# Covariance Sketching via Quadratic Sampling

Yuejie Chi
Department of ECE and BMI
The Ohio State University

Tsinghua University
June 2015

THE OHIO STATE UNIVERSITY

# Acknowledgement


- Thanks to my academic collaborators on some of the reported work:

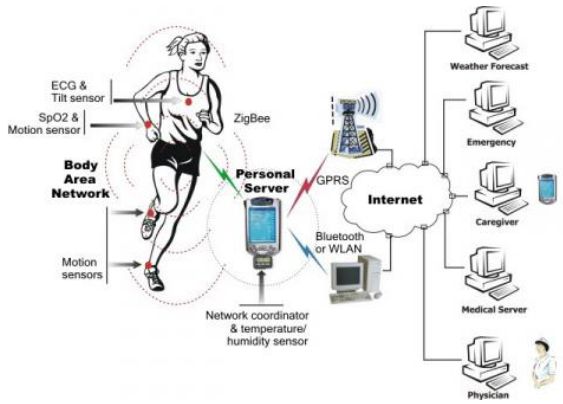
Yuxin Chen

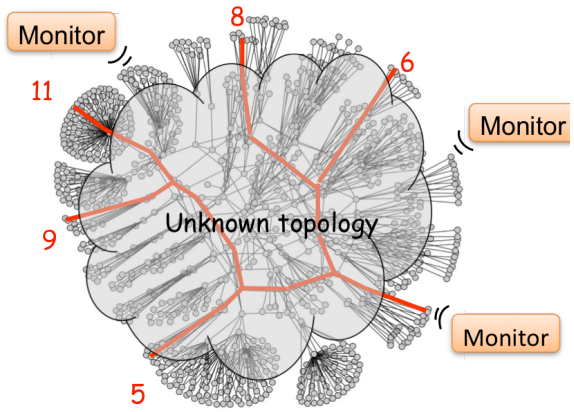
Andrea Goldsmith

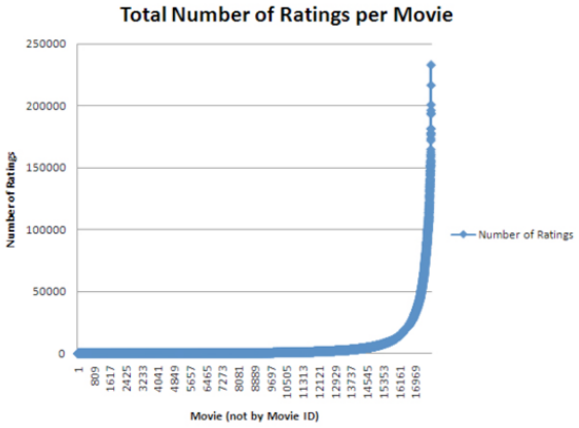
- This work is supported by NSF and AFOSR.

# High-dimensional Streaming Data

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




**Wireless health monitoring:** transmission loss and power-hungry sensors

**Internet monitoring:** limited measurements and massive data

**Netflix Ratings:** incomplete (and skewed) ratings

- Need to **learn and track** structures of minimally observed data streams.

---


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.

Netflix statistics source: `http://www.hackedexistence.com/project-netflix.html`

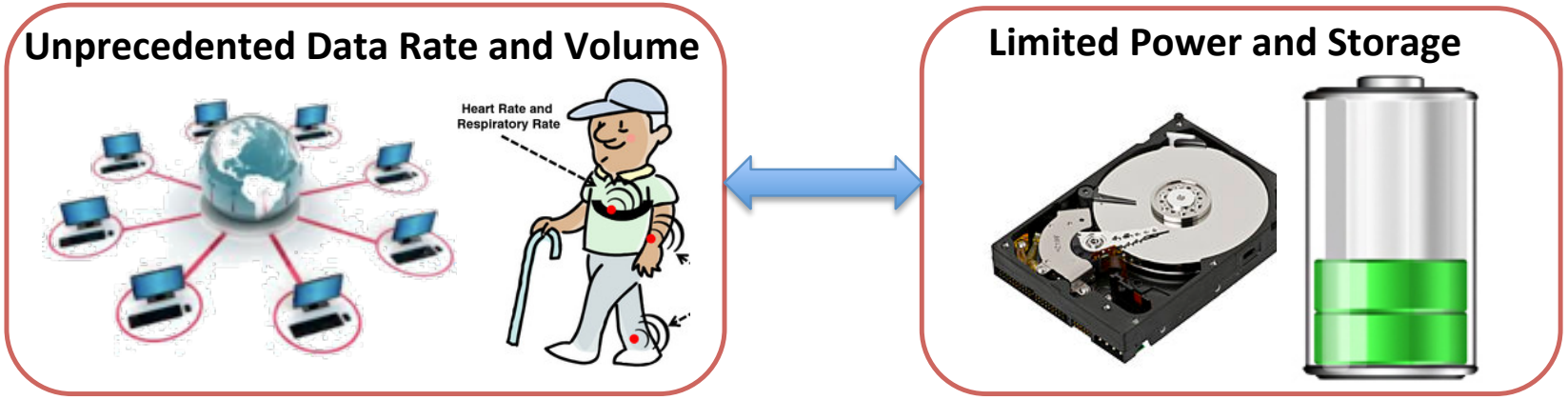
# Challenges in Modern Data Acquisition

**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 access 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.



# Covariance Sketching

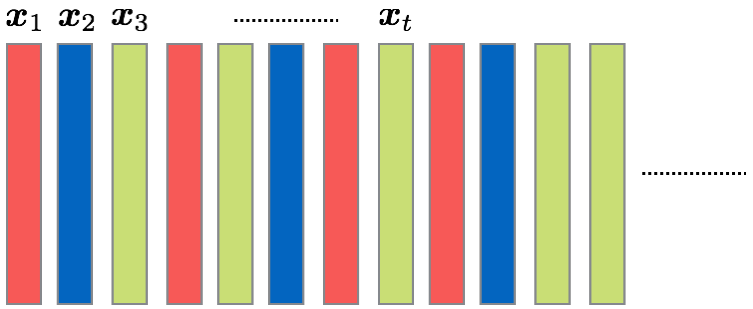
**Key Observation:** the *covariance structure* can be recovered without measuring the whole data stream.

