Sample Efficient Toeplitz Covariance Estimation

Christopher Musco (New York University)

With Yonina Eldar (Weizmann Institute), Jerry Li (Microsft Research), and Cameron Musco (UMass Amherst).

The second simplest statistical problem:

How many samples $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^d \sim \mathcal{D}$ required to learn covariance matrix $C = \mathbb{E}_{x \sim \mathcal{D}}[xx^T]$?

- $C \in \mathbb{R}^{d \times d}$. $C_{j,k}$ is the covariance between x_j and x_k.

The second simplest statistical problem:

How many samples $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^d \sim \mathcal{D}$ required to learn covariance matrix $C = \mathbb{E}_{x \sim \mathcal{D}}[xx^T]$?

- $C \in \mathbb{R}^{d \times d}$. $C_{j,k}$ is the covariance between x_j and x_k.

Reasonable goal: Find $\tilde{C}$ with $\|C - \tilde{C}\|_2 \leq \epsilon \|C\|_2$.[1]

[1] Lots of other possible metrics.

Assuming $\mathcal{D}$ is Gaussian, subgaussian, subexponential:

Assuming $\mathcal{D}$ is Gaussian, subgaussian, subexponential:

Known bound: $n = \Theta\left(\frac{d}{\epsilon^2}\right)$ samples are necessary and sufficient.

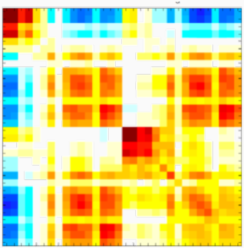
Estimator: Simple sample covariance.

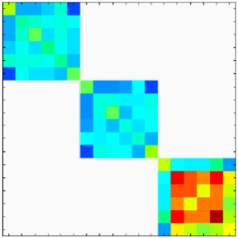
$$\tilde{C} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)} x^{(i)T}.$$

Analysis: Standard matrix concentration (e.g., Vershynin, 2019).

What is we know C has additional structure?

- Block structure.
- Low-rank, low-rank + diagonal.
- Diagonal, banded.
- Many other possibilities.

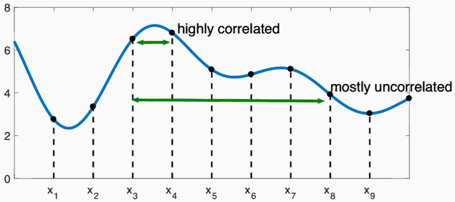




This work: Covariance matrix is Toeplitz.

$$T = \begin{bmatrix} a & b & c & d & e \\ b & a & b & c & d \\ c & b & a & b & c \\ d & c & b & a & b \\ e & d & c & b & a \end{bmatrix}$$

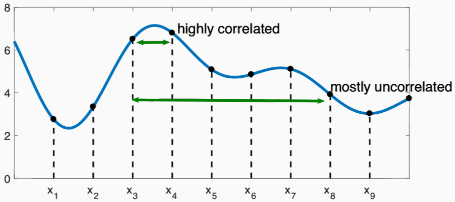
Arises when measurements taken on a spatial or temporal grid.
Covariance depends on distance between them: $\mathbb{E}[x_j \cdot x_k] = f(|j-k|)$.

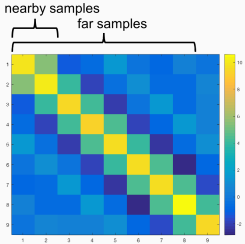


TOEPLITZ COVARIANCE ESTIMATION

Arises when measurements taken on a spatial or temporal grid.

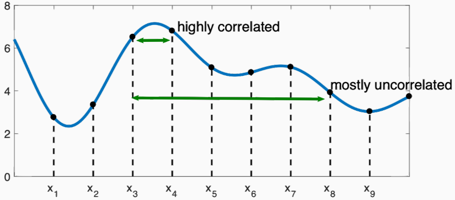
Covariance depends on distance between them: $\mathbb{E}[x_j \cdot x_k] = f(|j-k|)$.

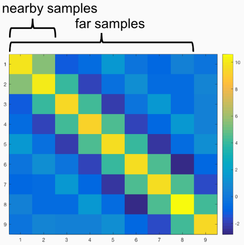




TOEPLITZ COVARIANCE ESTIMATION

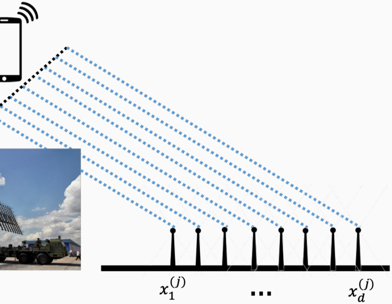
Arises when measurements taken on a spatial or temporal grid.
Covariance depends on distance between them: $\mathbb{E}[x_j \cdot x_k] = f(|j-k|)$.

