Sparse Support Recovery via Covariance Estimation

Lekshmi Ramesh

(Joint work with Prof. Chandra R. Murthy)

Indian Institute of Science
Bangalore

July 13, 2018

Outline

- Setup
  - Multiple measurement vector setting
  - Support recovery problem
- Support recovery as covariance estimation
  - Covariance matching, Gaussian approximation
  - Maximum likelihood-based estimation
  - Solution using non negative quadratic programming
  - Simulation results
- Remarks on non negative sparse recovery
- Conclusion

Problem setup

- Multiple measurement vector (MMV) model:
  Observations $\{\mathbf{y}_i\}_{i=1}^L$ are generated from the following linear model:
  $$\mathbf{y}_i = \Phi\mathbf{x}_i + \mathbf{w}_i, \quad i \in [L],$$
  where $\Phi \in \mathbb{R}^{m\times N}$ $(m < N)$, $\mathbf{x}_i \in \mathbb{R}^N$ unknown, random and noise $\mathbf{w}_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2 I)$

- $\mathbf{x}_i$ are k-sparse with common support
  $\mathrm{supp}(\mathbf{x}_i) = T$ for some $T \subset [N]$ with $|T| \leq k$, $\forall i \in [L]$

- Goal: Recover the common support T given $\{\mathbf{y}_i\}_{i=1}^L$, Φ

- Applications in hyperspectral imaging, sensor networks

Problem setup

- Generative model for $\mathbf{x}_i$
  Assumption: Non-zero entries uncorrelated

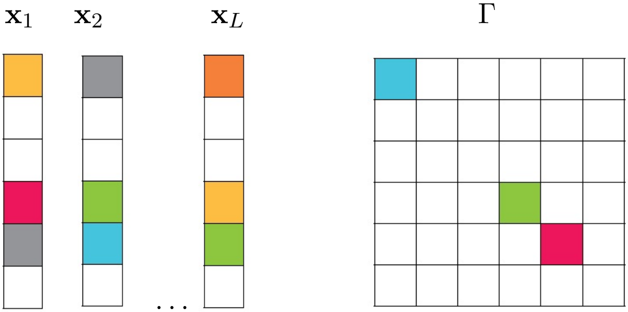
$$p(\mathbf{x}_i; \boldsymbol{\gamma}) = \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi\gamma_j}} \exp\left(-\frac{\mathbf{x}_{ij}^2}{2\gamma_j}\right)$$

$$\text{i.e.,} \quad \mathbf{x}_i \overset{iid}{\sim} \mathcal{N}(0, \Gamma) \text{ where } \Gamma = \mathrm{diag}(\boldsymbol{\gamma})$$

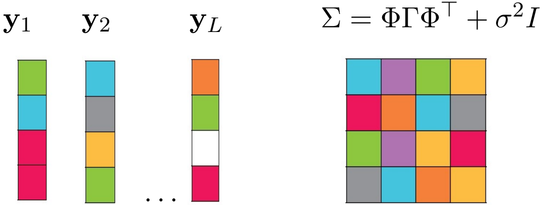
- Note:
  - $\mathrm{supp}(\mathbf{x}_i) = \mathrm{supp}(\boldsymbol{\gamma}) = T \quad (\text{since } \gamma_j = 0 \Leftrightarrow x_{ij} = 0 \ \text{ a.s.})$
  - $\mathbf{y}_i \sim \mathcal{N}(0, \underbrace{\Phi\Gamma\Phi^\top + \sigma^2 I}_{\Sigma \in \mathbb{R}^{m \times m}})$

- Equivalent problem: Recover $\mathrm{supp}(\boldsymbol{\gamma})$ given $\{\mathbf{y}_i\}_{i=1}^{L}$, Φ

- $\mathbf{x}_i \overset{iid}{\sim} \mathcal{N}(0, \Gamma)$



- $\mathbf{y}_i \overset{iid}{\sim} \mathcal{N}(0, \Sigma)$



Support recovery as covariance estimation

- Use the sample covariance matrix $\hat{\Sigma} = \frac{1}{L}\sum_{i=1}^{L} \mathbf{y}_i \mathbf{y}_i^\top$ to estimate Γ
- Express $\hat{\Sigma}$ as

$$\hat{\Sigma} = \Sigma + E,$$

where E: Noise/Error matrix

For the noiseless case $(\sigma^2 = 0)$[1]

$$\hat{\Sigma} = \Phi\Gamma\Phi^\top + E$$

$$\Big\downarrow \textit{vectorize}$$

$$\mathbf{r} = \underbrace{(\Phi \odot \Phi)}_{A \in \mathbb{R}^{m^2 \times N}} \boldsymbol{\gamma} + \mathbf{e}$$

where $\odot$ denotes the Khatri-Rao product

- Use Gaussian approximation for $\mathbf{e}$
- Find the maximum likelihood estimate of $\boldsymbol{\gamma}$

