# Rényi Divergence based Covariance Matching Algorithm for Joint Sparse Signal Recovery

Saurabh Khanna,

Signal Processing for Communication, ECE, IISc

# Outline

- Joint sparse signal recovery problem
- Covariance matching approach for support recovery
- Covariance matching using Rènyi matrix divergence
- Sub-Sup procedure for minimizing Rènyi matrix divergence
- Demo

# Joint sparse signal recovery problem

Multiple measurement vector (MMV) model:

$$\mathbf{y}_j = \mathbf{\Phi}\mathbf{x}_j + \mathbf{w}_j \qquad j = 1 \text{ to } L$$

$\mathbf{x}_j \in \mathbb{R}^n$ are unknown $k-$sparse vectors
$\mathbf{y}_j \in \mathbb{R}^m$ are the noisy linear measurements
$\mathbf{\Phi} \in \mathbb{R}^{m \times n}$ is the meas matrix with $m < n$
$\mathbf{w}_j \sim \mathcal{N}(0, \sigma_n^2 \mathbf{I})$ is the meas noise

Vectors $\mathbf{x}_1, \mathbf{x}_2 \ldots \mathbf{x}_L$ follow the JSM-2 sparsity model [Duarte, ??].

- $\mathbf{x}_j$ have a common nonzero support
- Nonzero entries are uncorrelated

Goal is to recover the joint sparse vectors $\mathbf{X} = \mathbf{x}_1, \mathbf{x}_2 \ldots \mathbf{x}_L$ from their noisy linear measurements $\mathbf{Y} = \mathbf{y}_1, \mathbf{y}_2 \ldots \mathbf{y}_L$.

Existing JSM-2 algorithms: M-OMP, M-FOCUSS, Row-LASSO, CRL-1/2, M-SBL ...

# A Bayesian approach

Assume a Gaussian-mixture prior on the unknown vectors $\mathbf{x}_j$.

$$\mathbf{x}_j(i) \sim (1 - s_i)\mathcal{N}(0, \sigma_z^2) + s_i \mathcal{N}(0, \sigma_s^2), \quad j = 1 \text{ to } L, \ i \in [n].$$

- $\mathbf{s} \in \{0,1\}^n$ denotes the common support of $\mathbf{x}_j$.
- $\sigma_s^2$ is the common signal variance of the active coefficients.
- $\sigma_z^2$ is the common signal variance of the inactive coefficients.

For $\sigma_z^2 = 0$, the prior simplifies to

$$\mathbf{x}_j \sim \mathcal{N}(0, \sigma_s^2 \text{diag}(\mathbf{s})), \qquad j = 1 \text{ to } L.$$

Given the model parameters $\theta = \{\sigma_s^2, \sigma_n^2, \mathbf{s}\}$, the LMMSE estimate of $\mathbf{x}_j$ is computed as:

$$\hat{\mathbf{x}}_j^{\text{MMSE}} = \left(\sigma_s^2 \text{diag}(\mathbf{s}) \mathbf{\Phi}^T\right) \left(\sigma_n^2 \mathbf{I}_m + \sigma_s^2 \mathbf{\Phi}_{\mathbf{s}} \mathbf{\Phi}_{\mathbf{s}}^T\right)^{-1} \mathbf{y}_j$$

Question: How to find the model $\theta = \{\sigma_s^2, \sigma_n^2, \mathbf{s}\}$ from the observations $\mathbf{Y}$?

# ML estimation of the model paramters

Goal is to find the ML estimate of the model $\theta = \{\sigma_s^2, \sigma_n^2, \mathbf{s}\}$ given the observations $\mathbf{Y}$.

$$(\hat{\sigma}_n^2, \hat{\sigma}_s^2, \hat{\mathbf{s}}) = \underset{\sigma_s^2, \sigma_n^2, \mathbf{s}}{\text{arg min}} \; -\log p(\mathbf{Y}; \sigma_s^2, \sigma_n^2, \mathbf{s})$$

ML cost:

$$-\log p(\mathbf{Y}; \sigma_s^2, \sigma_n^2, \mathbf{s}) \;\; \propto \;\; L \log |\sigma_n^2 \mathbf{I} + \sigma_s^2 \boldsymbol{\Phi}_\mathbf{s} \boldsymbol{\Phi}_\mathbf{s}| + Tr\left( \left(\sigma_n^2 \mathbf{I} + \sigma_s^2 \boldsymbol{\Phi}_\mathbf{s} \boldsymbol{\Phi}_\mathbf{s}\right)^{-1} \mathbf{Y}\mathbf{Y}^T \right)$$

To simplify exposition, assume $\sigma_s^2$ and $\sigma_n^2$ to be known. Only $\mathbf{s}$ needs to be estimated.

