Optimal Compressive Covariance Sketching via Rank-One Sampling

Wenbin Wang and **Ziping Zhao**
School of Information Science and Technology, ShanghaiTech University, Shanghai, China

Background and Motivation

- **High-dimensional Streaming Data:** Each time a snapshot of a data vector $\boldsymbol{x} \in \mathbb{R}^d$ is generated, with d large.
- **Challenges in Modern Data Acquisition:**
  - **Data generation at unprecedented rate:** data samples are 1) not observable due to privacy or security constraints; 2) distributed at multiple locations; 3) online generated on the fly and can only be accessed once.
  - **Limited processing power at sensor platforms:** 1) **time-sensitive:** impossible to obtain a complete snapshot of the system; 2) **storage-limited:** cannot store the whole data set; 3) **power-hungry:** minimize the number of observations.
- **Covariance Sketching:** Key Observation: The covariance structure can be recovered without measuring the whole data stream.
- **Contribution:** We propose a **nonconvex problem** with an **efficient algorithm**, which can be proved to attain the **oracle statistical rate**.

Sampling Model

- Consider n independent observations $\{\boldsymbol{x}_t\}_{t=1}^n$, each drawn from a zero-mean random vector $\boldsymbol{x}$ with covariance matrix $\boldsymbol{\Sigma}^*$. Given m sensing vectors $\{\boldsymbol{a}_i\}_{i=1}^m$, the quadratic measurement measurement y_i, $i = 1, \ldots, m$, is given by

$$\boldsymbol{y} = \boldsymbol{A}_{\otimes} \mathrm{vec}\,(\boldsymbol{S}) + \boldsymbol{\eta} = \mathcal{A}(\boldsymbol{S}) + \boldsymbol{\eta},$$

  where $\boldsymbol{y}\,[y_1, \cdots, y_m]^\top$, $\boldsymbol{\eta}\,[\eta_1, \cdots, \eta_m]^\top$ are additive measurement noises, $\boldsymbol{A}_{\otimes} = \left[(\boldsymbol{a}_1 \otimes \boldsymbol{a}_1) \cdots (\boldsymbol{a}_m \otimes \boldsymbol{a}_m)\right]^\top$ and $\mathrm{vec}\,(\boldsymbol{S})$ denotes the vectorization of $\boldsymbol{S}$ obtained by stacking its columns, and $\mathcal{A} : \mathbb{R}^{d\times d} \mapsto \mathbb{R}^m$ is a linear operator.
- Assumptions:
  - The sensing vectors $\{\boldsymbol{a}_i\}_{i=1}^m$ are i.i.d. sub-Gaussian random variables with zero mean and identity covariance.
  - The measurement noises $\{\eta_i\}_{i=1}^m$ are i.i.d. sub-exponential random variables with mean 0 and variance proxy σ^2.

Problem Formulation

- We propose to estimate the sparse covariance matrices from quadratic measurements using the non-convex penalty

$$\min_{\boldsymbol{\Sigma}\succ\boldsymbol{0}} \left\{ \frac{1}{2m}\|\boldsymbol{y} - \mathcal{A}(\boldsymbol{\Sigma})\|_{\mathrm{F}}^2 - \tau \log\det \boldsymbol{\Sigma} + \sum_{i,j} p_\lambda(|\Sigma_{ij}|) \right\}.$$

- Assumptions on the non-convex penalty function $p_\lambda(\cdot)$:

  (i) $p_\lambda(t)$ is non-decreasing on $[0, +\infty)$ with $p_\lambda(0) = 0$ and is differentiable on $(0, +\infty)$;

  (ii) $0 \le p'_\lambda(t_1) \le p'_\lambda(t_2) \le \lambda$ for all $t_1 \ge t_2 \ge 0$ and $\lim_{t\to 0} p'_\lambda(t) = \lambda$;

  (iii) There exists an $\alpha > 0$ such that $p'_\lambda(t) = 0$ for $t \ge \alpha\lambda$.

  Prototypical examples of the non-convex penalty function $p_\lambda(\cdot)$:

$$\text{SACD:}\quad p'_\lambda(t) = \begin{cases} \lambda, & \text{for } 0 < t \le \lambda, \\ \frac{b\lambda - t}{b-1}, & \text{for } \lambda \le t \le b\lambda, \\ 0, & \text{for } t \ge b\lambda, \end{cases}$$

  where $b > 2$ is an additional tuning parameter.

$$\text{MCP:}\quad p_\lambda(t) = \mathrm{sign}\,(t)\,\lambda \cdot \int_0^{|t|} \left(1 - \frac{z}{\lambda b}\right)_+ dz$$

  for some $b > 0$.

