Achieving Oracle Rate for Large Covariance Matrix Estimation From Quadratic Measurements

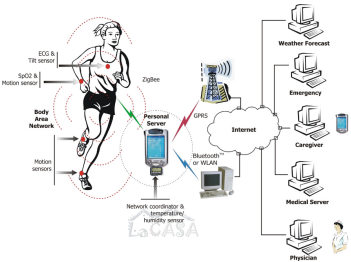
Wenbin Wang and Ziping Zhao

School of Information Science and Technology
ShanghaiTech University

High-dimensional Streaming Data

Each time a snapshot of a data vector $\boldsymbol{x} \in \mathbb{R}^d$ is generated, with d large.

(a) Internet monitoring[1]



(b) Wireless health monitoring[2]

(c) YouTube ranking[3]

[1] https://www.epitiro.com/

[2] Jovanov, Emil, et al. "A wireless body area network of intelligent motion sensors for computer assisted physical rehabilitation." Journal of NeuroEngineering and rehabilitation 2.1 (2005): 6.

[3] https://incartmarketing.com/youtube-ranking-how-to-get-more-views-on-youtube/

Challenges in Modern Data Acquisition

Data generation at unprecedented rate: data samples are

- not observable due to privacy or security constraints;
- distributed at multiple locations;
- online generated on the fly and can only be accessed once.

Limited processing power at sensor platforms:

- **time-sensitive:** impossible to obtain a complete snapshot of the system;
- **storage-limited:** cannot store the whole data set;
- **power-hungry:** minimize the number of observations.

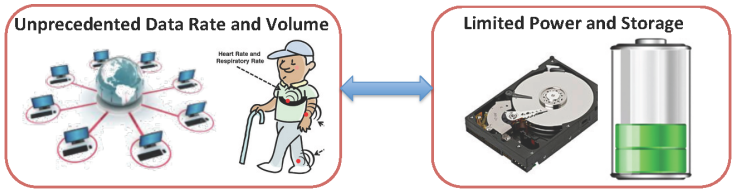


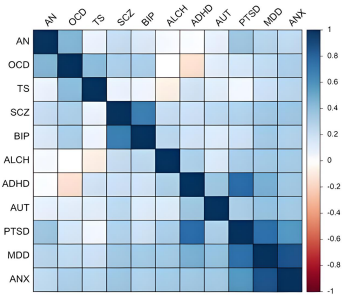
Figure 1: Mismatch streaming[4]

[4] https://users.ece.cmu.edu/ yuejiec/GeometricConstraints.html

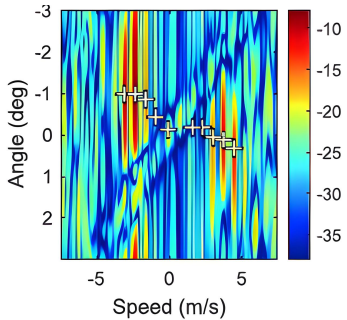
Covariance Sketching

Key Observation: The covariance structure can be recovered without measuring the whole data stream.

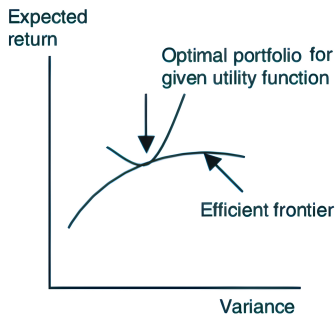
Applications of covariance sketching:



(a) Genetic Biology



(b) Radar System



(c) Portfolio Optimization

Quadratic Sketching for Covariance Estimation

Consider a data stream possibly distributedly observed at m sensors:

Quadratic Sketching: For each sensor $i = 1, \cdots, m$:

- randomly select a sketching vector $\boldsymbol{a}_i \in \mathbb{R}^d$ with i.i.d. sub-Gaussian entries;
- Sketching n independent observations $\{\boldsymbol{x}_t\}_{t=1}^n$ with an energy measurement $|\boldsymbol{a}^\top \boldsymbol{x}|^2$ and aggregate the average energy measurement[5]:

$$y_i = \frac{1}{n}\sum_{t=1}^{n} \left|\boldsymbol{a}_i^\top \boldsymbol{x}_t\right|^2 + \eta_i = \left\langle \boldsymbol{a}_i\boldsymbol{a}_i^\top, \boldsymbol{S}\right\rangle + \eta_i$$

[5]Since only finite samples are available, we have $\boldsymbol{S} = \boldsymbol{\Sigma}^\star + \boldsymbol{E}$ with $\boldsymbol{E}$ a bias term.

Related Work I

The *sparsity* assumption: a majority of the off-diagonal elements of $\boldsymbol{\Sigma}^*$ are zero or nearly so.

- Positive definite non-convex penalized covariance estimator (Quan 2023)

$$\widehat{\boldsymbol{\Sigma}} = \arg\min_{\boldsymbol{\Sigma} \succ \mathbf{0}} \left\{ \frac{1}{2} \|\boldsymbol{\Sigma} - \boldsymbol{S}\|_{\mathrm{F}}^2 - \tau \log\det \boldsymbol{\Sigma} + \sum_{i \neq j} p_\lambda(|\Sigma_{ij}|) \right\}$$

☺ is *always* positive definite,

☺ is *easy* to obtain by iterative algorithm due to its convexity,

☹ need the full data samples.

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 $\boldsymbol{y} \coloneqq [y_1, \cdots, y_m]^\top$, $\boldsymbol{\eta} \coloneqq [\eta_1, \cdots, \eta_m]^\top$ are additive measurement noises, $\boldsymbol{A}_{\otimes} = \left[\begin{array}{ccc} (\boldsymbol{a}_1 \otimes \boldsymbol{a}_1) & \cdots & (\boldsymbol{a}_m \otimes \boldsymbol{a}_m) \end{array}\right]^\top$ 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.

Additional 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

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

- 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 $p_\lambda(\cdot)$: SCAD and MCP

Challenges

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

- It is a non-convex problem.
- If we directly apply an iterative algorithm, e.g., coordinate descent (Rothman 2012),
  - the global optimum may not be obtainable,
  - the local optimums are in general hard to be characterized.

Question: how to develop numerical algorithms with provable statistical guarantees?

Optimization 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

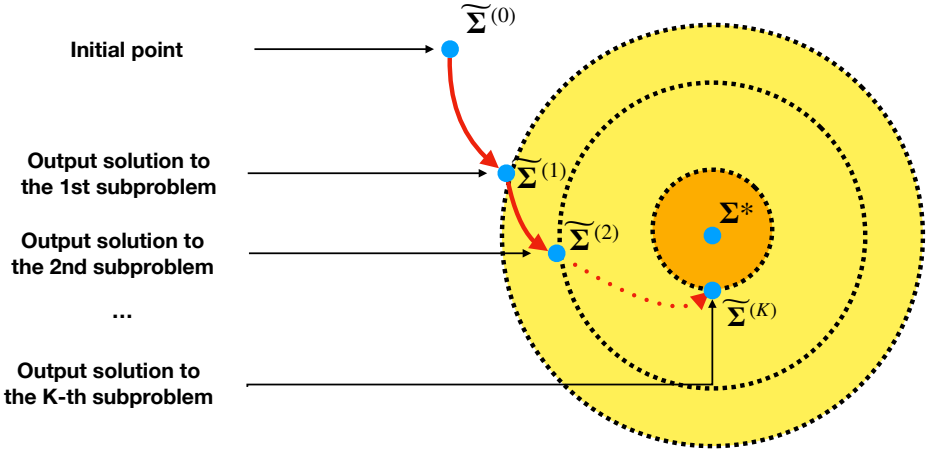
$$\min_{\boldsymbol{\Sigma} \succ \mathbf{0}} \left\{ \frac{1}{2m} \|\boldsymbol{y} - \mathcal{A}(\boldsymbol{\Sigma})\|_{\mathrm{F}}^2 - \tau \log\det \boldsymbol{\Sigma} + \sum_{i,j} p_{\lambda}(|\Sigma_{ij}|) \right\},$$

$k = k + 1$;

end for.

