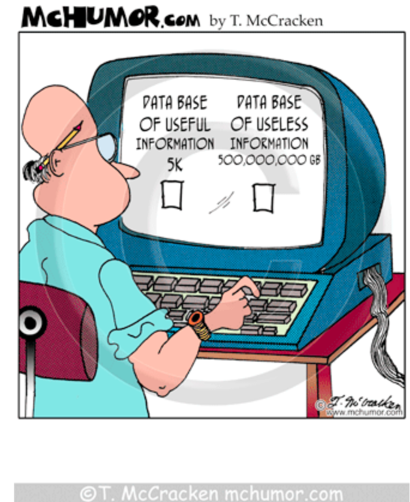
Exact and Stable Covariance Estimation from Quadratic Sampling via Convex Programming

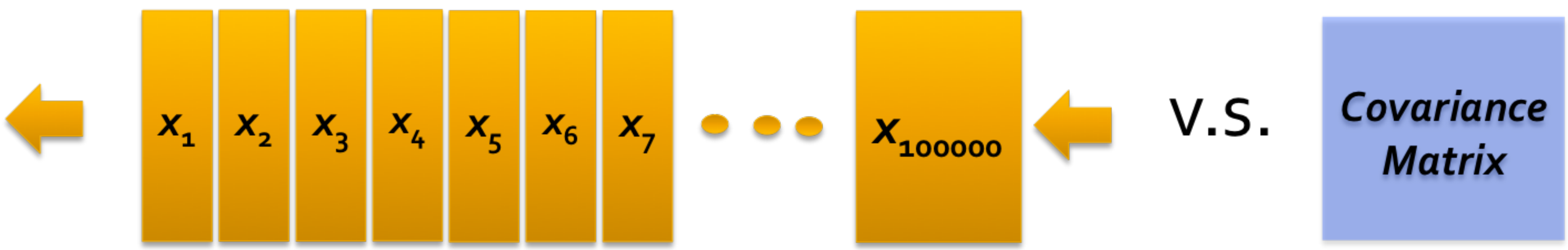
Yuxin Chen[†], Yuejie Chi[*], Andrea J. Goldsmith[†]

Stanford University[†], Ohio State University[*]

High-Dimensional Sequential Data / Signals

- **Data Stream / Stochastic Processes**
  - Each data instance can be high-dimensional
  - We're interested in information in the data rather than the data themselves



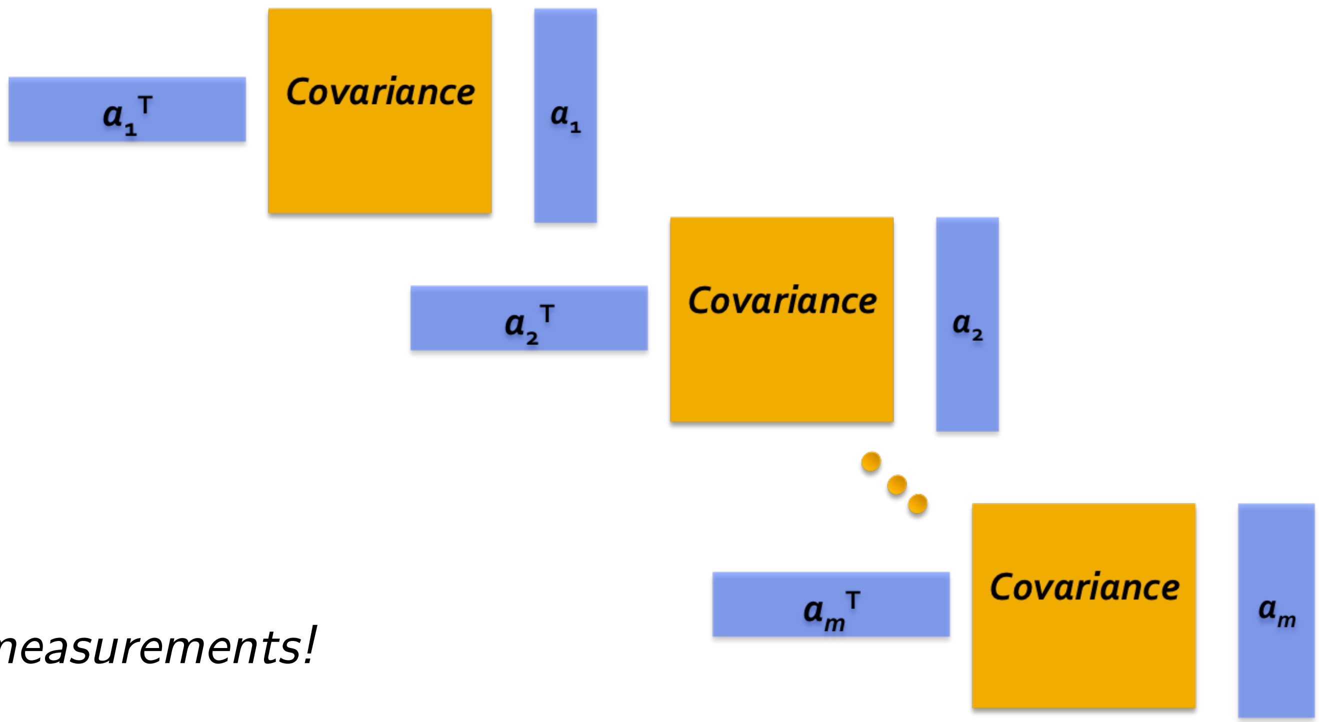


- **Covariance Estimation**
  - second-order statistics $\mathbf{\Sigma} \in \mathbb{R}^{n \times n}$
  - cornerstone of many information processing tasks

What are Quadratic Measurements?

- **Quadratic Measurements**
  - obtain m measurements of $\boldsymbol{\Sigma}$ taking the form

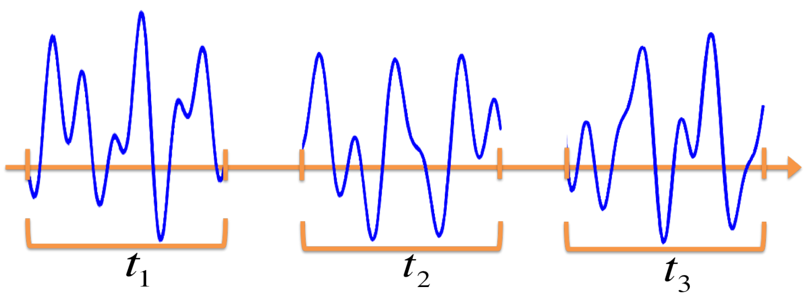
$$\boldsymbol{y}_i \approx \boldsymbol{a}_i^\top \boldsymbol{\Sigma} \boldsymbol{a}_i \qquad (1 \le i \le m)$$