[1] details for noisy case can be found in the paper

Noise statistics

- Mean

$$\mathbb{E}(E) = \frac{1}{L}\sum_{i=1}^{L}\mathbb{E}\mathbf{y}_i\mathbf{y}_i^\top - \Sigma = 0$$

- Covariance

$$\mathrm{cov}(\mathrm{vec}(E)) = \frac{1}{L}(\Phi\otimes\Phi)(\Gamma^{\frac{1}{2}}\otimes\Gamma^{\frac{1}{2}})\underbrace{\mathrm{cov}(\mathrm{vec}(\mathbf{z}\mathbf{z}^\top))}_{B\in\mathbb{R}^{N^2\times N^2}}(\Gamma^{\frac{1}{2}}\otimes\Gamma^{\frac{1}{2}})(\Phi\otimes\Phi)^\top,$$

where $\mathbf{z}\sim\mathcal{N}(0, I_N)$

Example: N=3

- Let $\mathbf{z} = [z_1, z_2, z_3]^\top$ with $z_i \overset{iid}{\sim} \mathcal{N}(0,1)$. Then,

$$\mathbf{z}\mathbf{z}^\top = \begin{bmatrix} z_1^2 & z_1 z_2 & z_1 z_3 \\ z_1 z_2 & z_2^2 & z_2 z_3 \\ z_1 z_3 & z_2 z_3 & z_3^2 \end{bmatrix} \xrightarrow{vectorize} \begin{bmatrix} z_1^2 \\ z_1 z_2 \\ z_1 z_3 \\ z_1 z_2 \\ z_2^2 \\ z_2 z_3 \\ z_1 z_3 \\ z_2 z_3 \\ z_3^2 \end{bmatrix}$$

Example: N=3

- The covariance matrix B of $\text{vec}(\mathbf{z}\mathbf{z}^\top)$ will be of size 9×9 with $B_{i,j} \in \{0, 1, 2\},\ 1 \leq i, j \leq 3$.

- For e.g.,

$$\begin{aligned}
B_{1,1} &= \text{cov}(z_1^2, z_1^2) = \mathbb{E}z_1^4 - (\mathbb{E}z_1^2)^2 = 3 - 1 = 2 \\
B_{1,2} &= \text{cov}(z_1^2, z_1 z_2) = \mathbb{E}z_1^3 z_2 - \mathbb{E}z_1^2 \mathbb{E}z_1 z_2 = 0 \\
B_{2,4} &= \text{cov}(z_1 z_2, z_1 z_2) = \mathbb{E}z_1^2 z_2^2 - \mathbb{E}z_1 z_2 \mathbb{E}z_1 z_2 = 1
\end{aligned}$$

Example: N=3

$$B = \mathrm{cov}(\mathrm{vec}(\mathbf{z}\mathbf{z}^\top)) = \begin{bmatrix} 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \end{bmatrix}$$

- We now have the following model

$$\mathbf{r} = A\boldsymbol{\gamma} + \mathbf{e}, \tag{1}$$

  where

$$\begin{aligned} A &= (\Phi \odot \Phi), \\ \mathbb{E}[\mathbf{e}] &= 0, \\ \operatorname{cov}(\mathbf{e}) = W &= \frac{1}{L}(\Phi \otimes \Phi)(\Gamma^{\frac{1}{2}} \otimes \Gamma^{\frac{1}{2}})B(\Gamma^{\frac{1}{2}} \otimes \Gamma^{\frac{1}{2}})(\Phi \otimes \Phi)^\top. \end{aligned}$$

- Remarks
  - The noise term vanishes as $L \to \infty$
  - The noise covariance depends on the parameter to be estimated
  - $\mathbf{r}$, $\Phi \odot \Phi$ and $\mathbf{e}$ have redundant entries – restrict to the $\frac{m(m+1)}{2}$ distinct entries

New model, Gaussian approximation

- Pre-multiply (1) by $P \in \mathbb{R}^{\frac{m(m+1)}{2} \times m^2}$, formed using a subset of the rows of I_{m^2}, that picks the relevant entries. Thus,

$$\mathbf{r}_P = A_P \boldsymbol{\gamma} + \mathbf{e}_P,$$

  where $\mathbf{r}_P := P\mathbf{r}$, $A_P := PA$, and $\mathbf{e}_P := P\mathbf{e}$.

