Sparse Support Recovery via Covariance Estimation

Lekshmi Ramesh

Indian Institute of Science
Bangalore

June 30, 2017

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
- Non negative sparse recovery problem, Guarantees
- Conclusions, Future work

Problem setup

- Multiple measurement vector 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 $\mathbf{w}_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2 I)$

- Assumptions:
  - $\mathbf{x}_i$ are k-sparse with common support
    $\mathrm{supp}(\mathbf{x}_i) = T$ for some $T \subset [N]$ with $|T| \leq k$, $\forall i \in [L]$
  - Non-zero entries uncorrelated
    $\mathbb{E}[\mathbf{x}_{t,i}\mathbf{x}_{t,j}] = 0$, $t \in [L]$, $i, j \in T$

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

Problem setup

- We impose the following prior on $\mathbf{x}_i$

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

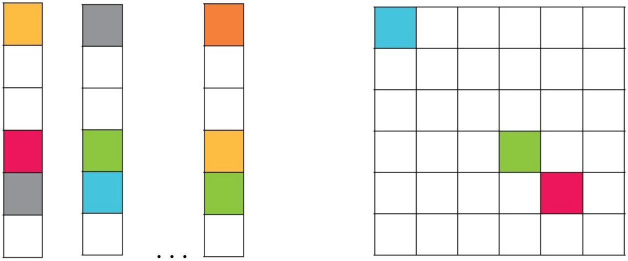
$$\text{i.e.,} \quad \mathbf{x}_i \overset{iid}{\sim} \mathcal{N}(0, \Gamma) \text{ where } \Gamma = \operatorname{diag}(\boldsymbol{\gamma})$$

- Note:
  - $\operatorname{supp}(\mathbf{x}_i) = \operatorname{supp}(\boldsymbol{\gamma}) = T$ (since $\gamma_j = 0 \Leftrightarrow x_{ij} = 0$ 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 Γ from (an estimate of) Σ

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

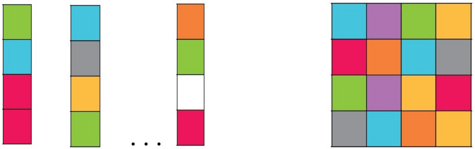
$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_L$ Γ



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

$\mathbf{y}_1$ $\mathbf{y}_2$ $\mathbf{y}_L$

$$\Sigma = \Phi\Gamma\Phi^\top + \sigma^2 I$$



Support recovery as covariance estimation

- We work with the sample covariance matrix $\hat{\Sigma} = \frac{1}{L}\sum_{i=1}^{L} \mathbf{y}_i \mathbf{y}_i^\top$

- Express $\hat{\Sigma}$ as

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

  where E: Noise/Error matrix

- Noiseless case ($\sigma^2 = 0$)

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

$$\Big\downarrow \text{vectorize}$$

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

  where $\odot$ denotes the Khatri-Rao product

- We will find the maximum likelihood estimate of $\boldsymbol{\gamma}$
  For that, we first derive the noise statistics

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

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

Noise statistics

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

- Represent $\mathbf{y}$ as

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

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

- For $\sigma^2 = 0$, $\Sigma = \Phi\Gamma\Phi^\top$; can take $C = \Phi\Gamma^{\frac{1}{2}}$

- Using properties of Kronecker products:

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

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 $B_{i,j} \in \{0, 1, 2\}, \; 1 \le i, j \le 3$.

- For e.g.,

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

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

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

Observations

- 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,

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

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

- 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 $\boldsymbol{\gamma}_{\mathrm{ML}}$

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

where

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

ML estimation of $\boldsymbol{\gamma}$

- Simplifying (2), we get

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

- To solve (3)
  - Initialize $\boldsymbol{\gamma}$, compute W_P
  - Solve (for fixed W_P)