**Approach:** distributed data sketching and aggregation to recover the covariance structure or principal components.

- access each data sample via linear or quadratic (energy) sketches;
- aggregate the sketches into linear observations of the covariance matrix.

# Quadratic Sketching for Covariance Estimation

Consider a data stream possible distributively observed at $m$ sensors:




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

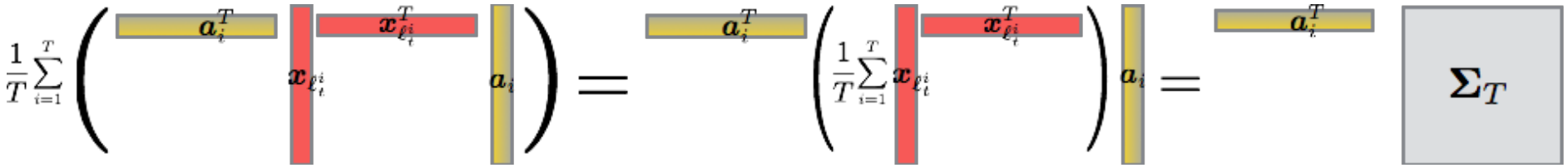
- *randomly* select a sketching vector $\boldsymbol{a}_i \in \mathbb{R}^n$ with i.i.d. sub-Gaussian entries;
- Sketch an arbitrary substream indexed by $\{\ell_t^i\}_{t=1}^T$ with an energy measurement $\left|\left\langle \boldsymbol{a}_i, \boldsymbol{x}_{\ell_t^i} \right\rangle\right|^2$ and aggregate the average **energy** measurement:

$$y_{i,T} = \frac{1}{T}\sum_{t=1}^{T}\left|\left\langle \boldsymbol{a}_i, \boldsymbol{x}_{\ell_t^i} \right\rangle\right|^2 = \frac{T-1}{T} y_{i,T-1} + \frac{1}{T}\left|\left\langle \boldsymbol{a}_i, \boldsymbol{x}_{\ell_T^i} \right\rangle\right|^2.$$

# Quadratic Sketching for Covariance Estimation

- For each sketch:

$$y_{i,T} = \frac{1}{T}\sum_{t=1}^{T}\left|\left\langle \boldsymbol{a}_i, \boldsymbol{x}_{\ell_t^i}\right\rangle\right|^2 = \frac{T-1}{T}y_{i,T-1} + \frac{1}{T}\left|\left\langle \boldsymbol{a}_i, \boldsymbol{x}_{\ell_T^i}\right\rangle\right|^2.$$




- As $T \to \infty$, $\boldsymbol{\Sigma}_T \to \boldsymbol{\Sigma}$,

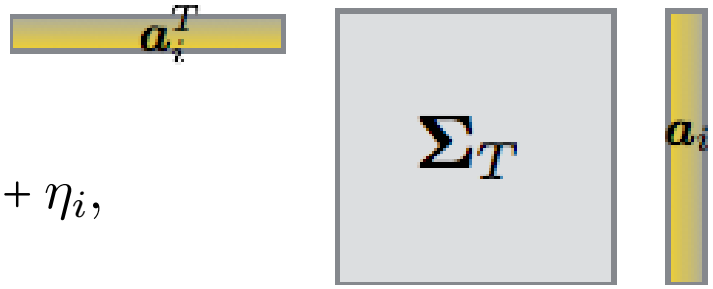
$$y_{i,T} \to \boldsymbol{a}_i^T \boldsymbol{\Sigma} \boldsymbol{a}_i$$

  yields a linear measurement of $\boldsymbol{\Sigma}$!

- All sketches can be obtained in a fully distributed manner.

# Covariance Estimation with Rank-One Measurements

- **Quadratic Measurement Model:**




$$y_{i,T} = \boldsymbol{a}_i^T \boldsymbol{\Sigma}_T \boldsymbol{a}_i := \boldsymbol{a}_i^T \boldsymbol{\Sigma} \boldsymbol{a}_i + \eta_i,$$

where $\eta_i = \boldsymbol{a}_i^T (\boldsymbol{\Sigma}_T - \boldsymbol{\Sigma}) \boldsymbol{a}_i$ is the additive noise.

- More generally, we assume the following measurement model:

$$z_i = \boldsymbol{a}_i^T \boldsymbol{\Sigma} \boldsymbol{a}_i + \eta_i, \quad i = 1, \ldots, m;$$

or more concisely,

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

- The measurements are quadratic in $\boldsymbol{a}_i$ and linear in the rank-one matrix $\boldsymbol{a}_i \boldsymbol{a}_i^T$;

# Sampling Model

We need some additional assumptions on the sampling model:

- sub-Gaussian i.i.d. sketching vectors: each $\boldsymbol{a}_i$ is i.i.d. copies of $\boldsymbol{a} = [a_1, a_2, \ldots, a_n]^T$ satisfying:

$$\mathbb{E}[a_i] = 0, \quad \mathbb{E}[a_i^2] = 1, \quad \text{and} \quad \mu_4 := \mathbb{E}a_i^4 > 1.$$

- noise model: deterministically bounded $\ell_1$-norm noise:

$$\|\boldsymbol{\eta}\|_1 \leq \epsilon$$

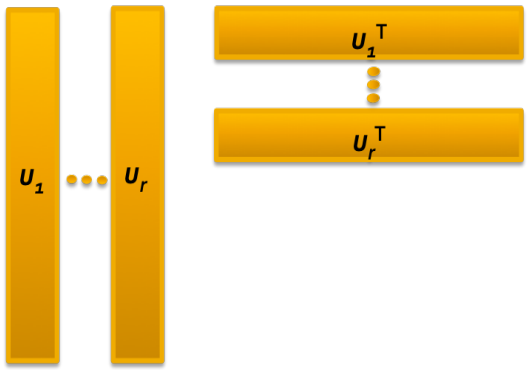
  where $\epsilon$ is known *a priori*.