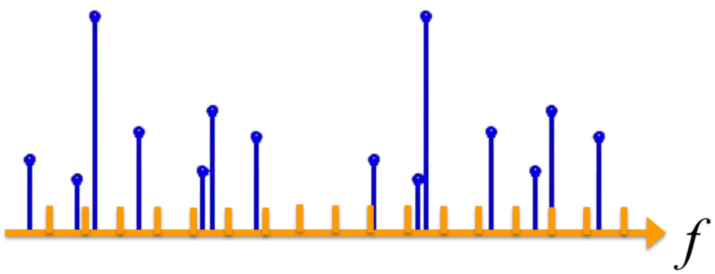


  - *rank-1 measurements!*

Example: Applications in Spectral Estimation

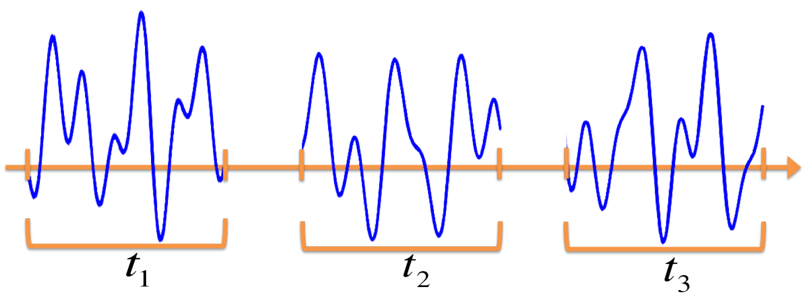
- **High-frequency wireless and signal processing** (Energy Measurements)
  - Spectral estimation of stationary processes (*possibly sparse*)

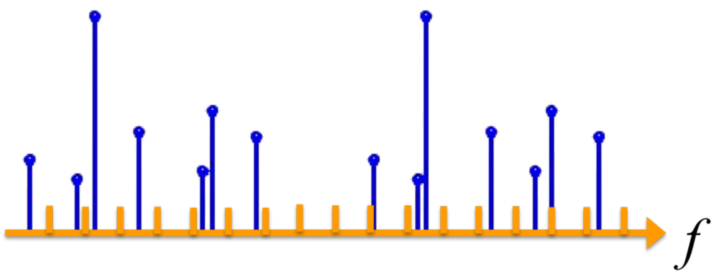




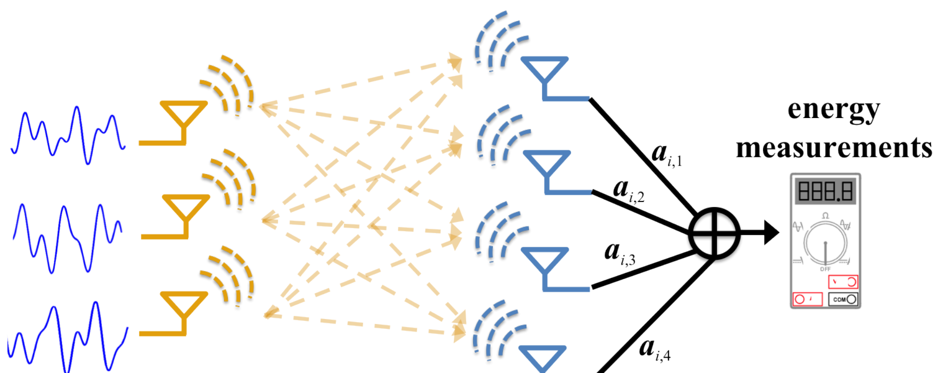
Example: Applications in Spectral Estimation

- **High-frequency wireless and signal processing** (Energy Measurements)
  - Spectral estimation of stationary processes (*possibly sparse*)





  - Channel Estimation in MIMO Channels



Example: Applications in Optics

- **Phase Space Tomography**
  - measure correlation functions of a wave field

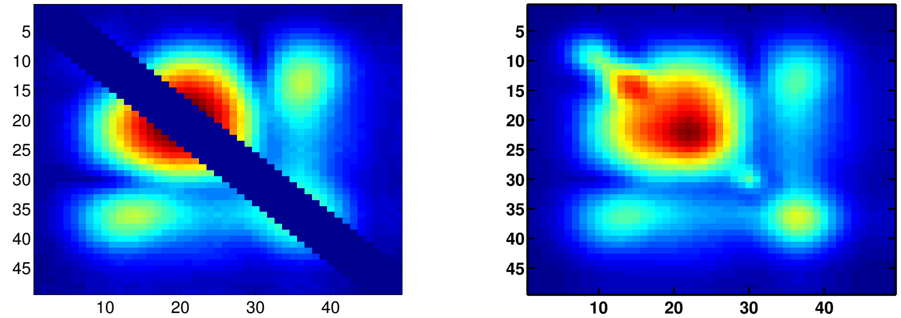
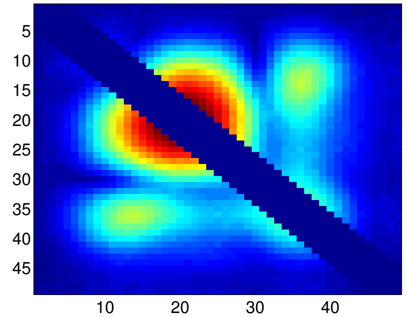
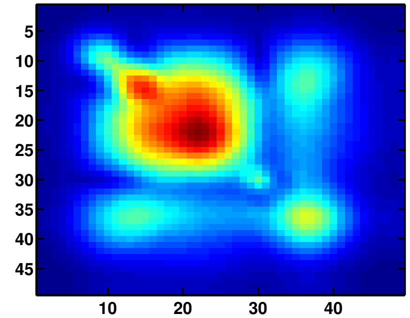


Fig credit: Chi et al

Example: Applications in Optics

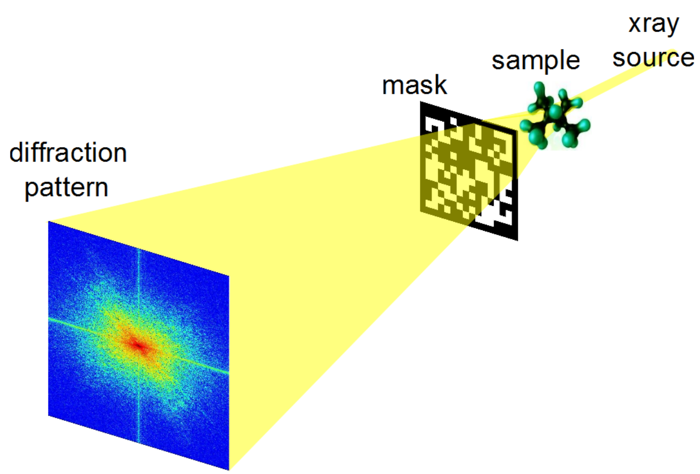
- **Phase Space Tomography**
  - measure correlation functions of a wave field





courtesy of Chi et al

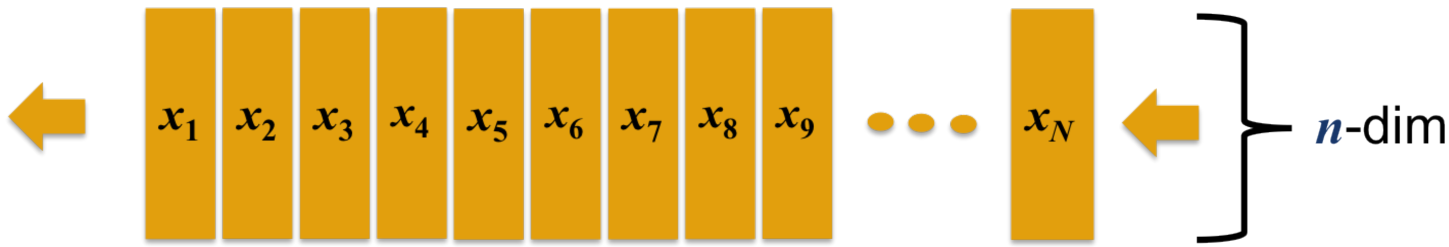
- **Phase Retrieval**
  - signal recovery from magnitude measurements



