the distributed Robbins-Monro stochastic approximation algorithm proposed in [131].

We begin by observing that the EM iterations (7.4), (7.7) in centralized MSBL can be

rewritten as a fixed point iteration:

$$\begin{aligned} \boldsymbol{\gamma}^{k+1} &= \sum_{j=1}^{L} G_j(\boldsymbol{\gamma}^k, \mathbf{y}_j) \\ &= \langle G(\mathbf{1}_L \otimes \boldsymbol{\gamma}^k, \mathbf{y}_\ominus) \rangle \end{aligned} \tag{7.22}$$

where $\langle \mathbf{x} \rangle \triangleq (\mathbf{1}_L^T \otimes \mathbf{I}_n)\mathbf{x}$ is a vector in $\mathbb{R}^n$ for any vector $\mathbf{x}$ in $\mathbb{R}^{nL}$. From the convergence property of EM [133], the above fixed point iteration (7.22) converges to one of the stationary points of the log-likelihood $\log p(\mathbf{Y}, \boldsymbol{\gamma})$.

We now setup a distributed stochastic approximation algorithm to find the fixed point of $\langle G(\mathbf{1}_L \otimes \boldsymbol{\gamma}^k, \mathbf{y}_\ominus) \rangle$. As proposed in [131], let us consider a distributed algorithm in which each node j implements a "local step" followed by a "gossip step" as described below:

$$\text{Local step:} \quad \tilde{\boldsymbol{\gamma}}_j^k = \boldsymbol{\gamma}_j^{k-1} + a_k \left(G_j(\boldsymbol{\gamma}_j^{k-1}, \mathbf{u}_j^k) - \boldsymbol{\gamma}_j^{k-1} \right) \tag{7.23}$$

$$\text{Gossip step:} \quad \boldsymbol{\gamma}^k = \frac{1}{|\mathcal{N}_j|} \sum_{j \in \mathcal{N}_j} \tilde{\boldsymbol{\gamma}}_l^k \tag{7.24}$$

where a_k is an iteration dependent positive step size satisfying $\sum_{k=1}^{\infty} a_k = \infty$ and $\sum_{k=1}^{\infty} a_k^2 < \infty$. In (7.23), the observation vectors $\mathbf{u}_j^k$s are drawn independently by node j according to some distribution $p_j(\mathbf{u})$ in every iteration. The distribution $p_j(\mathbf{u})$ depends on the measurement model and measurement noise distribution at node j. Once again, we can rewrite the update equations (7.23) and (7.24) together in a compact vector form as

$$\boldsymbol{\gamma}_\ominus^k = (\mathbf{W} \otimes \mathbf{I}_n) \left[\boldsymbol{\gamma}_\ominus^{k-1} + a_k \left(G(\boldsymbol{\gamma}_\ominus^{k-1}, \mathbf{u}_\ominus^k) - \boldsymbol{\gamma}_\ominus^{k-1} \right) \right] \tag{7.25}$$

where $\mathbf{u}_{\ominus} = \left(\mathbf{u}_1^T, \mathbf{u}_2^T, \ldots, \mathbf{u}_L^T\right)^T$ is $mL \times 1$ sized concatenated vector representing the combined observations across the nodes. For the above distributed Robbins-Monro type update [134], we introduce the associated *mean field function* $h(\boldsymbol{\gamma}) : \mathbb{R}_+^n \to \mathbb{R}_+^n$ as

$$h(\boldsymbol{\gamma}) = \mathbb{E}_{\mathbf{1}_n \otimes \boldsymbol{\gamma}} \left[\frac{1}{L} \left(\langle G(\mathbf{1}_L \otimes \boldsymbol{\gamma}, \mathbf{u}_{\ominus}) \rangle - \langle \mathbf{1}_L \otimes \boldsymbol{\gamma} \rangle \right) \right]. \tag{7.26}$$

where the expectation is evaluated as a conditional mean given the past observations [132]. In [131], the authors have shown that under certain assumptions related to the step size a_k, the weight matrix $\mathbf{W}$ and the function G, the distributed update in (7.25) converges to one of the zeros/roots of the mean field function h. Further, there is network wide consensus between the nodes with respect to their local hyperparameters $\boldsymbol{\gamma}_j$ upon convergence [131, 132].

In our distributed setup, the nodes have access to only a single observation, and hence are forced to use the same observation repeatedly in all iterations of the stochastic update (7.25). This repeated use of the observations is modeled as each node j drawing its local observations $\mathbf{u}_j^k$ independently according to the degenerate Dirac-delta distribution, i.e., $p_j(\mathbf{u}) = \delta(\mathbf{u} - \mathbf{y}_j)$. Under this modeling assumption, the stochastic update in (7.25) follows a deterministic trajectory and converges to the zero of the mean field function $\frac{1}{L}\langle G(\mathbf{1}_L \otimes \boldsymbol{\gamma}, \mathbf{y}_{\ominus}) \rangle - \langle \mathbf{1}_L \otimes \boldsymbol{\gamma} \rangle$, or equivalently, to one of the fixed points of $\langle G(\mathbf{1}_L \otimes \boldsymbol{\gamma}, \mathbf{y}_{\ominus}) \rangle$. From (7.22), we recall that every fixed point of $\langle G(\mathbf{1}_L \otimes \boldsymbol{\gamma}, \mathbf{y}_{\ominus}) \rangle$ is also a stationary point of the log-likelihood cost considered in MSBL. Hence, the pseudo-stochastic[3] updates proposed in (7.25) converge to the centralized MSBL solution.

[3]The prefix pseudo here emphasizes the fact that the stochastic updates are driven by repeated use of the same observations at every node.

It is interesting to note that for a constant step size $a_k = 1$ and $p_j(\mathbf{u}) = \delta(\mathbf{u} - \mathbf{y}_j)$, the distributed Robbins-Monro stochastic updates in (7.23)-(7.24) revert to the FB-DSBL updates in (7.18)-(7.19) when the censoring threshold is set to zero (i.e., $\alpha = 1$). In fact, in Section 7.6, we demonstrate through simulations that the proposed FB-DSBL updates with a constant step size ($a_k = 1$) converge significantly faster compared to its stochastic approximation based variant which uses (7.23) instead of (7.4)-(7.5). In the sequel, we refer to the FB-DSBL variant with the local EM step (7.4)-(7.5) replaced with the Robbins-Monro update (7.23) as FB-DSBL†. A detailed performance comparison of FB-DSBL and FB-DSBL† is presented via simulations in the next section.

7.6 Simulations

We now present simulation results to illustrate the efficacy of the proposed FB-DSBL algorithm and compare its performance against the following decentralized joint sparse signal recovery algorithms.

1. DRL-1 - Decentralized Re-weighted ℓ_1 Norm Minimization algorithm [84].
2. DCOMP - Distributed and Collaborative Orthogonal Matching Pursuit [92].
3. DCSP - Decentralized and Collaborative Subspace Pursuit [98].
4. CB-DSBL - Consensus Based Distributed Sparse Bayesian Learning [128].
5. FB-DSBL† - The FB-DSBL variant with the local M-step (7.5) replaced with the stochastic approximation inspired update given by (7.23).

For each trial, the node connectivity in the network is dictated by a randomly generated Erdős-Renyi graph with a connection probability of $p = 0.8$. The joint sparse vectors

$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L$ to be estimated are assumed to be of length $n = 50$ and sharing a common nonzero support which is obtained by randomly selecting $s = 5$ distinct indices out of the set $[n]$. Unless specified otherwise, the nonzero coefficients of the joint sparse vectors are drawn independently from the Rademacher distribution.

Among the algorithms compared here, DCOMP and DCSP require prior knowledge of s, i.e., the size of nonzero support. In the final step of CB-DSBL and FB-DSBL algorithms, the active support is identified by element-wise thresholding the local hyperparameter vector $\boldsymbol{\gamma}_j$ at node j using the thresholds $4\sigma_j^2$ and $0.25\sigma_j^2$, respectively, where σ_j^2 denotes the local measurement noise variance. For FB-DSBL, the per index probability of false detection, α, is set to 10^{-4} for all the nodes. In FB-DSBL†, the step size a_k is set to $k^{-0.51}$, where k is the iteration index.

7.6.1 Performance vs. SNR

In the first set of experiments, we compare the average normalized mean squared error (NMSE) of the signals reconstructed by different algorithms over a wide range of SNRs. The performance benchmark is set by the support-aware linear minimum mean squared error (SA-LMMSE) estimator which assumes knowledge of the true support $\mathcal{S}$. We define the NMSE as

$$\text{NMSE} = \frac{1}{L}\sum_{j=1}^{L}\frac{||\mathbf{x}_j - \hat{\mathbf{x}}_j||_2^2}{||\mathbf{x}_j||_2^2}. \tag{7.27}$$

Here, the size of the network is fixed to $L = 10$ nodes. Fig. 7.4 compares the NMSE achieved by the different algorithms, averaged over 200 trials. Both CB-DSBL and the proposed FB-DSBL closely match the benchmark performance of SA-LMMSE at moderate to high SNRs. It is interesting to note that despite exchanging only $\mathcal{O}(s \log n)$ sized

messages between the nodes, FB-DSBL is able to outperform DCSP, DRL-1 and DCOMP, which are of similar or higher communication complexity. Fig. 7.5 shows a similar trend in the relative performances of the algorithms when the nonzero coefficients of the unknown sparse vectors are drawn from the standard Gaussian distribution. The plots also highlight that even when the number of available measurements is not sufficient for independent signal reconstruction as depicted by the complete breakdown of the standalone SBL algorithm, the decentralized algorithms are able to recover the signals by exploiting their joint sparsity. Although not shown in the plots to avoid clutter, FB-DSBL† matches the performance of FB-DSBL in case of both the source distributions.

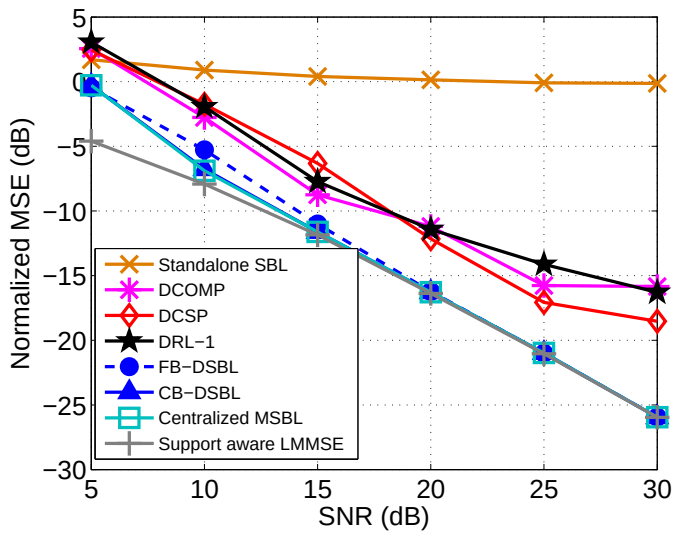


Figure 7.4: Normalized mean squared error of the signals (nonzero coefficients from the Rademacher distribution) reconstructed by different algorithms versus the measurement SNR. Other simulation parameters: $n = 50$, $m = 10$, $s = 5$ and $L = 10$ nodes, and 200 trials.

7.6.2 Support Recovery Performance

In the second set of experiments, we compare the support recovery performance of the decentralized algorithms. From Fig. 7.6, it is evident that FB-DSBL is able to recover

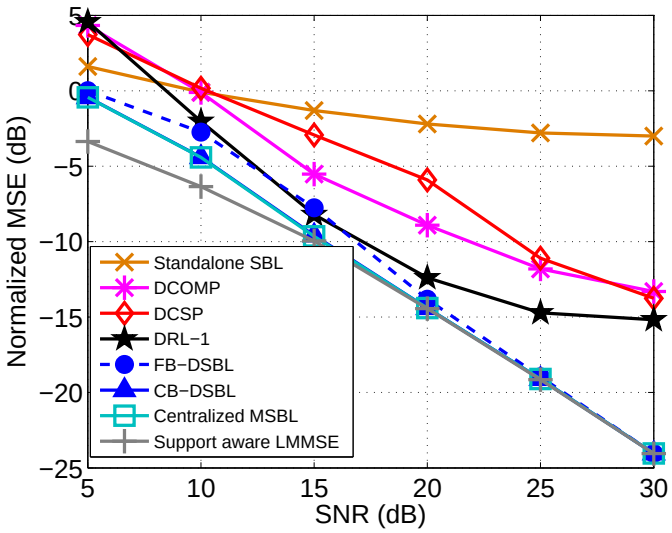


Figure 7.5: Normalized mean squared error of the signals (nonzero coefficients from the standard Gaussian distribution) reconstructed by different algorithms versus the measurement SNR. Other simulation parameters: $n = 50$, $m = 10$, $s = 5$, $L = 10$ nodes, and 100 trials.

the correct support of the unknown sparse vectors using fewer number of measurements per node compared to DRL-1, DCOMP and DCSP. Its stochastic approximation inspired variant FB-DSBL† performs equally well. CB-DSBL has the best support recovery performance among all the decentralized algorithms discussed here, but also has a much higher communication cost compared to FB-DSBL (see Table 7.2).

7.6.3 Phase Transition Characteristics

Here, we compare the phase transition characteristics [135] of the different algorithms under MSE and support recovery based pass/fail criteria. Fig. 7.7 plots the MSE phase transition curves of different algorithms. Any point below the phase transition curve represents a sparsity rate (k/n) and measurement rate (m/n) tuple which results in less than 1% signal reconstruction error. Similarly, in Fig. 7.8, points below the support recovery

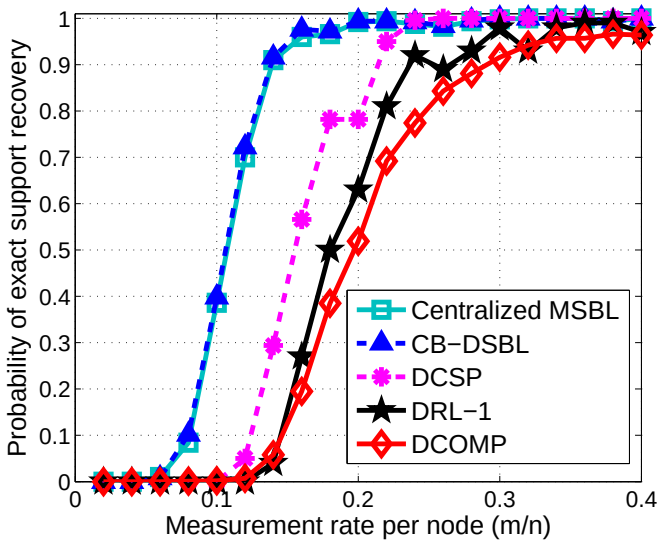


Figure 7.6: Probability of exact support recovery versus the measurement rate (m/n) for different decentralized algorithms. Other simulation parameters: $n = 50$, $s = 5$, $L = 10$ nodes, SNR $= 15$ dB, $\alpha = 10^{-4}$ and number of trials $= 400$.

phase transition curve represent $(k/n, m/n)$ tuples which result in more than 90% accurate nonzero support reconstruction across all the nodes. From their phase transition behaviors, we conclude that the proposed FB-DSBL is able to recover the support and nonzero signal coefficients from fewer measurements per node compared to DRL-1, DCSP and DCOMP. As before, the CB-DSBL algorithm has the best phase transition characteristics, at the cost of the $\mathcal{O}(nL^2)$ communication complexity per iteration. An interesting observation is that FB-DSBL† has slightly better NMSE and support recovery phase transition characteristics compared to FB-DSBL. This is not surprising, as the filtered updates used in FB-DSBL† are more robust in the presence of measurement noise.

7.6.4 Communication Complexity

We also compare the communication overhead of the aforementioned decentralized algorithms. From Fig. 7.9, the overall communication complexity of FB-DSBL is lower than

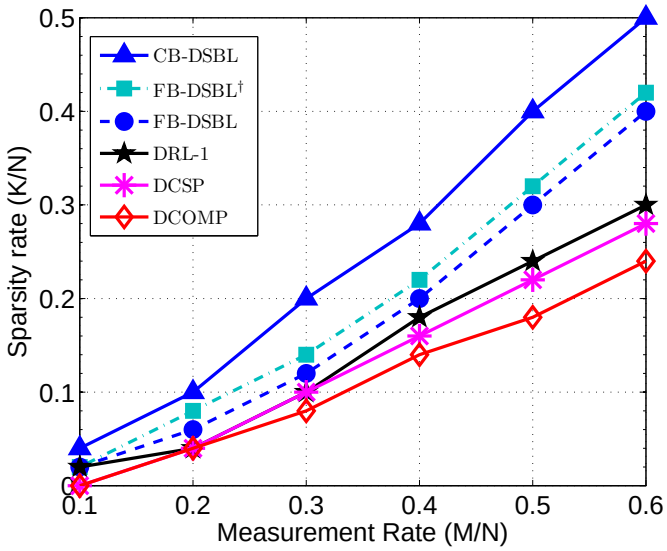


Figure 7.7: MSE phase transition for the different algorithms. For a given measurement rate, the phase transition curves represent the maximum sparsity rate of the unknown sparse vectors that can be recovered with at most 1% reconstruction error. Other simulation parameters: $n = 50$, $L = 5$ nodes, SNR = 30 dB, $\alpha = 10^{-4}$ and number of trials = 200.

CB-DSBL, DRL-1, DCOMP and FB-DSBL†, while DCSP still remains the most communication efficient algorithm amongst all the algorithms compared here. As pointed out in Section 7.4.4, in the proposed FB-DSBL algorithm, the nonzero coefficients of the censored vector $\mathbf{g}_j$ broadcast by node j can be represented using a finite number of bits. We study the impact of quantization of the nonzero coefficients of $\mathbf{g}_j$ on the performance of the FB-DSBL algorithm. Here, we have assumed uniform quantization of $\mathbf{g}_j$ in the logarithmic domain in the range 10^{-10} to 10^{5}. From the MSE phase transition plot in Fig. 7.10, we observe that there is negligible drop in signal reconstruction performance when 4 or more bits are used to represent the nonzero coefficients of the censored vector.

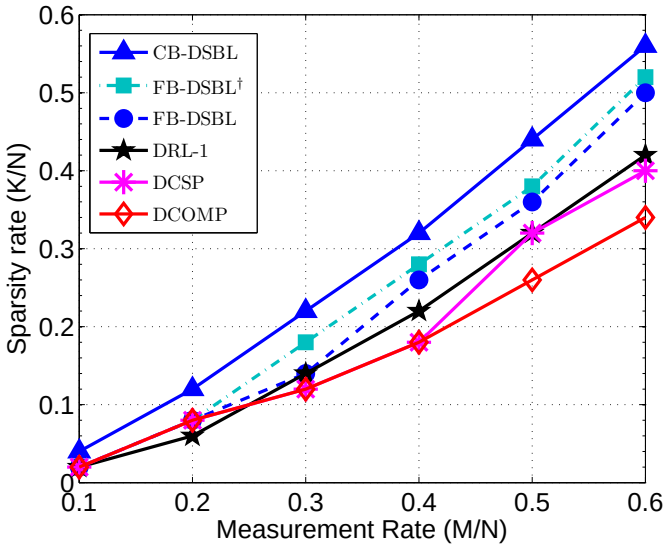


Figure 7.8: Support recovery phase transition for the different algorithms. For a given measurement rate, the phase transition curves represent the maximum sparsity rate of the unknown sparse vectors whose nonzero support can be recovered with at least 90% accuracy. Other simulation parameters: $n = 50$, $L = 5$ nodes, SNR = 30 dB, $\alpha = 10^{-4}$ and number of trials = 200.

7.6.5 Comparison of Convergence Rates

For a decentralized algorithm, the total cost of inter node communication also depends on the number of iterations required for convergence. The centralized MSBL on which the proposed FB-DSBL is based inherits the convergence guarantees of the underlying EM iterations, which always converge to a local minimum of the log-likelihood objective. Analyzing the convergence of FB-DSBL is non-trivial due to use of censored hyperparameter estimates in the decentralized implementation of (7.7). However, in practice, it converges within 10-30 iterations, as shown in Fig. 7.11. FB-DSBL†, on the other hand, converges slower than FB-DSBL, but is faster than CB-DSBL and DRL-1. In comparison to CB-DSBL, the variance hyperparameters belonging to the non-active support set converge to zero significantly faster in the FB-DSBL, resulting in its faster overall convergence. As expected, due to their greedy approach towards support estimation, both DCSP and

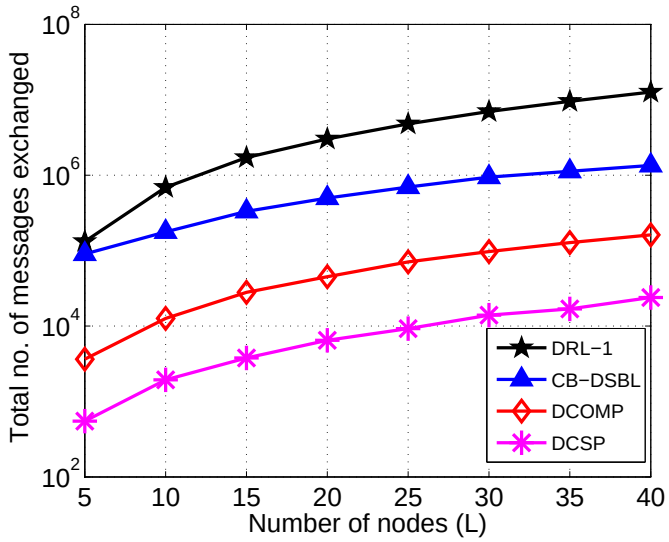


Figure 7.9: Average number of messages exchanged between the nodes versus the size of the network. Simulation parameters: $n = 50$, $m = 10$, $s = 5$, SNR $= 20$ dB.

DCOMP require the least number of iterations to converge.

From Fig. 7.12, we observe a small improvement in FB-DSBL's convergence speed as the network becomes increasingly densely connected, with fastest convergence observed in the case of a fully connected network. More importantly, the reconstruction error (measured in NMSE) remains constant independent of the density of node connections in the network, which is a highly desirable attribute for a distributed algorithm. From these experiments, we conclude that FB-DSBL exhibits stable performance and converges under wide variations in network size and node connection density, provided the network remains connected.

7.6.6 Selection of parameter α

The FB-DSBL parameter α has a direct impact on the size of messages exchanged between the nodes. As described in section 7.4.1, α represents the probability of false alarm used in the index-wise LLRTs for generation of hard support estimates at each node. Choosing

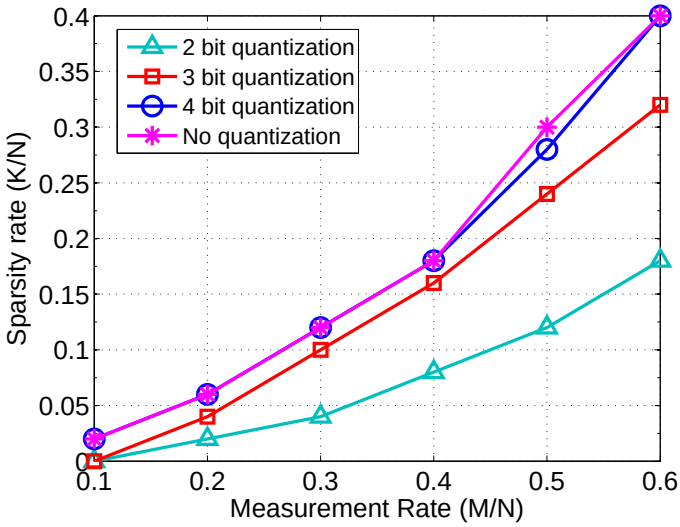


Figure 7.10: MSE phase transition for FB-DSBL variants using $2, 3, 4$ bit quantization and analog transmission to encode the nonzero coefficients of soft support estimates exchanged between the nodes. Other simulation parameters: $n = 50$, $L = 5$ nodes, SNR $= 30$ dB, $\alpha = 10^{-4}$ and number of trials $= 1000$.

a lower value for α biases the index-wise LLRTs towards the $\mathcal{H}_0$ hypothesis, leading to sparser hard/soft support estimates, and consequently, smaller message size. Likewise, higher values of α lead to larger message size. In fact, for $\alpha = 1$, there is no censoring of the locally estimated hyperparameters $\boldsymbol{\gamma}$ exchanged between the nodes. In this case, the FB-DSBL algorithm reverts to a local averaging based distributed implementation of the MSBL update rule (7.7). Fig. 7.13 illustrates the effect of α on the MSE performance and the communication cost of the FB-DSBL algorithm. As seen in the plots, there exists a stable range of α between 10^{-2} to 10^{-4}, where one can obtain the twin benefits of low communication complexity and good reconstruction performance. A good rule-of-thumb value of α is $0.01/n$, which corresponds to approximately 1% chance of observing a false alarm in one of the n hypothesis tests performed at each node.

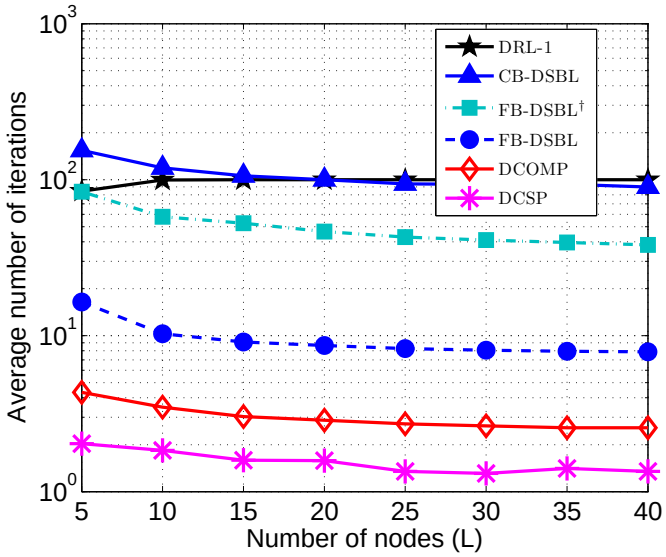


Figure 7.11: Average number of iterations versus network size for different decentralized algorithms. Simulation parameters: $n = 50$, $m = 10$, $s = 5$ and SNR= 20 dB.