# ML cost - an interesting interpretation

ML cost:

$$-\log p(\mathbf{Y};\mathbf{s}) \;\propto\; L\log|\sigma_n^2\mathbf{I}+\sigma_s^2\mathbf{\Phi_s}\mathbf{\Phi_s}| \;+\; Tr\left(\left(\sigma_n^2\mathbf{I}+\sigma_s^2\mathbf{\Phi_s}\mathbf{\Phi_s}\right)^{-1}\mathbf{Y}\mathbf{Y}^T\right)$$

Bregman matrix divergence with respect to $\phi(.) = -\log|.|$, is defined as:

$$\mathcal{D}(\mathbf{X},\mathbf{Y}) = \text{trace}(\mathbf{X}\mathbf{Y}^{-1}) - \log|\mathbf{X}\mathbf{Y}^{-1}| - N$$

ML cost can be interpreted as a matrix divergence:

$$-\log p(\mathbf{Y};\boldsymbol{\gamma}) \;=\; L\mathcal{D}_\phi\left(\frac{1}{L}\mathbf{Y}\mathbf{Y}^T\,,\;\sigma_n^2\mathbf{I}+\sigma_s^2\mathbf{\Phi_s}\mathbf{\Phi_s}^T\right) + \underbrace{m - \frac{L}{2}\log|\frac{1}{L}\mathbf{Y}\mathbf{Y}^T|}_{\text{constant}}$$

We want to find $\hat{\mathbf{s}}$ which minimizes $\mathcal{D}_\phi\left(\underbrace{\frac{1}{L}\mathbf{Y}\mathbf{Y}^T}_{\text{emp. cov mat}}\,,\;\underbrace{\sigma_n^2\mathbf{I}+\sigma_s^2\mathbf{\Phi_s}\mathbf{\Phi_s}^T}_{\text{param. cov mat}}\right)$.

# Generalizing the ML cost using Rényi divergence

We want to find an $\hat{\mathbf{s}}$ which minimizes $\mathcal{D}_\phi\left(\underbrace{\frac{1}{L}\mathbf{Y}\mathbf{Y}^T}_{\text{emp. cov mat}}, \underbrace{\sigma_n^2\mathbf{I} + \sigma_s^2\boldsymbol{\Phi}_{\mathbf{s}}\boldsymbol{\Phi}_{\mathbf{s}}^T}_{\text{param. cov mat}}\right)$.

Minimizing $\mathcal{D}_\phi$ with respect to $\mathbf{s}$ is a combinatorial problem.

Replace $\mathcal{D}_\phi$ with a convenient matrix divergence, which we call $\alpha$-Rényi matrix divergence,

$$\mathcal{D}_\alpha(\mathbf{X}, \mathbf{Y}) = \frac{1}{2(1-\alpha)} \log \frac{|\alpha\mathbf{X} + (1-\alpha)\mathbf{Y}|}{|\mathbf{X}|^\alpha\,|\mathbf{Y}|^{1-\alpha}}.$$

# $\alpha$-Rényi matrix divergence - interesting facts

For any two matrices $\mathbf{X}, \mathbf{Y} \in S_+^n$., we define $\alpha$-Rényi matrix divergence as:

$$\mathcal{D}_\alpha(\mathbf{X}, \mathbf{Y}) \triangleq \frac{1}{2(1-\alpha)} \log \frac{|\alpha\mathbf{X} + (1-\alpha)\mathbf{Y}|}{|\mathbf{X}|^\alpha \, |\mathbf{Y}|^{1-\alpha}}$$