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

  - Recompute W_P and iterate

Non-negative quadratic program

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

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

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

Support recovery performance

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

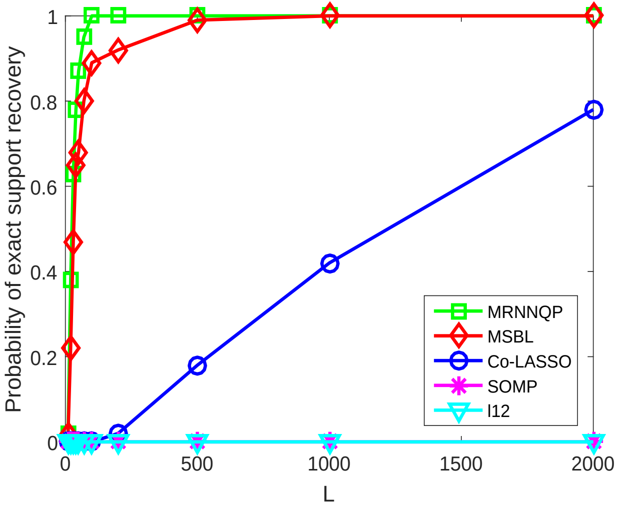


Figure 1: Support recovery performance of the NNQP-based approach

Support recovery performance

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

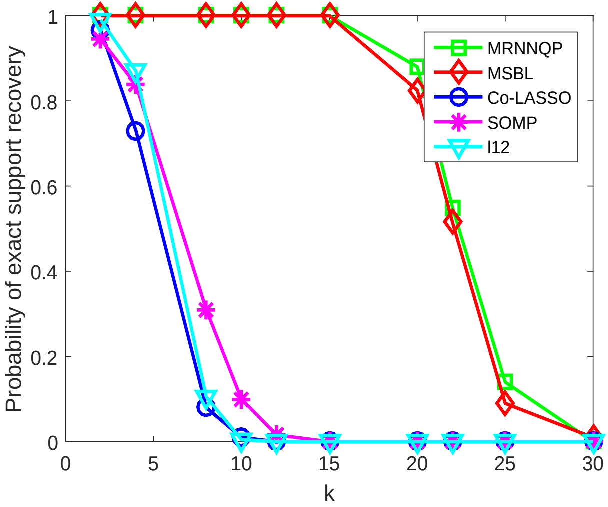


Figure 2: Support recovery performance of the NNQP-based approach

Observations

- Exact support recovery possible for $k < m$ regime with 'small' L
- For $m \leq k \leq \alpha m$ for some $1 \leq \alpha < \frac{N}{m}$, recovery possible with 'large' L
- Dependence of computational complexity on parameters
  - L: in computing $\hat{\Sigma}$ (offline)
  - m, N: scales as $m^4 N^2$

Non negative least squares (NNLS)

- 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

- Canonical NNLS problem

$$\arg\min_{\mathbf{x}\geq 0} \|\mathbf{y} - \Phi\mathbf{x}\|_2^2 \qquad \text{(NNLS)}$$

  Question: When does (NNLS) return a sparse solution?

Non negative sparse recovery

- Canonical problem

$$\begin{aligned} &\arg\min_{\mathbf{x}} \|\mathbf{x}\|_0 \\ &\text{s.t.} \quad \Phi\mathbf{x} = \mathbf{y}, \quad \mathbf{x} \geq 0, \end{aligned} \tag{P_0^+}$$

where $\|\mathbf{x}\|_0$: number of non-zero entries in $\mathbf{x}$

Question: Given $\mathbf{y} \in \mathbb{R}^m$ generated by $\mathbf{x}_0 \in \mathbb{R}^N$ that is non negative and k-sparse, when does (P_0^+) return $\mathbf{x}_0$?

Uniqueness condition–I

- Let $F := \{\mathbf{x} \in \mathbb{R}^N : \mathbf{x} \geq 0, \Phi\mathbf{x} = \mathbf{y}\}$ (feasible set for (P_0^+))
  $S_k := \{\mathbf{x} \in \mathbb{R}^N : \|\mathbf{x}\|_0 \leq k\}$
  If $F \cap S_k = \{\mathbf{x}_0\}$ then (P_0^+) returns $\mathbf{x}_0$.

Theorem

Let $\mathbf{x}_0 \in \mathbb{R}^N$ be a non negative k-sparse vector such that $\Phi\mathbf{x}_0 = \mathbf{y}$. Then $\mathbf{x}_0$ is the only k-sparse $\mathbf{x}$ satisfying $\mathbf{x} \geq 0$ and $\Phi\mathbf{x} = \mathbf{y}$ if and only if every $\mathbf{v} \in \ker(\Phi)\backslash\{0\}$ has at least $(k+1)$ positive *or* $(k+1)$ negative entries.

- Sufficient to guarantee that (P_0^+) returns the true solution

Uniqueness condition–I

- Proof

  (Sufficiency) Suppose that there exists $\mathbf{x}' \neq \mathbf{x}_0$ such that $\mathbf{x}' \geq 0$, $\|\mathbf{x}'\|_0 \leq k$ and $\Phi\mathbf{x}' = \mathbf{y}$.

  Then, $\Phi(\mathbf{x}' - \mathbf{x}_0) = 0$ which implies

  $$\mathbf{v} := \mathbf{x}' - \mathbf{x}_0 \in \ker(\Phi)\backslash\{0\}.$$

  Since both $\mathbf{x}_0$ and $\mathbf{x}'$ are non-negative and k-sparse, $\mathbf{v}$ has at most k positive and at most k negative entries, violating the sign-pattern condition.