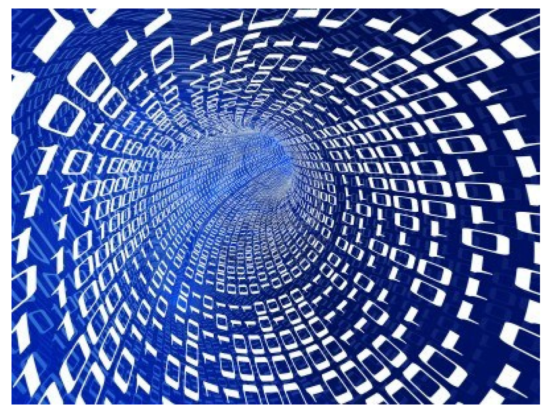
courtesy of Candes et al

Example: Applications in Data Streams

- **Covariance Sketching**
  - data stream: real-time data $\{\boldsymbol{x}_t\}_{t=1}^{\infty}$ arriving sequentially at a high rate...



- **Challenges**
  - limited memory
  - computational efficiency
  - hopefully a single pass over the data

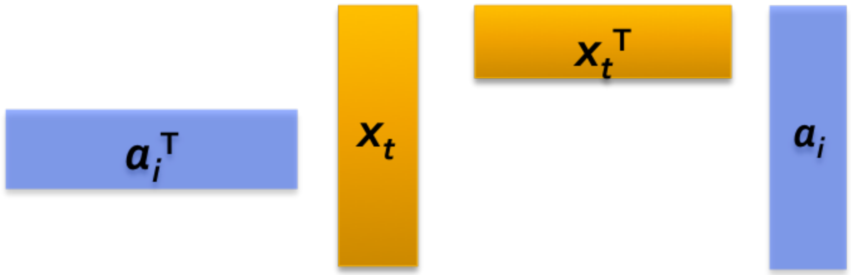


binary data stream by Kazmin

Proposed Quadratic Sketching Method

1) **Sketching**:

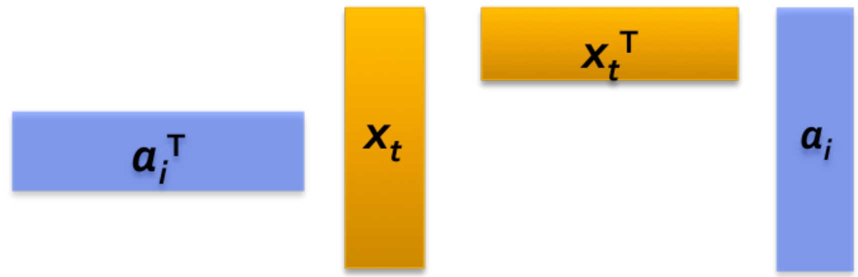
- at each time t, obtain a *quadratic* sketch $(\boldsymbol{a}_i^\top \boldsymbol{x}_t)^2$
  - $\boldsymbol{a}_i$: sketching vector



Proposed Quadratic Sketching Method

1) Sketching:

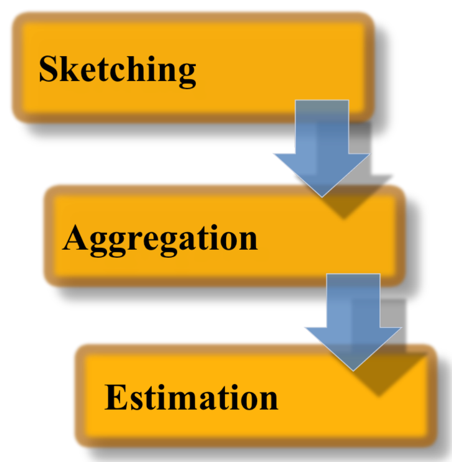
- at each time t, obtain a *quadratic* sketch $(\boldsymbol{a}_i^\top \boldsymbol{x}_t)^2$
  - — $\boldsymbol{a}_i$: sketching vector



2) Aggregation:

- all sketches are aggregated into m measurements

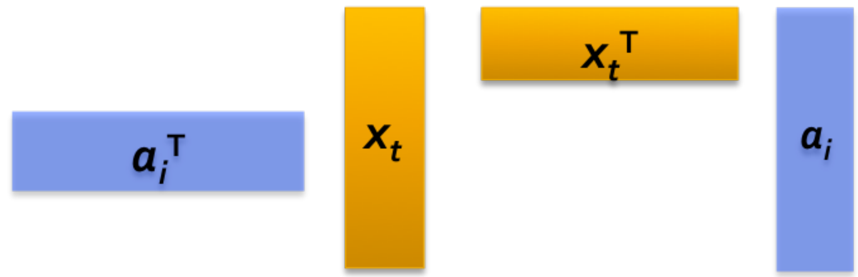
$$y_i = \boldsymbol{a}_i^\top \left(\frac{1}{T} \sum_{t=1}^{T} \boldsymbol{x}_t \boldsymbol{x}_t^\top \right) \boldsymbol{a}_i \approx \boldsymbol{a}_i^\top \boldsymbol{\Sigma} \boldsymbol{a}_i \quad (1 \le i \le m)$$



Proposed Quadratic Sketching Method

1) **Sketching**:

- at each time t, obtain a *quadratic* sketch $(\boldsymbol{a}_i^\top \boldsymbol{x}_t)^2$
  — $\boldsymbol{a}_i$: sketching vector

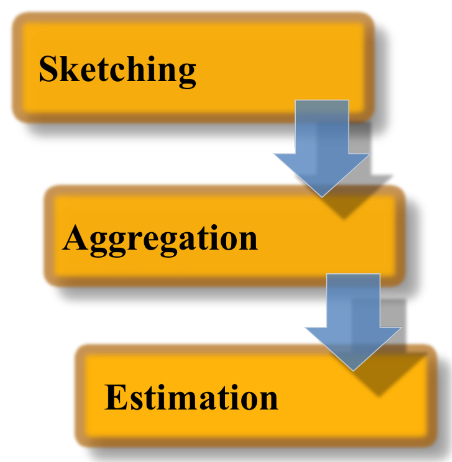


2) **Aggregation:**

- all sketches are aggregated into m measurements

$$y_i = \boldsymbol{a}_i^\top \left(\frac{1}{T} \sum_{t=1}^{T} \boldsymbol{x}_t \boldsymbol{x}_t^\top \right) \boldsymbol{a}_i \approx \boldsymbol{a}_i^\top \boldsymbol{\Sigma} \boldsymbol{a}_i \quad (1 \le i \le m)$$

- **Benefits:**
  - one pass
  - **minimal storage** (*as will be shown*)