While we can solve for $\boldsymbol{\Sigma}$ via least-squares estimation if $m \geq n^2$ (the size of $\boldsymbol{\Sigma}$), we can greatly reduce the number of $m$ by exploiting the low-dimensional structure of $\boldsymbol{\Sigma}$.

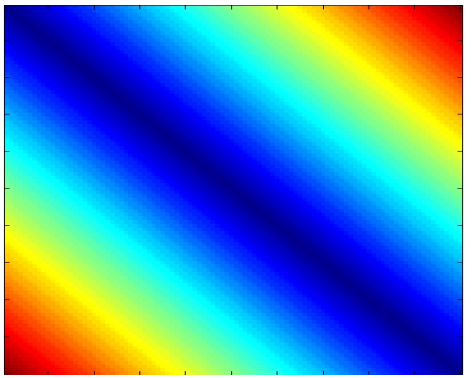
# Geometry of Covariance Structures

Many high-dimensional data lie in a low-dimensional subspace, resulting in a low-rank covariance matrix:

- Low-Rankness: the covariance matrix is low-rank, which occurs when a small number of components accounts for most of the variability in the data.
- Stationary Low-Rankness: the covariance matrix is simultaneously Toeplitz and low rank, which has many applications in array signal processing.




low-rank



Toeplitz low-rank

# Low-Rank Covariance Estimation via Convex Relaxation

- We would like to seek the covariance matrix satisfying the observations with the minimal rank:

$$\hat{\mathbf{\Sigma}} = \underset{\mathbf{\Sigma} \succeq 0}{\operatorname{argmin}} \ \operatorname{rank}(\mathbf{\Sigma}) \quad \text{s.t.} \quad \|\boldsymbol{z} - \mathcal{A}(\mathbf{\Sigma})\|_1 \leq \epsilon.$$

- However this is non-convex and NP-hard. Therefore, we replace it by the trace minimization, which is the tightest convex relaxation with respect to the rank function, over all matrices compatible with the measurements:

$$\hat{\mathbf{\Sigma}} = \underset{\mathbf{\Sigma} \succeq 0}{\operatorname{argmin}} \ \operatorname{Tr}(\mathbf{\Sigma}) \quad \text{s.t.} \quad \|\boldsymbol{z} - \mathcal{A}(\mathbf{\Sigma})\|_1 \leq \epsilon.$$

- Additionally, if $\mathbf{\Sigma}$ is Toeplitz, we add the additional structural constraint:

$$\hat{\mathbf{\Sigma}} = \underset{\mathbf{\Sigma} \succeq 0}{\operatorname{argmin}} \ \operatorname{Tr}(\mathbf{\Sigma}) \quad \text{s.t.} \quad \|\boldsymbol{z} - \mathcal{A}(\mathbf{\Sigma})\|_2 \leq \epsilon, \ \text{and} \ \ \mathbf{\Sigma} \ \text{is Toeplitz.}$$

---

Our theoretical results considered $\ell_2$ noise for Toeplitz covariance matrix recovery.

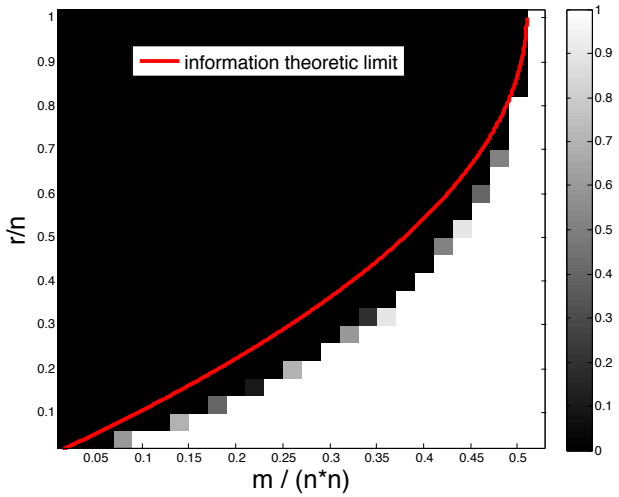
# Near-Optimal Covariance Estimation

**Theorem 1** (Chen, Chi and Goldsmith). *Consider the sub-Gaussian sampling model, then with probability exceeding* $1-\exp(-c_1 m)$, *the solution* $\hat{\mathbf{\Sigma}}$ *satisfies*

$$\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\mathrm{F}} \leq \underbrace{C_1 \frac{\|\mathbf{\Sigma} - \mathbf{\Sigma}_r\|_*}{\sqrt{r}}}_{\text{due to imperfect structure}} + \underbrace{C_2 \frac{\epsilon}{m}}_{\text{due to noise}},$$

*where* $\mathbf{\Sigma}_r$ *is the best rank-r approximation of* $\mathbf{\Sigma}$, *provided that* $m > c_0 nr$ , *where* $c_0$, $c_1$, $C_1$ *and* $C_2$ *are universal constants.*

- **Exact** with $\Theta(nr)$ measurements;
- **Universal** for all low-rank matrices;
- **Robust** against approximate low-rankness and bounded noise;
- Results hold for *i.i.d. bilinear measurements* $\boldsymbol{a}_i^T \mathbf{\Sigma} \boldsymbol{b}_i$ as well.