Algorithm

Algorithm 1: Majorization-Minimization Based Multistage Convex Relaxation

Initialize $\widetilde{\boldsymbol{\Sigma}}^{(0)} = \boldsymbol{I}$

for $k = 1, 2, \ldots, K$ **do**

update $\Lambda_{ij}^{(k-1)} = p'_\lambda(|\widetilde{\Sigma}_{ij}^{(k-1)}|)$;

obtain $\widetilde{\boldsymbol{\Sigma}}^{(k)}$ by solving

$$\min_{\boldsymbol{\Sigma}\succ\boldsymbol{0}} \left\{ \frac{1}{2m}\|\boldsymbol{y} - \mathcal{A}(\boldsymbol{\Sigma})\|_{\mathrm{F}}^2 - \tau \log\det \boldsymbol{\Sigma} + \sum_{i,j} p_\lambda(|\Sigma_{ij}|) \right\},$$

$k = k + 1$;

end for.

Algorithm 2: Proximal Newton Algorithm With Back-tracking Line Search

Input $\boldsymbol{\Sigma}^{k-1}$, $\boldsymbol{\Lambda}^k$, ε

Initialize $t = 0$, $\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}^{k-1}$, $\mu = 0.8$, $\alpha = 0.3$

Repeat

$\boldsymbol{\Sigma}_{t+\frac{1}{2}} \in \arg\min_{\boldsymbol{\Sigma}\succ\boldsymbol{0}} \widetilde{f}_t(\boldsymbol{\Sigma}) + \|\boldsymbol{\Lambda} \odot \boldsymbol{\Sigma}\|_1$;

$\boldsymbol{\Delta}_t = \boldsymbol{\Sigma}_{t+\frac{1}{2}} - \boldsymbol{\Sigma}_t$;

$\delta_t = \langle \nabla f(\boldsymbol{\Sigma}_t), \boldsymbol{\Delta}_t \rangle - \|\boldsymbol{\Lambda}^k \odot \boldsymbol{\Sigma}_t\|_1 + \|\boldsymbol{\Lambda}^k \odot (\boldsymbol{\Sigma}_t + \boldsymbol{\Delta}_t)\|_1$;

$\beta = 1$, $q = 0$;

Repeat

$\beta = \mu^q$, $q = q + 1$;

If $\boldsymbol{\Sigma}_t + \beta\boldsymbol{\Delta}_t \preceq \boldsymbol{0}$ **then**

continue;

end

until $\bar{F}(\boldsymbol{\Sigma}_t + \beta\boldsymbol{\Delta}_t) \le \bar{F}(\boldsymbol{\Sigma}_t) + \alpha\beta\delta_t$

$\boldsymbol{\Sigma}_{t+1} = \boldsymbol{\Sigma}_t + \beta\boldsymbol{\Delta}_t$;

$t = t + 1$;

until $\max_{i,j} \left| \left(\nabla f(\boldsymbol{\Sigma}_{t+1}) + \boldsymbol{\Lambda}^k \odot \boldsymbol{\Xi}^k\right)_{ij} \right| \le \varepsilon$

Output: $\boldsymbol{\Sigma}^K = \boldsymbol{\Sigma}_{t+1}$

Main Results

Essential Assumptions:

- The true covariance matrix $\boldsymbol{\Sigma}^\star$ satisfies

$$0 < \frac{1}{\kappa} \le \lambda_{\min}(\boldsymbol{\Sigma}^\star) \le \lambda_{\max}(\boldsymbol{\Sigma}^\star) \le \kappa < \infty,$$

  for some constant $\kappa \ge 1$.
- There exist universal constants α and μ such that

$$\|\boldsymbol{\Sigma}^\star_{\mathcal{S}}\|_{\min} = \min_{(i,j)\in\mathcal{S}} |\Sigma^\star_{ij}| \ge (\alpha + \mu)\,\lambda,$$

  where $\mu \in (0, \alpha)$ satisfies $p'_\lambda(\mu\lambda) \ge \frac{\lambda}{2}$.

Theorem 1: Let $\mathcal{S} = \{(i,j) \mid \Sigma^\star_{ij} \ne 0\}$ and $|\mathcal{S}| = s$.
Define $f(\boldsymbol{\Sigma}) = \frac{1}{2m}\|\boldsymbol{y} - \mathcal{A}(\boldsymbol{\Sigma})\|_{\mathrm{F}}^2 - \tau \log\det \boldsymbol{\Sigma}$.
Under some standard assumptions, the ε-optimal solution $\widetilde{\boldsymbol{\Sigma}}^{(k)}$ $(1 \le k \le K)$ satisfies the following contraction property:

$$\left\|\widetilde{\boldsymbol{\Sigma}}^{(k)} - \boldsymbol{\Sigma}^\star\right\|_{\mathrm{F}} \le \frac{1}{\rho}\left(\underbrace{\|(\nabla f(\boldsymbol{\Sigma}^\star))_{\mathcal{S}}\|_{\mathrm{F}}}_{\text{oracle rate}} + \underbrace{\varepsilon\sqrt{s}}_{\text{optimization error}}\right) + \underbrace{\delta\left\|\widetilde{\boldsymbol{\Sigma}}^{(k-1)} - \boldsymbol{\Sigma}^\star\right\|_{\mathrm{F}}}_{\text{contraction}},$$

where $\delta \in (0, 1)$ is the contraction parameter.