- Proof (contd.)

  (Necessity) Assume that the sign-pattern condition does not hold. That is, there exists $\mathbf{v} \in \ker(\Phi) \setminus \{0\}$ with at most k negative and k positive entries. We will show that we can find another non-negative k-sparse vector $\mathbf{x}'$ such that $\Phi\mathbf{x}' = \mathbf{y}$.

  Let $T := \{i \in [N] : \mathbf{v}_i < 0\}$. If $\mathbf{x}_0$ is of the form

$$(\mathbf{x}_0)_i = \begin{cases} -\mathbf{v}_i, & i \in T \\ 0, & \text{otherwise,} \end{cases}$$

  then $\mathbf{x}' = \mathbf{x}_0 + \mathbf{v}$ is a non-negative k-sparse vector satisfying $\Phi\mathbf{x}' = \Phi\mathbf{x}_0$.

  This contradicts the uniqueness of $\mathbf{x}_0$ as a non-negative k-sparse solution of $\Phi\mathbf{x} = \mathbf{y}$.

Uniqueness condition–II

- Let $F := \{\mathbf{x} \in \mathbb{R}^N : \mathbf{x} \geq 0, \Phi\mathbf{x} = \mathbf{y}\}$ (feasible set for (P_0^+))
  $S_k := \{\mathbf{x} \in \mathbb{R}^N : \|\mathbf{x}\|_0 \leq k\}$
  If $F = \{\mathbf{x}_0\}$ then (NNLS) returns $\mathbf{x}_0$.

Theorem

Let $\mathbf{x}_0 \in \mathbb{R}^N$ be a non negative k-sparse vector such that $\Phi\mathbf{x}_0 = \mathbf{y}$. Then $\mathbf{x}_0$ is the only $\mathbf{x}$ satisfying $\mathbf{x} \geq 0$ and $\Phi\mathbf{x} = \mathbf{y}$ if and only if every $\mathbf{v} \in \ker(\Phi)\backslash\{0\}$ has at least $(k+1)$ positive *and* $(k+1)$ negative entries.

- Sufficient to guarantee that (NNLS) returns the true solution (Any program of the form $\arg\min_{\mathbf{x}\geq 0} \|\mathbf{y} - \Phi\mathbf{x}\|_p$ with $p \geq 1$ will work)

Matrices satisfying uniqueness conditions

Every $\mathbf{v} \in \ker(\Phi) \backslash \{0\}$ has at least $(k+1)$ positive or/and $(k+1)$ negative entries

Question: Which matrices satisfy these sign pattern conditions?

- Consider Φ such that $\|\mathbf{v}\|_0 > 2k$, $\forall \mathbf{v} \in \ker(\Phi) \backslash \{0\}$
  Then, every $\mathbf{v} \in \ker(\Phi) \backslash \{0\}$ has at least $(k+1)$ positive or $(k+1)$ negative entries

- Every set of $2k$ columns of $\Phi \in \mathbb{R}^{m \times N}$ linearly independent
  Requires $m \geq 2k$
  Example: $\Phi_{ij} \overset{iid}{\sim} \mathcal{N}(0,1)$

Are there matrices satisfying the condition when $m < 2k$?

An equivalent characterization

- Define

  $U = \{\mathbf{x} \in \mathbb{R}^N : \mathbf{x} \text{ has at most } k \text{ positive \& at most } k \text{ negative entries}\}$

- Condition I: $\ker(\Phi) \cap U = \{0\}$
- Bad event: There exists $\mathbf{v} \in \ker(\Phi)$ such that

$$\sum_{i \in T_p} v_i \Phi_i + \sum_{j \in T_n} (-v_j)(-\Phi_j) = 0,$$

  where $T_p = \{i \in [N] : v_i > 0\}$, $T_n = \{i \in [N] : v_i < 0\}$, and $|T_p| \leq k$, $|T_n| \leq k$

An equivalent characterization

- Bad event: There exist indices $\{i_1, i_2, \ldots, i_{2k}\} \subset [N]$ such that

$$0 \in \text{conv}(\Phi_{i_1}, \ldots, \Phi_{i_k}, -\Phi_{i_{k+1}}, \ldots, -\Phi_{i_{2k}})$$

- Assume random Φ with entries drawn from some distribution P. For which P is the probability of the above event small?

Conclusions, Future work

Conclusions

- Sparse support recovery using maximum likelihood based covariance estimation, no regularization parameter needed
- Support recovery possible even when $k > m$
- Guarantees for non negative sparse recovery

Future work

- Characterization of the uniqueness conditions in terms of N, m, k
- Explore implications of uniqueness conditions for the covariance estimation problem

Thank you