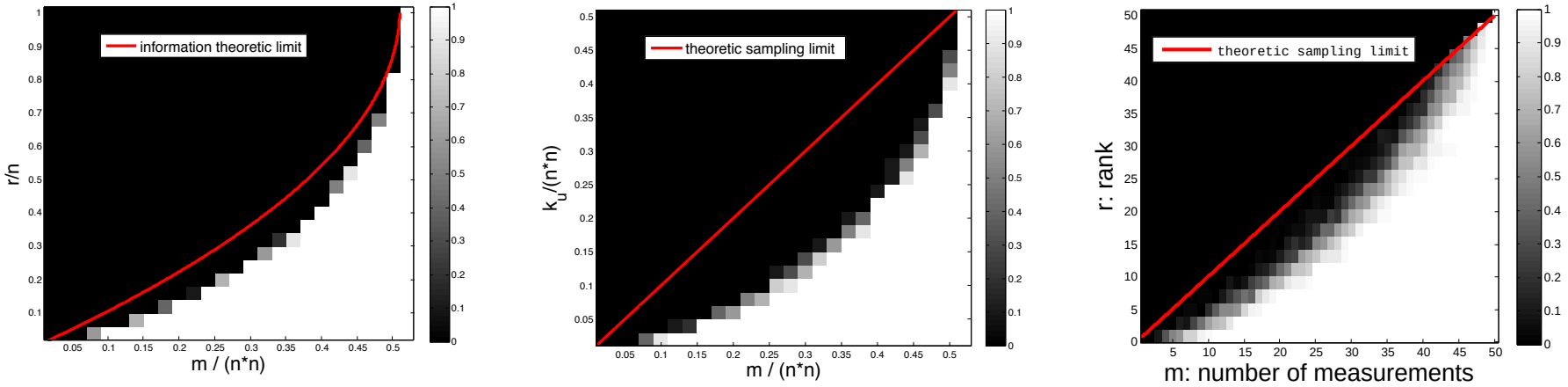



Y. Chen, Y. Chi, and A. J. Goldsmith, "Exact and Stable Covariance Estimation from Quadratic Sampling via Convex Programming," IEEE Trans. on Information Theory, vol. 61, no. 7, pp. 4034-4059, 2015.

# Phase Transition

Our covariance sketching scheme is

- universal
- works with sparse and Toeplitz low-rank covariance matrices as well;
- robust to noise and imperfect model assumptions




(a) Low-rank; (b) Sparse; (c) Toeplitz low-rank

Figure 1: Phase transition

# Proof Ingredient: Mixed-Norm RIP

- **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.*

- We proposed **a Mixed-norm Variant:** **RIP-$\ell_2/\ell_1$**

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

  - *does NOT hold for $\mathcal{A}$, but hold after $\mathcal{A}$ is debiased*:

$$\mathcal{B}_i(\boldsymbol{\Sigma}) = \left\langle \boldsymbol{\Sigma}, \boldsymbol{a}_{2i}\boldsymbol{a}_{2i}^T - \boldsymbol{a}_{2i+1}\boldsymbol{a}_{2i+1}^T \right\rangle$$

E. J. Candès, "The restricted isometry property and its implications for compressed sensing". Compte Rendus de l'Academie des Sciences, Paris, Serie I, 346 589–592.

# Comparisons with Related Work

- low-rank/sparse matrix recovery with linear measurements:

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

  - compressed sensing: $\mathcal{A}_i(\boldsymbol{\Sigma}) = \mathrm{Tr}(\boldsymbol{M}_i\boldsymbol{\Sigma})$
  - matrix completion: $\mathcal{A}_i(\boldsymbol{\Sigma}) = (\boldsymbol{\Sigma})_{i_1 i_2}$
- Sketching sparse matrices [Dasarathy et.al.]:

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T$$

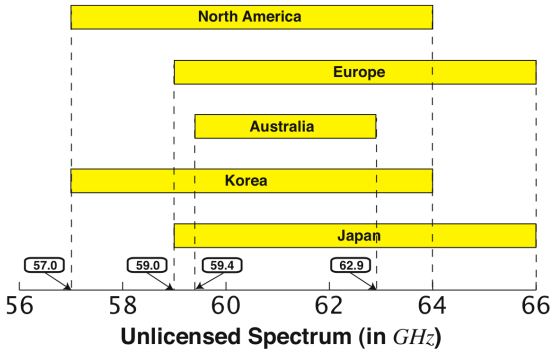
  - $A \in \mathbb{R}^{m\times n}$ with $m < n$, cannot estimate low-rank models;
  - no universal guarantees over sparse models;

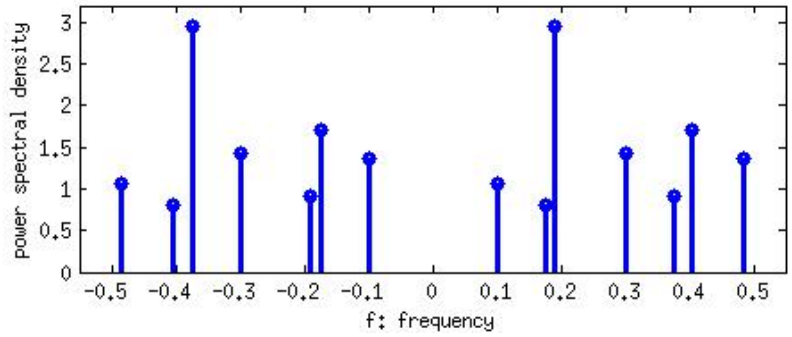
- Phaselift: recover $\boldsymbol{x} \in \mathbb{C}^n$ from $\{|\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle|^2\}_{i=1}^m$.

  - our algorithm is the same form of Phaselift when rank is one.
  - our algorithm extends the best performance guarantees of Phaselift to $O(n)$ $sub\text{-}Gaussian$ measurements.