Theorem 2: Under some standard assumptions, let $\boldsymbol{x}$ be a sub-Gaussian random vector with mean zero and covariance $\boldsymbol{\Sigma}^\star$ and $\{\boldsymbol{x}_i\}_{i=1}^n$ be a collection of i.i.d. samples drawn from $\boldsymbol{x}$, if $\lambda \asymp \sqrt{\frac{\log d}{mn}}$, $\tau \lesssim \sqrt{\frac{1}{mn}}\left\|(\boldsymbol{\Sigma}^\star)^{-1}\right\|_{\max}^{-1}$, $\varepsilon \lesssim \sqrt{\frac{1}{mn}}$, and $K \gtrsim \log(\lambda\sqrt{mn}) \gtrsim \log\log d$, then the ε-optimal solution $\widetilde{\boldsymbol{\Sigma}}^{(K)}$ satisfies

$$\left\|\widetilde{\boldsymbol{\Sigma}}^{(K)} - \boldsymbol{\Sigma}^\star\right\|_F \lesssim \sqrt{\frac{s}{mn}}$$

with high probability.

Numerical Simulations

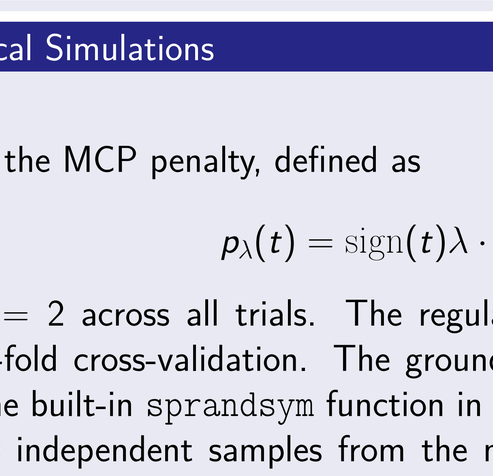
We use the MCP penalty, defined as

$$p_\lambda(t) = \mathrm{sign}(t)\lambda \cdot \int_0^{|t|} \left(1 - \frac{z}{\lambda b}\right)_+ dz,$$

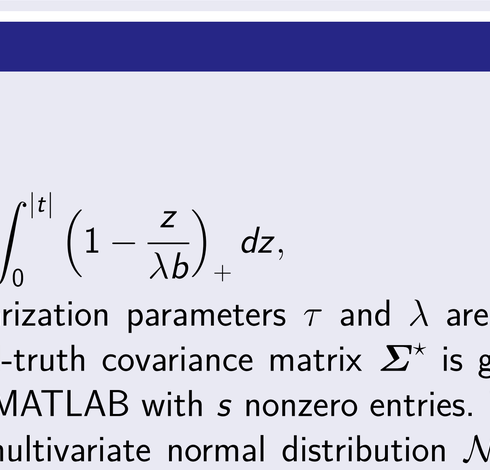
with $b = 2$ across all trials. The regularization parameters τ and λ are selected via five-fold cross-validation. The ground-truth covariance matrix $\boldsymbol{\Sigma}^\star$ is generated using the built-in `sprandsym` function in MATLAB with s nonzero entries. We draw $n = 50$ independent samples from the multivariate normal distribution $\mathcal{N}(0, \boldsymbol{\Sigma}^\star)$, and the noise variables η_i are sampled from a sub-exponential distribution with scale parameter γ, i.e., $\eta_i \sim \gamma \cdot \mathcal{N}(0, 1)$. To evaluate recovery performance, we measure the success probability as visualized in the color-coded matrix in Fig. (a). To reduce the impact of limited sample size, we directly apply sketching to the true covariance matrix $\boldsymbol{\Sigma}^\star$. A recovery is considered successful if the relative Frobenius error satisfies

$$\frac{\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^\star\|_{\mathrm{F}}}{\|\boldsymbol{\Sigma}^\star\|_{\mathrm{F}}} \le 10^{-3}.$$

Fig. (b) compares the proposed estimator with the ℓ_1-norm-based method under a consistent noise level $\gamma = 10^{-1}$. As the number of measurements increases, the recovery error decreases, and our method consistently outperforms the ℓ_1-based estimator.



(a) The rate of successful covariance reconstruction when $d = 100$.



(b) The FRE of the estimated covariance matrices for different sparsity levels with $\gamma = 10^{-1}$

Email: {wangwb2023@shanghaitech.edu.cn; zipingzhao@shanghaitech.edu.cn}