Interesting facts about $\mathcal{D}_\alpha(.,.)$:

- For $\alpha < 1$, $\mathcal{D}_\alpha$ lower bounds $\mathcal{D}_{-\log|.|}$.
- For $\alpha \to 1$, we have $\mathcal{D}_\alpha \to \mathcal{D}_{-\log|.|}$.
- For $\alpha = 1/2$, $\mathcal{D}_\alpha$ is symmetric in arguments and is called the Jensen-Bregman-Log-Det divergence.

$$\mathcal{D}_{1/2}(\mathbf{X}, \mathbf{Y}) = \log\left|\frac{\mathbf{X}+\mathbf{Y}}{2}\right| - \frac{1}{2}\log|\mathbf{X}| - \frac{1}{2}\log|\mathbf{Y}|$$

- $\mathcal{D}_\alpha$ is a type of Jensen difference divergence.
- $\mathcal{D}_\alpha$ appears as an error exponent while analyzing the error probability in multi class hypothesis testing.

# Modified support recovery problem

We formulate support recovery as the below optimization problem:

$$\hat{\mathbf{s}} = \arg\max_{\mathbf{s}} \ \log \left| \alpha \mathbf{R}_{\mathbf{Y}} + (1-\alpha)\left(\sigma_n^2 \mathbf{I} + \sigma_s^2 \mathbf{\Phi}_{\mathbf{s}} \mathbf{\Phi}_{\mathbf{s}}^T\right) \right| - (1-\alpha) \log \left| \sigma_n^2 \mathbf{I} + \sigma_s^2 \mathbf{\Phi}_{\mathbf{s}} \mathbf{\Phi}_{\mathbf{s}}^T \right|.$$

The objective can be interpreted as a difference of two submodular functions in $\mathbf{s}$.

Claim:
For any positive definite matrix $\mathbf{A}$, a generic $n \times p$ matrix $\mathbf{B}$ and constant $\beta > 0$, the set function $f(S) = \log|\mathbf{A} + \beta \mathbf{B}_S \mathbf{B}_S^T|$ is submodular.

Why the submodularity property is interesting ?
Any submodular function can be minimized adequately by a fast greedy algorithm.

# Submodular functions

Let $f : U \rightarrow \mathbb{R}^+$ be a set function.

- Then, $f$ is called monotone if $f(S \cup \{a\}) \geq f(S)$, for all $S \subset U, a \in U \backslash S$.

- Further, $f$ is called a submodular function if it satisfies

$$f(S \cup \{a\}) - f(S) \quad \geq \quad f(T \cup \{a\}) - f(T) \qquad \text{(Law of diminishing returns)}$$

for all elements $a \in U \backslash T$ and all pairs of subsets $S, T$ such that $S \subseteq T \subseteq U$.

- If above always holds with equality, then $f$ is called a modular function.

# Submodularity

Submodular functions exhibit the "diminishing returns" property.

"For a submodular function, the incremental gain from adding an extra element in the set decreases with the size of the set".

Examples of submodular functions:

i. Column rank of a matrix
ii. Cardinality of a set
iii. Joint entropy of a set of random variables
iv. Capacity of a MIMO channel w.r.t. the set of active transmitter antennas [Vaze and Ganapathy, 12]

Question: What makes submodular functions interesting ?

# Optimizing submodular functions

[Nemhauser and Wolsey, 1978, An analysis of approximations for maximizing submodular set functions]

For a non negative, monotone submodular set function $f : 2^V \to \mathbb{R}^+$, let $S \subseteq V$ be a subset of size $k$ obtained by selecting elements one at a time, each time choosing an an element that provides the largest marginal increase in the functional value.

Let $S^*$ be a set that maximizes the value of $f$ over all $k$-sized subsets of $V$.

Then, $f(S) \geq (1 - \frac{1}{e})f(S^*)$.

In other words, $S$ provides a $(1 - \frac{1}{e})$ approximation of $f(S^*)$.

# Submodularity of $\log|.|$

For any positive definite matrix $\mathbf{A}$, a generic $n \times p$ matrix $\mathbf{B}$ and constant $\beta > 0$, the set function $f(S) = \log|\mathbf{A} + \beta\mathbf{B}_S\mathbf{B}_S^T|$ is submodular.