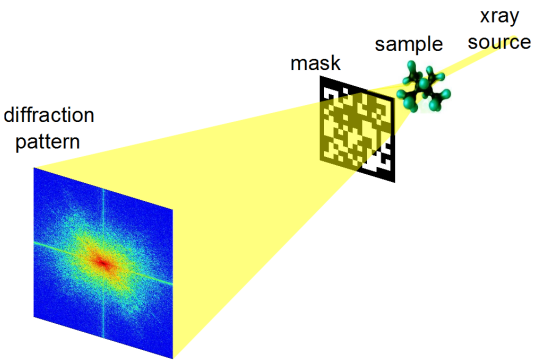
# Other Applications of Quadratic Sensing

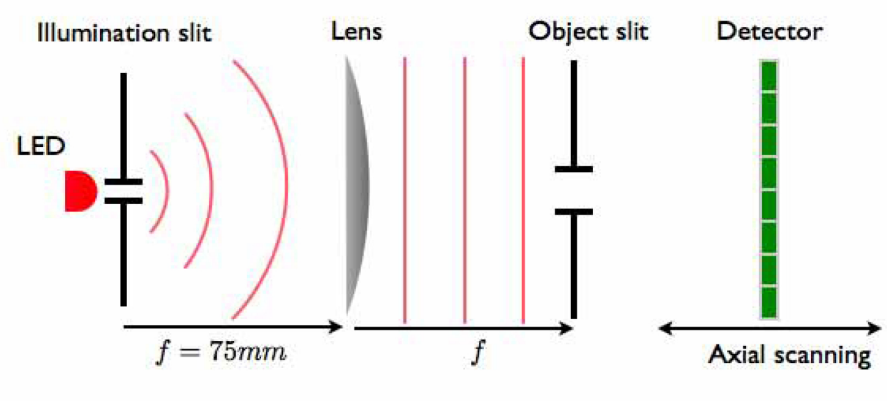
- Energy measurements are often more reliable with high-frequency applications for estimating *power spectral density*.




- Quadratic measurements arise in practical applications such as phase retrieval and phase space tomography.



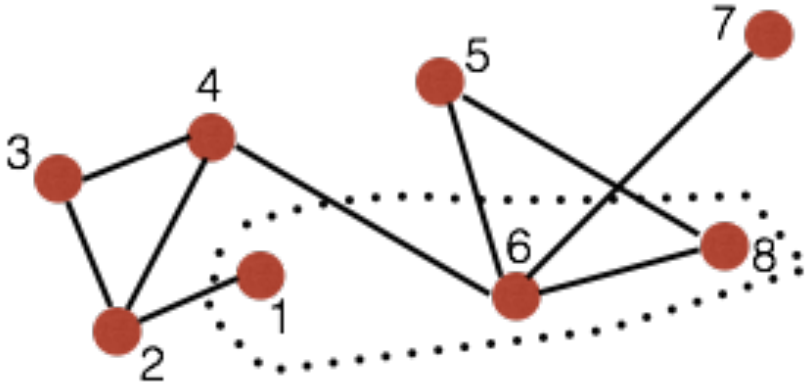

E. J. Candès, Y. C. Eldar, T. Strohmer and V. Voroninski, "Phase retrieval via matrix completion," SIAM J. on Imaging Sciences.
L. Tian, J. Lee, S. Oh, and G. Barbastathis, "Experimental compressive phase space tomography," Opt. Express.

# Graph Sketching

Consider an undirected graph with bounded degree $d$ and number of nodes $n$ with adjacency matrix $\boldsymbol{A}$.

- Define the $i$th sketching vector $\boldsymbol{x}_i \in \{0,1\}^n$ as composed of i.i.d. Bernoulli entries, then the sketch $y_i = \boldsymbol{x}_i^T \boldsymbol{A} \boldsymbol{x}_i$ amounts to counting twice the number of edges in random subgraphs.

- Example: $\boldsymbol{x} = [1,0,0,0,0,1,0,1]^T$ whose support is $\mathcal{I} = \{1,6,8\}$, then $\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} = 2$.




- Our results implies the graph can be perfectly reconstructed from $O(nd)$ quadratic sketches.

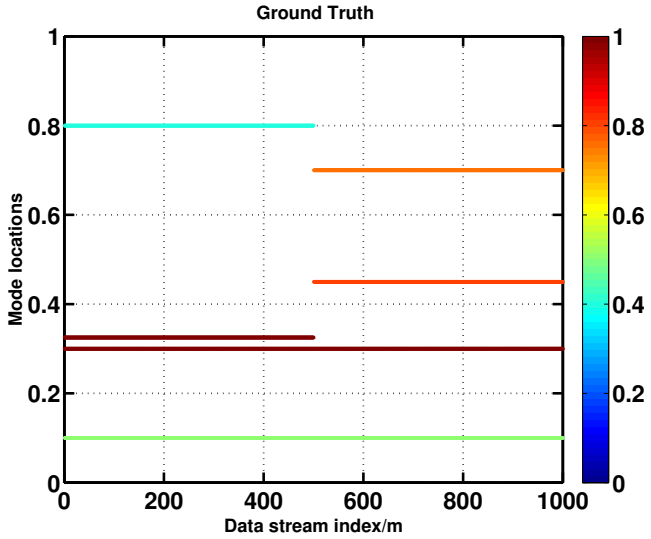
# Covariance Tracking for DOA Estimation

The aggregation step in the sketching scheme can be easily implemented in an online manner to allow tracking.

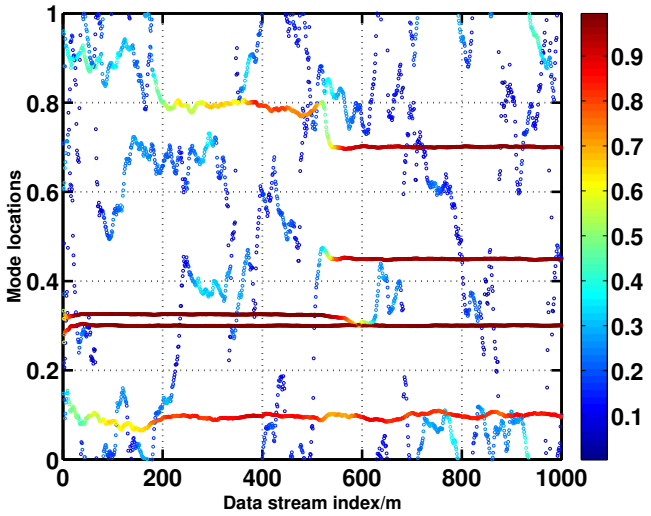
- Aggregation: The aggregates $y_{i,T}$ can be modified with a discounting factor $\lambda$ for tracking:

$$y_{i,T} = \lambda y_{i,T-1} + \left|\langle \boldsymbol{a}_i, \boldsymbol{x}_T \rangle\right|^2$$

- Estimation: replace the trace minimization by a Projection onto Convex Sets (POCS) procedure ($n = 40$, $m = 600$).