7.7 Chapter Summary

Most of the existing decentralized algorithms for joint sparse signal recovery entail inter-node exchange of messages whose size is proportional to n, the ambient signal dimension, which is typically very high. Thus the nodes have to expend a significant share of their limited time/energy resources for inter node communications. In this chapter, we have addressed this issue, by proposing a highly communication efficient, decentralized joint sparse signal recovery algorithm called FB-DSBL which requires exchange of only $\mathcal{O}(s \log n)$ sized messages between the network nodes. We also showed that the proposed algorithm can be seen as a degenerate case of a distributed consensus based stochastic approximation algorithm. From the extensive simulation results presented in this chapter, we can conclude that the proposed FB-DSBL algorithm outperforms existing decentralized algorithms DRL-1, DCSP, and DCOMP while at the same time has an attractively low communication complexity.

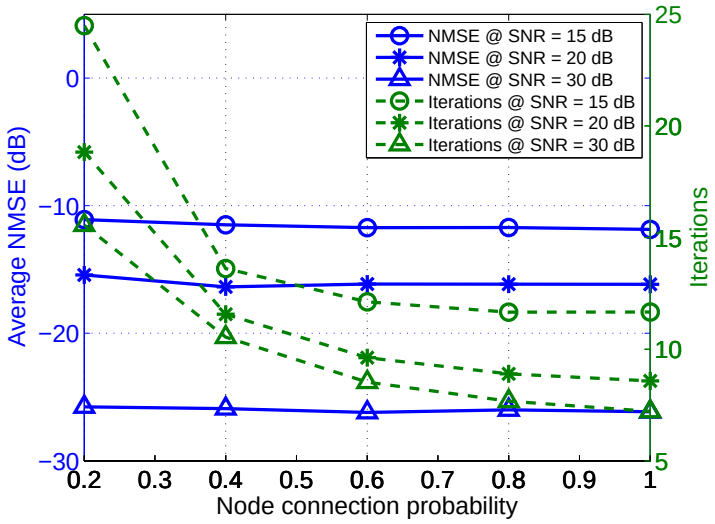


Figure 7.12: Illustration of the robustness of FB-DSBL's performance and convergence speed with varying node connection density. Connection probability of one corresponds to a fully connected network. Simulation parameters: $n = 100$, $m = 10$, $s = 5$, $L = 20$ nodes, trials $= 100$.

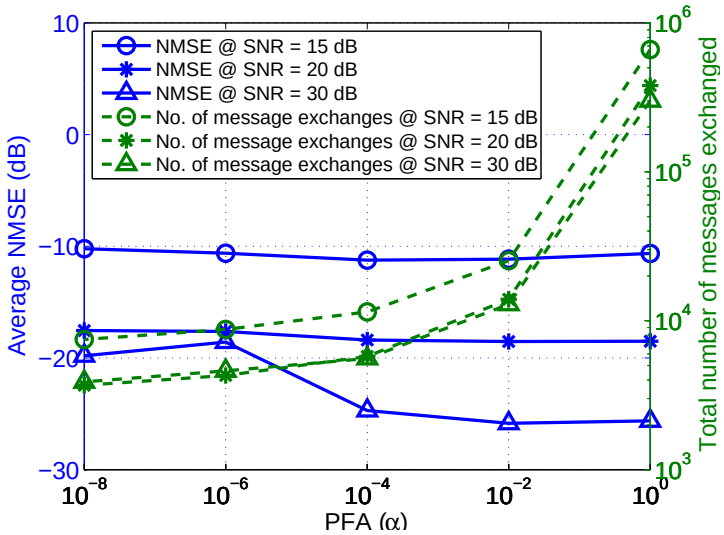


Figure 7.13: Plot illustrating the sensitivity of FB-DSBL's performance and communication complexity with respect to the PFA parameter α. Simulation parameters: $n = 50$, $m = 10$, $s = 5$ and $L = 10$ nodes.

Chapter 8

Conclusion

In which we conclude & look ahead...

8.1 Summary of Main Results

This thesis contributes new theoretical results, solution concepts, and algorithms concerning the estimation of joint sparse signals and their common nonzero support from linear compressive measurements using the Bayesian approach. The following is a summary of the main results in the thesis.

New Perspective on Support Recovery using Sparse Bayesian Learning (Chapter 2)

- MSBL's log-likelihood objective can be interpreted as negative Log-Det Bregman matrix divergence between the empirical and parametrized MMV covariance matrices.

- MSBL optimization minimizes the gap between the empirical and parameterized covariance matrices of the observations, where pointwise distances between matrices

are measured in the Log-Det Bregman matrix divergence sense. Thus, MSBL is essentially a covariance matching algorithm for sparse support recovery.

- In covariance matching based sparse support recovery, the maximum size of uniquely recoverable sparse support is determined by the Kruskal-Rank of columnwise Khatri-Rao product of the measurement matrix with itself.

Restricted Isometry Of Khatri-Rao Product (Chapter 3)

- Sensing matrices expressible as a Khatri-Rao product of two matrices appear in many fundamental inverse problems.
- The restricted isometry property (RIP) of a matrix is characterized in terms of upper bounds on its restricted isometry constants (RICs). RICs of Khatri-Rao matrices feature in sample complexity analysis of several machine learning and signal processing problems including the MMV problem.
- For $m \times n$ sized column normalized, RIP compliant matrices $\mathbf{A}$ and $\mathbf{B}$, when $k \leq m$, we have $k\text{-RIC}(\mathbf{A} \odot \mathbf{B}) \leq \left(\max\left(k\text{-RIC}(\mathbf{A}), k\text{-RIC}(\mathbf{B})\right)\right)^2$.
- For an $m \times n$ sized column normalized, RIP compliant matrix $\mathbf{A}$, when $k \leq m$, we have $k\text{-RIC}(\mathbf{A} \odot \mathbf{A}) \leq \left(k\text{-RIC}(\mathbf{A})\right)^2$ for $k \leq m$. Hence, taking Khatri-Rao product of a RIP compliant matrix with itself strengthens the restricted isometry property.
- For $m \times n$ sized random matrices $\mathbf{A}$ and $\mathbf{B}$ containing i.i.d. subgaussian entries, $\mathbf{A} \odot \mathbf{B}$ satisfies the restricted isometry property of order k with high probability provided $m \geq \mathcal{O}(k \log n)$. A similar orderwise result exists for the self Khatri-Rao product $\mathbf{A} \odot \mathbf{A}$.

New Sufficient Conditions for Perfect Support Recovery in Sparse Bayesian Learning (Chapter 4)

- For Gaussian sources, the support error probability in MSBL decays exponentially fast with the number of joint sparse vectors.
- The exponent of MSBL's exponentially decaying support error probability depends on the restricted null space property of the Khatri-Rao product of the measurement matrix with itself.
- For a random $m \times n$ sized measurement matrix with i.i.d $\mathcal{N}\left(0, \frac{1}{\sqrt{m}}\right)$ and $m = \mathcal{O}(K \log n)$ rows, any K-sparse support can perfectly recovered with arbitrarily high probability using finitely many MMVs by maximizing the MSBL's log-likelihood cost function.
- In the noiseless case, MSBL can perfectly recover any $k < \text{spark}(\mathbf{A}) - 1$ sized support even from a single measurement vector.

Rényi Divergence Based Joint Sparse Support Recovery (Chapter 5)

- The α-Rényi divergence objective generalizes MSBL's log-likelihood objective.
- The α-Rényi divergence between empirical and parameterized multivariate Gaussian densities can be expressed as a difference of submodular set functions of the unknown signal support.
- The above α-Rényi divergence can be optimized by a fast majorization-minimization procedure to recover the unknown support. The algorithm is named *Rényi Divergence based Covariance Matching Pursuit* or RD-CMP.

- Due to its recursive, rank one updates, and overall greedy approach towards support reconstruction, RD-CMP is significantly faster than existing covariance matching based support recovery algorithms.

Two Decentralized Extensions of MSBL (Chapters 6 & 7)

- We proposed the CB-DSBL algorithm which is a decentralized implementation of MSBL for in-network estimation of multiple joint sparse signals from linear compressive measurements using a network of processing nodes or sensors. CB-DSBL uses Alternating Directions Method of Multipliers (ADMM) to optimize the MSBL log-likelihood objective in a decentralized fashion.

- The inter-node communication is controlled by adopting a bridge node based network topology which allows convenient trade-off between communication cost and robustness to node failures.

- CB-DSBL is empirically shown to match the performance of the centralized MSBL algorithm.

- We proposed the FB-DSBL algorithm, another decentralized extension of MSBL, with low communication complexity as a key feature. In FB-DSBL, sparse censored messages are exchanged between the network nodes resulting in a significant reduction in the overall communication complexity of the decentralized algorithm.

- Using Monte-Carlo simulations, we showed that FB-DSBL offers Bayesian inference like signal reconstruction performance while at the same time its communication complexity is comparable to greedy reconstruction methods.

- The convergence behavior of FB-DSBL is analyzed by interpreting it as a distributed consensus based stochastic approximation algorithm.

Convergence Of Decentralized ADMM In Bridge Node Network Topology (Chapter 6)

- We proved that decentralized ADMM iterations converge R-linearly in a bridge node based network topology.

- Our convergence analysis reveals the optimal choice for the ADMM parameter ρ and how the ADMM convergence rate is impacted by the selection of the bridge nodes. These results are of independent interest and applicable whenever decentralized ADMM is used to minimize a strongly convex separable objective in a bridge node based network topology.

8.2 Future Work

Some possible avenues for future work are:

- **Covariance Matching Based Support Recovery In High Dimensions**

  The existing algorithms based on the covariance matching principle are still not computationally fast enough to handle large signal dimensions encountered in big data applications. Development of fast scalable covariance matching based support recovery methods is an important problem.

- **Open Questions On Sparse Bayesian Learning**

  There remain a few open questions about MSBL's support recovery performance:

i. What are necessary conditions for exact support recovery in terms of the number of required MMVs?

ii. Under what conditions do all local maxima of the MSBL objective yield the correct support estimate?

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

- **Restricted Null Space Property of Khatri-Rao product**

  For an $m \times n$ subgaussian matrix $\mathbf{A}$, when does the self Khatri-Rao product $\mathbf{A} \odot \mathbf{A}$ satisfies the NN-RNSP condition of order k (Definition 4.1). We conjecture that for $\mathbf{A}$ with i.i.d subgaussian entries and n fixed, if m scales as $\mathcal{O}(\sqrt{k})$, then $\mathbf{A} \odot \mathbf{A}$ will satisfy NN-RNSP of order k with high probability.

- **Support Recovery Sample Complexity of RD-CMP**

  Sample complexity of RD-CMP algorithm is currently unknown. The optimal value for α, the Rényi divergence order, is also unknown. Studying the support recovery behavior of RD-CMP for $\alpha > 1$ would also be interesting.

Appendix A

Mathematical Review for Chapter 3 (Part 1)

In this appendix, we present some preliminary concepts and results which are necessary for proving the deterministic k-RIC bound in Theorem 3.1. For the sake of brevity, we provide proofs only for claims that have not been explicitly shown in their cited sources.

A.0.1 Properties of the Kronecker and Hadamard product

For any two matrices $\mathbf{A}$ and $\mathbf{B}$ of dimensions $m \times n$ and $p \times q$, the kronecker product $\mathbf{A} \otimes \mathbf{B}$ is the $mp \times nq$ matrix

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2n}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{pmatrix}. \tag{A.1}$$

The following Proposition relates the spectral properties of the Kronecker product and its constituent matrices.

Proposition A.1 (7.1.10 in [55]). *Let* $\mathbf{A} \in \mathbb{C}^{n\times n}$ *and* $\mathbf{B} \in \mathbb{C}^{p\times p}$ *admit eigenvalue decompositions* $\mathbf{U}_A\Lambda_A\mathbf{U}_A^H$ *and* $\mathbf{U}_B\Lambda_B\mathbf{U}_B^H$*, respectively. Then,*

$$(\mathbf{U}_A \otimes \mathbf{U}_B)(\Lambda_A \otimes \Lambda_B)(\mathbf{U}_A \otimes \mathbf{U}_B)^H$$

yields the eigenvalue decomposition for $\mathbf{A} \otimes \mathbf{B}$*.*

For any two matrices of matching dimensions, say $m \times n$, their Hadamard product $\mathbf{A} \circ \mathbf{B}$ is obtained by elementwise multiplication of the entries of the input matrices, i.e.,

$$[\mathbf{A} \circ \mathbf{B}]_{i,j} = a_{ij}b_{ij} \quad \text{for } i \in [m], j \in [n]. \tag{A.2}$$

The Hadamard product $\mathbf{A} \circ \mathbf{B}$ is a principal submatrix of the Kronecker product $\mathbf{A} \otimes \mathbf{B}$ [56, 137]. For $n \times n$ sized square matrices $\mathbf{A}$ and $\mathbf{B}$, one can write,

$$\mathbf{A} \circ \mathbf{B} = \mathbf{J}^T(\mathbf{A} \otimes \mathbf{B})\mathbf{J}, \tag{A.3}$$

where $\mathbf{J}$ is an $n^2 \times n$ sized selection matrix constructed from 0's and 1's and satisfying $\mathbf{J}^T\mathbf{J} = \mathbf{I}_n$.

In Proposition A.2, we present an upper bound on the spectral radius of a generic Hadamard product.

Proposition A.2. *For every* $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{m\times n}$*, we have*

$$\sigma_{max}(\mathbf{A} \circ \mathbf{B}) \leq r_{max}(\mathbf{A})c_{max}(\mathbf{B}) \tag{A.4}$$

where $\sigma_{max}(\cdot)$*,* $r_{max}(\cdot)$ *and* $c_{max}(\cdot)$ *are the largest singular value, the largest row* ℓ_2*-norm and the largest column* ℓ_2*-norm of the input matrix, respectively.*

Proof. See Theorem 5.5.3 in [138]. □

We now state an important result about the Hadamard product of two positive semidefinite matrices.

Proposition A.3 (Mond and Pečarić [139])**.** *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be positive semidefinite* $n \times n$ *Hermitian matrices and let* r *and* s *be two nonzero integers such that* $s > r$*. Then,*

$$(\mathbf{A}^s \circ \mathbf{B}^s)^{1/s} \geq (\mathbf{A}^r \circ \mathbf{B}^r)^{1/r} . \tag{A.5}$$

In Propositions A.4 and A.5, we state some spectral properties of correlation matrices and their Hadamard products. Correlation matrices are Hermitian positive semidefinite matrices with diagonal entries equal to one. Later on, we will exploit the fact that the singular values of the columnwise KR product are related to the singular values of the Hadamard product of certain correlation matrices.

Proposition A.4. *If* $\mathbf{A}$ *is an* $n \times n$ *sized Hermitian correlation matrix, then* $\mathbf{A}^{1/2} \circ \bar{\mathbf{A}}^{1/2}$ *is a doubly stochastic matrix. Here,* $\bar{\mathbf{A}}$ *denotes the elementwise complex conjugate of* $\mathbf{A}$*.*

Proof. Since $\mathbf{A}$ is a Hermitian correlation matrix, it admits the Schur decomposition, $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H$, with unitary $\mathbf{U}$ and eigenvalue matrix $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$. Since $\mathbf{A}$ is positive semi-definite, its nonnegative square-root exists and is given by $\mathbf{A}^{1/2} = \mathbf{U}\Lambda^{1/2}\mathbf{U}^H$. Consider

$$\begin{aligned} \mathbf{A}^{1/2} \circ \bar{\mathbf{A}}^{1/2} &= \sum_{i=1}^{n} \lambda_i^{1/2} \mathbf{u}_i \mathbf{u}_i^H \circ \sum_{j=1}^{n} \lambda_j^{1/2} \overline{\mathbf{u}_j \mathbf{u}_j^H} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i^{1/2} \lambda_j^{1/2} \left(\mathbf{u}_i \bar{\mathbf{u}}_i^T\right) \circ \left(\bar{\mathbf{u}}_j \mathbf{u}_j^T\right) \end{aligned}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i^{1/2}\lambda_j^{1/2} \left(\mathbf{u}_i \circ \bar{\mathbf{u}}_j\right)\left(\bar{\mathbf{u}}_i \circ \mathbf{u}_j\right)^T. \tag{A.6}$$

The second equality above follows from the distributive property of the Hadamard product and the last step follows from Fact 7.6.2 in [55]. Using (A.6), we can show that the rows and columns of $\mathbf{A}^{1/2} \circ \bar{\mathbf{A}}^{1/2}$ sum to one, as follows:

$$\begin{aligned}
(\mathbf{A}^{1/2} \circ \bar{\mathbf{A}}^{1/2})\mathbf{1} &= \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i^{1/2}\lambda_j^{1/2} \left(\mathbf{u}_i \circ \bar{\mathbf{u}}_j\right)\left(\bar{\mathbf{u}}_i \circ \mathbf{u}_j\right)^T \mathbf{1} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i^{1/2}\lambda_j^{1/2} \left(\mathbf{u}_i \circ \bar{\mathbf{u}}_j\right)\left(\bar{\mathbf{u}}_i \circ \mathbf{1}\right)^T \mathbf{u}_j \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i^{1/2}\lambda_j^{1/2} \left(\mathbf{u}_i \circ \bar{\mathbf{u}}_j\right)\mathbf{u}_i^H \mathbf{u}_j \\
&= \sum_{i=1}^{n} \lambda_i \left(\mathbf{u}_i \circ \bar{\mathbf{u}}_i\right) = \mathbf{d} \ \text{(say)}.
\end{aligned}$$

The above arguments follow from the orthonormality of the columns of $\mathbf{U}$, and repeated application of Fact 7.6.1 in [55]. Note that for $k \in [n]$, $\mathbf{d}(k) = \sum_{i=1}^{n} \lambda(i) \left|\mathbf{u}_i(k)\right|^2 = [\mathbf{U}\mathbf{\Lambda}\mathbf{U}^H]_{kk} = \mathbf{A}_{kk} = 1$. Thus, we have shown that $(\mathbf{A}^{1/2} \circ \bar{\mathbf{A}}^{1/2})\mathbf{1} = \mathbf{1}$. Likewise, it can be shown that $\mathbf{1}^T(\mathbf{A}^{1/2} \circ \bar{\mathbf{A}}^{1/2}) = \mathbf{1}^T$. Thus, $\mathbf{A}^{1/2} \circ \bar{\mathbf{A}}^{1/2}$ is doubly stochastic. □

Proposition A.5 (Werner [140])**.** *For any Hermitian correlation matrices* $\mathbf{A}$ *and* $\mathbf{B}$ *of the same size, we have* $\mathbf{A}^{1/2} \circ \mathbf{B}^{1/2} \leq \mathbf{I}$*, where* $\mathbf{A}^{1/2}$ *and* $\mathbf{B}^{1/2}$ *are the positive square roots of* $\mathbf{A}$ *and* $\mathbf{B}$*, respectively.*

Proof. Since $\mathbf{A}$ is a correlation matrix, from Proposition A.4, it follows that $\mathbf{A}^{1/2} \circ \mathbf{A}^{1/2}$ is doubly stochastic. Since the rows and columns of $\mathbf{A}^{1/2} \circ \mathbf{A}^{1/2}$ sum to unity, we have $r_{\max}(\mathbf{A}^{1/2}) = c_{\max}(\mathbf{A}^{1/2}) = 1$. Similarly, $r_{\max}(\mathbf{B}^{1/2}) = c_{\max}(\mathbf{B}^{1/2}) = 1$. Then, from Proposition A.2, it follows that the largest eigenvalue of $\mathbf{A}^{1/2} \circ \mathbf{B}^{1/2}$ is at most unity. □

A.0.2 Matrix Kantorovich Inequalities

Matrix Kantorovich inequalities relate positive definite matrices by inequalities in the sense of the Löwner partial order.[1] These inequalities can be used to extend the Löwner partial order to the Hadamard product of positive definite matrices. Our proposed RIC bound relies on the tightness of these Kantorovich inequalities and their extensions.

A matrix version of the Kantorovich inequality was first proposed by Marshall and Olkin in [141]. It is stated below as Proposition A.6.

Proposition A.6 (Marshall and Olkin [141])**.** *Let* $\mathbf{A}$ *be an* $n \times n$ *positive definite Hermitian matrix. Let* $\mathbf{A}$ *admit the Schur decomposition* $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H$ *with unitary* $\mathbf{U}$ *and* $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ *such that* $\lambda_i \in [m, M]$*. Then, we have*

$$\mathbf{A}^2 \leq (M+m)\mathbf{A} - mM\mathbf{I}_n. \tag{A.7}$$

The above inequality (A.7) is the starting point for obtaining a variety of forward and reverse Kantorovich-type matrix inequalities for positive definite matrices. In Propositions A.7 and A.8, we state specific forward and reverse inequalities, respectively, which are relevant to us.

Proposition A.7 (Liu and Neudecker [142])**.** *Let* $\mathbf{A}$ *be an* $n \times n$ *positive definite Hermitian matrix, with eigenvalues in* $[m, M]$*. Let* $\mathbf{V}$ *be an* $n \times n$ *matrix such that* $\mathbf{V}^H\mathbf{V} = \mathbf{I}_n$*. Then,*

$$\mathbf{V}^H\mathbf{A}^2\mathbf{V} - \left(\mathbf{V}^H\mathbf{A}\mathbf{V}\right)^2 \leq \frac{1}{4}(M-m)^2\mathbf{I}_n. \tag{A.8}$$

[1] The Löwner partial order here refers to the relation "$\leq$". For positive definite matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A} \leq \mathbf{B}$ if and only if $\mathbf{B} - \mathbf{A}$ is a positive semi-definite matrix.

Proposition A.8 (Liu and Neudecker [143]). *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be* $n \times n$ *positive definite Hermitian matrices. Let* m *and* M *be the minimum and maximum eigenvalues of* $\mathbf{B}^{1/2}\mathbf{A}^{-1}\mathbf{B}^{1/2}$. *Let* $\mathbf{X}$ *be an* $n \times p$ *matrix. Then, we have*

$$(\mathbf{X}^H\mathbf{B}\mathbf{X})(\mathbf{X}^H\mathbf{A}\mathbf{X})^\dagger(\mathbf{X}^H\mathbf{B}\mathbf{X}) \geq \frac{4mM}{(M+m)^2}\mathbf{X}^H\mathbf{B}\mathbf{A}^{-1}\mathbf{B}\mathbf{X}. \tag{A.9}$$

Proposition A.7 can be proved using (A.7) by pre- and post-multiplying by $\mathbf{V}^H$ and $\mathbf{V}$, respectively, followed by completion of squares for the right hand side terms. The proof of Proposition A.8 is given in [143].

A.0.3 Matrix Kantorovich Inequalities for the Hadamard Products of Positive Definite Matrices

Lemmas A.1 and A.2 stated below extend the Kantorovich inequalities from the previous section to Hadamard products.

Lemma A.1 (Liu and Neudecker [142]). *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be* $n \times n$ *positive definite Hermitian matrices, with* m *and* M *denoting the minimum and maximum eigenvalues of* $\mathbf{A} \otimes \mathbf{B}$. *Then, we have*

$$\mathbf{A}^2 \circ \mathbf{B}^2 \leq (\mathbf{A} \circ \mathbf{B})^2 + \frac{1}{4}(M-m)^2\mathbf{I}_n. \tag{A.10}$$

Proof. Let $\mathbf{J}$ be the selection matrix such that $\mathbf{J}^T\mathbf{J} = \mathbf{I}_n$ and $\mathbf{A} \circ \mathbf{B} = \mathbf{J}^T(\mathbf{A} \otimes \mathbf{B})\mathbf{J}$. Then, by applying Proposition A.7 with $\mathbf{A}$ replaced with $\mathbf{A} \otimes \mathbf{B}$, and $\mathbf{V}$ replaced with $\mathbf{J}$, we obtain

$$\mathbf{J}^T(\mathbf{A} \otimes \mathbf{B})^2\mathbf{J} - (\mathbf{J}^T(\mathbf{A} \otimes \mathbf{B})\mathbf{J})^2 \leq \frac{1}{4}(M-m)^2\mathbf{I}_n.$$

Using Fact 8.21.29 in [55], i.e., $(\mathbf{A} \otimes \mathbf{B})^2 = \mathbf{A}^2 \otimes \mathbf{B}^2$, we get

$$\mathbf{J}^T(\mathbf{A}^2 \otimes \mathbf{B}^2)\mathbf{J} - (\mathbf{A} \circ \mathbf{B})^2 \leq \frac{1}{4}(M-m)^2\mathbf{I}_n,$$

Finally, by observing that $\mathbf{J}^T(\mathbf{A}^2 \otimes \mathbf{B}^2)\mathbf{J} = \mathbf{A}^2 \circ \mathbf{B}^2$, we obtain the desired result. □

Lemma A.2 (Liu [144])**.** *Let* $\mathbf{A}$*,* $\mathbf{B}$ *be* $n \times n$ *positive definite Hermitian correlation matrices. Then,*

$$\mathbf{A}^{1/2} \circ \mathbf{B}^{1/2} \geq \frac{2\sqrt{mM}}{m+M}\mathbf{I}_n \tag{A.11}$$

where the eigenvalues of $\mathbf{A}^{1/2}$ *and* $\mathbf{B}^{1/2}$ *lie inside* $[m, M]$*.*

Lemma A.2 follows from Proposition A.8, by replacing $\mathbf{A}$ with $\mathbf{I}_n \otimes \mathbf{B}$, $\mathbf{B}$ with $\mathbf{A}^{1/2} \otimes \mathbf{B}^{1/2}$, and $\mathbf{X}$ with $\mathbf{J}$, where $\mathbf{J}$ is the $n^2 \times n$ binary selection matrix such that $\mathbf{J}^T\mathbf{J} = \mathbf{I}_n$ and $\mathbf{A} \circ \mathbf{B} = \mathbf{J}^T(\mathbf{A} \otimes \mathbf{B})\mathbf{J}$.