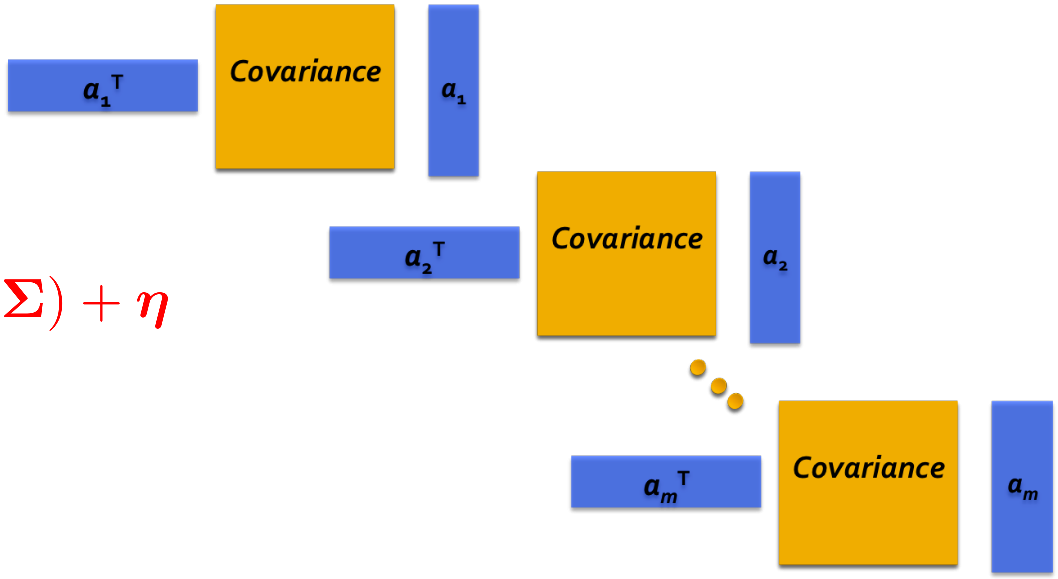
Problem Formulation

- **Given:** m ($\ll n^2$) quadratic measurements $\boldsymbol{y} = \{y_i\}_{i=1}^m$

$$y_i = \boldsymbol{a}_i^\top \boldsymbol{\Sigma} \boldsymbol{a}_i + \eta_i, \quad i = 1, \cdots, m,$$

  - $\boldsymbol{a}_i$: sampling vectors
  - $\boldsymbol{\eta} = \{\eta_i\}_{i=1}^m$: noise terms
  - more concise operator form:

$$\boldsymbol{y} = \mathcal{A}(\boldsymbol{\Sigma}) + \boldsymbol{\eta}$$

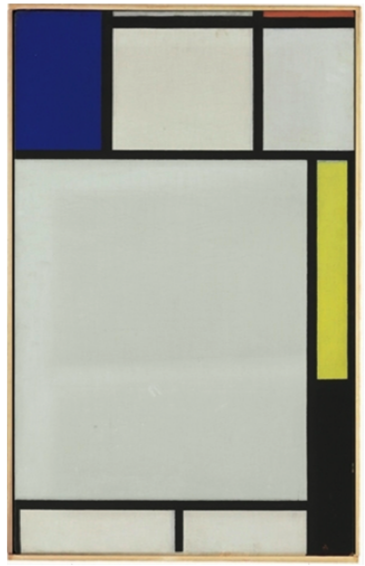


- **Goal:** recover $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$.

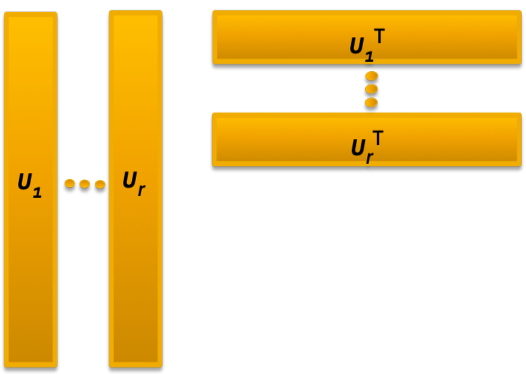
- **Sampling model**
  - sub-Gaussian i.i.d. sampling vectors

Geometry of Covariance Structure

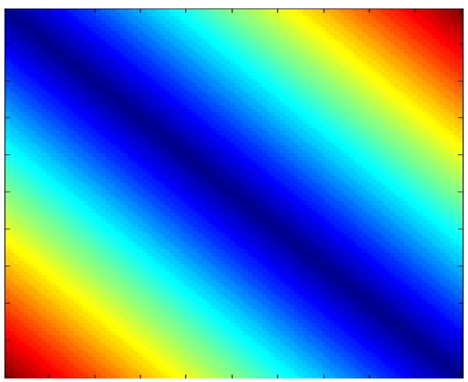
- **# unknown $>$ # stored measurements**
  - exploit low-dimensional structures!
- **Structures considered in this talk:**
  - low rank
  - Toeplitz low rank
  - simultaneously sparse and low-rank



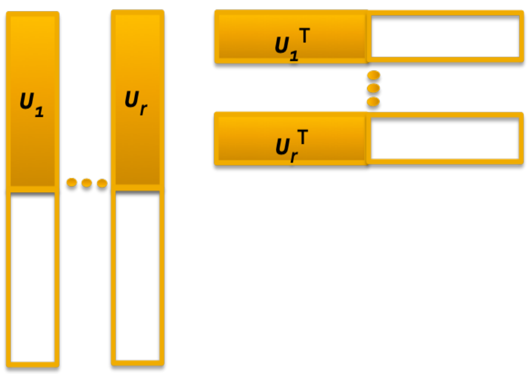
Piet Mondrian



1) low rank



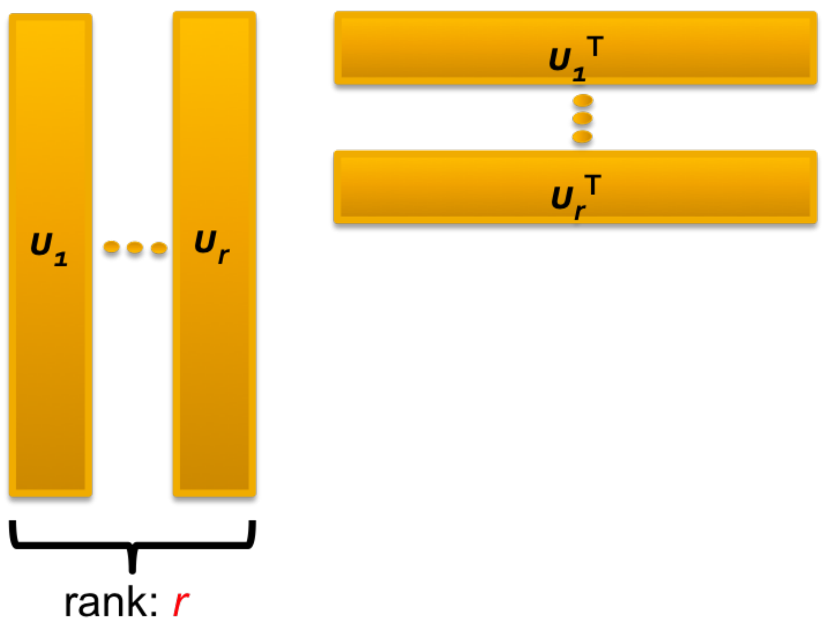
2) Toeplitz low rank



3) jointly sparse and low rank

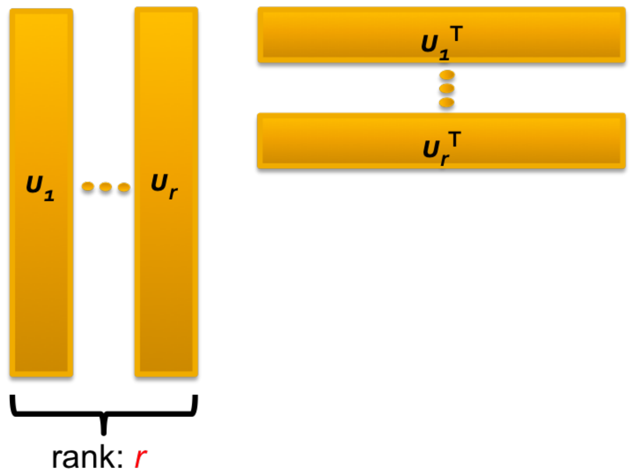
Low Rank

- **Low-Rank Structure:**
  - A few components explains most of the data variability
  - metric learning, array signal processing, collaborative filtering ...