(a) Ground Truth

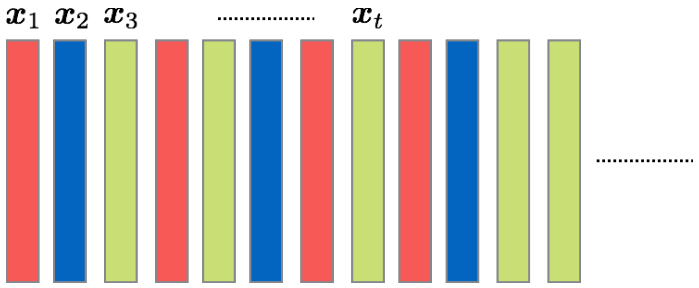



(b) Covariance Tracking

# Can we estimate the covariance from Bits?

Assumption: Let $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{x}_t \boldsymbol{x}_t^H] = \boldsymbol{U}\boldsymbol{U}^H$ be a rank-$r$ matrix with $\boldsymbol{U} \in \mathbb{C}^{n\times r}$.

One-Bit PCA: For each sensor $i = 1, \ldots, m$:




- *randomly* select two sketching vectors $\boldsymbol{a}_i, \boldsymbol{b}_i \in \mathbb{C}^n$ with i.i.d. Gaussian entries;
- Sketch an arbitrary substream indexed by $\{\ell_t^i\}_{t=1}^T$ with two energy measurements $\left|\left\langle \boldsymbol{a}_i, \boldsymbol{x}_{\ell_t^i} \right\rangle\right|^2$, $\left|\left\langle \boldsymbol{b}_i, \boldsymbol{x}_{\ell_t^i} \right\rangle\right|^2$, and transmit a binary bit indicating the energy comparison outcome to the fusion center:

$$y_{i,T} = \text{sign}\left( \frac{1}{T}\sum_{t=1}^{T} \left|\left\langle \boldsymbol{a}_i, \boldsymbol{x}_{\ell_t^i} \right\rangle\right|^2 - \frac{1}{T}\sum_{t=1}^{T} \left|\left\langle \boldsymbol{b}_i, \boldsymbol{x}_{\ell_t^i} \right\rangle\right|^2 \right)$$

- **Estimation:** The fusion center recovers the principal components $\hat{\boldsymbol{U}} \in \mathbb{R}^{n\times r}$ by computing the top $r$ eigenvectors of the surrogate matrix:

$$\boldsymbol{J}_m = \frac{1}{m}\sum_{i=1}^{m} y_{i,T}\left(\boldsymbol{a}_i\boldsymbol{a}_i^H - \boldsymbol{b}_i\boldsymbol{b}_i^H\right).$$

# Bit Comparisons are Robust

- With finite samples, the *numerical* energy difference measures the sample covariance $\boldsymbol{\Sigma}_T$:

$$z_{i,T} = \frac{1}{T}\sum_{t=1}^{T}\left|\left\langle \boldsymbol{a}_i, \boldsymbol{x}_{\ell_t^i}\right\rangle\right|^2 - \frac{1}{T}\sum_{t=1}^{T}\left|\left\langle \boldsymbol{b}_i, \boldsymbol{x}_{\ell_t^i}\right\rangle\right|^2 = \left\langle \boldsymbol{\Sigma}_T, \boldsymbol{a}_i\boldsymbol{a}_i^H - \boldsymbol{b}_i\boldsymbol{b}_i^H\right\rangle.$$

  The discrepancy $z_{i,T} - z_i = \langle \boldsymbol{\Sigma}_T - \boldsymbol{\Sigma}, \boldsymbol{a}_i\boldsymbol{a}_i^H - \boldsymbol{b}_i\boldsymbol{b}_i^H\rangle \neq 0$.

- The *ordinal* energy difference measures **the exact covariance** $\boldsymbol{\Sigma}$ with high probability as soon as $T$ is not too small:

$$y_{i,T} = \text{sign}\left(\left\langle \boldsymbol{\Sigma}_T, \boldsymbol{a}_i\boldsymbol{a}_i^H - \boldsymbol{b}_i\boldsymbol{b}_i^H\right\rangle\right) = \text{sign}\left(\left\langle \boldsymbol{\Sigma}, \boldsymbol{a}_i\boldsymbol{a}_i^H - \boldsymbol{b}_i\boldsymbol{b}_i^H\right\rangle\right) = y_i.$$

**Theorem 2** (Chi 2014). *Let $\boldsymbol{x}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$. Let $0 < \delta \leq 1$, then with probability at least $1 - \delta$ all bit measurements are exact, given that the number of samples observed by each sensor satisfies $T > cTr(\boldsymbol{\Sigma})/\|\boldsymbol{\Sigma}\|_F \log\left(\frac{m}{\delta}\right)$ for some sufficiently large constant $c$.*