Appendix B

Mathematical Review for Chapter 3

In this appendix, we present the celebrated Hanson-Wright inequality [145, 146] and its corollary characterizing the tail probability of functions of independent subgaussian random vectors. These results are used in the proofs of Theorems 3.2 and 3.3.

B.1 The Hanson-Wright Inequality

The following theorem states the Hanson-Wright inequality.

Theorem B.1 (Rudelson and Vershynin [146]). *Let* $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) \in \mathbb{R}^n$ *be a random vector with independent components* $\mathbf{x}_i$ *satisfying* $\mathbb{E}\mathbf{x}_i = 0$ *and* $||\mathbf{x}_i||_{\psi_2} \leq \kappa$. *Let* $\mathbf{A}$ *be an* $n \times n$ *matrix. Then, for every* $t \geq 0$,

$$\mathbb{P}\left\{\left|\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbb{E}\mathbf{x}^T\mathbf{A}\mathbf{x}\right| > t\right\} \leq 2\exp\left[-c\min\left(\frac{t^2}{\kappa^4\,||\mathbf{A}||_{HS}^2}, \frac{t}{\kappa^2\,|||\mathbf{A}|||_2}\right)\right]$$

where c *is a universal positive constant.*

The following corollary of the Hanson-Wright inequality characterizes the tail probability of weighted inner product between two independent subgaussian vectors.

Corollary B.1. *Let* $\mathbf{u} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n) \in \mathbb{R}^n$ *and* $\mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n) \in \mathbb{R}^n$ *be independent random vectors with independent subgaussian components satisfying* $\mathbb{E}\mathbf{u}_i = \mathbb{E}\mathbf{v}_i = 0$ *and* $||\mathbf{u}_i||_{\psi_2} \leq \kappa$, $||\mathbf{v}_i||_{\psi_2} \leq \kappa$. *Let* $\mathbf{D}$ *be an* $n \times n$ *matrix. Then, for every* $t \geq 0$,

$$\mathbb{P}\left\{\left|\mathbf{u}^T\mathbf{D}\mathbf{v}\right| > t\right\} \leq 2\exp\left[-c\min\left(\frac{t^2}{\kappa^4\left|\left|\mathbf{D}\right|\right|_{HS}^2}, \frac{t}{\kappa^2\left|\left|\left|\mathbf{D}\right|\right|\right|_2}\right)\right]$$

where c *is a universal positive constant.*

Proof. Use Hanson-Wright inequality in Theorem B.1 with $\mathbf{x} = \left[\mathbf{u}^T\ \mathbf{v}^T\right]^T$ and $\mathbf{A} = [\mathbf{0}_{n\times n} \mid \mathbf{D}; \mathbf{0}_{n\times n} \mid \mathbf{0}_{n\times n}]$ to obtain the desired tail bound. □

Appendix C

Mathematical review for chapter 4

In this appendix, we introduce some key definitions and results which are used in chapter 4.

C.1 Restricted Isometry Property

A matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ is said to satisfy the restricted isometry property (RIP) of order k if there exists a constant $\delta_k^{\mathbf{A}} \in (0, 1)$ such that

$$(1 - \delta_k^{\mathbf{A}}) \, ||\mathbf{x}||_2^2 \leq ||\mathbf{A}\mathbf{x}||_2^2 \leq (1 + \delta_k^{\mathbf{A}}) \, ||\mathbf{x}||_2^2 \tag{C.1}$$

holds for any k-sparse vector $\mathbf{x} \in \mathbb{R}^n$. The smallest such $\delta_k^{\mathbf{A}}$ is called the k^{th} order restricted isometry constant (RIC) of $\mathbf{A}$. The RIP condition $\delta_k^{\mathbf{A}} < 1$ implies that spark$(\mathbf{A}) > k$ and K-rank$(\mathbf{A}) \geq k$.

C.2 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 which can cover $\mathcal{A}$ is called the ϵ-covering number of $\mathcal{A}$, and is given by

$$N_{\text{cov}}^{\epsilon}(\mathcal{A}, d) = \min\{|\mathcal{A}^{\epsilon}| : \mathcal{A}^{\epsilon} \text{ is an } \epsilon\text{-net of } \mathcal{A}\} \tag{C.2}$$

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.

Proposition C.1 ([147]). *Let $\mathcal{B}(0,1)$ be a unit ball in $\mathbb{R}^n$ centered at 0. Then, its ϵ-covering number with respect to the standard Euclidean metric is bounded as*

$$N_{cov}^{\epsilon}(\mathcal{B}(0,1), ||\cdot||_2) \leq \left(\frac{3}{\epsilon}\right)^n. \tag{C.3}$$

C.3 α-Rényi Divergence

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 $\mathcal{F}$.

For $\alpha \in \mathbb{R}^+\backslash 1$, the *Rényi divergence* of order α between P and Q, denoted $\mathcal{D}_\alpha(p||q)$, is defined as

$$\mathcal{D}_\alpha(p||q) = \frac{1}{\alpha - 1} \log \int_{\mathcal{X}} p(x)^\alpha q(x)^{1-\alpha} \mu(dx). \tag{C.4}$$

$\mathcal{D}_\alpha(p||q)$ is a nondecreasing function of α. For $\alpha \in [0,1)$, $\mathcal{D}_\alpha(p||q) < \mathcal{D}_{\text{KL}}(p||q)$, with $\lim_{\alpha\to 1} \mathcal{D}_\alpha(p||q) = \mathcal{D}_{\text{KL}}(p||q)$, where $\mathcal{D}_{\text{KL}}$ is the Kullback-Leibler divergence [85].

For $p = \mathcal{N}(0, \mathbf{\Sigma}_1)$ and $q = \mathcal{N}(0, \mathbf{\Sigma}_2)$, the α-Rényi divergence $\mathcal{D}_\alpha(p||q)$ is available in closed form [86],

$$\mathcal{D}_\alpha(p||q) = \frac{1}{2(1-\alpha)} \log \frac{|(1-\alpha)\mathbf{\Sigma}_1 + \alpha\mathbf{\Sigma}_2|}{|\mathbf{\Sigma}_1|^{1-\alpha} |\mathbf{\Sigma}_2|^\alpha}. \tag{C.5}$$

C.3.1 Lower Bound for α-Rényi Divergence Between Multivariate Gaussian Distributions

Theorem C.1 provides a lower bound for the Rényi divergence between two multivariate Gaussian distributions.

Theorem C.1. *Let p_1 and p_2 be two multivariate Gaussian distributions with zero mean and positive definite covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$, respectively. Then, the $\frac{1}{2}$-Rényi divergence between p_1 and p_2 satisfies*

$$\mathcal{D}_{1/2}(p_1, p_2) \geq \frac{1}{2} tr\left(\left(\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2\right)\left(\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2\right)^{-1}\left(\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2\right)\left(\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2\right)^{-1}\right).$$

Proof. From (C.5), by using the property: $\log|\mathbf{XY}| = \log|\mathbf{X}| + \log|\mathbf{Y}|$ for any positive definite matrices $\mathbf{X}$ and $\mathbf{Y}$, we have

$$\begin{aligned}
\mathcal{D}_{1/2}(p_1, p_2) &= \log\left|\frac{\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2}{2}\right| - \frac{1}{2}\log|\mathbf{\Sigma}_1| - \frac{1}{2}\log|\mathbf{\Sigma}_2| \\
&= \log\left|\frac{\mathbf{I}_m + \mathbf{\Sigma}_1^{-1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{-1/2}}{2}\right| + \frac{1}{2}\log|\mathbf{\Sigma}_1| - \frac{1}{2}\log|\mathbf{\Sigma}_2| \\
&= \log\left|\frac{\mathbf{I}_m + \mathbf{\Sigma}_1^{-1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{-1/2}}{2}\right| + \frac{1}{2}\log\left|\mathbf{\Sigma}_1^{1/2}\mathbf{\Sigma}_2^{-1}\mathbf{\Sigma}_1^{1/2}\right| \\
&= \log\left|\frac{\mathbf{I}_m + \mathbf{H}^{-1}}{2}\right| + \frac{1}{2}\log|\mathbf{H}| = \log\left|\frac{\mathbf{H}^{1/2} + \mathbf{H}^{-1/2}}{2}\right|, \qquad \text{(C.6)}
\end{aligned}$$

where $\mathbf{H} \triangleq \mathbf{\Sigma}_1^{1/2}\mathbf{\Sigma}_2^{-1}\mathbf{\Sigma}_1^{1/2}$. Suppose $\mathbf{H}$ admits the eigenvalue decomposition $\mathbf{H} = \mathbf{U}\Lambda\mathbf{U}^T$ with the real and positive diagonal entries of $\Lambda = \mathrm{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_m)$ denoting the eigenvalues of $\mathbf{H}$, and the orthonormal columns of $\mathbf{U}$ denoting the eigenvectors of $\mathbf{H}$. By nonnegativity of KL divergence, for any positive definite matrix $\mathbf{X}$, $\log|\mathbf{X}| \geq \mathrm{tr}\,(\mathbf{I} - \mathbf{X}^{-1})$.

Therefore, using (C.6), we have

$$\begin{aligned}
D_{1/2}\left(p_1,p_2\right) &\geq \operatorname{tr}\left(\mathbf{I}_m-\left(\frac{\mathbf{H}^{1/2}+\mathbf{H}^{-1/2}}{2}\right)^{-1}\right)\\
&= \sum_{i=1}^{m}\left(1-\left(\frac{\lambda_i^{1/2}+\lambda_i^{-1/2}}{2}\right)^{-1}\right)\\
&= \sum_{i=1}^{m}\left(1-\frac{2\sqrt{\lambda_i}}{\lambda_i+1}\right)\\
&= \sum_{i=1}^{m}\frac{(\sqrt{\lambda_i}-1)^2}{\lambda_i+1}\\
&= \sum_{i=1}^{m}\frac{(\lambda_i-1)^2}{(1+\lambda_i)(1+\sqrt{\lambda_i})^2}\\
&\geq \sum_{i=1}^{m}\frac{(\lambda_i-1)^2}{2(1+\lambda_i)^2}\\
&= \frac{1}{2}\operatorname{tr}\left((\Lambda-\mathbf{I}_m)(\mathbf{I}_m+\Lambda)^{-1}(\Lambda-\mathbf{I}_m)(\mathbf{I}_m+\Lambda)^{-1}\right)\\
&= \frac{1}{2}\operatorname{tr}\left((\mathbf{H}-\mathbf{I}_m)(\mathbf{I}_m+\mathbf{H})^{-1}(\mathbf{H}-\mathbf{I}_m)(\mathbf{I}_m+\mathbf{H})^{-1}\right). \qquad \text{(C.7)}
\end{aligned}$$

Plugging $\mathbf{H}=\boldsymbol{\Sigma}_1^{1/2}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1^{1/2}$ back in (C.7), we obtain the desired lower bound for $\mathcal{D}_{1/2}(p_1,p_2)$ as shown below.

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

□

C.4 Concentration of Sample Covariance Matrix

Proposition C.2 (Vershynin [79]). *Let* $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_L \in \mathbb{R}^m$ *be* L *independent samples from* $\mathcal{N}(0, \mathbf{\Sigma})$*, and let* $\mathbf{\Sigma}_L = \frac{1}{L}\sum_{j=1}^{L} \mathbf{y}_j \mathbf{y}_j^T$ *denote the sample covariance matrix. Then, for any* $\epsilon > 0$*,*

$$|||\mathbf{\Sigma}_L - \mathbf{\Sigma}|||_2 \leq \epsilon \, |||\mathbf{\Sigma}|||_2 \,, \tag{C.8}$$

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

C.5 Spectral Norm Bound for Gaussian Matrices

Proposition C.3 (Corollary 5.35 in [79]). *Let* $\mathbf{A}$ *be an* $m \times n$ *matrix whose entries are independent standard normal random variables. Then for every* $t \geq 0$*, with probability at least* $1 - 2e^{-t^2/2}$*, one has*

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

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

Corollary C.1. *Let* $\mathbf{A}$ *be an* $m \times n$ *sized matrix whose entries are independent standard normal random variables. Then, for any* $\mathcal{S} \subseteq [n], |\mathcal{S}| \leq k$*, the submatrix* $\mathbf{A}_\mathcal{S}$ *obtained by sampling the columns of* $\mathbf{A}$ *indexed by* $\mathcal{S}$ *satisfies*

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

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

Proof. For a fixed support $\mathcal{S}, |\mathcal{S}| \leq k$, from Proposition C.3, we have

$$\mathbb{P}\left(|||\mathbf{A}_{\mathcal{S}}|||_2 \geq \sqrt{m} + \sqrt{k} + \sqrt{6k \log n}\right) \leq 2e^{-3k \log n}. \tag{C.9}$$

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

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{\mathcal{S}\subset[n]:|\mathcal{S}|\leq k} \left\{|||\mathbf{A}_{\mathcal{S}}|||_2 \geq \sqrt{m} + \sqrt{k} + \sqrt{6k \log n}\right\}\right) \\
&\quad \leq \left(\frac{3ne}{2}\right)^k \mathbb{P}\left(|||\mathbf{A}_{\mathcal{S}}|||_2 \geq \sqrt{m} + \sqrt{k} + \sqrt{6k \log n}\right) \\
&\quad \leq 2e^{-3k \log n + k \log(3ne/2)} \leq \frac{2}{n^k}.
\end{aligned} \tag{C.10}$$

for $n > 5$.

□

Appendix D

Appendix for Chapter 2

D.1 Derivation of the M-step Cost Function

The conditional expectation in (2.7) can be simplified as:

$$\begin{aligned}\mathbb{E}_{\mathbf{X}}\left[\log p(\mathbf{Y},\mathbf{X};\boldsymbol{\gamma})|\mathbf{Y};\boldsymbol{\gamma}^k\right] &= \mathbb{E}_{\mathbf{X}|\mathbf{Y};\boldsymbol{\gamma}^k}[\log p(\mathbf{Y}|\mathbf{X}) + \log p(\mathbf{X};\boldsymbol{\gamma})] \\ &= \mathbb{E}_{\mathbf{X}|\mathbf{Y};\boldsymbol{\gamma}^k}\log p(\mathbf{Y}|\mathbf{X}) + \sum_{j\in\mathcal{J}}\mathbb{E}_{[\mathbf{x}_j|\mathbf{y}_j;\boldsymbol{\gamma}^k]}\log p(\mathbf{x}_j;\boldsymbol{\gamma}).\end{aligned} \tag{D.1}$$

Using (2.1), and discarding the terms independent of $\boldsymbol{\gamma}$ in (D.1), the M-step objective function $Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^k)$ is given by

$$\begin{aligned}Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^k) &= \sum_{j\in\mathcal{J}}\mathbb{E}_{\mathbf{x}_j|\mathbf{y}_j,\boldsymbol{\gamma}^k}\left(-\frac{1}{2}\log|\boldsymbol{\Gamma}| - \frac{1}{2}\mathbf{x}_j^T\boldsymbol{\Gamma}^{-1}\mathbf{x}_j\right) \\ &= -\frac{1}{2}\sum_{j\in\mathcal{J}}\left(\log|\boldsymbol{\Gamma}| + \sum_{i=1}^{n}\frac{\mathbb{E}_{\mathbf{x}_j\sim\mathcal{N}(\boldsymbol{\mu}_j^{k+1},\boldsymbol{\Sigma}^{k+1})}\mathbf{x}_j(i)^2}{\boldsymbol{\gamma}(i)}\right) \\ &= -\frac{1}{2}\sum_{j\in\mathcal{J}}\sum_{i=1}^{n}\left(\log\boldsymbol{\gamma}(i) + \frac{\boldsymbol{\Sigma}^{k+1}(i,i) + \boldsymbol{\mu}_j^{k+1}(i)^2}{\boldsymbol{\gamma}(i)}\right).\end{aligned} \tag{D.2}$$

Appendix E

Appendix for Chapter 3

E.1 Proof of the Probabilistic k-RIC Bound (Theorem 3.2)

Proof. The proof of Theorem 3.2 starts with a variational definition of the k-RIC, $\delta_k\left(\frac{\mathbf{A}}{\sqrt{m}} \odot \frac{\mathbf{B}}{\sqrt{m}}\right)$ given below.

$$\delta_k\left(\frac{\mathbf{A}}{\sqrt{m}} \odot \frac{\mathbf{B}}{\sqrt{m}}\right) = \sup_{\substack{\mathbf{z}\in\mathbb{R}^n,\\ ||\mathbf{z}||_2=1, ||\mathbf{z}||_0\leq k}} \left| \left|\left| \left(\frac{\mathbf{A}}{\sqrt{m}} \odot \frac{\mathbf{B}}{\sqrt{m}}\right)\mathbf{z} \right|\right|_2^2 - 1 \right|. \tag{E.1}$$

In order to find a probabilistic upper bound for δ_k, we intend to find a constant $\delta \in (0,1)$ such that $\mathbb{P}(\delta_k\left(\frac{\mathbf{A}}{\sqrt{m}} \odot \frac{\mathbf{B}}{\sqrt{m}}\right) \geq \delta)$ is arbitrarily close to zero. We therefore consider the tail event

$$\mathcal{E} \triangleq \left\{ \sup_{\substack{\mathbf{z}\in\mathbb{R}^n,\\ ||\mathbf{z}||_2=1, ||\mathbf{z}||_0\leq k}} \left| \left|\left| \left(\frac{\mathbf{A}}{\sqrt{m}} \odot \frac{\mathbf{B}}{\sqrt{m}}\right)\mathbf{z} \right|\right|_2^2 - 1 \right| \geq \delta \right\}, \tag{E.2}$$

and show that for m sufficiently large, $\mathbb{P}(\mathcal{E})$ can be driven arbitrarily close to zero. In other words, the constant δ serves as a probabilistic upper bound for $\delta_k\left(\frac{\mathbf{A}}{\sqrt{m}} \odot \frac{\mathbf{B}}{\sqrt{m}}\right)$. Let $\mathcal{U}_k$ denote the set of all k or less sparse unit norm vectors in $\mathbb{R}^n$. Then, using Proposition 3.1,

the tail event in (E.2) can be rewritten as

$$\begin{aligned}\mathbb{P}(\mathcal{E}) &= \mathbb{P}\left(\sup_{\mathbf{z}\in\mathcal{U}_k}\left|\mathbf{z}^T\left(\mathbf{A}\odot\mathbf{B}\right)^T\left(\mathbf{A}\odot\mathbf{B}\right)\mathbf{z}-m^2\right|\geq\delta m^2\right)\\ &= \mathbb{P}\left(\sup_{\mathbf{z}\in\mathcal{U}_k}\left|\mathbf{z}^T\left(\mathbf{A}^T\mathbf{A}\circ\mathbf{B}^T\mathbf{B}\right)\mathbf{z}-m^2\right|\geq\delta m^2\right)\\ &= \mathbb{P}\left(\sup_{\mathbf{z}\in\mathcal{U}_k}\left|\sum_{i=1}^n\sum_{j=1}^n z_iz_j\left(\mathbf{a}_i^T\mathbf{a}_j\right)\left(\mathbf{b}_i^T\mathbf{b}_j\right)-m^2\right|\geq\delta m^2\right) \qquad \text{(E.3)}\end{aligned}$$

where $\mathbf{a}_i$ and $\mathbf{b}_i$ denote the ith column of $\mathbf{A}$ and $\mathbf{B}$, respectively. Further, by applying the triangle inequality and the union bound, the above tail probability splits as

$$\begin{aligned}\mathbb{P}(\mathcal{E}) \leq\ & \mathbb{P}\left(\sup_{\mathbf{z}\in\mathcal{U}_k}\left|\sum_{i=1}^n z_i^2\left|\left|\mathbf{a}_i\right|\right|_2^2\left|\left|\mathbf{b}_i\right|\right|_2^2-m^2\right|\geq\alpha\delta m^2\right)\\ &+\mathbb{P}\left(\sup_{\mathbf{z}\in\mathcal{U}_k}\left|\sum_{i=1}^n\sum_{j=1,j\neq i}^n z_iz_j\mathbf{a}_i^T\mathbf{a}_j\mathbf{b}_i^T\mathbf{b}_j\right|\geq(1-\alpha)\delta m^2\right). \qquad \text{(E.4)}\end{aligned}$$

In the above, $\alpha\in(0,1)$ is a variational union bound parameter which can be optimized at a later stage. We now proceed to find separate upper bounds for each of the two probability terms in (E.4).

The first probability term in (E.4) admits the following sequence of relaxations.

$$\begin{aligned}&\mathbb{P}\left(\sup_{\mathbf{z}\in\mathcal{U}_k}\left|\sum_{i=1}^n z_i^2\left|\left|\mathbf{a}_i\right|\right|_2^2\left|\left|\mathbf{b}_i\right|\right|_2^2-m^2\right|\geq\alpha\delta m^2\right)\\ \overset{(a)}{\leq}\ &\mathbb{P}\left(\sup_{\mathbf{z}\in\mathcal{U}_k}\sum_{i=1}^n z_i^2\left|\left|\left|\mathbf{a}_i\right|\right|_2^2\left|\left|\mathbf{b}_i\right|\right|_2^2-m^2\right|\geq\alpha\delta m^2\right)\\ \overset{(b)}{\leq}\ &\mathbb{P}\left(\max_{1\leq i\leq n}\left|\left|\left|\mathbf{a}_i\right|\right|_2^2\left|\left|\mathbf{b}_i\right|\right|_2^2-m^2\right|\geq\alpha\delta m^2\right)\\ \overset{(c)}{=}\ &\mathbb{P}\left(\bigcup_{1\leq i\leq n}\left\{\left|\left|\left|\mathbf{a}_i\right|\right|_2^2\left|\left|\mathbf{b}_i\right|\right|_2^2-m^2\right|\geq\alpha\delta m^2\right\}\right)\\ \overset{(d)}{=}\ &n\mathbb{P}\left(\left|\left|\left|\mathbf{a}_1\right|\right|_2^2\left|\left|\mathbf{b}_1\right|\right|_2^2-m^2\right|\geq\alpha\delta m^2\right)\\ \overset{(e)}{\leq}\ &n\mathbb{P}\left(\left|\left|\left|\mathbf{a}_1\right|\right|_2^2-m\right|\left|\left|\left|\mathbf{b}_1\right|\right|_2^2-m\right|\geq\alpha\beta\delta m^2\right)+2n\mathbb{P}\left(\left|\left|\left|\mathbf{a}_1\right|\right|_2^2-m\right|\geq\frac{\alpha(1-\beta)\delta m}{2}\right)\end{aligned}$$

$$\begin{aligned}
&\overset{(f)}{\leq} \quad 2n\mathbb{P}\left(\left|||\mathbf{a}_1||_2^2 - m\right| \geq \sqrt{\alpha\beta\delta}m\right) + 2n\mathbb{P}\left(\left|||\mathbf{a}_1||_2^2 - m\right| \geq \frac{\alpha(1-\beta)\delta m}{2}\right). \\
&\overset{(g)}{\leq} \quad 4n\mathbb{P}\left(\left|||\mathbf{a}_1||_2^2 - m\right| \geq \frac{\alpha\delta m}{2}\left(1 - \frac{\alpha\delta}{4}\right)\right) \\
&\overset{(h)}{\leq} \quad 8ne^{-cm\frac{\alpha^2\delta^2}{4\kappa^4}(1-\alpha\delta/4)^2} = 8n^{-\left(\frac{cm\alpha^2\delta^2(1-\alpha\delta/4)^2}{4\kappa^4\log n} - 1\right)}.
\end{aligned} \tag{E.5}$$

In the above, step (a) follows from the triangle inequality combined with the fact that z_i^2's sum to one. The inequality in step (b) is a consequence of the fact that any nonnegative convex combination of n arbitrary numbers is at most the maximum among the n numbers. Step (c) is obtained by simply rewriting the tail event for the maximum of n random variables as the union of tail events for the individual random variables. Step (d) is the application of the union bound over values of index $i \in [n]$ and exploiting the i.i.d. nature of the columns of $\mathbf{A}$ and $\mathbf{B}$. Step (e) is the union bound combined with the fact that for any two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^m$, the following triangle inequality holds:

$$\left|||\mathbf{a}||_2^2\,||\mathbf{b}||_2^2 - m^2\right| \leq \left|\left(||\mathbf{a}||_2^2 - m\right)\left(||\mathbf{b}||_2^2 - m\right)\right| + m\left|||\mathbf{a}||_2^2 - m\right| + m\left|||\mathbf{b}||_2^2 - m\right|.$$

In step (e), $\beta \in (0,1)$ is a variational union bound parameter. Step (f) is once again the union bound which exploits the fact that the columns $\mathbf{a}_1$ and $\mathbf{b}_1$ are identically distributed. Step (g) is obtained by setting $\beta = \alpha\delta/4$. Lastly, step (h) is the Hanson-Wright inequality (Theorem B.1) applied to the subgaussian vector $\mathbf{a}_1$.