- $\text{rank}(\boldsymbol{\Sigma}) = r \ll n$.

Trace Minimization for Low-Rank Structure



- **Trace Minimization**

$$
\begin{aligned}
\text{(TraceMin)} \quad \text{minimize}_{M} \quad & \underbrace{\text{trace}\,(M)}_{\text{low rank}} \\
\text{s.t.} \quad & \|\mathcal{A}(M) - y\|_1 \le \underbrace{\epsilon}_{\text{noise bound}}, \\
& M \succeq 0.
\end{aligned}
$$

  - inspired by *Candes et. al.* for phase retrieval

Near-Optimal Recovery for Low-Rank Structure

$$\text{minimize} \quad \text{tr}\,(\boldsymbol{M}) \quad \text{s.t.} \; \|\mathcal{A}\,(\boldsymbol{M}) - \boldsymbol{y}\|_1 \le \epsilon, \quad \boldsymbol{M} \succeq 0$$

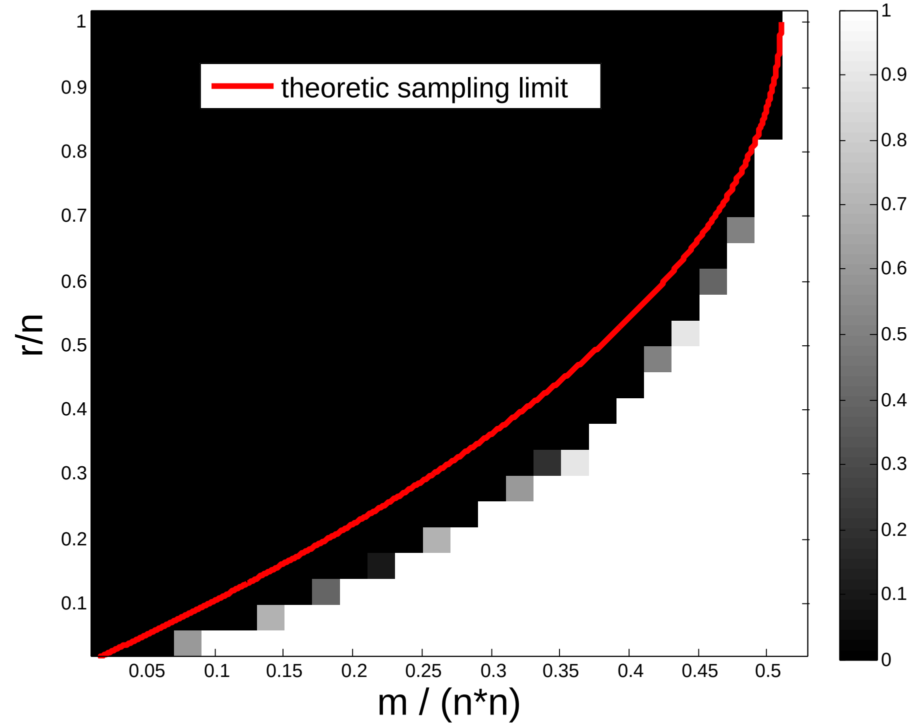
Theorem 1 (Low Rank). With high prob, for *all* $\boldsymbol{\Sigma}$ with rank$(\boldsymbol{\Sigma}) \le r$, the solution $\hat{\boldsymbol{\Sigma}}$ to TraceMin obeys

$$\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathrm{F}} \lesssim \underbrace{\frac{\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_r\|_*}{\sqrt{r}}}_{\text{due to imperfect structure}} + \underbrace{\frac{\epsilon}{m}}_{\text{due to noise}},$$

provided that $\boldsymbol{m} \gtrsim \boldsymbol{rn}$. ($\boldsymbol{\Sigma_r}$: rank-$r$ approx of $\boldsymbol{\Sigma}$)

- **Exact recovery** in the noiseless case
- **Universal recovery**: simultaneously works for all low-rank matrices
- **Robust recovery** when $\boldsymbol{\Sigma}$ is *approximately* low-rank
- **Stable recovery** against bounded noise