Proof:

i $f(S) \geq 0$ for all for $S \subseteq [n]$.

$$\begin{aligned} f(S) = \log|\mathbf{A} + \beta\mathbf{B}_S\mathbf{B}_S^T| &= \log|\mathbf{A}| + \log|\mathbf{I} + \beta\mathbf{A}^{-1}\mathbf{B}_S\mathbf{B}_S^T| \\ &= \log|\mathbf{A}| + \log|\mathbf{I} + \beta\mathbf{B}_S^T\mathbf{A}^{-1}\mathbf{B}_S| \end{aligned}$$

The rest follows from positive definiteness of $\mathbf{A}$ and $\mathbf{B}_S^T\mathbf{A}^{-1}\mathbf{B}_S$.

ii $f$ is monotone. Let $S \subset T \subseteq [n]$.

$$\begin{aligned} f(T) - f(S) &= \log|\mathbf{A} + \beta\mathbf{B}_T\mathbf{B}_T^T| - \log|\mathbf{A} + \beta\mathbf{B}_S\mathbf{B}_S^T| \\ &= \log|\mathbf{A} + \beta\mathbf{B}_S\mathbf{B}_S^T + \beta\mathbf{B}_{T\setminus S}\mathbf{B}_{T\setminus S}^T| - \log|\mathbf{A} + \beta\mathbf{B}_S\mathbf{B}_S^T| \\ &= \log\left|\mathbf{I} + \beta\left(\mathbf{A} + \beta\mathbf{B}_S\mathbf{B}_S^T\right)^{-1}\mathbf{B}_{T\setminus S}\mathbf{B}_{T\setminus S}^T\right| \\ &= \log\left|\mathbf{I} + \beta\underbrace{\mathbf{B}_{T\setminus S}^T\left(\mathbf{A} + \beta\mathbf{B}_S\mathbf{B}_S^T\right)^{-1}\mathbf{B}_{T\setminus S}}_{\text{positive definite}}\right| \geq 0. \end{aligned}$$

# Submodularity of $\log|.|$

For any positive definite matrix $\mathbf{A}$, a generic $n \times p$ matrix $\mathbf{B}$ and constant $\beta > 0$, the set function $f(S) = \log|\mathbf{A} + \beta \mathbf{B}_S \mathbf{B}_S^T|$ is submodular.

Proof:

- i $f(S) \geq 0$ for all for $S \subseteq [n]$.
- ii $f$ is monotone.
- iii $f$ satisfies "diminishing returns" property.

Let $S, T$ be arbitrary subsets of $[n]$ such that $S \subseteq T$. Let $a \in ([n] \backslash T)$

$$\begin{aligned}
f(S \cup \{a\}) - f(S) &= \log|\mathbf{A} + \beta \mathbf{B}_S \mathbf{B}_S^T + \beta \mathbf{b}_a \mathbf{b}_a^T| - \log|\mathbf{A} + \beta \mathbf{B}_S \mathbf{B}_S^T| \\
&= \log\left|\mathbf{I} + \beta \left(\mathbf{A} + \beta \mathbf{B}_S \mathbf{B}_S^T\right)^{-1} \mathbf{b}_a \mathbf{b}_a^T\right| \\
&= \log\left|1 + \beta \mathbf{b}_a^T \left(\mathbf{A} + \beta \mathbf{B}_S \mathbf{B}_S^T\right)^{-1} \mathbf{b}_a\right|
\end{aligned}$$

Likewise, we can show that

$$f(T \cup \{a\}) - f(T) = \log\left|1 + \beta \mathbf{b}_a^T \left(\mathbf{A} + \beta \mathbf{B}_T \mathbf{B}_T^T\right)^{-1} \mathbf{b}_a\right|$$