Next, we turn our attention to the second tail probability term in (E.4). We note that

$$\begin{aligned}
\sup_{\mathbf{z}\in\mathcal{U}_k}\left|\sum_{i=1}^n\sum_{j=1,j\neq i}^n z_iz_j\mathbf{a}_i^T\mathbf{a}_j\mathbf{b}_i^T\mathbf{b}_j\right| &\leq \sup_{\mathbf{z}\in\mathcal{U}_k}\sum_{i=1}^n\sum_{j=1,j\neq i}^n |z_iz_j|\left|\mathbf{a}_i^T\mathbf{a}_j\right|\left|\mathbf{b}_i^T\mathbf{b}_j\right| \\
&\leq \sup_{\mathbf{z}\in\mathcal{U}_k}\left(\sum_{i=1}^n\sum_{j=1,j\neq i}^n |z_iz_j|\right)\left(\max_{\substack{i,j\in\mathrm{supp}(\mathbf{u}),\\ i\neq j}}\left|\mathbf{a}_i^T\mathbf{a}_j\right|\left|\mathbf{b}_i^T\mathbf{b}_j\right|\right)
\end{aligned}$$

$$\leq \quad k \left(\max_{\substack{i,j \in [n], \\ i \neq j}} \left| \mathbf{a}_i^T \mathbf{a}_j \right| \left| \mathbf{b}_i^T \mathbf{b}_j \right| \right), \tag{E.6}$$

where the second step is the application of the Hölders inequality. The last step uses the fact that $||\mathbf{z}||_1 \leq \sqrt{k}$ for $\mathbf{z} \in \mathcal{U}_k$. Using (E.6), and by applying the union bound over $\binom{n}{2}$ possible distinct (i, j) pairs, the second probability term in (E.4) can be bounded as

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\mathbf{z} \in \mathcal{U}_k} \left| \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} z_i z_j \mathbf{a}_i^T \mathbf{a}_j \mathbf{b}_i^T \mathbf{b}_j \right| \geq (1-\alpha)\delta m^2 \right) \\
&\quad \leq \frac{n^2}{2} \mathbb{P}\left(\left| \mathbf{a}_1^T \mathbf{a}_2 \right| \left| \mathbf{b}_1^T \mathbf{b}_2 \right| \geq \frac{(1-\alpha)\delta m^2}{k} \right) \\
&\quad \leq n^2 \mathbb{P}\left(\left| \mathbf{a}_1^T \mathbf{a}_2 \right| \geq \frac{\sqrt{(1-\alpha)\delta} m}{\sqrt{k}} \right) \\
&\quad \leq 2n^2 e^{-\frac{c(1-\alpha)\delta m}{\kappa^2 k}} = 2n^{-\left(\frac{c(1-\alpha)\delta m}{\kappa^2 k \log n} - 2 \right)}.
\end{aligned} \tag{E.7}$$

The last inequality in the above is obtained by using the tail bound for $|\mathbf{a}_1^T \mathbf{a}_2|$ from Corollary B.1. Finally, by combining (E.4), (E.5) and (E.7), and setting $\alpha = 1/2$, we obtain the following simplified tail bound,

$$\mathbb{P}(\mathcal{E}) \leq 8n^{-\left(\frac{cm\delta^2(1-\delta/8)^2}{16\kappa^4 \log n} - 1 \right)} + 2n^{-\left(\frac{c\delta m}{2\kappa^2 k \log n} - 2 \right)}. \tag{E.8}$$

From (E.8), for $m > \max\left(\frac{4\gamma\kappa^2 k \log n}{c\delta}, \frac{32\gamma\kappa^4 \log n}{c\delta^2(1-\delta/8)^2} \right)$ and any $\gamma > 1$, we have $\mathbb{P}(\mathcal{E}) < 10n^{-2(\gamma-1)}$. Note that, in terms of k and n, the first term in the inequality for m scales as $k \log n$; it dominates the second term, which scale as $\log n$. This ends our proof. □

E.2 Proof of Theorem 3.3

Proof. The proof of Theorem 3.3 is along similar lines as that of Theorem 3.2. We consider the tail event

$$\mathcal{E}_1 \triangleq \left\{ \sup_{\mathbf{z}\in\mathcal{U}_k} \left| \left|\left| \left(\frac{\mathbf{A}}{\sqrt{m}} \odot \frac{\mathbf{A}}{\sqrt{m}} \right) \mathbf{z} \right|\right|_2^2 - 1 \right| \geq \delta \right\}, \tag{E.9}$$

and show that for sufficiently large m, $\mathbb{P}(\mathcal{E}_1)$ can be driven arbitrarily close to zero, thereby implying that δ is a probabilistic upper bound for $\delta_k\left((\mathbf{A}/\sqrt{m} \odot \mathbf{A}/\sqrt{m})\right)$. Once again, $\mathcal{U}_k$ denotes the set of all k or less sparse unit norm vectors in $\mathbb{R}^m$. We note that $\mathbb{P}(\mathcal{E}_1)$ admits the following union bound:

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_1) &= \mathbb{P}\left(\sup_{\mathbf{z}\in\mathcal{U}_k} \left| \mathbf{z}^T \left(\mathbf{A}\odot\mathbf{A}\right)^T \left(\mathbf{A}\odot\mathbf{A}\right) \mathbf{z} - m^2 \right| \geq \delta m^2 \right) \\
&= \mathbb{P}\left(\sup_{\mathbf{z}\in\mathcal{U}_k} \left| \mathbf{z}^T \left(\mathbf{A}^T\mathbf{A}\circ\mathbf{A}^T\mathbf{A}\right) \mathbf{z} - m^2 \right| \geq \delta m^2 \right) \\
&= \mathbb{P}\left(\sup_{\mathbf{z}\in\mathcal{U}_k} \left| \sum_{i=1}^n \sum_{j=1}^n z_i z_j \left(\mathbf{a}_i^T \mathbf{a}_j\right)^2 - m^2 \right| \geq \delta m^2 \right) \\
&\leq \mathbb{P}\left(\sup_{\mathbf{z}\in\mathcal{U}_k} \left| \sum_{i=1}^n z_i^2 \left|\left|\mathbf{a}_i\right|\right|_2^4 - m^2 \right| \geq \alpha\delta m^2 \right) \\
&\qquad + \mathbb{P}\left(\sup_{\mathbf{z}\in\mathcal{U}_k} \left| \sum_{i=1}^n \sum_{j=1,j\neq i}^n z_i z_j \left(\mathbf{a}_i^T \mathbf{a}_j\right)^2 \right| \geq (1-\alpha)\delta m^2 \right).
\end{aligned} \tag{E.10}$$

In the above, the second identity follows from Proposition 3.1. The last inequality uses the triangle inequality followed by the union bound, with $\alpha \in (0,1)$ being a variational parameter to be optimized later. Similar to the proof of Theorem 3.2, we now derive separate upper bounds for each of the two probability terms in (E.10).

The first term in (E.10) admits the following series of relaxations.

$$\mathbb{P}\left(\sup_{\mathbf{z}\in\mathcal{U}_k} \left| \sum_{i=1}^n z_i^2 \left|\left|\mathbf{a}_i\right|\right|_2^4 - m^2 \right| \geq \alpha\delta m^2 \right)$$

$$
\begin{aligned}
&\overset{(a)}{\leq} \mathbb{P}\left(\sup_{\mathbf{z}\in\mathcal{U}_k}\sum_{i=1}^{n} z_i^2\left|||\mathbf{a}_i||_2^4 - m^2\right| \geq \alpha\delta m^2\right)\\
&\overset{(b)}{\leq} \mathbb{P}\left(\max_{1\leq i\leq n}\left|||\mathbf{a}_i||_2^4 - m^2\right| \geq \alpha\delta m^2\right)\\
&\overset{(c)}{\leq} n\mathbb{P}\left(\left|||\mathbf{a}_1||_2^4 - m^2\right| \geq \alpha\delta m^2\right).\\
&\overset{(d)}{\leq} n\mathbb{P}\left(\left|||\mathbf{a}_1||_2^2 - m\right|^2 \geq \alpha\beta\delta m^2\right) + n\mathbb{P}\left(\left|||\mathbf{a}_1||_2^2 - m\right| \geq \frac{\alpha(1-\beta)\delta m}{2}\right)\\
&\overset{(e)}{\leq} n\mathbb{P}\left(\left|||\mathbf{a}_1||_2^2 - m\right| \geq \sqrt{\alpha\beta\delta}m\right) + n\mathbb{P}\left(\left|||\mathbf{a}_1||_2^2 - m\right| \geq \frac{\alpha(1-\beta)\delta m}{2}\right).\\
&\overset{(f)}{\leq} 2n\mathbb{P}\left(\left|||\mathbf{a}_1||_2^2 - m\right| \geq \frac{\alpha\delta m}{2}\left(1-\frac{\alpha\delta}{4}\right)\right)\\
&\overset{(g)}{\leq} 4ne^{-cm\frac{\alpha^2\delta^2}{4\kappa^4}(1-\alpha\delta/4)^2} = 4n^{-\left(\frac{cm\alpha^2\delta^2(1-\alpha\delta/4)^2}{4\kappa^4\log n}-1\right)}.
\end{aligned}
\tag{E.11}
$$

In the above, step (a) is the triangle inequality. The inequality in step (b) follows from the fact that nonnegative convex combination of n arbitrary numbers is at most the maximum among the n numbers. Step (c) is a union bound. Step (d) is also a union bounding argument with $\beta \in (0,1)$ as a variational parameter, combined with the fact that for any vector $\mathbf{a}$, the triangle inequality $\left|||\mathbf{a}||_2^4 - m^2\right| \leq \left|||\mathbf{a}||_2^2 - m\right|^2 + 2m\left|||\mathbf{a}||_2^2 - m\right|$ is always true. Step (f) is obtained by choosing the union bound parameter $\beta = \alpha\delta/4$. Finally, step (g) is the Hanson-Wright inequality (Theorem B.1) applied to the subgaussian vector $\mathbf{a}_1$.

Next, we derive an upper bound for the second tail probability term in (E.4). We observe that

$$
\begin{aligned}
\sup_{\mathbf{z}\in\mathcal{U}_k}\left|\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} z_i z_j\left(\mathbf{a}_i^T\mathbf{a}_j\right)^2\right| &\leq \sup_{\mathbf{z}\in\mathcal{U}_k}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}|z_i z_j|\left(\mathbf{a}_i^T\mathbf{a}_j\right)^2\\
&\leq \sup_{\mathbf{z}\in\mathcal{U}_k}\left(\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}|z_i z_j|\right)\left(\max_{\substack{i,j\in\mathrm{supp}(\mathbf{u}),\\ i\neq j}}\left|\mathbf{a}_i^T\mathbf{a}_j\right|^2\right)\\
&\leq k\left(\max_{\substack{i,j\in[n],\\ i\neq j}}\left|\mathbf{a}_i^T\mathbf{a}_j\right|^2\right),
\end{aligned}
\tag{E.12}
$$

where the second step is the application of the Hölders inequality. The last step uses the fact that $||\mathbf{z}||_1 \leq \sqrt{k}$ for $\mathbf{z} \in \mathcal{U}_k$. Using (E.12), and by applying the union bound over $\binom{n}{2}$ possible distinct (i, j) pairs, the second probability term in (E.4) can be bounded as

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\mathbf{z}\in\mathcal{U}_k}\left|\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} z_i z_j \left(\mathbf{a}_i^T\mathbf{a}_j\right)^2\right| \geq (1-\alpha)\delta m^2\right) \\
&\quad\leq \frac{n^2}{2}\mathbb{P}\left(\left|\mathbf{a}_1^T\mathbf{a}_2\right|^2 \geq \frac{(1-\alpha)\delta m^2}{k}\right) \\
&\quad\leq n^2\mathbb{P}\left(\left|\mathbf{a}_1^T\mathbf{a}_2\right| \geq \frac{\sqrt{(1-\alpha)\delta}m}{\sqrt{k}}\right) \\
&\quad\leq n^2 e^{-\frac{c(1-\alpha)\delta m}{\kappa^2 k}} = n^{-\left(\frac{c(1-\alpha)\delta m}{\kappa^2 k\log n}-2\right)}. \qquad \text{(E.13)}
\end{aligned}$$

The last inequality in the above is obtained by using the tail bound for $|\mathbf{a}_1^T\mathbf{a}_2|$ from Corollary B.1. Finally, by combining (E.10), (E.11) and (E.13), and setting $\alpha = 1/2$, we obtain the following tail bound.

$$\mathbb{P}(\mathcal{E}) \leq 4n^{-\left(\frac{cm\delta^2(1-\delta/8)^2}{16\kappa^4\log n}-1\right)} + n^{-\left(\frac{c\delta m}{2\kappa^2 k\log n}-2\right)} \qquad \text{(E.14)}$$

From (E.14), we observe that for $m > \max\left(\frac{4\gamma\kappa^2 k\log n}{c\delta}, \frac{32\gamma\kappa^4\log n}{c\delta^2(1-\delta/8)^2}\right)$ and any $\gamma > 1$, we have $\mathbb{P}(\mathcal{E}) < 5n^{-2(\gamma-1)}$. Note that, in terms of k and n, the first term in the inequality for m scales as $\frac{k\log n}{\delta}$; it dominates the second term, which scale as $\log n$. □

Appendix F

Appendix for Chapter 4

F.1 Proof of Proposition 4.1

Proof. 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 the δ-blow up of $\Theta(\mathcal{S})$, denoted by $\Theta_{\uparrow\delta}(\mathcal{S})$, and defined as

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

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

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

4. Then, $\Theta^{\delta}(\mathcal{S}) \triangleq \left(\Theta^{\delta}_{\uparrow\delta}(\mathcal{S}) \cap \Theta(\mathcal{S})\right) \cup \mathcal{P}$ 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 $\boldsymbol{\gamma} \in \Theta(\mathcal{S})$, there exists an element $\mathbf{a}$ in $\Theta^{\delta}(\mathcal{S})$ such that $\|\boldsymbol{\gamma} - \mathbf{a}\|_2 \leq \delta$. Let $\boldsymbol{\gamma}$ be an arbitrary

element in $\Theta(\mathcal{S})$. Then, $\boldsymbol{\gamma}$ also belongs to the larger set $\Theta_{\uparrow\delta}(\mathcal{S})$, and consequently, there exists $\boldsymbol{\gamma}' \in \Theta^{\delta}_{\uparrow\delta}(\mathcal{S})$ such that $\|\boldsymbol{\gamma} - \boldsymbol{\gamma}'\|_2 \leq \delta$. Now, there are two cases. (i) $\boldsymbol{\gamma}' \in \Theta(\mathcal{S})$, and (ii) $\boldsymbol{\gamma}' \notin \Theta(\mathcal{S})$.

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

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

$$\begin{aligned} \delta &\geq ||\boldsymbol{\gamma} - \boldsymbol{\gamma}'||_2^2 = ||(\boldsymbol{\gamma} - \boldsymbol{\gamma}'') + (\boldsymbol{\gamma}'' - \boldsymbol{\gamma}')||_2^2 \\ &= ||\boldsymbol{\gamma} - \boldsymbol{\gamma}''||_2^2 + ||\boldsymbol{\gamma}'' - \boldsymbol{\gamma}'||_2^2 + 2\langle \boldsymbol{\gamma} - \boldsymbol{\gamma}'', \boldsymbol{\gamma}'' - \boldsymbol{\gamma}' \rangle \\ &\geq ||\boldsymbol{\gamma} - \boldsymbol{\gamma}''||_2^2 . \end{aligned} \tag{F.1}$$

The last inequality is obtained by dropping the last two nonnegative terms in the RHS. From (F.1), $\mathbf{a} = \boldsymbol{\gamma}''$ 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 $\sqrt{|\mathcal{S}|}(\boldsymbol{\gamma}_{\max} - \boldsymbol{\gamma}_{\min})$. Based on the construction in step-4, the cardinality of $\Theta^{\delta}(\mathcal{S})$ can be upper bounded as:

$$\begin{aligned} |\Theta^{\delta}(\mathcal{S})| &= |\Theta^{\delta}_{\uparrow\delta}(\mathcal{S}) \cap \Theta(\mathcal{S})| + |\Theta^{\delta}_{\uparrow\delta}(\mathcal{S}) \cap \Theta(\mathcal{S})^c| \\ &= |\Theta^{\delta}_{\uparrow\delta}(\mathcal{S})| \leq \left|\delta\text{-net of } \ell_2 \text{ ball of radius } \sqrt{|\mathcal{S}|}(\boldsymbol{\gamma}_{\max} - \boldsymbol{\gamma}_{\min}) \text{ in } \mathbb{R}^{|\mathcal{S}|}\right| \end{aligned}$$

$$\leq \max\left(1, \left(\frac{3\sqrt{|\mathcal{S}|}(\boldsymbol{\gamma}_{\max} - \boldsymbol{\gamma}_{\min})}{\delta}\right)^{|\mathcal{S}|}\right). \tag{F.2}$$

The last step is an extension of the volumetric arguments in [147] 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 $\Theta(\mathcal{S})$.

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 $\mathcal{L}(\mathbf{Y}, \boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma} \in \Theta(\mathcal{S})$. We claim that $\Theta_{\mathcal{S}}^{\epsilon/C_{\mathcal{L},\mathcal{S}}}$ is the desired set which simultaneously satisfies conditions (i) and (ii) stated in Proposition 4.1.

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

Condition (ii) follows from (F.2) by setting $\delta = \epsilon/C_{\mathcal{L},\mathcal{S}}$. □

F.2 Proof of Theorem 4.1

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

$$\begin{aligned}
\mathbb{P}\left(\log\frac{p_{\boldsymbol{\gamma}}(\mathbf{Y})}{p_{\boldsymbol{\gamma}^*}(\mathbf{Y})} \geq -\epsilon\right) &= \mathbb{P}\left(\sum_{j=1}^{L}\log\frac{p_{\boldsymbol{\gamma}}(\mathbf{y}_j)}{p_{\boldsymbol{\gamma}^*}(\mathbf{y}_j)} \geq -\epsilon\right) \\
&\leq \mathbb{E}_{p_{\boldsymbol{\gamma}^*}}\left[\exp\left(t\sum_{j=1}^{L}\log\frac{p_{\boldsymbol{\gamma}}(\mathbf{y}_j)}{p_{\boldsymbol{\gamma}^*}(\mathbf{y}_j)}\right)\right]\exp\left(t\epsilon\right)
\end{aligned}$$

$$
\begin{aligned}
&= \left(\mathbb{E}_{p_{\boldsymbol{\gamma}^*}}\left[\exp\left(t\log\frac{p_{\boldsymbol{\gamma}}(\mathbf{y})}{p_{\boldsymbol{\gamma}^*}(\mathbf{y})}\right)\right]\right)^L \exp\left(t\epsilon\right) \\
&= \left(\mathbb{E}_{p_{\boldsymbol{\gamma}^*}}\left[\left(\frac{p_{\boldsymbol{\gamma}}(\mathbf{y})}{p_{\boldsymbol{\gamma}^*}(\mathbf{y})}\right)^t\right]\right)^L \exp\left(t\epsilon\right) \\
&= \left(\int_{\mathbf{y}\in\mathcal{Y}} p_{\boldsymbol{\gamma}}(\mathbf{y})^t p_{\boldsymbol{\gamma}^*}(\mathbf{y})^{1-t} d\mathbf{y}\right)^L \exp\left(t\epsilon\right) \\
&= \exp\left(-L\left[t\left(-\frac{\epsilon}{L}\right) - (t-1)\mathcal{D}_t(p_{\boldsymbol{\gamma}}, p_{\boldsymbol{\gamma}^*})\right]\right). \qquad \text{(F.3)}
\end{aligned}
$$

In the above, the first and third steps follow from the independence of $\mathbf{y}_j$. The second step is the application of Chernoff bound. The last step is obtained by using the definition of the Rényi divergence from (C.4) and rearranging the terms in the exponent.

By introducing the function $\psi(t) = (t-1)\mathcal{D}_t(p_{\boldsymbol{\gamma}}, p_{\boldsymbol{\gamma}^*})$, the Chernoff bound (F.3) can be restated as

$$
\mathbb{P}\left(\log\frac{p_{\boldsymbol{\gamma}}(\mathbf{Y})}{p_{\boldsymbol{\gamma}^*}(\mathbf{Y})} \geq -\epsilon\right) \leq \exp\left(-L\left[t\left(-\frac{\epsilon}{L}\right) - \psi(t)\right]\right). \qquad \text{(F.4)}
$$

For $t = \arg\sup_{t>0}\left(t\left(-\frac{\epsilon}{L}\right) - \psi(t)\right)$, the upper bound in (F.4) attains its tightest value $\exp\left(-L\psi^*(-\epsilon/L)\right)$, where ψ^* is the Legendre transform of ψ. □

F.3 Proof of Theorem 4.3

Proof. Consider the unit norm m^2 length vector $\mathbf{w} \triangleq \frac{\mathbf{1}_{\{1,m+2,2m+3,\ldots,m^2\}}}{\sqrt{m}}$. 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

$$
\mathbf{b}(i) = \frac{(\mathbf{a}_i \circ \mathbf{a}_i)^T\mathbf{1}_m}{\sqrt{m}} = \frac{||\mathbf{a}_i||_2^2}{\sqrt{m}} \;\; \forall i \in [n], \qquad \text{(F.5)}
$$

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

$$\frac{1-\alpha}{\sqrt{m}}\mathbf{1}_n \preceq (\mathbf{A} \odot \mathbf{A})^T \mathbf{w} \preceq \frac{1+\alpha}{\sqrt{m}}\mathbf{1}_n. \tag{F.6}$$

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

$$\begin{aligned}
\mathbf{v}^T\mathbf{X}^T\mathbf{X}\mathbf{v} &= \mathbf{v}^T\mathbf{X}^T\mathbf{w}\mathbf{w}^T\mathbf{X}\mathbf{v} + \mathbf{v}^T\mathbf{X}^T\Pi\mathbf{X}\mathbf{v} \\
&\geq \mathbf{v}^T\mathbf{X}^T\mathbf{w}\mathbf{w}^T\mathbf{X}\mathbf{v} \\
&= \mathbf{v}_+^T\mathbf{X}^T\mathbf{w}\mathbf{w}^T\mathbf{X}\mathbf{v}_+ + \mathbf{v}_-^T\mathbf{X}^T\mathbf{w}\mathbf{w}^T\mathbf{X}\mathbf{v}_- - 2\mathbf{v}_+^T\mathbf{X}^T\mathbf{w}\mathbf{w}^T\mathbf{X}\mathbf{v}_- \\
&\overset{(a)}{\geq} \frac{(1-\alpha)^2}{m}\left(\mathbf{v}_+^T\mathbf{1}_n\mathbf{1}_n^T\mathbf{v}_+\right) + \frac{(1-\alpha)^2}{m}\left(\mathbf{v}_-^T\mathbf{1}_n\mathbf{1}_n^T\mathbf{v}_-\right) - 2\frac{(1+\alpha)^2}{m}\left(\mathbf{v}_+^T\mathbf{1}_n\mathbf{1}_n^T\mathbf{v}_-\right) \\
&= \frac{(1-\alpha)^2}{m}\left(||\mathbf{v}_+||_1^2 + ||\mathbf{v}_-||_1^2\right) - 2\,||\mathbf{v}_+||_1\,||\mathbf{v}_-||_1\,\frac{(1+\alpha)^2}{m} \\
&= \frac{(1-\alpha)^2}{m}\left(||\mathbf{v}_+||_1^2 + ||\mathbf{v}_-||_1^2\right)\left[1 - \frac{2\,||\mathbf{v}_+||_1\,||\mathbf{v}_-||_1}{||\mathbf{v}_+||_1^2 + ||\mathbf{v}_-||_1^2}\left(\frac{1+\alpha}{1-\alpha}\right)^2\right].
\end{aligned} \tag{F.7}$$

In above, step (a) follows from (F.6). 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

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

□

F.4 Proof of Proposition 4.2

Proof. From Theorem C.1, we have the following pointwise lower bound.

$$\mathcal{D}_{1/2}\left(p_{\boldsymbol{\gamma}}, p_{\boldsymbol{\gamma}^*}\right) \geq \operatorname{tr}\left(\left(\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} - \boldsymbol{\Sigma}_{\boldsymbol{\gamma}^*}\right)\left(\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} + \boldsymbol{\Sigma}_{\boldsymbol{\gamma}^*}\right)^{-1}\left(\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} - \boldsymbol{\Sigma}_{\boldsymbol{\gamma}^*}\right)\left(\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} + \boldsymbol{\Sigma}_{\boldsymbol{\gamma}^*}\right)^{-1}\right). \tag{F.8}$$

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

$$\begin{aligned}
\mathcal{D}_{1/2}(p_{\boldsymbol{\gamma}}, p_{\boldsymbol{\gamma}^*}) &\geq \text{tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\gamma}+\boldsymbol{\gamma}^*}^{-1}(\mathbf{A}\Delta\Gamma\mathbf{A}^T)\boldsymbol{\Sigma}_{\boldsymbol{\gamma}+\boldsymbol{\gamma}^*}^{-1}(\mathbf{A}\Delta\Gamma\mathbf{A}^T)\right) \\
&\geq \frac{\text{tr}\left((\mathbf{A}\Delta\Gamma\mathbf{A}^T)\boldsymbol{\Sigma}_{\boldsymbol{\gamma}+\boldsymbol{\gamma}^*}^{-1}(\mathbf{A}\Delta\Gamma\mathbf{A}^T)\right)}{|||\boldsymbol{\Sigma}_{\boldsymbol{\gamma}+\boldsymbol{\gamma}^*}|||_2} \\
&\geq \frac{\text{tr}\left((\mathbf{A}\Delta\Gamma\mathbf{A}^T)(\mathbf{A}\Delta\Gamma\mathbf{A}^T)\right)}{|||\boldsymbol{\Sigma}_{\boldsymbol{\gamma}+\boldsymbol{\gamma}^*}|||_2^2} \\
&= \frac{\left|\left|\mathbf{A}\Delta\Gamma\mathbf{A}^T\right|\right|_F^2}{|||\boldsymbol{\Sigma}_{\boldsymbol{\gamma}+\boldsymbol{\gamma}^*}|||_2^2} = \frac{||(\mathbf{A}\odot\mathbf{A})\Delta\boldsymbol{\gamma}||_2^2}{|||\boldsymbol{\Sigma}_{\boldsymbol{\gamma}+\boldsymbol{\gamma}^*}|||_2^2}.
\end{aligned} \tag{F.9}$$