Phase Transition for Low-Rank Recovery

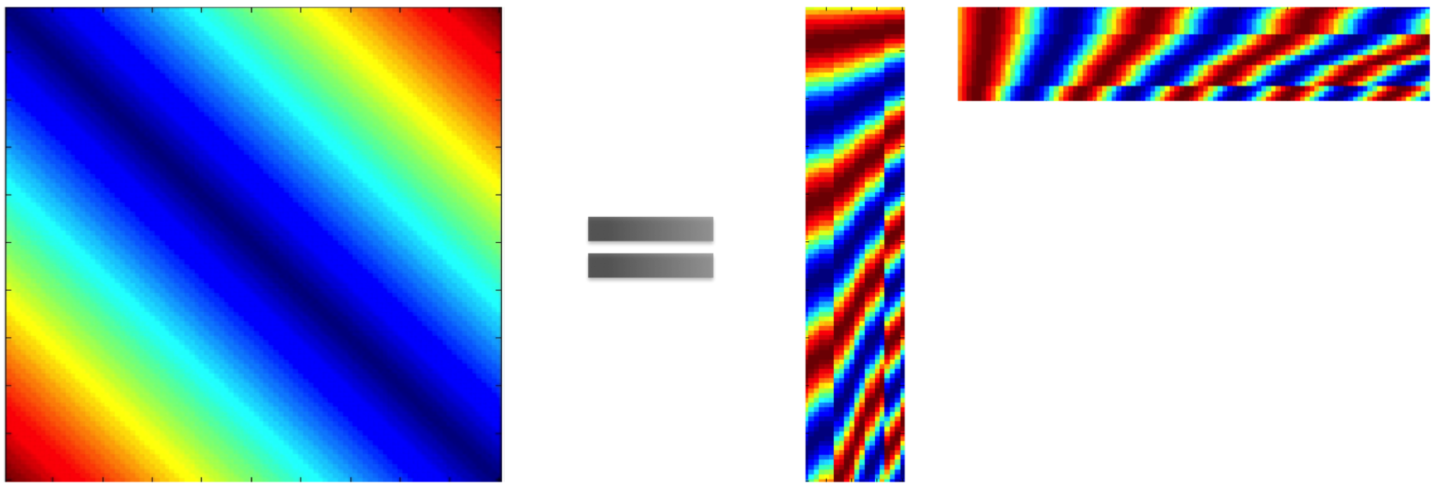


empirical success probability of Monte Carlo trials: $n = 50$

- **Near-Optimal** Storage Complexity!
  - degrees of freedom $\approx rn$

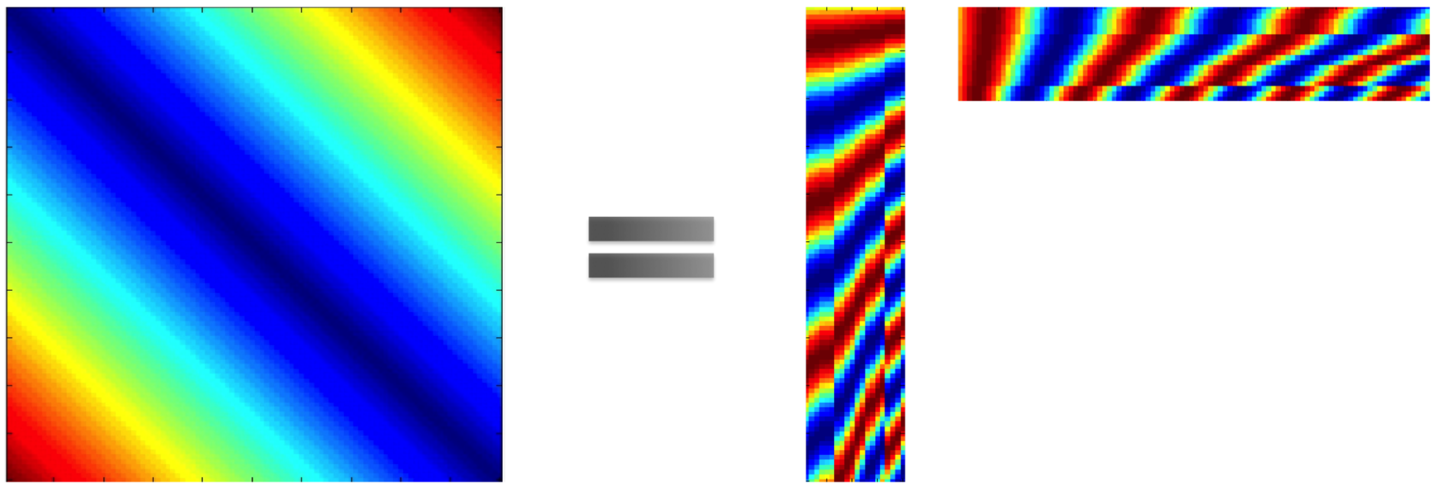
Toeplitz Low Rank

- **Toeplitz Low-Rank Structure:**
  - **Spectral sparsity!**
    - possibly *off-the-grid* frequency spikes (Vandemonde decomposition)
  - wireless communication, array signal processing ...



- $\mathrm{rank}(\mathbf{\Sigma}) = r \ll n$.

Trace Minimization for Toeplitz Low-Rank Structure



- **Trace Minimization**

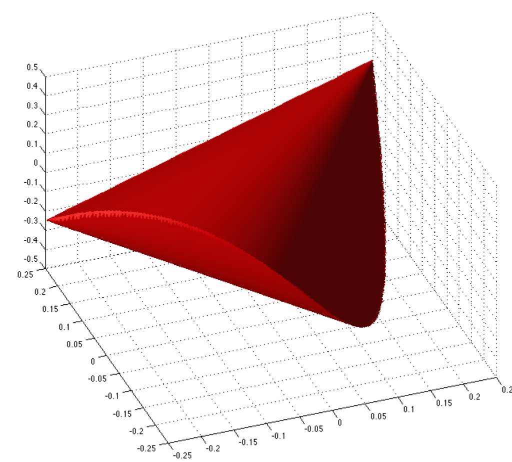
$$
\begin{aligned}
(\text{ToepTraceMin}) \quad \text{minimize}_{\boldsymbol{M}} \quad & \underbrace{\text{trace}\left(\boldsymbol{M}\right)}_{\text{low rank}} \\
\text{s.t.} \quad & \left\|\mathcal{A}\left(\boldsymbol{M}\right) - \boldsymbol{y}\right\|_2 \leq \underbrace{\epsilon_2}_{\text{noise bound}}, \\
& \boldsymbol{M} \succeq \boldsymbol{0}, \\
& \boldsymbol{M} \text{ is Toeplitz.}
\end{aligned}
$$

Near-Optimal Recovery for Toeplitz Low-Rank Structure

$$\text{minimize} \quad \mathrm{tr}\,(\boldsymbol{M}) \quad \text{s.t.}\ \|\mathcal{A}\,(\boldsymbol{M}) - \boldsymbol{y}\|_2 \le \epsilon_2, \quad \boldsymbol{M} \succeq 0, \quad \boldsymbol{M} \text{ is Toeplitz}$$

Theorem 2 (Toeplitz Low Rank). With high prob, for *all* Toeplitz $\boldsymbol{\Sigma}$ with rank$(\boldsymbol{\Sigma}) \le r$, the solution $\hat{\boldsymbol{\Sigma}}$ to ToepTraceMin obeys

$$\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathrm{F}} \lesssim \underbrace{\frac{\epsilon_2}{\sqrt{m}}}_{\text{due to noise}},$$

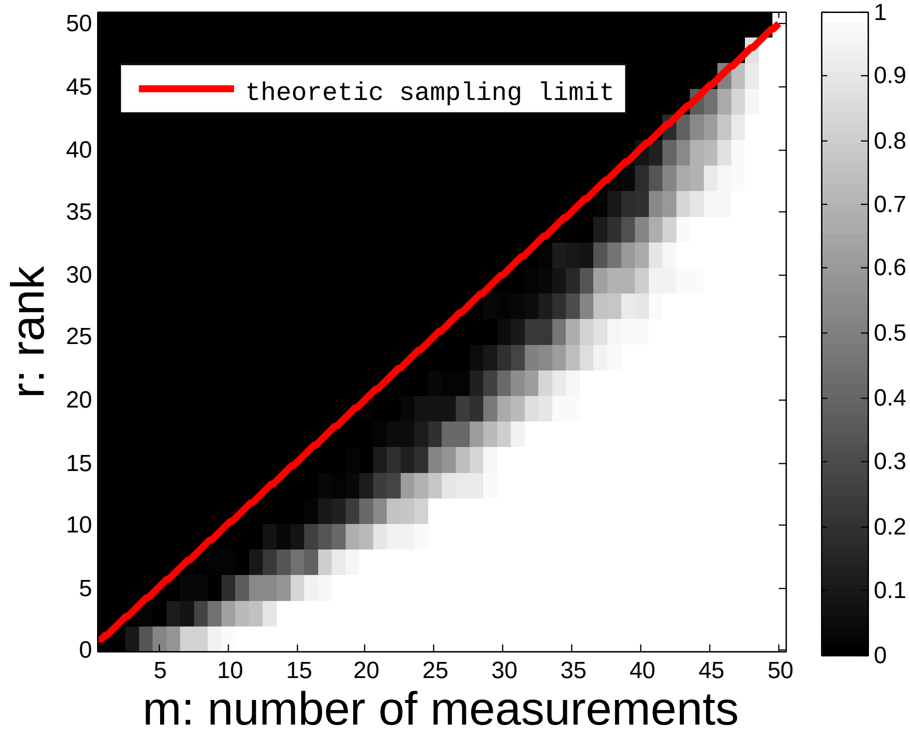


Toeplitz ball

provided that $\boldsymbol{m} \gtrsim \boldsymbol{r}\,\text{poly}\,\mathbf{log}(\boldsymbol{n})$.