- Due to numerical optimization error in practice, we can only compute an **approximate** solution (ε-optimal solution) $\widetilde{\boldsymbol{\Sigma}}$ instead of the **optimal** solution to each subproblem.

Algorithm Illustration



- Our algorithm can guarantee that an **approximate local optimum** enjoys the **optimal** statistical property.

Essential Assumptions

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

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

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

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

  where $\mu \in (0, \alpha)$ satisfies $p'_\lambda(\mu\lambda) \geq \frac{\lambda}{2}$.
- **Restricted Strong Convexity (RSC) & Restricted Strong Smoothness (RSS)**: For the function f, there exists some $\infty > \rho^+ > \rho^- > 0$ such that, for all $\boldsymbol{\Delta} \in \mathcal{B}\left(\boldsymbol{\Sigma}^\star, \frac{\rho^-}{4\tau\kappa}\right)$,

$$f\left(\boldsymbol{\Sigma} + \boldsymbol{\Delta}\right) \geq f\left(\boldsymbol{\Sigma}\right) + \left\langle \nabla f\left(\boldsymbol{\Sigma}\right), \boldsymbol{\Delta}\right\rangle + \frac{\rho^-}{2}\left\|\boldsymbol{\Delta}\right\|_{\mathrm{F}}, \tag{1}$$

$$f\left(\boldsymbol{\Sigma} + \boldsymbol{\Delta}\right) \leq f\left(\boldsymbol{\Sigma}\right) + \left\langle \nabla f\left(\boldsymbol{\Sigma}\right), \boldsymbol{\Delta}\right\rangle + \frac{\rho^+}{2}\left\|\boldsymbol{\Delta}\right\|_{\mathrm{F}}. \tag{2}$$

Theoretical Results I

Let $\mathcal{S} = \left\{(i,j) \mid \Sigma^\star_{ij} \neq 0\right\}$ and $|\mathcal{S}| = s$. Define $f(\boldsymbol{\Sigma}) = \frac{1}{2m}\|\boldsymbol{y} - \mathcal{A}(\boldsymbol{\Sigma})\|_{\mathrm{F}}^2 - \tau \log\det \boldsymbol{\Sigma}$.

Theorem 1 (contraction property)

Under some standard assumptions, with probability exceeding $1 - c_1\exp(-c_2 m)$ for some $c_1, c_2 > 0$, then 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, provided that $m = \mathcal{O}((s+s^\diamond)\log^2(d/(s+s^\diamond)))$ with $s^\diamond \ge \beta s$ for some universal constant β.

Theoretical Results II

Corollary 2 (oracle rate)

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.

- The *oracle rate* refers to the statistical convergence rate of the *oracle estimator*, defined as $\widehat{\boldsymbol{\Sigma}}^O = \arg\min_{\boldsymbol{\Sigma}:\, \boldsymbol{\Sigma}_{\overline{\mathcal{S}}}=\mathbf{0}} f(\boldsymbol{\Sigma})$, which knows the true coefficients in advance. By the mean value theorem, it is easy to obtain that $\|\widehat{\boldsymbol{\Sigma}}^O - \boldsymbol{\Sigma}^\star\|_{\mathrm{F}} \lesssim \|(\nabla f(\boldsymbol{\Sigma}^\star))_{\mathcal{S}}\|_{\mathrm{F}} \lesssim \sqrt{\frac{s}{mn}}$.