Further, using matrix inversion lemma,

$$\begin{aligned}
\mathbf{b}_a^T \left(\mathbf{A} + \beta \mathbf{B}_T \mathbf{B}_T^T\right)^{-1} \mathbf{b}_a &= \mathbf{b}_a^T \left(\mathbf{A} + \beta \mathbf{B}_S \mathbf{B}_S^T\right)^{-1} \mathbf{b}_a \\
-\mathbf{b}_a^T \left(\mathbf{A} + \beta \mathbf{B}_S \mathbf{B}_S^T\right)^{-1} \mathbf{B}_{T \backslash S} &\left(\frac{1}{\beta}\mathbf{I} + \mathbf{B}_{T \backslash S}^T \left(\mathbf{A} + \beta \mathbf{B}_S \mathbf{B}_S^T\right)^{-1} \mathbf{B}_{T \backslash S}\right)^{-1} \mathbf{B}_{T \backslash S}^T \left(\mathbf{A} + \beta \mathbf{B}_S \mathbf{B}_S^T\right)^{-1} \mathbf{b}
\end{aligned}$$

Rest follows from the monotonicity of $\log(1 + x)$.

# Proposed support recovery scheme

Recover support $\mathbf{s}$ by solving the below optimization:

$$\hat{\mathbf{s}} = \arg\max_{\mathbf{s}} \underbrace{\log\left|\alpha\mathbf{R}_{\mathbf{Y}} + (1-\alpha)\left(\sigma_n^2\mathbf{I} + \sigma_s^2\mathbf{\Phi}_{\mathbf{s}}\mathbf{\Phi}_{\mathbf{s}}^T\right)\right|}_{\text{submodular in } \mathbf{s}} - (1-\alpha)\underbrace{\log\left|\sigma_n^2\mathbf{I} + \sigma_s^2\mathbf{\Phi}_{\mathbf{s}}\mathbf{\Phi}_{\mathbf{s}}^T\right|}_{\text{submodular in } \mathbf{s}}.$$

The objective is a difference of two submodular functions in $\mathbf{s}$.

Can we minimize the difference of two submodular functions in a computationally efficient manner?

Supermodular-submodular (SupSub) procedure[1]

[1] Rishabh Iyer and Jeff Bilmes, Algorithms for approximate minimization of difference between submodular functions, with applications.

# Sub-Sup algorithm

Let $V$ be the base set. Let $f : 2^V \to \mathbb{R}$ and $g : 2^V \to \mathbb{R}$ be two submodular functions. Then, we want to solve:

$$\min_{X \subseteq V} f(X) - g(X).$$

Sub-Sup procedure: (a majorization-minimization approach)

i Construct a tight modular lower bound $h(.)$ for $g(.)$ such that $h(X_t) = g(X_t)$ and $h(X) \leq g(X)$ for $X \neq X_t$.

ii Minimize the submodular upper bound for $f(X) - g(X)$, i.e. $X_{t+1} = \arg\min_{X \subseteq V} f(X) - h(X)$.

iii Repeat steps (i) and (ii) until convergence (i.e., $X_{t+1} = X_t$).

The Sub-Sup procedure monotonically reduces the objective in each iteration.

$$f(X_t) - g(X_t) = f(X_t) - h(X_t) \geq f(X_{t+1}) - h(X_{t+1}) \geq f(X_{t+1}) - g(X_{t+1}).$$

# Modular lower bound for a submodular function

[Narasimhan & Bilmes, '12][2]

A tight modular lower bound h(.) for the submodular $g(.)$:

Suppose that $g : 2^V \to \mathbb{R}$ is a submodular function.

Let $\pi$ be any permutation of the set $V$.

Let $W_i = \{\pi(1), \pi(2), \ldots, \pi(i)\}$, so that $W_{|V|} = V$.

We define a function $h : V \to \mathbb{R}$ as follows:

$$h(\pi(i)) = \begin{cases} g(W_1) & if \quad i = 1 \\ g(W_i) - g(W_{i-1}) & otherwise \end{cases}$$

Extend elementwise function $h$ to all subsets of $V$ by defining

$$h(A) = \sum_{x \in A} h(x) \qquad \text{for every } A \subseteq V.$$

Then,

1. $h(A) \le g(A)$ for every $A \subseteq V$.
2. $h(W_m) = g(W_m)$ for every $1 \le m \le |V|$.

[2] Mukund Narasimhan and Jeff Bilmes, A submodular-supermodular procedure with applications to discriminative structure learning.