- **Exact recovery** in the absence of noise
- **Universal recovery**: simultaneously works for all Toeplitz low-rank matrices
- **Stable recovery** against bounded noise

Phase Transition for Toeplitz Low-Rank Recovery

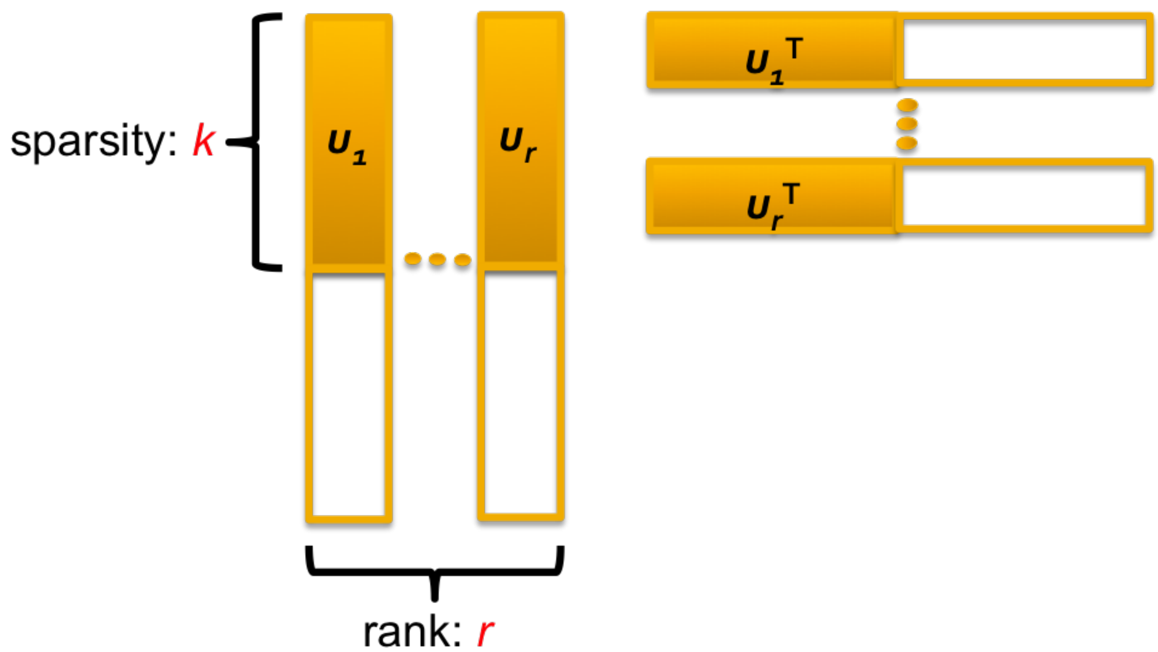


empirical success probability of Monte Carlo trials: $n = 50$

- **Near-Optimal** Storage Complexity!
  - degrees of freedom $\approx r$

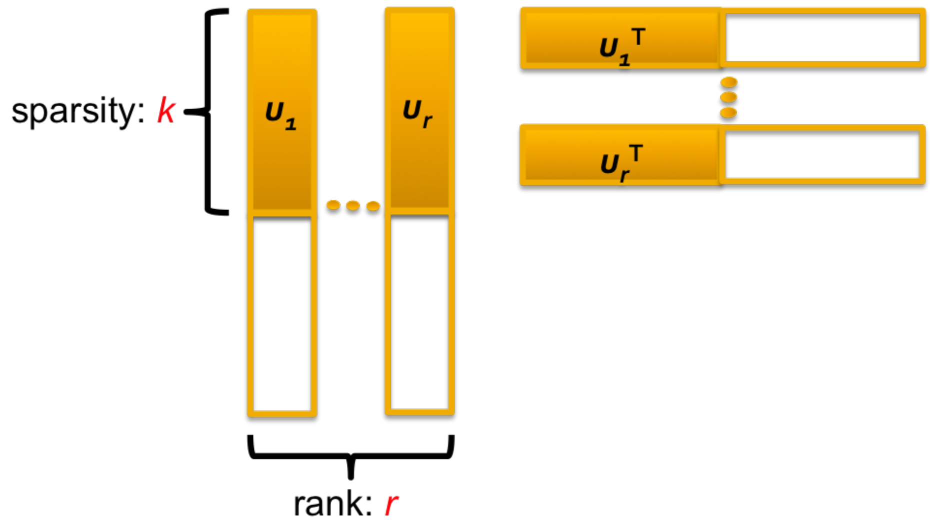
Simultaneous Structure

- **Joint Structure:** $\boldsymbol{\Sigma}$ is *simultaneously* sparse and low-rank.
  - **rank:** r
  - **sparsity:** k



  - SVD: $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$, where $\boldsymbol{U} = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r]$

Convex Relaxation for Simultaneous Structure



- **Convex Relaxation**

$$
\begin{aligned}
\text{minimize}_{\boldsymbol{M}} \quad & \underbrace{\text{trace}\,(\boldsymbol{M})}_{\text{low rank}} + \boldsymbol{\lambda} \underbrace{\|\boldsymbol{M}\|_1}_{\text{sparsity}} \\
\text{s.t.} \quad & \|\mathcal{A}\,(\boldsymbol{M}) - \boldsymbol{y}\|_1 \le \underbrace{\epsilon}_{\text{noise bound}}, \\
& \boldsymbol{M} \succeq \boldsymbol{0}.
\end{aligned}
$$

  - coincides with *Li and Voroninski* for rank-1 cases

Exact Recovery for Simultaneous Structure

$$\text{minimize} \quad \text{tr}\left(\boldsymbol{M}\right) + \lambda \left\| \boldsymbol{M} \right\|_1 \quad \text{s.t.} \;\; \mathcal{A}\left(\boldsymbol{M}\right) = \boldsymbol{y}, \quad \boldsymbol{M} \succeq 0$$

Theorem 3 (Simultaneous Structure). SDP with $\lambda \in \left[\frac{1}{n}, \frac{1}{N_{\boldsymbol{\Sigma}}}\right]$ is exact with high probability, provided that

$$m \gtrsim \frac{r \log n}{\lambda^2} \tag{1}$$

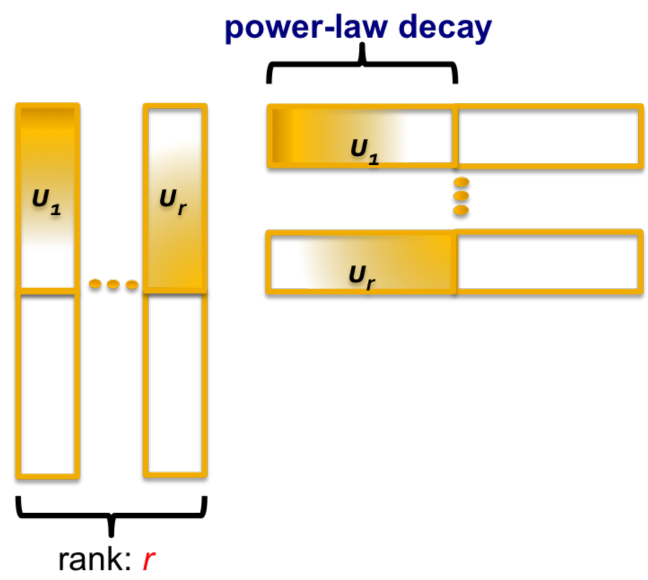
where $N_{\boldsymbol{\Sigma}} := \max\left\{ \left\|\text{sign}\left(\boldsymbol{\Sigma}_{\Omega}\right)\right\|, \sqrt{\frac{k \sum_{i=1}^{r} \|\boldsymbol{u}_i\|_1^2}{r}} \right\}$.