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

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

$$\begin{aligned}
|||\boldsymbol{\Sigma}_{\boldsymbol{\gamma}+\boldsymbol{\gamma}^*}|||_2 &= \left|\left|\left|\sigma^2\mathbf{I}_m + \mathbf{A}(\boldsymbol{\Gamma}+\boldsymbol{\Gamma}^*)\mathbf{A}^T\right|\right|\right|_2 \\
&\leq \sigma^2 + 2\boldsymbol{\gamma}_{\max}\left|\left|\left|\mathbf{A}_{\mathcal{S}\cup\mathcal{S}^*}^T\mathbf{A}_{\mathcal{S}\cup\mathcal{S}^*}\right|\right|\right|_2.
\end{aligned} \tag{F.10}$$

Finally, using (F.10) in (F.9), we obtain the desired lower bound for $\mathcal{D}_{1/2}(p_{\boldsymbol{\gamma}}, p_{\boldsymbol{\gamma}^*})$. □

F.5 Proof of Proposition 4.3

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

$$\Delta\boldsymbol{\gamma} = \Delta\boldsymbol{\gamma}_+ - \Delta\boldsymbol{\gamma}_-, \tag{F.11}$$

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, that differ in exactly $k_d^{\mathcal{S},\mathcal{S}^*}$ locations. By construction of $\Delta\boldsymbol{\gamma}_+$ and $\Delta\boldsymbol{\gamma}_-$, we have

$$||\Delta\boldsymbol{\gamma}||_2^2 \quad \geq k_d^{\mathcal{S},\mathcal{S}^*}\boldsymbol{\gamma}_{\min}^2 \tag{F.12}$$

$$\left(|\mathcal{S}^*| - k_d^{\mathcal{S},\mathcal{S}^*}\right)_+ \boldsymbol{\gamma}_{\min} \leq \quad \left|\left|\Delta\boldsymbol{\gamma}_-\right|\right|_1 \quad \leq |\mathcal{S}^*|\boldsymbol{\gamma}_{\max} \tag{F.13}$$

$$\left(k_d^{\mathcal{S},\mathcal{S}^*} - |\mathcal{S}^*|\right)_+ \boldsymbol{\gamma}_{\min} \leq \quad \left|\left|\Delta\boldsymbol{\gamma}_+\right|\right|_1 \quad \leq k_d^{\mathcal{S},\mathcal{S}^*}\boldsymbol{\gamma}_{\max} + |\mathcal{S}^*|(\boldsymbol{\gamma}_{\max} - \boldsymbol{\gamma}_{\min}) \tag{F.14}$$

We introduce $K_{\text{threshold}} \triangleq \left(1 + 4\frac{\boldsymbol{\gamma}_{\max}}{\boldsymbol{\gamma}_{\min}}\left(\frac{1+\alpha}{1-\alpha}\right)^2\right)K$, and $\mathcal{B} \triangleq \left\{\mathcal{S} \in [n] : k_d^{\mathcal{S},\mathcal{S}^*} \leq K_{\text{threshold}}\right\}$. Then, from (4.14), we have

$$\eta = \min_{\mathcal{S}\subseteq[n]} \frac{\mathcal{D}_{\mathcal{S}}^*}{k_d^{\mathcal{S},\mathcal{S}^*}} = \min\left(\min_{\mathcal{S}\in\mathcal{B}} \frac{\mathcal{D}_{\mathcal{S}}^*}{k_d^{\mathcal{S},\mathcal{S}^*}},\ \min_{\mathcal{S}\in\mathcal{B}^c} \frac{\mathcal{D}_{\mathcal{S}}^*}{k_d^{\mathcal{S},\mathcal{S}^*}}\right). \tag{F.15}$$

Note that for $\text{supp}(\boldsymbol{\gamma}) = \mathcal{S}$ and $\mathcal{S} \in \mathcal{B}$, 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 4.2, we can bound $\min_{\mathcal{S}\in\mathcal{B}} \frac{\mathcal{D}_{\mathcal{S}}^*}{k_d^{\mathcal{S},\mathcal{S}^*}}$ as follows.

$$\begin{aligned}
\min_{\mathcal{S}\in\mathcal{B}} \frac{\mathcal{D}_{\mathcal{S}}^*}{k_d^{\mathcal{S},\mathcal{S}^*}} &= \min_{\mathcal{S}\in\mathcal{B}} \frac{\beta\,||\Delta\boldsymbol{\gamma}||_2^2}{k_d^{\mathcal{S},\mathcal{S}^*}\left(\sigma^2 + 2\boldsymbol{\gamma}_{\max}\sigma_{\max}^2(\mathbf{A}_{\mathcal{S}\cup\mathcal{S}^*})\right)^2} \\
&\geq \min_{\mathcal{S}\in\mathcal{B}} \frac{\beta k_d^{\mathcal{S},\mathcal{S}^*}\boldsymbol{\gamma}_{\min}^2}{k_d^{\mathcal{S},\mathcal{S}^*}\left(\sigma^2 + 2\boldsymbol{\gamma}_{\max}\sigma_{\max}^2(\mathbf{A}_{\mathcal{S}\cup\mathcal{S}^*})\right)^2} \\
&\geq \frac{\boldsymbol{\gamma}_{\min}^2}{\left(\sigma^2 + 2\boldsymbol{\gamma}_{\max}\right)^2}\left(\frac{\beta}{\delta_{(K+K_{\text{threshold}})}^2}\right).
\end{aligned} \tag{F.16}$$

For the case where $\mathcal{S} \in \mathcal{B}^c$, i.e., $k_d^{\mathcal{S},\mathcal{S}^*} > K_{\text{threshold}}$, it follows from (F.13) and (F.14) that $\left|\left|\Delta\boldsymbol{\gamma}_+\right|\right|_1 \geq 4\left(\frac{1+\alpha}{1-\alpha}\right)\left|\left|\Delta\boldsymbol{\gamma}_-\right|\right|_1$. Therefore, we can invoke the restricted null space property of $\mathbf{A}\odot\mathbf{A}$ from Theorem 4.3 to bound $\min_{\mathcal{S}\in\mathcal{B}^c} \frac{\mathcal{D}_{\mathcal{S}}^*}{k_d^{\mathcal{S},\mathcal{S}^*}}$ as follows.

$$\min_{\mathcal{S}\in\mathcal{B}^c} \frac{\mathcal{D}_{\mathcal{S}}^*}{k_d^{\mathcal{S},\mathcal{S}^*}} \quad \geq \quad \min_{\mathcal{S}\in\mathcal{B}^c} \frac{\left(\left|\left|\Delta\boldsymbol{\gamma}_+\right|\right|_1^2 + \left|\left|\Delta\boldsymbol{\gamma}_-\right|\right|_1^2\right)}{2mk_d^{\mathcal{S},\mathcal{S}^*}\left(\sigma^2 + 2\boldsymbol{\gamma}_{\max}\sigma_{\max}^2(\mathbf{A}_{\mathcal{S}\cup\mathcal{S}^*})\right)^2}$$

$$
\begin{aligned}
&\geq \min_{\mathcal{S}\in\mathcal{B}^c} \frac{\boldsymbol{\gamma}_{\min}^2 \left(\left|\left|\Delta\boldsymbol{\gamma}_+\right|\right|_0^2 + \left|\left|\Delta\boldsymbol{\gamma}_-\right|\right|_0^2 \right)}{2mk_d^{\mathcal{S},\mathcal{S}^*} \left(\sigma^2 + 2\boldsymbol{\gamma}_{\max}\right)^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{\boldsymbol{\gamma}_{\min}^2 \left(k_d^{\mathcal{S},\mathcal{S}^*} - |\mathcal{S}^*| \right)^2}{2mk_d^{\mathcal{S},\mathcal{S}^*} \left(\sigma^2 + 2\boldsymbol{\gamma}_{\max}\right)^2 \left(\max\left(1, \sigma_{\max}^2(\mathbf{A}_{\mathcal{S}\cup\mathcal{S}^*})\right)\right)^2} \\
&\geq \frac{\boldsymbol{\gamma}_{\min}^2}{2m\left(\sigma^2 + 2\boldsymbol{\gamma}_{\max}\right)^2} \min_{\mathcal{S}\in\mathcal{B}^c} \left(1 - \frac{|\mathcal{S}^*|}{k_d^{\mathcal{S},\mathcal{S}^*}}\right) \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{\boldsymbol{\gamma}_{\min}^2}{2m\left(\sigma^2 + 2\boldsymbol{\gamma}_{\max}\right)^2} \left(1 - \frac{K}{K_{\text{threshold}}}\right) \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{2\boldsymbol{\gamma}_{\min}^2}{5m\left(\sigma^2 + 2\boldsymbol{\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) \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{2\boldsymbol{\gamma}_{\min}^2}{5m\left(\sigma^2 + 2\boldsymbol{\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) \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{\boldsymbol{\gamma}_{\min}^2}{4m\delta_n\left(\sigma^2 + 2\boldsymbol{\gamma}_{\max}\right)^2} \left(\min_{\mathcal{S}\in\mathcal{B}^c} \frac{|\mathcal{S}\cup\mathcal{S}^*|}{\delta_{|\mathcal{S}\cup\mathcal{S}^*|}} \right). \qquad \text{(F.17)}
\end{aligned}
$$

Using (F.16) and (F.17) in (F.15), we obtain the following lower bound for η.

$$
\eta \geq \frac{\boldsymbol{\gamma}_{\min}^2}{\left(\sigma^2 + 2\boldsymbol{\gamma}_{\max}\right)^2} \min\left(\frac{\beta}{\delta_{(K+K_{\text{threshold}})}^2}, \frac{1}{4m\delta_n} \min_{\mathcal{S}\in\mathcal{B}^c} \frac{|\mathcal{S}\cup\mathcal{S}^*|}{\delta_{|\mathcal{S}\cup\mathcal{S}^*|}} \right). \qquad \text{(F.18)}
$$

This concludes our proof. □

F.6 Proof of Proposition 4.4

Proof. By setting $\epsilon = L\mathcal{D}_{\mathcal{S}}^*/2$ in Proposition 4.1, and noting that $\mathcal{D}_{\mathcal{S}}^* \geq \eta k_d^{\mathcal{S},\mathcal{S}^*}$, we have

$$
|\Theta^{\epsilon}(\mathcal{S})| \leq \max\left\{1, \left(\frac{6\sqrt{|\mathcal{S}|}(\boldsymbol{\gamma}_{\max} - \boldsymbol{\gamma}_{\min})C_{\mathcal{L},\mathcal{S}}}{L\eta k_d^{\mathcal{S},\mathcal{S}*}} \right)^{|\mathcal{S}|} \right\}. \qquad \text{(F.19)}
$$

where $C_{\mathcal{L},\mathcal{S}}$ denotes the Lipschitz constant of $\mathcal{L}(\mathbf{Y};\boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$ over the bounded domain $\Theta(\mathcal{S})$. Proposition F.1 characterizes the Lipschitz property of $\mathcal{L}(\mathbf{Y},\boldsymbol{\gamma})$.

Proposition F.1. *For $\mathcal{S} \in \mathcal{S}_n$, the log-likelihood $\mathcal{L}(\mathbf{Y};\boldsymbol{\gamma}) : \Theta(\mathcal{S}) \rightarrow \mathbb{R}$ is Lipschitz continuous in $\boldsymbol{\gamma}$ as shown below.*

$$|\mathcal{L}(\mathbf{Y},\boldsymbol{\gamma}_2) - \mathcal{L}(\mathbf{Y},\boldsymbol{\gamma}_1)| \;\leq\; \frac{mL}{\boldsymbol{\gamma}_{min}}\left(1 + \frac{|||\mathbf{R}_{\mathbf{yy}}|||_2}{\sigma^2}\right)||\boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_1||_2\,,$$

for any $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2 \in \Theta(\mathcal{S})$. Here, $\mathbf{R}_{\mathbf{yy}} \triangleq \frac{1}{L}\mathbf{Y}\mathbf{Y}^T$.

Proof. See Appendix F.7. □

For $\mathbf{Y} \in \mathcal{G}$ (as defined in (4.6)), we have $|||\mathbf{R}_{\mathbf{yy}}|||_2 \leq 2\,(\sigma^2 + \boldsymbol{\gamma}_{\max}\delta_K)$. Using Proposition F.1 in (F.19), combined with the lower bound for η in (4.20) yields the following bound for κ_{cov} when $\boldsymbol{\gamma}_{\min} \neq \boldsymbol{\gamma}_{\max}$.

$$\begin{aligned}
\kappa_{\text{cov}} &= \max_{\mathcal{S}\subseteq[n]} \frac{\log\left|\Theta^{LD^*_{\mathcal{S}}}(\mathcal{S})|_{\mathcal{G}}\right|}{k_d^{\mathcal{S},\mathcal{S}^*}} \\
&\leq \max_{\mathcal{S}\subseteq[n]} \frac{|\mathcal{S}|}{k_d^{\mathcal{S},\mathcal{S}^*}} \log\left(\frac{6m|\mathcal{S}|^{1/2}(\boldsymbol{\gamma}_{\max} - \boldsymbol{\gamma}_{\min})\left(3 + 2\delta_K\frac{\boldsymbol{\gamma}_{\max}}{\sigma^2}\right)}{\boldsymbol{\gamma}_{\min}\eta k_d^{\mathcal{S},\mathcal{S}^*}}\right) \\
&\leq \max_{\mathcal{S}\subseteq[n]} \frac{k_d^{\mathcal{S},\mathcal{S}^*} + |\mathcal{S}^*|}{k_d^{\mathcal{S},\mathcal{S}^*}} \log\left(\frac{6m\sqrt{\frac{k_d^{\mathcal{S},\mathcal{S}^*}+|\mathcal{S}^*|}{k_d^{\mathcal{S},\mathcal{S}^*}}}(\boldsymbol{\gamma}_{\max} - \boldsymbol{\gamma}_{\min})\left(3 + 2\delta_K\frac{\boldsymbol{\gamma}_{\max}}{\sigma^2}\right)}{\boldsymbol{\gamma}_{\min}\eta\sqrt{k_d^{\mathcal{S},\mathcal{S}^*}}}\right) \\
&\leq (1 + |\mathcal{S}^*|)\log\left(\frac{6m\sqrt{1+|\mathcal{S}^*|}(\boldsymbol{\gamma}_{\max} - \boldsymbol{\gamma}_{\min})\left(3 + 2\delta_K\frac{\boldsymbol{\gamma}_{\max}}{\sigma^2}\right)}{\boldsymbol{\gamma}_{\min}\eta}\right) \\
&\leq (1 + K)\log\left(\frac{6\sqrt{2}m\sqrt{K}(\boldsymbol{\gamma}_{\max} - \boldsymbol{\gamma}_{\min})\,(3\sigma^2 + 2\boldsymbol{\gamma}_{\max})\,\delta_K}{\boldsymbol{\gamma}_{\min}\sigma^2\eta}\right) \qquad \text{(F.20)}
\end{aligned}$$

where η and δ_K are bounded as per Proposition 4.3. □

F.7 Proof of Proposition F.1

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

First, we derive an upper bound for the Lipschitz constant of $f(\boldsymbol{\gamma}) = -L\log|\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}|$ for $\boldsymbol{\gamma} \in \Theta(\mathcal{S})$. By the mean value theorem, any upper bound for $\|\nabla_{\boldsymbol{\gamma}} f(\boldsymbol{\gamma})\|_2$ also serves as an upper bound for the Lipschitz constant of f. So, we derive an upper bound for $\|\nabla_{\boldsymbol{\gamma}} f(\boldsymbol{\gamma})\|_2$. Note that $\left|\frac{\partial f(\boldsymbol{\gamma})}{\boldsymbol{\gamma}(i)}\right| = L(\mathbf{a}_i \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \mathbf{a}_i)$ for $i \in \mathcal{S}$, and 0 otherwise. Here, $\mathbf{a}_i$ denotes the i^{th} column of $\mathbf{A}$. Then, $\|\nabla_{\boldsymbol{\gamma}} f(\boldsymbol{\gamma})\|_2$ can be upper bounded as shown below.

$$\begin{aligned}
\|\nabla_{\boldsymbol{\gamma}} f(\boldsymbol{\gamma})\|_2 &\leq \|\nabla_{\boldsymbol{\gamma}} f(\boldsymbol{\gamma})\|_1 = L\sum_{u\in\mathcal{S}} \mathbf{a}_i^T \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \mathbf{a}_i \\
&= L\left(\text{tr}\left(\mathbf{A}_{\mathcal{S}}^T(\sigma^2\mathbf{I}_m + \mathbf{A}_{\mathcal{S}}\boldsymbol{\Gamma}_{\mathcal{S}}\mathbf{A}_{\mathcal{S}}^T)^{-1}\mathbf{A}_{\mathcal{S}}\right)\right) \\
&= L\left(\text{tr}\left(\boldsymbol{\Gamma}_{\mathcal{S}}^{-1/2}\tilde{\mathbf{A}}_{\mathcal{S}}^T(\sigma^2\mathbf{I}_m + \tilde{\mathbf{A}}_{\mathcal{S}}\tilde{\mathbf{A}}_{\mathcal{S}}^T)^{-1}\tilde{\mathbf{A}}_{\mathcal{S}}\boldsymbol{\Gamma}_{\mathcal{S}}^{-1/2}\right)\right) \\
&\overset{(a)}{=} L\left|\left|\left|\boldsymbol{\Gamma}_{\mathcal{S}}^{-1}\right|\right|\right|_2 \text{tr}\left(\tilde{\mathbf{A}}_{\mathcal{S}}^T(\sigma^2\mathbf{I}_m + \tilde{\mathbf{A}}_{\mathcal{S}}\tilde{\mathbf{A}}_{\mathcal{S}}^T)^{-1}\tilde{\mathbf{A}}_{\mathcal{S}}\right) \\
&\overset{(b)}{\leq} \frac{L\min(m,|\mathcal{S}|)}{\boldsymbol{\gamma}_{\min}}.
\end{aligned} \tag{F.21}$$

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 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 $\left|\frac{\partial g(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}(i)}\right| = L(\mathbf{a}_i^T \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \mathbf{R}_Y \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \mathbf{a}_i)$ for $i \in \mathcal{S}$, and 0 otherwise. Then, $\|\nabla_{\boldsymbol{\gamma}} g(\boldsymbol{\gamma})\|_2$ can be bounded as

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

$$
\begin{aligned}
&\leq \left(L \left|\left|\left|\mathbf{R}_{\mathbf{Y}}\right|\right|\right|_2 / \boldsymbol{\gamma}_{\min}\right) \operatorname{tr}\left(\tilde{\mathbf{A}}_{\mathcal{S}}^T \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \tilde{\mathbf{A}}_{\mathcal{S}}\right) \\
&\leq \left(L \left|\left|\left|\mathbf{R}_{\mathbf{Y}}\right|\right|\right|_2 \left|\left|\left|\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1}\right|\right|\right|_2 / \boldsymbol{\gamma}_{\min}\right) \operatorname{tr}\left(\tilde{\mathbf{A}}_{\mathcal{S}}^T \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \tilde{\mathbf{A}}_{\mathcal{S}}\right) \\
&\overset{(a)}{\leq} \left(L \left|\left|\left|\mathbf{R}_{\mathbf{Y}}\right|\right|\right|_2 \left|\left|\left|\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1}\right|\right|\right|_2 \min\left(m, |\mathcal{S}|\right)\right) / \boldsymbol{\gamma}_{\min} \\
&\overset{(b)}{\leq} \frac{L \left|\left|\left|\mathbf{R}_{\mathbf{Y}}\right|\right|\right|_2 \min\left(m, |\mathcal{S}|\right)}{\boldsymbol{\gamma}_{\min} \sigma^2}.
\end{aligned} \tag{F.22}
$$

where $\tilde{\mathbf{A}}_{\mathcal{S}} \triangleq \mathbf{A}_{\mathcal{S}} \boldsymbol{\Gamma}_{\mathcal{S}}^{1/2}$. The inequality in (F.22-a) follows from $\left(\tilde{\mathbf{A}}_{\mathcal{S}}^T \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \tilde{\mathbf{A}}_{\mathcal{S}}\right)$ having $\min\left(m, |\mathcal{S}|\right)$ nonzero eigenvalues, all of them less than unity. The last inequality in (F.22-b) is due to $\left|\left|\left|\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1}\right|\right|\right|_2 \leq 1/\sigma^2$. 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 (F.21) and (F.22), and noting that $\min(m, |\mathcal{S}|) \leq K$, we obtain the desired result. □

F.8 Proof of Corollary 4.2

Proof. Using (4.20) as a starting point, we first derive probabilistic bounds for the parameters α, β and $\delta_k, k \in [n]$ when $m \geq \mathcal{O}(K \log n)$.

Lower bound for α:

By the Hanson-Wright concentration (Theorem B.1), and taking the union bound over all columns of $\mathbf{A}$,

$$
\mathbb{P}\left(\bigcup_{i \in [n]} \left\{\left|\|\mathbf{a}_i\|_2^2 - 1\right| \geq \alpha\right\}\right) \leq n \mathbb{P}\left(\left|\mathbf{a}_1^T \mathbf{a}_1 - 1\right| \geq \alpha\right) \leq 2ne^{-c_2 m \alpha^2}. \tag{F.23}
$$

From (F.23), the ℓ_2 norm of columns of $\mathbf{A}$ lie inside the interval $\left[1 - \frac{1}{\sqrt{K}}, 1 + \frac{1}{\sqrt{K}}\right]$ with probability exceeding $1 - c_2 n^{-2}$, where c_2 is a universal constant.

Lower bound for β:

From Proposition 4.3, by definition, β refers to the minimum singular value among all submatrices obtained by sampling k or fewer columns of $\mathbf{A} \odot \mathbf{A}$. By Theorem (3.3), $\mathbf{A} \odot \mathbf{A}$ satisfies the restricted isometry property of order $K + K_{\text{threshold}}$ for $m \geq \mathcal{O}(K \log n)$, it follows that

$$\mathbb{P}\left(\beta < 1 - \frac{8c_1(K + K_{\text{threshold}}) \log n}{m}\right) \leq \frac{5}{n^2}, \tag{F.24}$$

where c_1 is a universal absolute constant. Thus, for $m \geq \mathcal{O}(K \log n)$,

$$\beta \geq \frac{1}{2} \tag{F.25}$$

with probability exceeding $1 - 5n^{-2}$.

Upper bound for δ_k for $k \in [n]$:

From Proposition 4.3, δ_k is defined as the maximum singular value among any k or fewer column submatrix of $\mathbf{A}$. By direct application of Corollary C.1, for any $k \leq n$,

$$\delta_k \leq \frac{\left(\sqrt{m} + \sqrt{k} + \sqrt{6k \log n}\right)^2}{m} \tag{F.26}$$

with probability exceeding $1 - 2n^{-k}$. From (F.26), for $m \geq \mathcal{O}(K \log n)$,

$$\begin{aligned} \delta_{K+K_{\text{threshold}}} &\leq \mathcal{O}(1), && \text{(F.27)} \\ \delta_n &\leq \mathcal{O}\left(\frac{n}{K \log n}\right), && \text{(F.28)} \\ \text{and } \delta_{|\mathcal{S} \cup \mathcal{S}^*|} &\leq \mathcal{O}\left(\frac{|\mathcal{S} \cup \mathcal{S}^*|}{K}\right), && \text{(F.29)} \end{aligned}$$

with overwhelming probability.

Finally, by using the above probabilistic bounds for α, β, δ_n, $\delta_{|\mathcal{S}\cup\mathcal{S}^*|}$ and $\delta_{K+K_{\text{threshold}}}$ together in Proposition 4.3, η can be bounded as

$$\begin{aligned}
\eta &\geq \frac{\boldsymbol{\gamma}_{\min}^2}{\left(\sigma^2+2\boldsymbol{\gamma}_{\max}\right)^2}\min\left(\frac{\beta}{\delta^2_{(K+K_{\text{threshold}})}}, \frac{1}{4m\delta_n}\min_{\substack{\mathcal{S}\subseteq[n],\\ |\mathcal{S}\setminus\mathcal{S}^*|+|\mathcal{S}^*\setminus\mathcal{S}|\leq K_{\text{threshold}}}}\frac{|\mathcal{S}\cup\mathcal{S}^*|}{\delta_{|\mathcal{S}\cup\mathcal{S}^*|}}\right)\\
&\geq \frac{c_3\boldsymbol{\gamma}_{\min}^2}{\left(\sigma^2+2\boldsymbol{\gamma}_{\max}\right)^2}\min\left(\frac{1}{2},\frac{K}{4n}\right)\\
&\geq \frac{c_4K\boldsymbol{\gamma}_{\min}^2}{n\left(\sigma^2+2\boldsymbol{\gamma}_{\max}\right)^2},
\end{aligned} \tag{F.30}$$

with probability exceeding $1-c_5n^{-2}$ for sufficiently large n. Here, the constants c_4 and c_5 do not depend on the problem dimensions.

By using the above probabilistic lower bound for η from (F.30) in Proposition 4.4, we obtain the following simplified upper bound for κ_{cov}.

$$\begin{aligned}
\kappa_{\text{cov}} &\leq (K+1)\log\left(\frac{6\sqrt{2}m\sqrt{K}(\boldsymbol{\gamma}_{\max}-\boldsymbol{\gamma}_{\min})\left(3\sigma^2+2\boldsymbol{\gamma}_{\max}\right)\delta_K}{\boldsymbol{\gamma}_{\min}\sigma^2\eta}\right)\\
&\leq (K+1)\log\left(\frac{c_5n\sqrt{K}\log n(\boldsymbol{\gamma}_{\max}-\boldsymbol{\gamma}_{\min})\left(3\sigma^2+2\boldsymbol{\gamma}_{\max}\right)^3}{\boldsymbol{\gamma}_{\min}^3\sigma^2}\right).
\end{aligned}$$

□

F.9 Proof of Proposition 4.5

Proof. 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,

$$\begin{aligned}
\mu^* \geq \frac{\text{tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_2}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_1}\right)}{m} &= \frac{1}{m}\left[\sigma^2\text{Tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_2}^{-1}\right)+\text{Tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_2}^{-1}\mathbf{A}\boldsymbol{\Gamma}_1\mathbf{A}^T\right)\right]\\
&\geq \frac{1}{m}\text{tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_2}^{-1}\mathbf{A}\boldsymbol{\Gamma}_1\mathbf{A}^T\right).
\end{aligned} \tag{F.31}$$

Here, the second step is because $\mathbf{\Sigma}_{\boldsymbol{\gamma}_1} = \sigma^2\mathbf{I}_m + \mathbf{A}\mathbf{\Gamma}_1\mathbf{A}^T$. The last inequality is obtained by dropping the strictly positive $\frac{\sigma^2}{m}\text{tr}\left(\mathbf{\Sigma}_{\boldsymbol{\gamma}_2}^{-1}\right)$ 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 $\mathbf{\Sigma}_{\boldsymbol{\gamma}_2}$ be $\mathbf{U}\Lambda\mathbf{U}^T$, where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_m)$, λ_i's are the eigenvalues of $\mathbf{\Sigma}_{\boldsymbol{\gamma}_2}$ and $\mathbf{U}$ is a unitary matrix with columns as the eigenvectors of $\mathbf{\Sigma}_{\boldsymbol{\gamma}_2}$. 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 Λ_2 and $\Lambda_{2^\perp}$ 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 Λ corresponding to the eigenvectors in $\mathbf{U}_2$ and $\mathbf{U}_{2^\perp}$, respectively. We observe that $\Lambda_{2^\perp} = \sigma^2\mathbf{I}_{m-|\mathcal{S}_2|}$.