- Further, we approximate the distribution of n_P by $\mathcal{N}(0, W_P)$, where $W_P = PWP^\top$

- Thus, $\mathbf{r}_P \sim \mathcal{N}(A_P\boldsymbol{\gamma}, W_P)$

ML estimation of $\boldsymbol{\gamma}$

- Denote the ML estimate of $\boldsymbol{\gamma}$ by $\boldsymbol{\gamma}_{\text{ML}}$

$$\boldsymbol{\gamma}_{\text{ML}} = \arg\max_{\boldsymbol{\gamma}\geq 0}\; p(\mathbf{r}_P;\boldsymbol{\gamma}), \tag{2}$$

where

$$p(\mathbf{r}_P;\boldsymbol{\gamma}) = \frac{1}{(2\pi)^{\frac{m(m+1)}{4}}|W_P|^{\frac{1}{2}}}\exp\left(\frac{-(\mathbf{r}_P - A_P\boldsymbol{\gamma})^\top W_P^{-1}(\mathbf{r}_P - A_p\boldsymbol{\gamma})}{2}\right).$$

- Simplifying (2), we get

$$\boldsymbol{\gamma}_{\text{ML}} = \arg\min_{\boldsymbol{\gamma}\geq 0}\;\log|W_P| + (\mathbf{r}_P - A_P\boldsymbol{\gamma})^\top W_P^{-1}(\mathbf{r}_P - A_p\boldsymbol{\gamma}). \tag{3}$$

- For a fixed W_P, (3) can be solved using Non Negative Quadratic Programming (NNQP)

NNQP-based algorithm

Algorithm 1 MRNNQP

Input: Measurement matrix Φ, vectorized sample covariance $\mathbf{r}$, initial value $\Gamma^{(0)} = \mathrm{diag}(\boldsymbol{\gamma}^{(0)})$, $i = 1$
While (not converged) **do**
  $W_P^{(i)} \leftarrow \frac{1}{L} P(\Phi \otimes \Phi) B(\Gamma^{(i-1)} \otimes \Gamma^{(i-1)})(\Phi \otimes \Phi)^\top P^\top$
  $\mathbf{b}^{(i)} \leftarrow -A_P^\top {W_P^{(i)}}^{-1} \mathbf{r}_P$
  $Q^{(i)} \leftarrow A_P^\top {W_P^{(i)}}^{-1} A_P$
  $\gamma^{(i)} \leftarrow \mathrm{NNQP}(Q^{(i)}, \mathbf{b}^{(i)})$
  $\Gamma^{(i)} \leftarrow \mathrm{diag}(\boldsymbol{\gamma}^{(i)})$
  $i \leftarrow i + 1$
end While
Output: support of $\boldsymbol{\gamma}^{(i)}$

The MSBL algorithm

- $X = [\mathbf{x}_1, \cdots, \mathbf{x}_L]$, $Y = [\mathbf{y}_1, \cdots, \mathbf{y}_L]$
- Posterior moments
  $R = \mathrm{cov}(\mathbf{x}_i | \mathbf{y}_i; \boldsymbol{\gamma})$; $M = [\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_L]$

Algorithm 2 MSBL[2]

1: Input: Measurement matrix Φ, observations Y, initial value $\Gamma^{(0)} = \mathrm{diag}(\boldsymbol{\gamma}^{(0)})$, $i = 1$
2: **While** (not converged) **do**
3: $\quad R^{(i)} \leftarrow \Gamma^{i-1} - \Gamma^{(i-1)}\Phi^\top(\Sigma^{(i-1)})^{-1}\Phi\Gamma^{(i-1)}$
4: $\quad M^{(i)} \leftarrow \Gamma^{(i-1)}\Phi^\top(\Sigma^{(i-1)})^{-1}Y$
5: $\quad \gamma_j^{(i)} \leftarrow \frac{1}{L}\|\boldsymbol{\mu}_j^{(i)}\|_2^2 + R_{jj}^{(i)}$
6: $\quad i \leftarrow i + 1$
7: **end While**
8: Output: $\hat{\mathbf{x}}_j = \boldsymbol{\mu}_j^{(i)}$

[2] David P. Wipf and Bhaskar D. Rao. "An Empirical Bayesian Strategy for Solving the Simultaneous Sparse Approximation Problem". In: *TSP* 55.7-2 (2007).