Numerical Experiments I

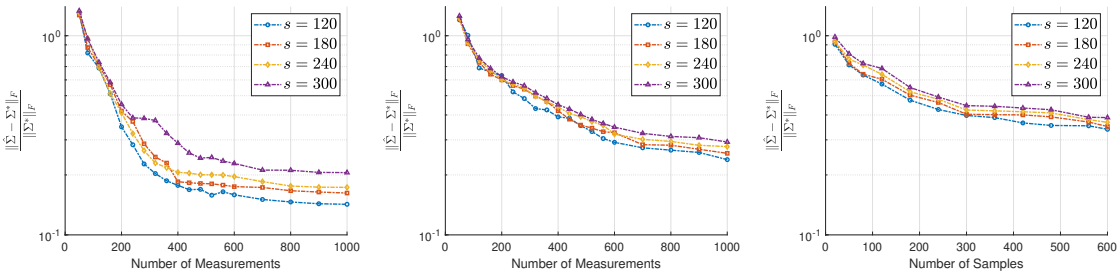


Figure 2: The FRE of the estimated covariance matrices is examined in three distinct scenarios: (a) the true covariance without added noise; (b) the sample covariance with a noise parameter of $\gamma = 0.1$ and $n = 50$; (c) the sample covariance with a noise parameter of $\gamma = 0.1$ and $m = 300$;

Numerical Experiments II

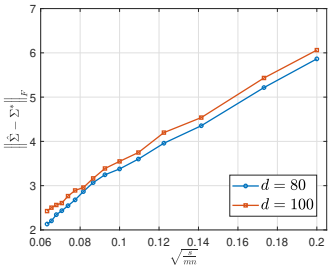


Figure 3: The oracle rate of "sprandsym" Matrix with $s = 120$ and $\gamma = 0.1$.

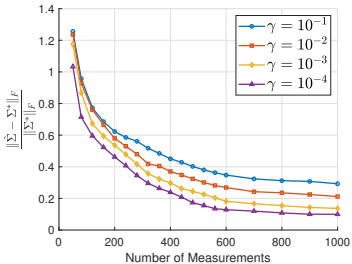


Figure 4: The FRE of the estimated covariance matrices for different noise levels when $s = 300$.

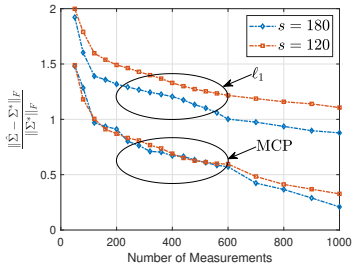


Figure 5: The FRE of the estimated covariance matrices for different sparsity levels with noise level $\gamma = 10^{-1}$(ℓ_1 v.s. MCP).

Numerical Experiments III

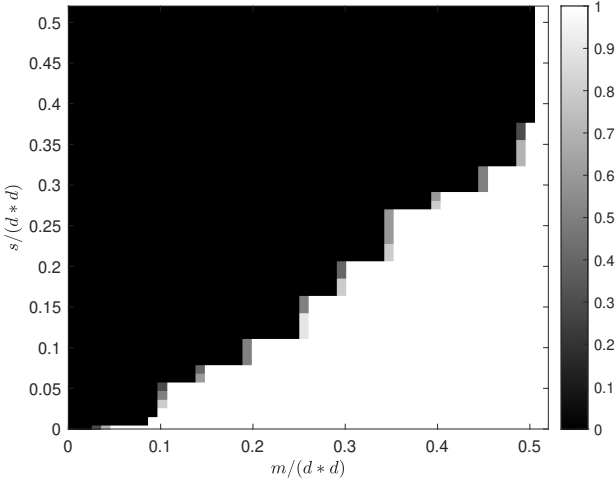


Figure 6: The rate of successful covariance reconstruction when $d = 100$.

Conclusion

- We have proposed a novel approach for large sparse covariance matrix estimation from quadratic measurements using the non-convex penalty and presented both the theoretical and empirical results.

- To the best of our knowledge, this is **the first work** to obtain the **optimal statistical rate** for large sparse covariance matrix estimation from quadratic measurements.

Thank you!