By setting $\mathbf{\Sigma}_{\boldsymbol{\gamma}_2}^{-1} = \mathbf{U}_2\Lambda_2^{-1}\mathbf{U}_2^T + \mathbf{U}_{2^\perp}\Lambda_{2^\perp}^{-1}\mathbf{U}_{2^\perp}^T$ in (F.31), we get

$$\begin{aligned}
\mu^* &\geq \frac{1}{m}\left(\text{tr}\left(\mathbf{U}_2\Lambda_2^{-1}\mathbf{U}_2^T\mathbf{A}\mathbf{\Gamma}_1\mathbf{A}^T\right) + \text{tr}\left(\mathbf{U}_{2^\perp}\Lambda_{2^\perp}^{-1}\mathbf{U}_{2^\perp}^T\mathbf{A}\mathbf{\Gamma}_1\mathbf{A}^T\right)\right) \\
&\geq \frac{1}{m}\text{tr}\left(\Lambda_{2^\perp}^{-1}\mathbf{U}_{2^\perp}^T\mathbf{A}\mathbf{\Gamma}_1\mathbf{A}^T\mathbf{U}_{2^\perp}\right),
\end{aligned}$$

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

$$\begin{aligned}
\mu^* &\geq \frac{1}{m}\text{tr}\left(\Lambda_{2^\perp}^{-1}\mathbf{U}_{2^\perp}^T\mathbf{A}_{\mathcal{S}_2^c}\mathbf{\Gamma}_{1,\mathcal{S}_2^c}\mathbf{A}_{\mathcal{S}_2^c}^T\mathbf{U}_{2^\perp}\right) \\
&= \frac{1}{m\sigma^2}\sum_{i=1}^{m-|\mathcal{S}_2|}(\mathbf{u}_{2^\perp,i})^T\mathbf{A}_{\mathcal{S}_2^c}\mathbf{\Gamma}_{1,\mathcal{S}_2^c}\mathbf{A}_{\mathcal{S}_2^c}^T\mathbf{u}_{2^\perp,i} \\
&= \frac{1}{m\sigma^2}\sum_{i=1}^{m-|\mathcal{S}_2|}(\mathbf{u}_{2^\perp,i})^T\mathbf{A}_{\mathcal{S}_1\backslash\mathcal{S}_2}\mathbf{\Gamma}_{1,\mathcal{S}_1\backslash\mathcal{S}_2}\mathbf{A}_{\mathcal{S}_1\backslash\mathcal{S}_2}^T\mathbf{u}_{2^\perp,i}. \qquad \text{(F.32)}
\end{aligned}$$

In the above, $\mathbf{u}_{2^\perp,i}$ denotes the i^{th} 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 (F.32) be equal to zero. This implies that the columns of $\mathbf{U}_{2^\perp}$ belong to $\text{Null}(\mathbf{\Gamma}^{1/2}_{1,\mathcal{S}_1\setminus\mathcal{S}_2}\mathbf{A}^T_{\mathcal{S}_1\setminus\mathcal{S}_2})$, which means that they also belong to $\text{Null}(\mathbf{A}_{\mathcal{S}_1\setminus\mathcal{S}_2}\mathbf{\Gamma}_{1,\mathcal{S}_1\setminus\mathcal{S}_2}\mathbf{A}^T_{\mathcal{S}_1\setminus\mathcal{S}_2})$. Since, for a symmetric matrix, the row and column spaces are equal and orthogonal to the null space of the matrix, it follows that $\text{Col}(\mathbf{A}_{\mathcal{S}_1\setminus\mathcal{S}_2}\mathbf{\Gamma}_{1,\mathcal{S}_1\setminus\mathcal{S}_2}\mathbf{A}^T_{\mathcal{S}_1\setminus\mathcal{S}_2})$ (same as $\text{Col}(\mathbf{A}_{\mathcal{S}_1\setminus\mathcal{S}_2}\mathbf{\Gamma}^{1/2}_{1,\mathcal{S}_1\setminus\mathcal{S}_2})$) is spanned by the columns of $\mathbf{U}_2$, or equivalently by columns of $\mathbf{A}_{\mathcal{S}_2}$. Thus, every column in $\mathbf{A}_{\mathcal{S}_1\setminus\mathcal{S}_2}$ can be expressed as a linear combination of columns in $\mathbf{A}_{\mathcal{S}_2}$. This contradicts our initial assumption that $K + 1 < \text{spark}(\mathbf{A})$, implying that any $K + 1$ or fewer columns of $\mathbf{A}$ are linearly independent. Therefore, we conclude that there is at least one strictly positive term in the summation in (F.32), and consequently there exists a constant $c_1 > 0$ such that $\mu^* \geq c_1/\sigma^2$. □

Appendix G

Appendix for Chapter 5

G.1 Proof of Proposition 5.1

Proof. Let $\mathcal{S}$ and $\mathcal{T}$ be arbitrary subsets of $[n]$ such that $\mathcal{S} \subseteq \mathcal{T}$. Then, we have

$$\begin{aligned} f(\mathcal{S}) &= \log|\mathbf{A} + \beta \mathbf{B}_{\mathcal{S}} \mathbf{B}_{\mathcal{S}}^T| \\ &= \log|\mathbf{A}| + \log|\mathbf{I} + \beta \mathbf{A}^{-1} \mathbf{B}_{\mathcal{S}} \mathbf{B}_{\mathcal{S}}^T| \\ &= \log|\mathbf{A}| + \log|\mathbf{I} + \beta \mathbf{B}_{\mathcal{S}}^T \mathbf{A}^{-1} \mathbf{B}_{\mathcal{S}}|, \end{aligned}$$

where the first equality follows from $\mathbf{A}$ being non-singular and the determinant property: $\det(\mathbf{XY}) = \det(\mathbf{X})\det(\mathbf{Y})$ for any two square matrices $\mathbf{X}, \mathbf{Y}$ of the same size. The second equality follows from the determinant property that $\det(\mathbf{I}_n + \mathbf{XY}^T) = \det(\mathbf{I}_p + \mathbf{Y}^T\mathbf{X})$ for any two $n \times p$ sized matrices $\mathbf{X}, \mathbf{Y}$. Using the above result, we can establish the monotonicity of $f(.)$ as shown below.

$$\begin{aligned} f(\mathcal{T}) - f(\mathcal{S}) &= \log|\mathbf{A} + \beta \mathbf{B}_{\mathcal{T}} \mathbf{B}_{\mathcal{T}}^T| - \log|\mathbf{A} + \beta \mathbf{B}_{\mathcal{S}} \mathbf{B}_{\mathcal{S}}^T| \\ &= \log|\mathbf{A} + \beta \mathbf{B}_{\mathcal{S}} \mathbf{B}_{\mathcal{S}}^T + \beta \mathbf{B}_{\mathcal{T}\setminus\mathcal{S}} \mathbf{B}_{\mathcal{T}\setminus\mathcal{S}}^T| - \log|\mathbf{A} + \beta \mathbf{B}_{\mathcal{S}} \mathbf{B}_{\mathcal{S}}^T| \end{aligned}$$

$$
\begin{aligned}
&= \log\left|\mathbf{I}_n + \beta\left(\mathbf{A}+\beta\mathbf{B}_{\mathcal{S}}\mathbf{B}_{\mathcal{S}}^T\right)^{-1}\mathbf{B}_{\mathcal{T}\backslash\mathcal{S}}\mathbf{B}_{\mathcal{T}\backslash\mathcal{S}}^T\right| \\
&= \log\left|\mathbf{I} + \beta\underbrace{\mathbf{B}_{\mathcal{T}\backslash\mathcal{S}}^T\left(\mathbf{A}+\beta\mathbf{B}_{\mathcal{S}}\mathbf{B}_{\mathcal{S}}^T\right)^{-1}\mathbf{B}_{\mathcal{T}\backslash\mathcal{S}}}_{\text{positive definite}}\right| \geq 0.
\end{aligned}
$$

To show submodularity of f, let us consider $i \in [n]\backslash\mathcal{T}$. The marginal increment $f(\mathcal{S}\cup\{x\}) - f(\mathcal{S})$ is evaluated as

$$
\begin{aligned}
f(\mathcal{S}\cup\{i\}) - f(\mathcal{S}) &= \log|\mathbf{A}+\beta\mathbf{B}_{\mathcal{S}}\mathbf{B}_{\mathcal{S}}^T + \beta\mathbf{b}_i\mathbf{b}_i^T| - \log|\mathbf{A}+\beta\mathbf{B}_{\mathcal{S}}\mathbf{B}_{\mathcal{S}}^T| \\
&= \log|\mathbf{I}_n + \beta\left(\mathbf{A}+\beta\mathbf{B}_{\mathcal{S}}\mathbf{B}_{\mathcal{S}}^T\right)^{-1}\mathbf{b}_i\mathbf{b}_i^T| \\
&= \log|1 + \beta\mathbf{b}_i^T\left(\mathbf{A}+\beta\mathbf{B}_{\mathcal{S}}\mathbf{B}_{\mathcal{S}}^T\right)^{-1}\mathbf{b}_i|. \qquad \text{(G.1)}
\end{aligned}
$$

Likewise, it can be shown that

$$
f(\mathcal{T}\cup\{i\}) - f(\mathcal{T}) = \log|1 + \beta\mathbf{b}_i^T\left(\mathbf{A}+\beta\mathbf{B}_{\mathcal{T}}\mathbf{B}_{\mathcal{T}}^T\right)^{-1}\mathbf{b}_i|. \qquad \text{(G.2)}
$$

Using Woodbury matrix identity, we observe that

$$
\begin{aligned}
\mathbf{b}_i^T\left(\mathbf{A}+\beta\mathbf{B}_T\mathbf{B}_T^T\right)^{-1}\mathbf{b}_i &= \mathbf{b}_i^T\left(\mathbf{A}+\beta\mathbf{B}_S\mathbf{B}_S^T\right)^{-1}\mathbf{b}_i \\
&\quad -\mathbf{b}_i^T\mathbf{P}_{\mathcal{S}}^{-1}\mathbf{B}_{T\backslash S}\left(\frac{1}{\beta}\mathbf{I} + \mathbf{B}_{T\backslash S}^T\mathbf{P}_{\mathcal{S}}^{-1}\mathbf{B}_{T\backslash S}\right)^{-1}\mathbf{B}_{T\backslash S}^T\mathbf{P}_{\mathcal{S}}^{-1}\mathbf{b}_i, \qquad \text{(G.3)}
\end{aligned}
$$

where $\mathbf{P}_{\mathcal{S}} = \mathbf{A} + \beta\mathbf{B}_T\mathbf{B}_T^T$ is a positive definite matrix. By dropping the positive second term in the RHS, we obtain the following inequality

$$
\mathbf{b}_i^T\left(\mathbf{A}+\beta\mathbf{B}_T\mathbf{B}_T^T\right)^{-1}\mathbf{b}_i \leq \mathbf{b}_i^T\left(\mathbf{A}+\beta\mathbf{B}_S\mathbf{B}_S^T\right)^{-1}\mathbf{b}_i \qquad \text{(G.4)}
$$

The inequality in (G.4) and the monotonic increase of $\log(1+\beta x)$ with $x(>0)$ together

imply that f satisfies the diminishing returns property that $f(\mathcal{S} \cup \{i\}) - f(\mathcal{S}) \geq f(\mathcal{T} \cup \{i\}) - f(\mathcal{T})$, and hence f is a submodular function. □

G.2 Derivation of modular upper bound $h^f_{\mathcal{S}_{k-1}}$

As per (5.15), the modular upper bound of submodular $f(\mathcal{S})$ at point $\mathcal{S}_{k-1}$ is evaluated as

$$h^f_{\mathcal{S}_{k-1}}(\mathcal{S}) \triangleq f(\mathcal{S}_{k-1}) - \sum_{j \in \mathcal{S}_{k-1} \setminus \mathcal{S}} [f(\mathcal{S}_{k-1}) - f(\mathcal{S}_{k-1} \setminus \{j\})] + \sum_{j \in \mathcal{S} \setminus \mathcal{S}_{k-1}} [f(j) - f(\phi)]\,. \quad \text{(G.5)}$$

We note that for $\mathcal{S} \subset [n]$, $f(\mathcal{S})$ can be decomposed as

$$\begin{aligned} f(\mathcal{S}) &= \log \left| (1-\alpha)\hat{\mathbf{R}}_{\mathbf{yy}} + \alpha \left(\sigma^2 \mathbf{I}_m + \boldsymbol{\gamma} \mathbf{A}_{\mathcal{S}} \mathbf{A}_{\mathcal{S}}^T \right) \right| \\ &= \log \left| \mathbf{H} + \alpha\rho \mathbf{A}_{\mathcal{S}} \mathbf{A}_{\mathcal{S}}^T \right| + m \log \sigma^2 \\ &= \log \left| \mathbf{I}_{|\mathcal{S}|} + \alpha\rho \mathbf{A}_{\mathcal{S}}^T \mathbf{H}^{-1} \mathbf{A}_{\mathcal{S}} \right| + \underbrace{m \log \sigma^2 + \log |\mathbf{H}|}_{\text{terms independent of } \mathcal{S}}, \end{aligned} \quad \text{(G.6)}$$

where $\mathbf{H} = \left(\alpha \mathbf{I}_m + (1-\alpha)\sigma^{-2} \hat{\mathbf{R}}_{\mathbf{yy}} \right)$. The last step is obtained by using the matrix identity $\log |\mathbf{I} + \mathbf{AB}| = \log |\mathbf{I} + \mathbf{BA}|$ for any two matrices $\mathbf{A}$ and $\mathbf{B}$ of matching dimensions.

Using (G.6), the difference term $f(\mathcal{S}_{k-1}) - f(\mathcal{S}_{k-1} \setminus \{j\})$ is evaluated as

$$\begin{aligned} f(\mathcal{S}_{k-1}) - f(\mathcal{S}_{k-1} \setminus \{j\}) &= \log \left| \mathbf{H} + \alpha\rho \mathbf{A}_{\mathcal{S}_{k-1}} \mathbf{A}_{\mathcal{S}_{k-1}}^T \right| - \log \left| \mathbf{H} + \alpha\rho \mathbf{A}_{\mathcal{S}_{k-1} \setminus \{j\}} \mathbf{A}_{\mathcal{S}_{k-1} \setminus \{j\}}^T \right| \\ &= \log |\mathbf{T}_{k-1}| - \log \left| \mathbf{T}_{k-1} - \alpha\rho \mathbf{a}_j \mathbf{a}_j^T \right| \\ &= -\log \left| 1 - \alpha\rho \mathbf{a}_j^T \mathbf{T}_{k-1}^{-1} \mathbf{a}_j \right|, \end{aligned} \quad \text{(G.7)}$$

where $T_{k-1} = \mathbf{H} + \alpha\rho \mathbf{A}_{\mathcal{S}_{k-1}} \mathbf{A}_{\mathcal{S}_{k-1}}^T$. Also, for $j \in [n]$, $f(j)$ is evaluated as

$$f(j) = \log \left| 1 + \alpha\rho \mathbf{a}_j^T \mathbf{H}^{-1} \mathbf{a}_j \right|. \quad \text{(G.8)}$$

Using the fact that $f(\phi) = 0$, and by using (G.7) and (G.8) in (G.5), we obtain the following modular upper bound of f which is tight at $\mathcal{S}_{k-1}$. For any $\mathcal{S} \subseteq [n]$, the upper bound $h^f_{\mathcal{S}_{k-1}}(\mathcal{S})$ is evaluated as

$$h^f_{\mathcal{S}_{k-1}}(\mathcal{S}) = f(\mathcal{S}_{k-1}) + \sum_{j \in \mathcal{S}_{k-1} \setminus \mathcal{S}} \log \left|1 - \alpha\rho \mathbf{a}_j^T \mathbf{T}_{k-1}^{-1} \mathbf{a}_j\right| + \sum_{j \in \mathcal{S} \setminus \mathcal{S}_{k-1}} \log \left|1 + \alpha\rho \mathbf{a}_j^T \mathbf{H}^{-1} \mathbf{a}_j\right|, \quad \text{(G.9)}$$

In above, by dropping the first term which is independent of $\mathcal{S}$, we obtain the final form of the modular upper bound $h^f_{\mathcal{S}_{k-1}}$ claimed in (5.16).

Appendix H

Appendix for Chapter 6

H.1 Derivation of the M-step Cost Function

The conditional expectation in (6.5) can be simplified as:

$$\begin{aligned}\mathbb{E}_{\mathbf{X}}\left[\log p(\mathbf{Y},\mathbf{X};\boldsymbol{\gamma})|\mathbf{Y};\boldsymbol{\gamma}^k\right] &= \mathbb{E}_{[\mathbf{X}|\mathbf{Y};\boldsymbol{\gamma}^k]}[\log p(\mathbf{Y}|\mathbf{X}) + \log p(\mathbf{X};\boldsymbol{\gamma})] \\ &= \mathbb{E}_{[\mathbf{X}|\mathbf{Y};\boldsymbol{\gamma}^k]} \log p(\mathbf{Y}|\mathbf{X}) + \sum_{j\in\mathcal{J}} \mathbb{E}_{[\mathbf{x}_j|\mathbf{y}_j;\boldsymbol{\gamma}^k]} \log p(\mathbf{x}_j;\boldsymbol{\gamma}).\end{aligned} \tag{H.1}$$

Using (6.2), and discarding the terms independent of $\boldsymbol{\gamma}$ in (H.1), the M-step objective function $Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^k)$ is given by

$$\begin{aligned}Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^k) &= \sum_{j\in\mathcal{J}} \mathbb{E}_{[\mathbf{x}_j|\mathbf{y}_j,\boldsymbol{\gamma}^k]} \left(-\frac{1}{2}\log|\boldsymbol{\Gamma}| - \frac{1}{2}\mathbf{x}_j^T\boldsymbol{\Gamma}^{-1}\mathbf{x}_j\right) \\ &= -\frac{1}{2}\sum_{j\in\mathcal{J}}\left(\log|\boldsymbol{\Gamma}| + \sum_{i=1}^{n}\frac{\mathbb{E}_{[\mathbf{x}_j\sim\mathcal{N}(\boldsymbol{\mu}_j^{k+1},\boldsymbol{\Sigma}_j^{k+1})]}\mathbf{x}_j(i)^2}{\boldsymbol{\gamma}(i)}\right) \\ &= -\frac{1}{2}\sum_{j\in\mathcal{J}}\sum_{i=1}^{n}\left(\log\boldsymbol{\gamma}(i) + \frac{\boldsymbol{\Sigma}_j^{k+1}(i,i) + \boldsymbol{\mu}_j^{k+1}(i)^2}{\boldsymbol{\gamma}(i)}\right).\end{aligned} \tag{H.2}$$

H.2 Derivation of the Simplified Update for $\boldsymbol{\gamma}_b$

By summing the dual variable update rule (6.15) across all nodes, the following holds for all $b \in \mathcal{B}$

$$\sum_{j \in \mathcal{N}_b} (\boldsymbol{\lambda}_j^b)^{r+1} = \sum_{j \in \mathcal{N}_b} (\boldsymbol{\lambda}_j^b)^r + \rho \sum_{j \in \mathcal{N}_b} \boldsymbol{\gamma}_j^{r+1} - \rho |\mathcal{N}_b| \boldsymbol{\gamma}_b^{r+1}. \tag{H.3}$$

Plugging (6.16) in (H.3), we obtain

$$\sum_{j \in \mathcal{N}_b} (\boldsymbol{\lambda}_j^b)^{r+1} = 0 \qquad \forall\, b \in \mathcal{B}. \tag{H.4}$$

Using (H.4) in (6.16), we obtain the simplified update for $\boldsymbol{\gamma}_b$.

H.3 Proof of Theorem 6.1

Proof. The proof of the convergence of ADMM discussed in the sequel is a based on the proof given in [114]. However, our proof differs from the one in [114] due to the different internode communication scheme adopted here, which uses the auxiliary/bridge nodes to enforce consensus between the nodes. We make the following assumptions about the objective function f in (6.19).

1. f is twice differentiable and strongly convex in $\boldsymbol{\gamma}_{\mathcal{J}}$. This implies that there exists $m_f \in \mathbb{R}_+ \backslash \{0\}$ such that, for all $\boldsymbol{\gamma}_{\mathcal{J}}, \boldsymbol{\gamma}_{\mathcal{J}}^{'}$, the following holds

$$\langle \nabla f(\boldsymbol{\gamma}_{\mathcal{J}})^T - \nabla f(\boldsymbol{\gamma}_{\mathcal{J}}^{'})^T, \boldsymbol{\gamma}_{\mathcal{J}} - \boldsymbol{\gamma}_{\mathcal{J}}^{'} \rangle \;\; \geq \;\; m_f ||\boldsymbol{\gamma}_{\mathcal{J}} - \boldsymbol{\gamma}_{\mathcal{J}}^{'}||_2^2. \tag{H.5}$$

2. ∇f is Lipschitz continuous, i.e., there exists a positive scalar M_f such that, for all

$\boldsymbol{\gamma}_{\mathcal{J}}, \boldsymbol{\gamma}_{\mathcal{J}}^{'}$, we have

$$||\nabla f(\boldsymbol{\gamma}_{\mathcal{J}}) - \nabla f(\boldsymbol{\gamma}_{\mathcal{J}}^{'})||_2 \leq M_f ||\boldsymbol{\gamma}_{\mathcal{J}} - \boldsymbol{\gamma}_{\mathcal{J}}^{'}||_2. \tag{H.6}$$

Let r denote the ADMM iteration count. From the zero subgradient optimality conditions corresponding to (6.13) and (6.14), we have

$$\nabla f(\boldsymbol{\gamma}_{\mathcal{J}}^{r+1})^T + \mathbf{E}_1^T \boldsymbol{\lambda}^r + \rho \mathbf{E}_1^T \mathbf{E}_1 \boldsymbol{\gamma}_{\mathcal{J}}^{r+1} + \rho \mathbf{E}_1^T \mathbf{E}_2 \boldsymbol{\gamma}_{\mathcal{B}}^{r} = 0 \tag{H.7}$$

$$\mathbf{E}_2^T \boldsymbol{\lambda}^r + \rho \mathbf{E}_2^T \mathbf{E}_2 \boldsymbol{\gamma}_{\mathcal{B}}^{r+1} + \rho \mathbf{E}_2^T \mathbf{E}_1 \boldsymbol{\gamma}_{\mathcal{J}}^{r+1} = 0. \tag{H.8}$$

From the dual variable update equation, we have,

$$\boldsymbol{\lambda}^{r+1} = \boldsymbol{\lambda}^r + \rho(\mathbf{E}_1 \boldsymbol{\gamma}_{\mathcal{J}}^{r+1} + \mathbf{E}_2 \boldsymbol{\gamma}_{\mathcal{B}}^{r+1}). \tag{H.9}$$

Premultiplying (H.9) with $\mathbf{E}_1^T$ and $\mathbf{E}_2^T$ followed by its summation to (H.7) and (H.8) respectively gives

$$\nabla f(\boldsymbol{\gamma}_{\mathcal{J}}^{r+1})^T + \mathbf{E}_1^T \boldsymbol{\lambda}^{r+1} + \rho \mathbf{E}_1^T \mathbf{E}_2 (\boldsymbol{\gamma}_{\mathcal{B}}^{r} - \boldsymbol{\gamma}_{\mathcal{B}}^{r+1}) = 0. \tag{H.10}$$

$$\mathbf{E}_2^T \boldsymbol{\lambda}^{r+1} = 0. \tag{H.11}$$

By initializing $\boldsymbol{\lambda}$ equal to zero, $\boldsymbol{\lambda}^r$ always lies in the nullspace $\mathcal{N}(\mathbf{E}_2^T)$, physically implying that the sum of the Lagrange multipliers of nodes connected to a given bridge node is always equal to zero. Let $\boldsymbol{\gamma}_{\mathcal{J}}^r \rightarrow \boldsymbol{\gamma}_{\mathcal{J}}^*$, $\boldsymbol{\gamma}_{\mathcal{B}}^r \rightarrow \boldsymbol{\gamma}_{\mathcal{B}}^*$ and $\boldsymbol{\lambda}^r \rightarrow \boldsymbol{\lambda}^*$ as $r \rightarrow \infty$, then putting $r \rightarrow \infty$ in (H.9), (H.10) and (H.11) gives

$$\nabla f(\boldsymbol{\gamma}_{\mathcal{J}}^*)^T + \mathbf{E}_1^T \boldsymbol{\lambda}^* = 0 \tag{H.12}$$

$$\mathbf{E}_2^T \boldsymbol{\lambda}^* = 0 \tag{H.13}$$

$$\mathbf{E}_1 \boldsymbol{\gamma}_{\mathcal{J}}^* + \mathbf{E}_2 \boldsymbol{\gamma}_{\mathcal{B}}^* = 0. \tag{H.14}$$