Support recovery performance

$N = 40, m = 20, k = 25$; exact recovery over 200 trials

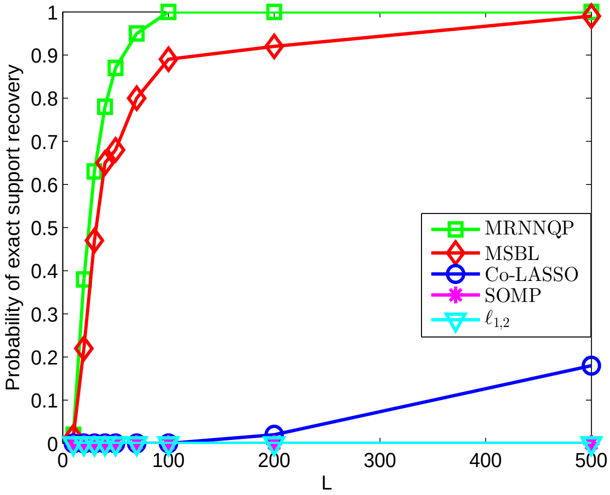


Figure 1 : Support recovery performance of the NNQP-based approach

Support recovery performance

$N = 70, m = 20, L = 50, 1000$; exact recovery over 200 trials

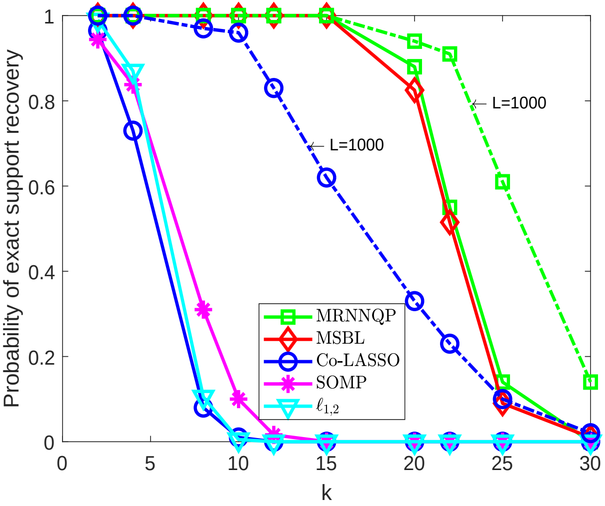


Figure 2 : Support recovery performance of the NNQP-based approach

Phase transition

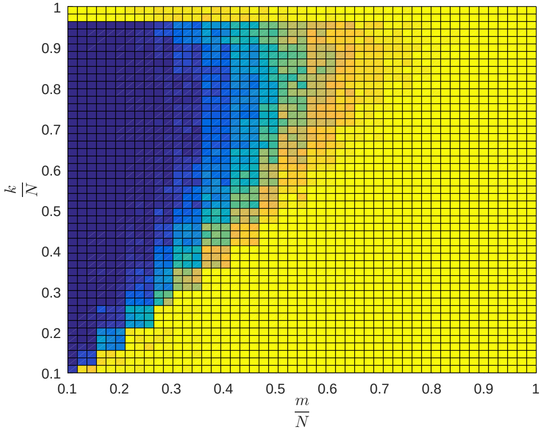


Figure 3 : Phase transition. $N = 20, L = 200$

Observations

- Exact support recovery possible for $k < m$ regime with "small" L
  For $k \geq m$, recovery possible with "large" L

- Dependence of computational complexity on parameters
  - L: in computing $\hat{\Sigma}$ (offline)
  - m, N: scales as $m^4 N^2$

- Comparison with Co-LASSO, MSBL
  - Improvement in performance by accounting for error due to $\hat{\Sigma}$
  - Only a one time computation of $\hat{\Sigma}$ is required whereas MSBL uses the entire set of measurements $\{\mathbf{y}_i\}_{i=1}^{L}$ in every iteration of EM

Remarks on non negative sparse recovery

- Inner loop in the ML estimation problem

$$\arg\min_{\gamma \geq 0} \; (\mathbf{r}_P - A_P \boldsymbol{\gamma})^\top W_P^{-1} (\mathbf{r}_P - A_p \boldsymbol{\gamma})$$

  Note: no sparsity-inducing regularizer

- Implicit regularization property of NNQP has been noted before[3,4]

- For successful recovery, require conditions on sign pattern of vectors in null space of A

[3] Martin Slawski and Matthias Hein. "Sparse Recovery by Thresholded Nonnegative Least Squares". In: *Advances in Neural Information Processing Systems.* 2011.