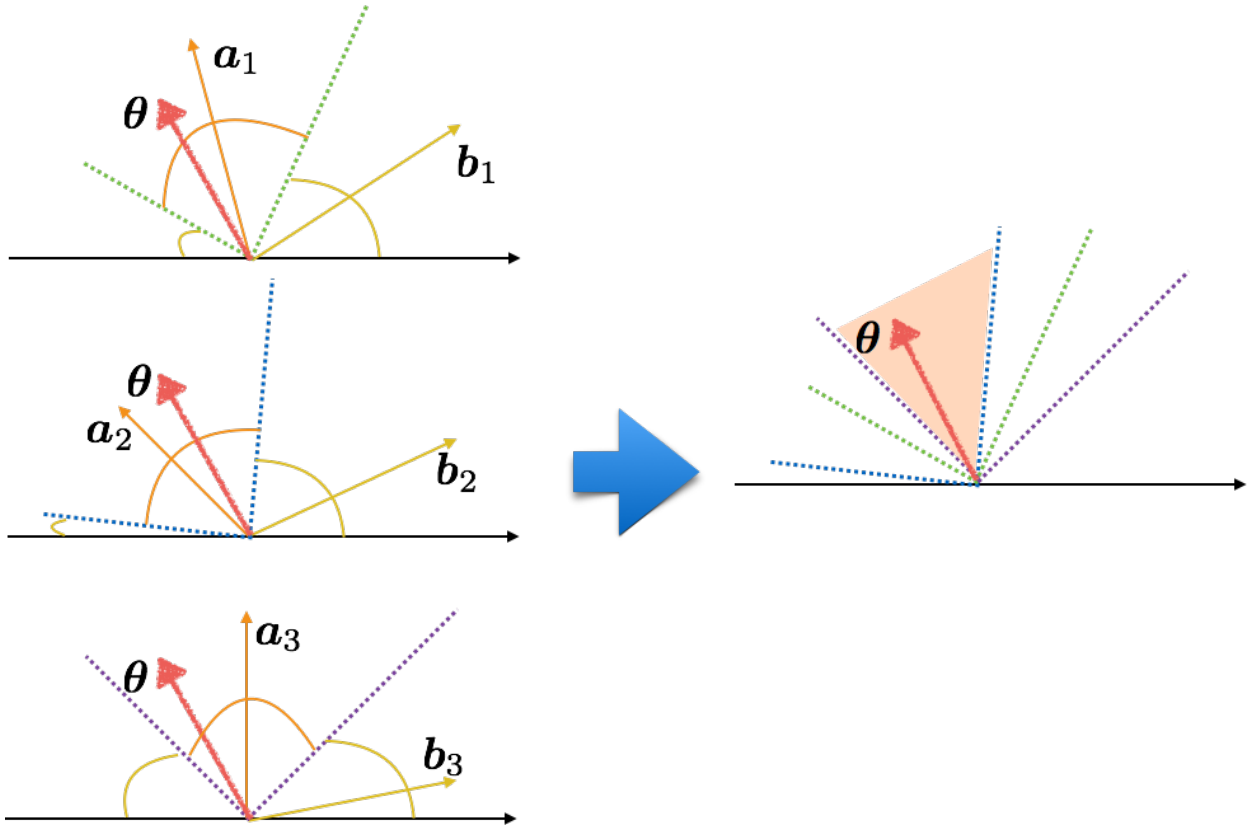
---

Y. Chi, "One-Bit Principal Subspace Estimation", IEEE Global Conference on Signal and Information Processing (GlobalSIP), Atlanta, GA, Dec. 2014.

# One-Bit PCA: Why does it work?

Consider a rank-one example $\boldsymbol{\Sigma} = \boldsymbol{\theta}\boldsymbol{\theta}^H$ with the eigenvector $\boldsymbol{\theta} \in \mathbb{C}^2$:

- Each bit $y_i = \text{sign}(|\langle \boldsymbol{a}_i, \boldsymbol{\theta} \rangle|^2 - |\langle \boldsymbol{b}_i, \boldsymbol{\theta} \rangle|^2)$ selects the halfspace towards the direction with a smaller angle with either $\boldsymbol{a}_i$ or $\boldsymbol{b}_i$.
- With enough bit measurements, we can trap the eigenvector $\boldsymbol{\theta}$ accurately up to a sign difference.



# One-Bit PCA: Performance Guarantee

Conditioned on that the bit measurements are exact, the principal subspace of $\boldsymbol{J}_m$ agrees with $\boldsymbol{U}$ with high probability given $m$ is sufficiently large.

**Theorem 3** (Chi 2014). *Denote $\boldsymbol{U} \in \mathbb{C}^{n\times r}$ as the principal subspace of $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{U}}$ as the principal subspace of $\boldsymbol{J}_m = \frac{1}{m}\sum_{i=1}^m y_i\left(\boldsymbol{a}_i\boldsymbol{a}_i^H - \boldsymbol{b}_i\boldsymbol{b}_i^H\right)$. Let $0<\delta<1$, then with probability at least $1-\delta$, there exists an $r\times r$ orthogonal matrix $\boldsymbol{Q}$ such that*

$$\|\hat{\boldsymbol{U}} - \boldsymbol{U}\boldsymbol{Q}\|_F \le c_1\sqrt{\frac{nr^2}{m}\log\left(\frac{2n}{\delta}\right)}$$

*for all rank-$r$ matrices $\boldsymbol{\Sigma}$, where $c_1$ is an absolute constant depending on $r$.*

- The subspace estimate is accurate as soon as $m = \Theta(nr^2\log n)$ which is near-optimal as the subspace requires at least $n$ measurements.

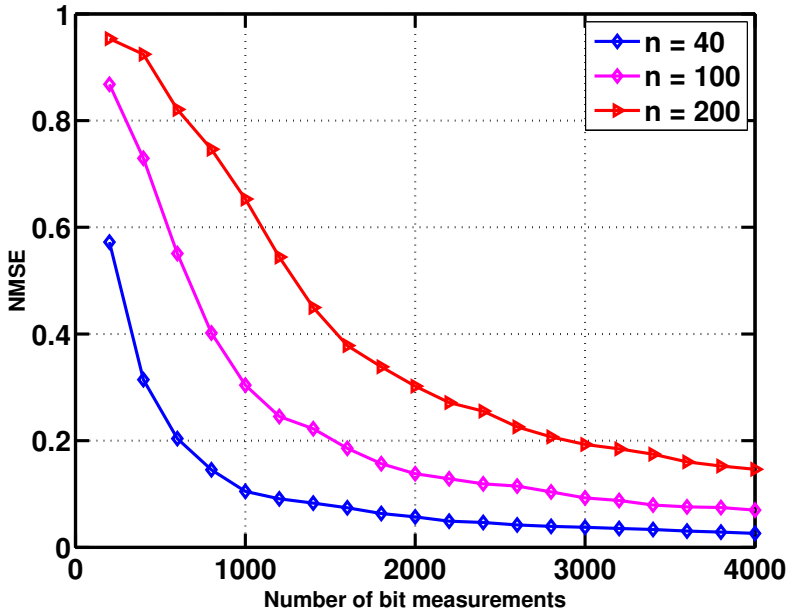
- Not an exact recovery guarantee.