Note that the condition (H.14) implies consensus among $\boldsymbol{\gamma}_j, j \in \mathcal{J}$, upon convergence. By subtracting (H.12), (H.13) and (H.14) from (H.10), (H.11) and (H.9), respectively, we get the desired difference terms needed for showing convergence results.

$$\nabla f(\boldsymbol{\gamma}_{\mathcal{J}}^{r+1})^T - \nabla f(\boldsymbol{\gamma}_{\mathcal{J}}^*)^T + \mathbf{E}_1^T(\boldsymbol{\lambda}^{r+1} - \boldsymbol{\lambda}^*) + \rho \mathbf{E}_1^T \mathbf{E}_2(\boldsymbol{\gamma}_{\mathcal{B}}^r - \boldsymbol{\gamma}_{\mathcal{B}}^{r+1}) = 0 \tag{H.15}$$

$$\mathbf{E}_2^T(\boldsymbol{\lambda}^{r+1} - \boldsymbol{\lambda}^*) = 0 \tag{H.16}$$

$$\boldsymbol{\lambda}^{r+1} - \boldsymbol{\lambda}^r = \rho \mathbf{E}_1(\boldsymbol{\gamma}_{\mathcal{J}}^{r+1} - \boldsymbol{\gamma}_{\mathcal{J}}^*) + \rho \mathbf{E}_2(\boldsymbol{\gamma}_{\mathcal{B}}^{r+1} - \boldsymbol{\gamma}_{\mathcal{B}}^*). \tag{H.17}$$

Premultiplying (H.17) with $\mathbf{E}_2^T$ and using (H.11), we obtain,

$$\mathbf{E}_2^T \mathbf{E}_1(\boldsymbol{\gamma}_{\mathcal{J}}^{r+1} - \boldsymbol{\gamma}_{\mathcal{J}}^*) = -\mathbf{E}_2^T \mathbf{E}_2(\boldsymbol{\gamma}_{\mathcal{B}}^{r+1} - \boldsymbol{\gamma}_{\mathcal{B}}^*). \tag{H.18}$$

From the strong convexity of f and using (H.15), we can write,

$$\begin{aligned}
& m_f ||\boldsymbol{\gamma}_{\mathcal{J}}^{r+1} - \boldsymbol{\gamma}_{\mathcal{J}}^*||_2^2 \\
\leq\ & \langle \mathbf{E}_1^T(\boldsymbol{\lambda}^* - \boldsymbol{\lambda}^{r+1}), \boldsymbol{\gamma}_{\mathcal{J}}^{r+1} - \boldsymbol{\gamma}_{\mathcal{J}}^* \rangle \ +\ \rho \langle \mathbf{E}_1^T \mathbf{E}_2(\boldsymbol{\gamma}_{\mathcal{B}}^{r+1} - \boldsymbol{\gamma}_{\mathcal{B}}^r), (\boldsymbol{\gamma}_{\mathcal{J}}^{r+1} - \boldsymbol{\gamma}_{\mathcal{J}}^*) \rangle \\
=\ & \langle (\boldsymbol{\lambda}^* - \boldsymbol{\lambda}^{r+1}), \mathbf{E}_1(\boldsymbol{\gamma}_{\mathcal{J}}^{r+1} - \boldsymbol{\gamma}_{\mathcal{J}}^*) \rangle \ +\ \rho \langle (\boldsymbol{\gamma}_{\mathcal{B}}^{r+1} - \boldsymbol{\gamma}_{\mathcal{B}}^r), \mathbf{E}_2^T \mathbf{E}_1(\boldsymbol{\gamma}_{\mathcal{J}}^{r+1} - \boldsymbol{\gamma}_{\mathcal{J}}^*) \rangle \\
=\ & \langle (\boldsymbol{\lambda}^* - \boldsymbol{\lambda}^{r+1}), \mathbf{E}_1(\boldsymbol{\gamma}_{\mathcal{J}}^{r+1} - \boldsymbol{\gamma}_{\mathcal{J}}^*) \rangle \ -\ \rho \langle (\boldsymbol{\gamma}_{\mathcal{B}}^{r+1} - \boldsymbol{\gamma}_{\mathcal{B}}^r), \mathbf{E}_2^T \mathbf{E}_2(\boldsymbol{\gamma}_{\mathcal{B}}^{r+1} - \boldsymbol{\gamma}_{\mathcal{B}}^*) \rangle \\
=\ & \langle (\boldsymbol{\lambda}^* - \boldsymbol{\lambda}^{r+1}), \frac{1}{\rho}(\boldsymbol{\lambda}^{r+1} - \boldsymbol{\lambda}^r) \ -\ \mathbf{E}_2(\boldsymbol{\gamma}_{\mathcal{B}}^{r+1} - \boldsymbol{\gamma}_{\mathcal{B}}^*) \rangle \ -\ \rho \langle (\boldsymbol{\gamma}_{\mathcal{B}}^{r+1} - \boldsymbol{\gamma}_{\mathcal{B}}^r), \mathbf{E}_2^T \mathbf{E}_2(\boldsymbol{\gamma}_{\mathcal{B}}^{r+1} - \boldsymbol{\gamma}_{\mathcal{B}}^*) \rangle \\
=\ & \frac{1}{\rho} \langle (\boldsymbol{\lambda}^* - \boldsymbol{\lambda}^{r+1}), (\boldsymbol{\lambda}^{r+1} - \boldsymbol{\lambda}^r) \rangle \ +\ \rho \langle \mathbf{E}_2(\boldsymbol{\gamma}_{\mathcal{B}}^{r+1} - \boldsymbol{\gamma}_{\mathcal{B}}^r), \mathbf{E}_2(\boldsymbol{\gamma}_{\mathcal{B}}^* - \boldsymbol{\gamma}_{\mathcal{B}}^{r+1}) \rangle.
\end{aligned} \tag{H.19}$$

Here, the first identity is obtained by using a property of the inner product. The second, third and fourth identities are obtained by using (H.18), (H.17) and (H.16) respectively. By defining $\mathbf{u} = [(\mathbf{E}_2\boldsymbol{\gamma}_{\mathcal{B}})^T \mid \boldsymbol{\lambda}^T]^T$, the RHS in (H.19) can be expressed as $(\mathbf{u}^r - \mathbf{u}^{r+1})^T\mathbf{G}(\mathbf{u}^{r+1} - \mathbf{u}^*)$, where $\mathbf{G}$ is given by

$$\mathbf{G} = \begin{bmatrix} \rho I_{n|\mathcal{B}|} & 0 \\ 0 & \frac{1}{\rho} I_{N_C} \end{bmatrix}.$$

Using the identity:

$$2(\mathbf{u}^r - \mathbf{u}^{r+1})^T\mathbf{G}(\mathbf{u}^{r+1} - \mathbf{u}^*) = ||\mathbf{u}^r - \mathbf{u}^*||^2_{\mathbf{G}} - ||\mathbf{u}^{r+1} - \mathbf{u}^*||^2_{\mathbf{G}} - ||\mathbf{u}^r - \mathbf{u}^{r+1}||^2_{\mathbf{G}}, \quad \text{(H.20)}$$

the inequality in (H.19) can be rewritten as

$$m_f||\boldsymbol{\gamma}^{r+1}_{\mathcal{J}} - \boldsymbol{\gamma}^*_{\mathcal{J}}||_2^2 \;\leq\; \frac{1}{2}\left(||\mathbf{u}^r - \mathbf{u}^*||^2_{\mathbf{G}} - ||\mathbf{u}^{r+1} - \mathbf{u}^*||^2_{\mathbf{G}} - ||\mathbf{u}^r - \mathbf{u}^{r+1}||^2_{\mathbf{G}}\right). \quad \text{(H.21)}$$

By discarding the non-positive terms in the LHS of (H.21), we obtain the following upper bound on the primal optimality gap.

$$||\boldsymbol{\gamma}^{r+1}_{\mathcal{J}} - \boldsymbol{\gamma}^*_{\mathcal{J}}||_2^2 \;\leq\; \frac{1}{2m_f}||\mathbf{u}^r - \mathbf{u}^*||^2_{\mathbf{G}}. \quad \text{(H.22)}$$

In Appendix H.4, we prove the monotonic convergence of $\mathbf{u}^r$ to $\mathbf{u}^*$. Thus, from the monotonic decay of the RHS in (H.22), we have R-linear convergence of $\boldsymbol{\gamma}^r_{\mathcal{J}}$ to $\boldsymbol{\gamma}^*_{\mathcal{J}}$. □

H.4 Proof of monotonic convergence of $\mathbf{u}^r$ to $\mathbf{u}^*$

Proof. In order to prove monotonic convergence of $\mathbf{u}^r$ to $\mathbf{u}^*$, it is sufficient to show that there exists a $\delta > 0$ such that

$$||\mathbf{u}^{r+1} - \mathbf{u}^*||^2_{\mathbf{G}} \;\leq\; \frac{1}{1+\delta}||\mathbf{u}^r - \mathbf{u}^*||^2_{\mathbf{G}}. \tag{H.23}$$

By rearranging the terms in (H.21), we have

$$||\mathbf{u}^{r+1} - \mathbf{u}^*||^2_{\mathbf{G}} \leq ||\mathbf{u}^r - \mathbf{u}^*||^2_{\mathbf{G}} - ||\mathbf{u}^{r+1} - \mathbf{u}^r||^2_{\mathbf{G}} - 2m_f||\boldsymbol{\gamma}^{r+1}_{\mathcal{J}} - \boldsymbol{\gamma}^*_{\mathcal{J}}||^2_2. \tag{H.24}$$

By comparing terms in (H.23) and (H.24), we observe that if

$$2m_f||\boldsymbol{\gamma}^{r+1}_{\mathcal{J}} - \boldsymbol{\gamma}^*_{\mathcal{J}}||^2_2 + ||\mathbf{u}^{r+1} - \mathbf{u}^r||^2_{\mathbf{G}} \geq \delta||\mathbf{u}^{r+1} - \mathbf{u}^*||^2_{\mathbf{G}}, \tag{H.25}$$

or equivalently,

$$\begin{aligned} 2m_f||\boldsymbol{\gamma}^{r+1}_{\mathcal{J}} - \boldsymbol{\gamma}^*_{\mathcal{J}}||^2_2 + \rho||\mathbf{E}_2(\boldsymbol{\gamma}^{r+1}_{\mathcal{B}} - \boldsymbol{\gamma}^r_{\mathcal{B}})||^2_2 + \frac{1}{\rho}||\boldsymbol{\lambda}^{r+1} - \boldsymbol{\lambda}^r||^2_2 \\ \geq \delta\left(\rho||\mathbf{E}_2(\boldsymbol{\gamma}^{r+1}_{\mathcal{B}} - \boldsymbol{\gamma}^*_{\mathcal{B}})||^2_2 + \frac{1}{\rho}||\boldsymbol{\lambda}^{r+1} - \boldsymbol{\lambda}^*||^2_2\right), \end{aligned} \tag{H.26}$$

holds, then $\mathbf{u}^r$ converges monotonically to $\mathbf{u}^*$. We now proceed to derive upper bounds for $||\mathbf{E}_2(\boldsymbol{\gamma}^{r+1}_{\mathcal{B}} - \boldsymbol{\gamma}^*_{\mathcal{B}})||_2$ and $||\boldsymbol{\lambda}^{r+1} - \boldsymbol{\lambda}^*||_2$ in terms of the LHS. These upper bounds will be used in the sequel to establish the inequality in (H.26).

- *An upper bound for* $\rho||\mathbf{E}_2(\boldsymbol{\gamma}^{k+1}_{\mathcal{B}} - \boldsymbol{\gamma}^*_{\mathcal{B}})||_2$

  Note that for any two vectors $\mathbf{a}$, $\mathbf{b}$ and a scalar $\mu > 1$

$$||\mathbf{a} + \mathbf{b}||^2_2 \;\geq\; (1-\mu)||\mathbf{a}||^2_2 + \left(1 - \frac{1}{\mu}\right)||\mathbf{b}||^2_2. \tag{H.27}$$

Applying inequality (H.27) to (H.17), we get the following upper bound.

$$\rho||\mathbf{E}_2(\boldsymbol{\gamma}_{\mathcal{B}}^{r+1}-\boldsymbol{\gamma}_{\mathcal{B}}^{*})||_2^2 \leq \left(\frac{\mu}{\mu-1}\right)\frac{1}{\rho}||\boldsymbol{\lambda}^{r+1}-\boldsymbol{\lambda}^{r}||_2^2+\left(\mu\rho\sigma_{\max}^2(\mathbf{E}_1)\right)||\boldsymbol{\gamma}_{\mathcal{J}}^{r+1}-\boldsymbol{\gamma}_{\mathcal{J}}^{*}||_2^2. \quad \text{(H.28)}$$

Here, $\sigma_{\max}(\mathbf{E}_1)$ is the largest singular value of $\mathbf{E}_1$.

- *An upper bound for* $\frac{1}{\rho}||\boldsymbol{\lambda}^{r+1}-\boldsymbol{\lambda}^{*}||_2$

  Similar application of inequality (H.27) to (H.15) results in an upper bound for $\frac{1}{\rho}||\boldsymbol{\lambda}^{r+1}-\boldsymbol{\lambda}^{*}||_2$ as shown below.

$$\begin{aligned} ||\mathbf{E}_1^T(\boldsymbol{\lambda}^{r+1}-\boldsymbol{\lambda}^{*})||_2^2 &\leq \frac{\nu}{(\nu-1)}||\nabla f(\boldsymbol{\gamma}_{\mathcal{J}}^{r+1})^T-\nabla f(\boldsymbol{\gamma}_{\mathcal{J}}^{*})^T||_2^2 \\ &\quad + \nu||\rho\mathbf{E}_1^T\mathbf{E}_2(\boldsymbol{\gamma}_{\mathcal{B}}^{r}-\boldsymbol{\gamma}_{\mathcal{B}}^{r+1})||_2^2 \\ \implies \frac{1}{\rho}||\boldsymbol{\lambda}^{r+1}-\boldsymbol{\lambda}^{*}||_2^2 &\leq \frac{\nu}{\rho(\nu-1)\sigma_{\min}^2(\mathbf{E}_1)}||\nabla f(\boldsymbol{\gamma}_{\mathcal{J}}^{r+1})^T-\nabla f(\boldsymbol{\gamma}_{\mathcal{J}}^{*})^T||_2^2 \\ &\quad + \frac{\nu\rho\sigma_{\max}^2(\mathbf{E}_1)}{\sigma_{\min}^2(\mathbf{E}_1)}||\mathbf{E}_2(\boldsymbol{\gamma}_{\mathcal{B}}^{r}-\boldsymbol{\gamma}_{\mathcal{B}}^{r+1})||_2^2. \end{aligned} \quad \text{(H.29)}$$

From the Lipschitz continuity of ∇f (H.6), we obtain the following modified upper bound.

$$\begin{aligned} \frac{1}{\rho}||\boldsymbol{\lambda}^{r+1}-\boldsymbol{\lambda}^{*}||_2^2 &\leq \frac{\nu M_f^2}{\rho(\nu-1)\sigma_{\min}^2(\mathbf{E}_1)}||\boldsymbol{\gamma}_{\mathcal{J}}^{r+1}-\boldsymbol{\gamma}_{\mathcal{J}}^{*}||_2^2 \\ &\quad + \frac{\nu\rho\sigma_{\max}^2(\mathbf{E}_1)}{\sigma_{\min}^2(\mathbf{E}_1)}||\mathbf{E}_2(\boldsymbol{\gamma}_{\mathcal{B}}^{r}-\boldsymbol{\gamma}_{\mathcal{B}}^{r+1})||_2^2. \end{aligned} \quad \text{(H.30)}$$

Here, $\sigma_{\min}(\mathbf{E}_1)$ denotes the smallest singular value of $\mathbf{E}_1$ and ν is a positive scalar greater than unity.

By summing the upper bounds in (H.28) and (H.30), we get

$$\rho||\mathbf{E}_2(\boldsymbol{\gamma}_{\mathcal{B}}^{r+1}-\boldsymbol{\gamma}_{\mathcal{B}}^{*})||_2^2 \;+\; \frac{1}{\rho}||\boldsymbol{\lambda}^{r+1}-\boldsymbol{\lambda}^{*}||_2^2$$
$$\leq \frac{1}{\delta}\left(2m_f||\boldsymbol{\gamma}_{\mathcal{J}}^{r+1}-\boldsymbol{\gamma}_{\mathcal{J}}^{*}||_2^2+\rho||\mathbf{E}_2(\boldsymbol{\gamma}_{\mathcal{B}}^{r}-\boldsymbol{\gamma}_{\mathcal{B}}^{r+1})||_2^2+\frac{1}{\rho}||\boldsymbol{\lambda}^{r+1}-\boldsymbol{\lambda}^{r}||_2^2\right) \quad \text{(H.31)}$$

where

$$\delta \triangleq \left(\max_{\mu,\nu\geq 1}\max\left(\frac{\frac{\nu M_f^2}{\rho(\nu-1)\sigma_{\min}^2(\mathbf{E}_1)}+\mu\rho\sigma_{\max}^2(\mathbf{E}_1)}{2m_f},\nu\kappa,\frac{\mu}{\mu-1}\right)\right)^{-1}. \quad \text{(H.32)}$$

Thus, for δ as defined above, the inequality (H.26) holds and consequently the inequality (H.23) also holds, thereby establishing the Q-linear convergence of $\mathbf{u}^k$ to $\mathbf{u}^*$. □

H.5 Proof of Theorem 6.2

Proof. Let δ_{opt} denote the maximum value of δ for any $\rho > 0$. Then, we can write

$$\begin{aligned}\delta_{\text{opt}} &= \max_{\rho>0}\left(\max_{\mu,\nu\geq 1}\left(\min\left(f_1(\mu,\nu,\rho),f_2(\nu),f_3(\mu)\right)\right)\right)\\ &= \max_{\mu,\nu\geq 1}\left(\max_{\rho>0}\left(\min\left(f_1(\mu,\nu,\rho),f_2(\nu),f_3(\mu)\right)\right)\right) \quad \text{(H.33)}\end{aligned}$$

where the scalar functions f_1, f_2 and f_3 represent the three terms inside the minimum operator in (6.22). The following two Lemmas summarize the optimization of δ in (H.33).

Lemma H.1. $\delta_{opt} = \underset{\mu,\nu\geq 1}{max}\left\{min\left(\bar{f}_1(\mu,\nu),f_2(\nu),f_3(\mu)\right)\right\}$ *where,* $\bar{f}_1(\mu,\nu) \triangleq \underset{\rho>0}{max}\, f_1(\mu,\nu,\rho)$.

Proof. See Appendix H.6. □

Lemma H.2. *There exists a unique* $(\mu,\nu)=(\mu^*,\nu^*)$ *which simultaneously satisfies: (i)* $\bar{f}_1 = f_2 = f_3$, *and (ii)* $\mu \geq 1, \nu \geq 1$. *Further, such a* (μ^*,ν^*) *maximizes* $g(\mu,\nu) = min\left(\bar{f}_1(\mu,\nu),f_2(\nu),f_3(\mu)\right)$ *over* $\mu,\nu \geq 1$.

Proof. See Appendix H.7. □

The scalar function f_1 in Lemma H.1 is maximized at $\rho = \dfrac{M_f}{\sigma_{\max}\sigma_{\min}}\sqrt{\dfrac{\nu}{\mu(\nu-1)}}$ to give $\bar{f}_1 = \dfrac{M_f}{\sigma_{\min}\sigma_{\max}}\sqrt{\dfrac{\nu}{\mu(\nu-1)}}$. Further, by solving for the unique tuple (μ^*, ν^*) which satisfies the two optimality conditions specified in Lemma H.2, the optimal augmented Lagrangian parameter ρ and corresponding optimal δ can be shown to be equal to the ρ_{opt} and δ_{opt} as defined in Theorem 6.2. □

H.6 Proof of Lemma H.1

Proof. Let $\rho_{\mu,\nu} \triangleq \arg\max_{\rho>0} f_1$. Then, by restricting the feasible set in (H.33), we have,

$$\begin{aligned}\delta_{\text{opt}} &\geq \max_{\mu,\nu\geq 1}\left[\max_{\rho=\rho_{\mu,\nu}}\left\{\min\left(f_1(\mu,\nu,\rho), f_2(\nu), f_3(\mu)\right)\right\}\right] \\ &= \max_{\mu,\nu\geq 1}\left\{\min\left(\tilde{f}_1(\mu,\nu), f_2(\nu), f_3(\mu)\right)\right\}. \end{aligned} \tag{H.34}$$

On the other hand, from (H.33) and using $\tilde{f}_1 \geq f_1$, we have,

$$\begin{aligned}\delta_{\text{opt}} &= \max_{\mu,\nu\geq 1}\left[\max_{\rho>0}\left\{\min\left(f_1(\mu,\nu,\rho), f_2(\nu), f_3(\mu)\right)\right\}\right] \\ &\leq \max_{\mu,\nu\geq 1}\left\{\min\left(\tilde{f}_1(\mu,\nu), f_2(\nu), f_3(\mu)\right)\right\}. \end{aligned} \tag{H.35}$$

Combining (H.34) and (H.35) establishes Lemma H.1. □

H.7 Proof of Lemma H.2

Proof. In order to prove the Lemma, we claim the following.

1. For any $\epsilon > 0$, there exist positive constants B_μ and B_ν such that $g(\mu,\nu) \leq \epsilon$ when either $\mu \geq B_\mu$ or $\nu \geq B_\nu$ holds.

2. Any point (μ, ν) which satisfies condition 2 but does not satisfy condition 1 cannot be a local maximum of g.

Note that claim (a) holds trivially for $B_\mu = \frac{m_f^2}{\kappa \mathcal{M}_f^2 \epsilon^2}$ and $B_\nu = \frac{1}{\kappa\epsilon}$. In order to verify claim (b), let us consider a point (μ_0, ν_0) which satisfies condition 2, but not condition 1. Then, we need to consider three cases.

- *Case-I:* $\tilde{f}_1$, f_2 and f_3 are distinct at (μ_0, ν_0). Without loss of generality (WLOG), let $g = \tilde{f}_1$ at (μ_0, ν_0). Then, from the continuity of $\tilde{f}_1, f_2, f_3$, there exists an ϵ (> 0) ball B_ϵ, centered at (μ_0, ν_0) and with radius ϵ inside which $g = \tilde{f}_1$ holds. Since, inside B_ϵ, g is strictly monotonic with respect to μ and ν, there exists $(\mu, \nu) \in \mathcal{B}_\epsilon$ such that $g(\mu, \nu) > g(\mu_0, \nu_0)$. Hence, (μ_0, ν_0) is not a local maximum.

- *Case-II:* At (μ_0, ν_0), two of $\tilde{f}_1$, f_2 and f_3 are equal and strictly greater than the remaining one. The same arguments as Case-I apply here as well.

- *Case-III:* At (μ_0, ν_0), two of $\tilde{f}_1$, f_2 and f_3 are equal and strictly less than the remaining one. WLOG, let $\tilde{f}_1 = f_2 < f_3$. Let $\mathcal{C}(\mu, \nu)$ denote the continuous curve in (μ, ν) plane whose each point satisfies $\tilde{f}_1 = f_2$. Clearly, (μ_0, ν_0) also lies on the curve $\mathcal{C}$. Moreover, there are an uncountably infinite number of points of $\mathcal{C}$ inside B_ϵ, with B_ϵ defined as in Case-I. Due to the monotonicity of g along $\mathcal{C}$, there exists $(\mu, \nu) \in \mathcal{B}_\epsilon$ such that $g(\mu, \nu) > g(\mu_0, \nu_0)$. Hence, (μ_0, ν_0) is not a local maximum.