- Exact recovery with appropriate regularization parameters
- Question: how good is the storage complexity (1)?

Compressible Covariance Matrices: Near-Optimal Recovery

Definition (Compressible Matrices)

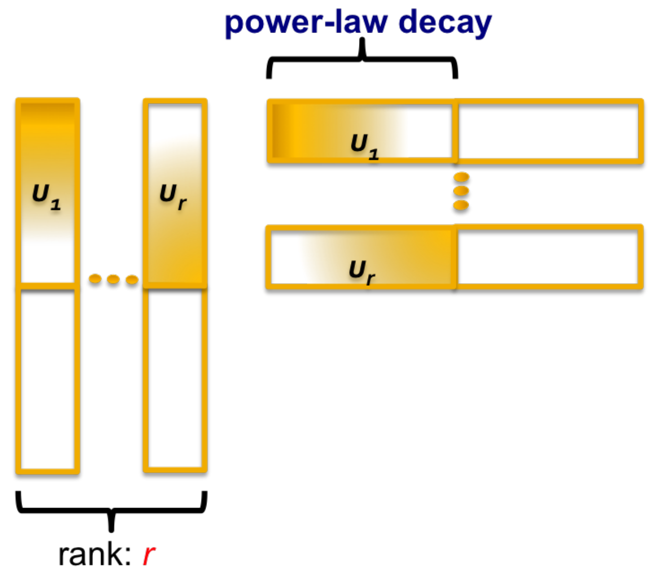
- *non-zero entries* of $\boldsymbol{u}_i$ exhibit power-law decays
  - $\|\boldsymbol{u}_i\|_1 = O(\text{poly}\log(n))$.



Compressible Covariance Matrices: Near-Optimal Recovery

Definition (Compressible Matrices)

- *non-zero entries* of $\boldsymbol{u}_i$ exhibit power-law decays
  - $\|\boldsymbol{u}_i\|_1 = O(\text{poly}\log(n))$.



Corollary 1 (Compressible Case). For compressible covariance matrices, SDP with $\lambda \approx \frac{1}{\sqrt{k}}$ is exact w.h.p., provided that

$$m \gtrsim kr \cdot \text{poly}\log(n).$$

- *Near-Minimal Measurements!*
  - degree-of-freedom: $\Theta(kr)$

Stability and Robustness

- **noise**: $\|\boldsymbol{\eta}\|_1 \leq \epsilon$
- **imperfect structural assumption**: $\boldsymbol{\Sigma} = \underbrace{\boldsymbol{\Sigma}_{\Omega}}_{\text{simultaneous sparse and low-rank}} + \underbrace{\boldsymbol{\Sigma}_{\mathsf{c}}}_{\text{residuals}}$

Stability and Robustness

- **noise**: $\|\boldsymbol{\eta}\|_1 \leq \epsilon$

- **imperfect structural assumption**: $\boldsymbol{\Sigma} = \underbrace{\boldsymbol{\Sigma}_{\Omega}}_{\text{simultaneous sparse and low-rank}} + \underbrace{\boldsymbol{\Sigma}_{\mathrm{c}}}_{\text{residuals}}$

Theorem 4. Under the same λ as in Theorem 1 or Corollary 1,

$$\left\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_{\Omega}\right\|_{\mathrm{F}} \lesssim \frac{1}{\sqrt{r}} \left(\underbrace{\left\|\boldsymbol{\Sigma}_{\mathrm{c}}\right\|_* + \lambda \left\|\boldsymbol{\Sigma}_{\mathrm{c}}\right\|_1}_{\text{due to imperfect structure}} \right) + \underbrace{\frac{\epsilon}{m}}_{\text{due to noise}}$$

- stable against bounded noise

- robust against imperfect structural assumptions

Mixed-Norm RIP (for Low-Rank and Joint Structure)

- **Restricted Isometry Property**: a powerful notion for compressed sensing

$$\forall \boldsymbol{X} \text{ in some class :} \qquad \left\|\mathcal{B}\left(\boldsymbol{X}\right)\right\|_{2} \approx \left\|\boldsymbol{X}\right\|_{\mathrm{F}}.$$

  - *unfortunately, it does NOT hold for quadratic models*

Mixed-Norm RIP (for Low-Rank and Joint Structure)

- **Restricted Isometry Property**: a powerful notion for compressed sensing

$$\forall \boldsymbol{X} \text{ in some class}: \qquad \|\mathcal{B}(\boldsymbol{X})\|_2 \approx \|\boldsymbol{X}\|_{\mathrm{F}}.$$

  - *unfortunately, it does NOT hold for quadratic models*

- **A Mixed-norm Variant:** **RIP-ℓ_2/ℓ_1**

$$\forall \boldsymbol{X} \text{ in some class}: \qquad \|\mathcal{B}(\boldsymbol{X})\|_1 \approx \|\boldsymbol{X}\|_{\mathrm{F}}.$$

Mixed-Norm RIP (for Low-Rank and Joint Structure)

- **Restricted Isometry Property**: a powerful notion for compressed sensing

$$\forall \boldsymbol{X} \text{ in some class}: \qquad \|\mathcal{B}(\boldsymbol{X})\|_2 \approx \|\boldsymbol{X}\|_{\mathrm{F}}.$$

  - *unfortunately, it does NOT hold for quadratic models*

- **A Mixed-norm Variant: RIP-ℓ_2/ℓ_1**

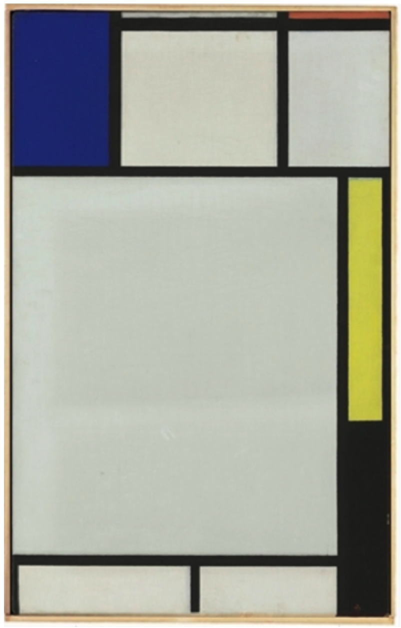
$$\forall \boldsymbol{X} \text{ in some class}: \qquad \|\mathcal{B}(\boldsymbol{X})\|_1 \approx \|\boldsymbol{X}\|_{\mathrm{F}}.$$

  - *does NOT hold for $\mathcal{A}$, but hold after $\mathcal{A}$ is* ***debiased***
  - A very simple proof for PhaseLift!

Concluding Remarks

- **Our approach / analysis works for other structural models**
  - Sparse covariance matrix
  - Low-Rank plus Sparse matrix

- **The way ahead**
  - Sparse *inverse* covariance matrix
  - Beyond sub-Gaussian sampling
  - Online recovery algorithms



Piet Mondrian

Q&A

Full-length version available at arXiv:

Exact and Stable Covariance Estimation from Quadratic Sampling
via Convex Programming

http://arxiv.org/abs/1310.0807

Thank You! Questions?