---

Y. Chi, "Principal Subspace Estimation and Tracking From One-Bit Energy Measurements", IEEE Global Conference on Signal and Information Processing (GlobalSIP), Atlanta, GA, Dec. 2014, submitted.

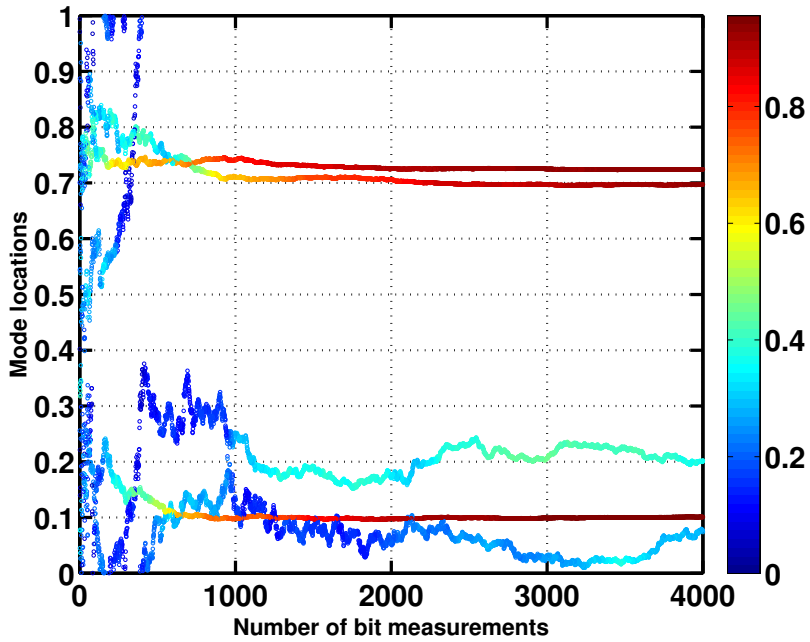
# How many bits do we need?

- We generate the covariance matrix as $\boldsymbol{\Sigma} = \boldsymbol{X}\boldsymbol{X}^T$, where $\boldsymbol{X} \in \mathbb{R}^{n\times 3}$ is composed of standard Gaussian entries. The sketching vectors $a_i$'s and $b_i$'s are also generated with standard Gaussian entries.
- The estimate $\hat{\boldsymbol{X}}$ is calculated via computing the top eigenvectors of $\boldsymbol{J}_m$.
- The error metric is calculated as $\|\boldsymbol{P}_{\hat{\boldsymbol{X}}^\perp}\boldsymbol{X}\|_{\mathrm{F}}^2/\|\boldsymbol{X}\|_{\mathrm{F}}^2$.



# Online DOA Estimation with Bit Measurements

- The covariance matrix $\mathbf{\Sigma}$ is a low-rank Toeplitz PSD matrix with $n = 40$ and $r = 3$. The set of modes is $\mathcal{F} = [0.1, 0.7, 0.725]$ (notice the last two modes are separated by the Rayleigh limit $1/n$), and their variance is $\sigma^2 = 1$.
- The subspace is updated online as new bit measurements arrive sequentially; and ESPRIT is applied to estimate $5$ modes using the subspace estimate.




1000 bits successfully distinguish two close modes separated by the Rayleigh limit.

# Recap: Comparing the Two Schemes

Covariance Sketching with Real Measurements:

- *Near-Optimal Sample Complexity* for a variety of covariance structures;
- *Energy* measurements are easier to obtain;
- Requires a *noise estimate* to performance the algorithm with finite sample size and additive noise;
- The convex optimization might still be computationally expensive;

Covariance Sketching with One-bit Measurements:

- Communication overhead is minimized with bit measurements;
- Simple algorithm via computing the top eigenvectors;
- Robust measurement with respect to (possibly heterogeneous) noise.

# Summary and Future Directions

## Summary:

- Covariance estimation is possible without observing and reconstructing the whole data stream.
- The sensing and estimation procedure can be jointly designed to minimize complexity by leveraging the low-dimensional structures of data.
- Many potential applications in network traffic monitoring, video surveillance, and covariance estimation in privacy-aware and crowdsourcing environments.

## Future Directions:

- Quality-Quantity-Computation Complexity Trade-offs for statistical inference.

# List of Related Publications


- Y. Chen, Y. Chi and A. J. Goldsmith, "Universal and Robust Covariance Estimation via Convex Programming," in International Symposium on Information Theory (ISIT), Honolulu, HI, Jun. 2014.
- Y. Chen, Y. Chi and A. J. Goldsmith, "Estimation of Simultaneously Structured Covariance Matrices from Quadratic Measurements," in International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Florence, Italy, May 2014.
- Y. Chen, Y. Chi, and A. J. Goldsmith, "Exact and Stable Covariance Estimation from Quadratic Sampling via Convex Programming," IEEE Trans. on Information Theory, vol. 61, no. 7, pp. 4034-4059, 2015.
- Y. Chi, "One-Bit Principal Subspace Estimation", IEEE Global Conference on Signal and Information Processing (GlobalSIP), Atlanta, GA, Dec. 2014.
- Y. Jiang and Y. Chi, "Covariance Tracking from Sketches of Rapid Data Streams", in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brisbane, Australia, April 2015.
- Y. Chi, "Compressive Graph Clustering via Semidefinite Programming," *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brisbane, Australia, April 2015.