From claim (a) and the fact that at the boundary points ($\mu = 1$ or $\nu = 1$), the objective g evaluates to zero, we may restrict our search for the global maximizer of g to the set $\mathcal{D} = \{(\mu, \nu) \mid 1 \leq \mu \leq B_\mu, 1 \leq \nu \leq B_\nu\}$. Then, from claim (b), uniqueness of $(\mu^*, \nu^*) \in D$

and Weierstrass theorem, it follows that (μ^*, ν^*) is indeed the unique global maximizer of the continuous function g. Thus, the proof is complete. □

Appendix I

Appendix for Chapter 7

I.1 Index-wise LLRT for Hard Support Estimation

Due to the zero mean Gaussian measurement noise and the Gaussian prior $\mathcal{N}(0, \mathbf{\Gamma}_j)$ for the unknown sparse vector $\mathbf{x}_j$ at node j, the likelihood $p(\mathbf{y}_j; \boldsymbol{\gamma}_j^k)$ is given by

$$p(\mathbf{y}_j; \boldsymbol{\gamma}_j) = \mathcal{N}(0, \sigma_j^2 \mathbf{I}_m + \mathbf{\Phi}_j \mathbf{\Gamma}_j \mathbf{\Phi}_j^T). \tag{I.1}$$

Using (I.1), we setup the LLRT for i^{th} index at node j as:

$$\text{Decide in favor of } \mathcal{H}_1 \text{ for index } i \text{ if } \log \frac{p(\mathbf{y}_j; \boldsymbol{\gamma}_j^k, \boldsymbol{\gamma}_j^k(i) \neq 0)}{p(\mathbf{y}_j; \boldsymbol{\gamma}_j^k, \boldsymbol{\gamma}_j^k(i) = 0)} \geq \theta \tag{I.2}$$

or equivalently,

$$\log \frac{\mathcal{N}(\mathbf{y}_j; 0, \sigma_j^2 \mathbf{I}_m + \mathbf{\Phi}_j \mathbf{\Gamma}_j^k \mathbf{\Phi}_j^T)}{\mathcal{N}(\mathbf{y}_j; 0, \sigma_j^2 \mathbf{I}_m + \mathbf{\Phi}_j \tilde{\mathbf{\Gamma}}_{j,i}^k \mathbf{\Phi}_j^T)} \geq \theta \tag{I.3}$$

where $\tilde{\mathbf{\Gamma}}_{j,i}^k$=diag$(\boldsymbol{\gamma}_j^k(1), \ldots, \boldsymbol{\gamma}_j^k(i-1), 0, \boldsymbol{\gamma}_j^k(i+1), \ldots, \boldsymbol{\gamma}_j^k(n))$. Using the determinant property: $\det(\mathbf{I} + \mathbf{AB}) = \det(\mathbf{I} + \mathbf{BA})$ and Woodbury matrix identity, (I.3) simplifies

to

$$-\frac{1}{2}\log\left(1+\boldsymbol{\gamma}_j(i)\boldsymbol{\Phi}_{j,i}^T\left(\sigma_j^2\mathbf{I}_m+\boldsymbol{\Phi}_j\tilde{\boldsymbol{\Gamma}}_{j,i}^k\boldsymbol{\Phi}_j^T\right)^{-1}\boldsymbol{\Phi}_{j,i}\right) + \frac{\left(\boldsymbol{\Phi}_{j,i}^T\left(\sigma_j^2\mathbf{I}_m+\boldsymbol{\Phi}_j\tilde{\boldsymbol{\Gamma}}_{j,i}^k\boldsymbol{\Phi}_j^T\right)^{-1}\mathbf{y}_j\right)^2}{2\left(\boldsymbol{\gamma}_j(i)^{-1}+\boldsymbol{\Phi}_{j,i}^T\left(\sigma_j^2\mathbf{I}_m+\boldsymbol{\Phi}_j\tilde{\boldsymbol{\Gamma}}_{j,i}^k\boldsymbol{\Phi}_j^T\right)^{-1}\boldsymbol{\Phi}_{j,i}\right)} \geq \theta$$

Moving terms independent of $\mathbf{y}_j$ to the RHS and dividing both sides by the term $\boldsymbol{\Phi}_{j,i}^T\left(\sigma_j^2\mathbf{I}_m+\boldsymbol{\Phi}_j\tilde{\boldsymbol{\Gamma}}_{j,i}^k\boldsymbol{\Phi}_j^T\right)^{-1}\boldsymbol{\Phi}_{j,i}$ yields the Neyman-Pearson test given below

$$\frac{\left(\boldsymbol{\Phi}_{j,i}^T\left(\sigma_j^2\mathbf{I}_m+\boldsymbol{\Phi}_j\tilde{\boldsymbol{\Gamma}}_{j,i}^k\boldsymbol{\Phi}_j^T\right)^{-1}\mathbf{y}_j\right)^2}{\boldsymbol{\Phi}_{j,i}^T\left(\sigma_j^2\mathbf{I}_m+\boldsymbol{\Phi}_j\tilde{\boldsymbol{\Gamma}}_{j,i}^k\boldsymbol{\Phi}_j^T\right)^{-1}\boldsymbol{\Phi}_{j,i}} \geq h(\theta,\boldsymbol{\gamma}_j^k(i))\left\{\frac{1}{\boldsymbol{\gamma}_j(i)}+\boldsymbol{\Phi}_{j,i}^T\left(\sigma_j^2\mathbf{I}_m+\boldsymbol{\Phi}_j\tilde{\boldsymbol{\Gamma}}_{j,i}^k\boldsymbol{\Phi}_j^T\right)^{-1}\boldsymbol{\Phi}_{j,i}\right\} \tag{I.4}$$

where

$$h(\theta,\boldsymbol{\gamma}_j^k(i)) \triangleq \left(\frac{2\theta+\log\left(1+\boldsymbol{\gamma}_j(i)\boldsymbol{\Phi}_{j,i}^T\left(\sigma_j^2\mathbf{I}_m+\boldsymbol{\Phi}_j\tilde{\boldsymbol{\Gamma}}_{j,i}^k\boldsymbol{\Phi}_j^T\right)^{-1}\boldsymbol{\Phi}_{j,i}\right)}{\boldsymbol{\Phi}_{j,i}^T\left(\sigma_j^2\mathbf{I}_m+\boldsymbol{\Phi}_j\tilde{\boldsymbol{\Gamma}}_{j,i}^k\boldsymbol{\Phi}_j^T\right)^{-1}\boldsymbol{\Phi}_{j,i}}\right).$$

Bibliography

[1] I. D. Schizas, A. Ribeiro, and G. B. Giannakis, "Consensus in ad hoc WSNs with noisy links; part I: Distributed estimation of deterministic signals," *IEEE Trans. Signal Process.*, vol. 56, no. 1, pp. 350–364, 2008.

[2] 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.

[3] J. F. C. Mota, J. M. F. Xavier, P. M. Q. Aguiar, and M. Puschel, "D-ADMM: A communication-efficient distributed algorithm for separable optimization," *IEEE Trans. Signal Process.*, vol. 61, no. 10, pp. 2718–2723, 2013.

[4] W. Shi, Q. Ling, G. Wu, and W. Yin, "EXTRA: An exact first-order algorithm for decentralized consensus optimization," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 944–966, 2015. [Online]. Available: http://dx.doi.org/10.1137/14096668X

[5] J. A. Tropp, "Greed is good: algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2231–2242, Oct 2004.

[6] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 6–18, Jan 2006.

[7] 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, 2005.

[8] M. Fornasier and F. Pitolli, "Adaptive iterative thresholding algorithms for magnetoencephalography (MEG)," *Journal of Computational and Applied Mathematics*, vol. 221, no. 2, pp. 386 – 395, 2008, special Issue: Recent Progress in Spline and Wavelet Approximation.

[9] R. Prasad, C. R. Murthy, and B. D. 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.

[10] 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.

[11] E. Lagunas, S. K. Sharma, S. Chatzinotas, and B. Ottersten, "Compressive sensing based target counting and localization exploiting joint sparsity," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 3231–3235.

[12] Z. Tan and A. Nehorai, "Sparse direction of arrival estimation using co-prime arrays with off-grid targets," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 26–29, Jan 2014.

[13] M. D. Iordache, J. M. Bioucas-Dias, and A. Plaza, "Collaborative sparse regression for hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 1, pp. 341–354, Jan 2014.

[14] 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, 2010.

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

[16] 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

[17] H. Zhang, V. M. Patel, and R. Chellappa, "Low-rank and joint sparse representations for multi-modal recognition," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4741–4752, Oct 2017.

[18] A. Adler, M. Elad, Y. Hel-Or, and E. Rivlin, "Sparse coding with anomaly detection," in *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Sept 2013, pp. 1–6.

[19] 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.

[20] 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," vol. 2, May 2004, pp. ii–369–72 vol.2.

[21] 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.

[22] 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.

[23] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Simultaneous sparse approximation via greedy pursuit," vol. 5, 2005, pp. v/721–v/724 Vol. 5.

[24] 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, April 2014.

[25] 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.

[26] R. Gribonval, H. Rauhut, K. Schnass, and P. Vandergheynst, "Atoms of all channels, unite! Average case analysis of multi-channel sparse recovery using greedy algorithms," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 655–687, 2008.

[27] P. Feng and Y. Bresler, "Spectrum-blind minimum-rate sampling and reconstruction of multiband signals," vol. 3, May 1996, pp. 1688–1691 vol. 3.

[28] 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.

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

[30] 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.

[31] ——, "Parameter identifiability in sparse Bayesian learning," May 2014, pp. 1851–1855.

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

[33] D. J. MacKay, "Bayesian interpolation," *Neural Computation*, vol. 4, pp. 415–447, 1991.

[34] J. Ziniel and P. Schniter, "Efficient high-dimensional inference in the multiple measurement vector problem," *IEEE Trans. Signal Process.*, vol. 61, no. 2, pp. 340–354, Jan 2013.

[35] B. D. Rao, Z. Zhang, and Y. Jin, "Sparse signal recovery in the presence of intra-vector and inter-vector correlation," in *2012 International Conference on Signal Processing and Communications (SPCOM)*, July 2012, pp. 1–5.

[36] R. Prasad, C. R. Murthy, and B. D. Rao, "Nested sparse Bayesian learning for block-sparse signals with intra-block correlation," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 7183–7187.

[37] G. Joseph and C. R. Murthy, "A noniterative online Bayesian algorithm for the recovery of temporally correlated sparse vectors," *IEEE Transactions on Signal Processing*, vol. 65, no. 20, pp. 5510–5525, Oct 2017.

[38] M. F. Duarte, S. Sarvotham, D. Baron, M. B. Wakin, and R. G. Baraniuk, "Distributed compressed sensing of jointly sparse signals," Oct 2005, pp. 1537–1541.

[39] H. Lu, X. Long, and J. Lv, "A fast algorithm for recovery of jointly sparse vectors based on the alternating direction methods," *J. Mach. Learn. Res*, pp. 461–469, 2011.

[40] J. Matamoros, S. M. Fosson, E. Magli, and C. Anton-Haro, "Distributed ADMM for in-network reconstruction of sparse signals with innovations," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 1, no. 4, pp. 225–234, Dec 2015.

[41] D. Wipf and B. Rao, "Sparse Bayesian learning for basis selection," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2153–2164, 2004.

[42] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[43] 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, 2014.

[44] J. Palmer, B. D. Rao, and D. P. Wipf, "Perspectives on sparse Bayesian learning," in *Advances in Neural Information Processing Systems*, 2004, pp. 249–256. [Online]. Available: http://papers.nips.cc/paper/2393-perspectives-on-sparse-bayesian-learning.pdf

[45] R. Neal and G. E. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models*. Kluwer Academic Publishers, 1998, pp. 355–368.

[46] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

[47] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.

[48] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, series B*, vol. 39, no. 1, pp. 1–38, 1977.

[49] 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.

[50] I. S. Dhillon and J. A. Tropp, "Matrix nearness problems with Bregman divergences," *SIAM Journal on Matrix Analysis and Applications*, vol. 29, no. 4, pp. 1120–1146, 2008.

[51] D. G. Luenberger and Y. Ye, *Linear and Nonlinear Programming, Third Edition*. Springer Publishing Company, Incorporated, 2010.

[52] W. K. Ma, T. H. Hsieh, and C. Y. Chi, "DOA estimation of quasi-stationary signals with less sensors than sources and unknown spatial noise covariance: A Khatri-Rao subspace approach," *IEEE Trans. Signal Process.*, vol. 58, no. 4, pp. 2168–2180, April 2010.

[53] S. Khanna and C. R. Murthy, "On the support recovery of jointly sparse Gaussian sources using sparse Bayesian learning," *CoRR*, vol. abs/1703.04930. [Online]. Available: http://arxiv.org/abs/1703.04930

[54] C. G. Khatri and C. R. Rao, "Solutions to some functional equations and their applications to characterization of probability distributions," *Sankhya: The Indian Journal of Statistics, Series A (1961-2002)*, vol. 30, no. 2, pp. 167–180, 1968.

[55] D. S. Bernstein, *Matrix Mathematics: Theory, Facts, and Formulas (Second Edition)*. Princeton University Press, 2009.

[56] S. Liu and G. Trenkler, "Hadamard, Khatri-Rao, Kronecker and other matrix products," *International Journal of Information and System Sciences*, vol. 4, no. 1, pp. 160–177, 2007.

[57] M. F. Duarte and R. G. Baraniuk, "Kronecker compressive sensing," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 494–504, Feb 2012.

[58] A. Koochakzadeh and P. Pal, "Sparse source localization in presence of co-array perturbations," in *International Conference on Sampling Theory and Applications (SampTA)*, May 2015, pp. 563–567.

[59] D. Romero, D. D. Arianandа, Z. Tian, and G. Leus, "Compressive covariance sensing: Structure-based compressive sensing beyond sparsity," *IEEE Signal Process. Mag.*, vol. 33, no. 1, pp. 78–93, Jan 2016.

[60] G. Dasarathy, P. Shah, B. N. Bhaskar, and R. D. Nowak, "Sketching sparse matrices, covariances, and graphs via tensor products," *IEEE Trans. Inf. Theory*, vol. 61, no. 3, pp. 1373–1388, March 2015.

[61] N. D. Sidiropoulos and A. Kyrillidis, "Multi-way compressed sensing for sparse low-rank tensors," *IEEE Signal Process. Lett.*, vol. 19, no. 11, pp. 757–760, Nov 2012.

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

[63] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[64] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec 2005.

[65] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation*, vol. 28, no. 3, pp. 253–263, 2008.

[66] T. T. Cai, L. Wang, and G. Xu, "New bounds for restricted isometry constants," *IEEE Trans. Inf. Theory*, vol. 56, no. 9, pp. 4388–4394, Sept 2010.

[67] A. M. Tillmann and M. E. Pfetsch, "The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing," *IEEE Trans. Inf. Theory*, vol. 60, no. 2, pp. 1248–1259, Feb. 2014.

[68] S. P. Chepuri and G. Leus, "Graph sampling for covariance estimation," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 3, no. 3, pp. 451–466, Sept 2017.

[69] N. D. Sidiropoulos and A. Kyrillidis, "Multi-way compressed sensing for sparse low-rank tensors," *IEEE Signal Processing Letters*, vol. 19, no. 11, pp. 757–760, Nov. 2012.

[70] D. Nion and N. D. Sidiropoulos, "A PARAFAC-based technique for detection and localization of multiple targets in a MIMO radar system," 2009, pp. 2077–2080.

[71] R. A. Horn and C. R. Johnson, Eds., *Matrix Analysis*. Cambridge University Press, 1986.

[72] S. Jokar, "Sparse recovery and Kronecker products," in *44th Annual Conference on Information Sciences and Systems (CISS)*, Mar 2010, pp. 1–4.

[73] 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.

[74] J. Anderson, M. Belkin, N. Goyal, L. Rademacher, and J. R. Voss, "The more, the merrier: the blessing of dimensionality for learning large Gaussian mixtures," in *27th Annual Conference on Learning Theory*, Jun 2014, pp. 1135–1164.

[75] N. D. Sidiropoulos and R. Bro, "On the uniqueness of multilinear decomposition of N-way arrays," *Journal of Chemometrics*, vol. 14, no. 3, pp. 229–239, 2000.

[76] P. Koiran and A. Zouzias, "Hidden cliques and the certification of the restricted isometry property," *IEEE Trans. Inf. Theory*, vol. 60, no. 8, pp. 4999–5006, Aug 2014.

[77] S. Khanna and C. R. Murthy, "Rényi divergence based covariance matching pursuit of joint sparse support," in *2017 IEEE 18th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, July 2017, pp. 1–5.

[78] C. R. Rao and M. B. Rao, *Matrix Algebra and Its Applications to Statistics and Econometrics*. Singapore: World Scientific, 1998.

[79] 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

[80] S. Khanna and C. R. Murthy, "On the restricted isometry of the columnwise Khatri-Rao product," *IEEE Trans. Signal Process.*, vol. abs/1709.05789, pp. 1–14, 2017. [Online]. Available: http://arxiv.org/abs/1709.05789

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

[82] 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.

[83] 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.

[84] Q. Ling, Z. Wen, and W. Yin, "Decentralized jointly sparse optimization by reweighted l-q minimization," *IEEE Trans. Signal Process.*, vol. 61, no. 5, pp. 1165–1170, 2013.

[85] T. van Erven and P. Harremos, "Rényi divergence and Kullback-Leibler divergence," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 3797–3820, July 2014.

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

[87] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions—i," *Mathematical Programming*, vol. 14, no. 1, pp. 265–294, 1978.

[88] M. Narasimhan and J. Bilmes, "A submodular-supermodular procedure with applications to discriminative structure learning," in *Uncertainty in Artificial Intelligence (UAI)*, Edinburgh, Scotland, July 2005.

[89] R. Iyer and J. Bilmes, "Algorithms for approximate minimization of the difference between submodular functions, with applications," in *Uncertainty in Artificial Intelligence (UAI)*, 2012, pp. 407–417.

[90] R. Horst and N. V. Thoai, "DC programming: Overview," *Journal of Optimization*

Theory and Applications, vol. 103, no. 1, pp. 1–43, Oct 1999. [Online]. Available: https://doi.org/10.1023/A:1021765131316

[91] D. Sundman, S. Chatterjee, and M. Skoglund, "Distributed greedy pursuit algorithms," *Signal Processing*, vol. 105, pp. 298 – 315, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S016516841400245X

[92] T. Wimalajeewa and P. Varshney, "Cooperative sparsity pattern recovery in distributed networks via distributed-OMP," May 2013, pp. 5288–5292.

[93] P. Stoica, P. Babu, and J. Li, "SPICE: A sparse covariance-based estimation method for array processing," *IEEE Transactions on Signal Processing*, vol. 59, no. 2, pp. 629–638, Feb 2011.

[94] M. Masood and T. Al-Naffouri, "Sparse reconstruction using distribution agnostic Bayesian matching pursuit," *IEEE Trans. Signal Process.*, vol. 61, no. 21, pp. 5298–5309, Nov. 2013.

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

[96] J. Le Roux, P. Boufounos, K. Kang, and J. Hershey, "Source localization in reverberant environments using sparse optimization," in *Proc. ICASSP*, May 2013, pp. 4310–4314.

[97] N. H. Nguyen, N. M. Nasrabadi, and T. D. Tran, "Robust multi-sensor classification

via joint sparse representation," in *Proc. 14th Int. Conf. Inform. Fusion*, July 2011, pp. 1–8.

[98] G. Li, T. Wimalajeewa, and P. Varshney, "Decentralized subspace pursuit for joint sparsity pattern recovery," in *Proc. ICASSP*, May 2014, pp. 3365–3369.

[99] D. Sundman, S. Chatterjee, and M. Skoglund, "Distributed greedy pursuit algorithms," *Signal Processing*, vol. 105, pp. 298 – 315, 2014.

[100] Q. Ling and Z. Tian, "Decentralized support detection of multiple measurement vectors with joint sparsity," May 2011, pp. 2996–2999.

[101] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Optimization with sparsity-inducing penalties," *Foundations and Trends in Machine Learning*, vol. 4, no. 1, pp. 1–106, Jan. 2012. [Online]. Available: http://dx.doi.org/10.1561/2200000015

[102] D. P. Wipf and S. S. Nagarajan, "A new view of automatic relevance determination," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2007, pp. 1625–1632.

[103] A. Makhzani and S. Valaee, "Distributed spectrum sensing in cognitive radios via graphical models," 2013, pp. 376–379.

[104] G. Tzagkarakis, J. L. Starck, and P. Tsakalides, "Joint sparse signal ensemble reconstruction in a wsn using decentralized Bayesian matching pursuit," in *Signal Processing Conference, 2011 19th European*, Aug 2011, pp. 338–342.

[105] D. Yang, H. Li, and G. Peterson, "Space-time turbo bayesian compressed sensing for UWB systems," May 2010, pp. 1–6.

[106] T. Wimalajeewa and P. K. Varshney, "Compressive sensing based signal processing in wireless sensor networks: A survey," *CoRR*, vol. arXiv:1709.10401. [Online]. Available: https://arxiv.org/abs/1709.10401

[107] G. Li, T. Wimalajeewa, and P. Varshney, "Decentralized subspace pursuit for joint sparsity pattern recovery," May 2014, pp. 3365–3369.

[108] P. A. Forero, A. Cano, and G. B. Giannakis, "Consensus-based distributed expectation-maximization algorithm for density estimation and classification using wireless sensor networks," Mar 2008, pp. 1989–1992.

[109] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2010.

[110] R. Zhang and J. Kwok, "Asynchronous distributed ADMM for consensus optimization," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. JMLR Workshop and Conference Proceedings, 2014, pp. 1701–1709. [Online]. Available: http://jmlr.org/proceedings/papers/v32/zhange14.pdf

[111] J. Matamoros, S. M. Fosson, E. Magli, and C. Anton-Haro, "Distributed ADMM for in-network reconstruction of sparse signals with innovations," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 1, no. 4, pp. 225–234, Dec 2015.

[112] 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.

[113] W. Deng and W. Yin, "On the global and linear convergence of the generalized alternating direction method of multipliers," *Rice University CAAM Technical Report TR12-14*, 2012.

[114] W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin, "On the linear convergence of the ADMM in decentralized consensus optimization," *IEEE Trans. Signal Process.*, vol. 62, no. 7, pp. 1750–1761, Apr. 2014.

[115] T. Erseghe, D. Zennaro, E. Dall'Anese, and L. Vangelista, "Fast consensus by the alternating direction multipliers method," *IEEE Trans. Signal Process.*, vol. 59, no. 11, pp. 5523–5537, Nov 2011.

[116] D. Jakovetic, J. Xavier, and J. Moura, "Fast distributed gradient methods," vol. 59, no. 5, pp. 1131–1146, May 2014.

[117] A. Chen and A. Ozdaglar, "A fast distributed proximal-gradient method," Oct 2012, pp. 601–608.

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

[119] Z. Quan, S. Cui, A. H. Sayed, and H. V. Poor, "Wideband spectrum sensing in cognitive radio networks," May 2008, pp. 901–906.

[120] H. Sun, A. Nallanathan, C. X. Wang, and Y. Chen, "Wideband spectrum sensing for cognitive radio networks: A survey," *IEEE Wireless Communications*, vol. 20, no. 2, pp. 74–81, Apr 2013.

[121] A. Sharma and C. R. Murthy, "Group testing-based spectrum hole search for cognitive radios," *IEEE Trans. Veh. Technol.*, vol. 63, no. 8, pp. 3794–3805, Oct 2014.

[122] J. Yoo, S. Becker, M. Monge, M. Loh, E. Candés, and A. Emami-Neyestanak, "Design and implementation of a fully integrated compressed-sensing signal acquisition system," Mar 2012, pp. 5325–5328.

[123] S. Khanna and C. Murthy, "Decentralized Bayesian learning of jointly sparse signals," in *Proc. GLOBECOM*, Dec 2014, pp. 3103–3108.

[124] M. Hurtado, C. Muravchik, and A. Nehorai, "Enhanced sparse Bayesian learning via statistical thresholding for signals in structured noise," *IEEE Trans. Signal Process.*, vol. 61, no. 21, pp. 5430–5443, 2013.

[125] M. E. Tipping, A. Faul, and J. J. T. Avenue, "Fast marginal likelihood maximisation for sparse Bayesian models," in *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, Jan. 2003, pp. 3–6.

[126] D. Shutin, S. Kulkarni, and H. Poor, "Incremental reformulated automatic relevance determination," *IEEE Trans. Signal Process.*, vol. 60, no. 9, pp. 4977–4981, Sep. 2012.

[127] S. A. Aldosari and J. M. F. Moura, "Fusion in sensor networks with communication constraints," in *Proceedings of the 3rd International Symposium on Information*

Processing in Sensor Networks, ser. IPSN '04. New York, NY, USA: ACM, Apr. 2004, pp. 108–115.

[128] S. Khanna and C. R. Murthy, "Decentralized joint-sparse signal recovery: A sparse Bayesian learning approach," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 3, no. 1, pp. 29–45, March 2017.

[129] A. Ribeiro and G. Giannakis, "Bandwidth-constrained distributed estimation for wireless sensor networks-part 1: Gaussian case," *IEEE Trans. Signal Process.*, vol. 54, no. 3, pp. 1131–1143, Mar. 2006.

[130] X. Wu and Z. Tian, "Optimized data fusion in bandwidth and energy constrained sensor networks," vol. 4, May 2006, pp. 713–716.

[131] P. Bianchi, G. Fort, W. Hachem, and J. Jakubowicz, "Convergence of a distributed parameter estimator for sensor networks with local averaging of the estimates," May 2011, pp. 3764–3767.

[132] P. Bianchi, G. Fort, and W. Hachem, "Performance of a distributed stochastic approximation algorithm," *IEEE Trans. Inf. Theory*, vol. 59, no. 11, pp. 7405–7418, Nov 2013.

[133] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[134] S. M. Herbert Robbins, "A stochastic approximation method," *The Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, 1951.

[135] D. Donoho and J. Tanner, "Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 367, no. 1906, pp. 4273–4293, 2009.

[136] S. Khanna and C. R. Murthy, "Communication-efficient decentralized sparse bayesian learning of joint sparse signals," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 3, no. 3, pp. 617–630, Sept 2017.

[137] G. Visick, "A quantitative version of the observation that the Hadamard product is a principal submatrix of the Kronecker product," *Linear Algebra and its Applications*, vol. 304, no. 13, pp. 45 – 68, 2000.

[138] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*. Cambridge; New York: Cambridge University Press, 1994.

[139] B. Mond and J. E. Pečarić, "Inequalities for the Hadamard product of matrices," *SIAM J. Matrix Anal. Appl.*, vol. 19, no. 1, pp. 66–70, Jan. 1998.

[140] H. J. Werner, "Hadamard product of square roots of correlation matrices," *Image*, vol. 26, pp. 1–32, April 2001.

[141] A. W. Marshall and I. Olkin, "Reversal of the Lyapunov, Holder, and Minkowski inequalities and other extensions of the kantorovich inequality," *Journal of Mathematical Analysis and Applications*, vol. 8, no. 3, pp. 503 – 514, 1964.

[142] S. Liu and H. Neudecker, "Several matrix Kantorovich-type inequalities," *Journal of Mathematical Analysis and Applications*, vol. 197, no. 1, pp. 23 – 26, 1996.

[143] ——, "Kantorovich inequalities and efficiency comparisons for several classes of estimators in linear models," *Statistica Neerlandica*, vol. 51, no. 3, pp. 345–355, 1997.

[144] S. Liu, "On the Hadamard product of square roots of correlation matrices," *Econometric Theory*, vol. 18, no. 4, p. 1007, August 2002.

[145] R. Adamczak, "A note on the Hanson-Wright inequality for random vectors with dependencies," *Electron. Commun. Probab.*, vol. 20, p. 13 pp., 2015.

[146] M. Rudelson and R. Vershynin, "Hanson-wright inequality and sub-gaussian concentration," *Electronic Communications in Probability*, vol. 18, p. 9 pp., 2013.

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