[4] Simon Foucart and David Koslicki. "Sparse Recovery by means of Nonnegative Least Squares". In: *IEEE Signal Proc. Letters* 21 (2014), pp. 498–502.

Concluding remarks

- Sparse support recovery can be done using maximum likelihood-based covariance estimation
- Support recovery possible even when $k > m$
- No explicit sparsity promoting regularizer needed
- Recovery guarantees depend on properties of null space of $\Phi \odot \Phi$

Thank you

Non-negative quadratic program[5]

$$\underset{\boldsymbol{\gamma}\geq 0}{\text{minimize}}\ \ (\mathbf{r}_P - A_P\boldsymbol{\gamma})^\top W_P^{-1}(\mathbf{r}_P - A_p\boldsymbol{\gamma})$$

Solution (entry-wise update equation for $\boldsymbol{\gamma}$):

$$\gamma_j^{(i+1)} = \gamma_j^{(i)} \left(\frac{-b_j + \sqrt{b_j^2 + 4(Q^+\boldsymbol{\gamma}^{(i)})_j (Q^-\boldsymbol{\gamma}^{(i)})_j}}{2(Q^+\boldsymbol{\gamma}^{(i)})_j} \right),$$

where $\mathbf{b} = -A_P^\top W_P^{-1}\mathbf{r}_P$, $Q = A_P^\top W_P^{-1} A_P$,

$$Q_{ij}^+ = \begin{cases} Q_{ij}, & \text{if} \quad Q_{ij} > 0, \\ 0, & \text{otherwise}, \end{cases} \qquad Q_{ij}^- = \begin{cases} -Q_{ij}, & \text{if} \quad Q_{ij} < 0, \\ 0, & \text{otherwise}. \end{cases}$$

[5] Fei Sha, Lawrence K. Saul, and Daniel D. Lee. "Multiplicative Updates for Nonnegative Quadratic Programming in Support Vector Machines". In: *Advances in Neural Information Processing Systems*. 2002, pp. 1041–1048.

Noise statistics

- Covariance

$$
\begin{aligned}
\mathrm{cov}(E) &= \mathrm{cov}\left(\sum_{i=1}^{L}\left(\frac{\mathbf{y}_i\mathbf{y}_i^\top}{L} - \frac{\Sigma}{L}\right)\right) \\
&= L\mathrm{cov}\left(\frac{\mathbf{y}_1\mathbf{y}_1^\top}{L} - \frac{\Sigma}{L}\right) \qquad \text{(sum of } L \text{ indep. random matrices)} \\
&= \frac{1}{L}\mathrm{cov}(\mathbf{y}_1\mathbf{y}_1^\top - \Sigma) \\
&= \frac{1}{L}\mathrm{cov}(\mathbf{y}\mathbf{y}^\top)
\end{aligned}
$$

- Represent $\mathbf{y}$ as

$$\mathbf{y} = C\mathbf{z},$$

where $\mathbf{z} \sim \mathcal{N}(0, I)$ and $\Sigma = CC^\top$

Noise statistics

$$\mathrm{cov}(E) = \frac{1}{L}\mathrm{cov}(\mathbf{y}\mathbf{y}^\top)$$

- For $\sigma^2 = 0$, $\Sigma = \Phi\Gamma\Phi^\top$; can take $C = \Phi\Gamma^{\frac{1}{2}}$
- Using properties of Kronecker products:

$$\begin{aligned}
\mathrm{cov}(\mathrm{vec}(E)) &= \frac{1}{L}\mathrm{cov}(\mathrm{vec}(C\mathbf{z}\mathbf{z}^\top C^\top)) \\
&= \frac{1}{L}\mathrm{cov}((C \otimes C)\mathrm{vec}(\mathbf{z}\mathbf{z}^\top)) \\
&= \frac{1}{L}(C \otimes C)\mathrm{cov}(\mathrm{vec}(\mathbf{z}\mathbf{z}^\top))(C \otimes C)^\top \\
&= \frac{1}{L}(\Phi \otimes \Phi)(\Gamma^{\frac{1}{2}} \otimes \Gamma^{\frac{1}{2}})\underbrace{\mathrm{cov}(\mathrm{vec}(\mathbf{z}\mathbf{z}^\top))}_{B \in \mathbb{R}^{N^2 \times N^2}}(\Gamma^{\frac{1}{2}} \otimes \Gamma^{\frac{1}{2}})(\Phi \otimes \Phi)^\top
\end{aligned}$$

- Last step: use $(A \otimes B)(C \otimes D) = AB \otimes CD$