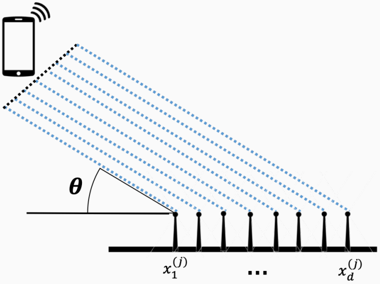




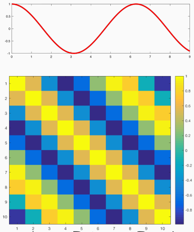
Applications in signal processing: spectrum sensing/cognitive radio, radar, prediction via Gaussian process regression, kriging etc.

EXAMPLE: DIRECTION OF ARRIVAL ESTIMATION

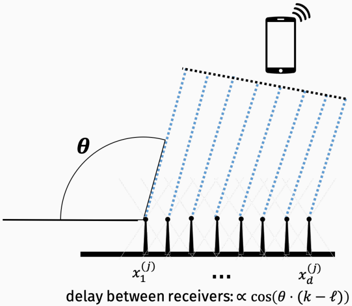


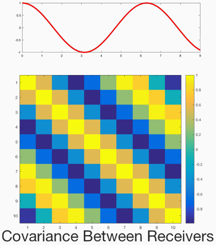


delay between receivers: $\propto \cos(\theta \cdot (k - \ell))$

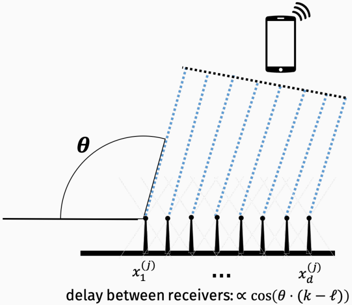


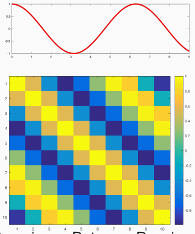
Covariance Between Receivers





Can back out direction of arrival θ from covariance structure.





Covariance Between Receivers

Can back out direction of arrival θ from covariance structure.

Additional structure: When just one transmitter, T is rank 1. When k transmitters, T is rank k.

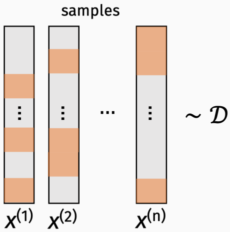
Goal: Minimize two types of sample complexity:

Goal: Minimize two types of sample complexity:

- **Vector sample complexity:** How many samples $x^{(1)}, \ldots, x^{(n)} \sim \mathcal{D}$ are required to estimate T?

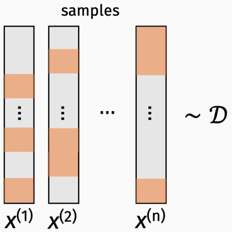
Goal: Minimize two types of sample complexity:

- **Vector sample complexity:** How many samples $x^{(1)}, \ldots, x^{(n)} \sim \mathcal{D}$ are required to estimate T?
- **Entry sample complexity:** How many entries s must be read from each sample $x^{(1)}, \ldots, x^{(n)}$?

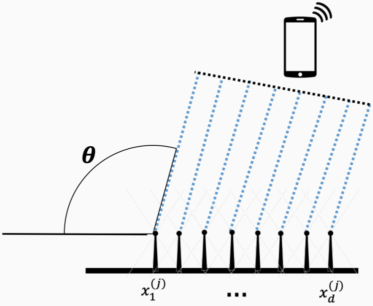


Goal: Minimize two types of sample complexity:

- **Vector sample complexity:** How many samples $x^{(1)}, \ldots, x^{(n)} \sim \mathcal{D}$ are required to estimate T?
- **Entry sample complexity:** How many entries s must be read from each sample $x^{(1)}, \ldots, x^{(n)}$?

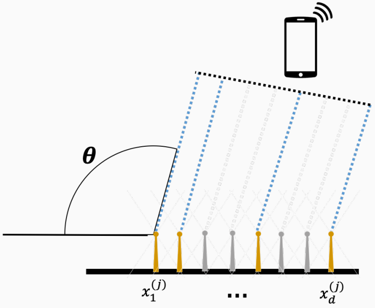


In different applications, these complexities correspond to different costs. Typically there is a tradeoff.



- **Vector sample complexity:** Estimation time (# snapshots).

EXAMPLE: DIRECTION OF ARRIVAL ESTIMATION



- **Vector sample complexity:** Estimation time (# snapshots).
- **Entry sample complexity:** Number of active receivers.

Total sample complexity: Total number of entries read, $n \cdot s$.

Total sample complexity: Total number of entries read, $n \cdot s$.

- For general covariance matrices, vector sample complexity is $\Theta(d/\epsilon^2)$, entry sample complexity is d, so total sample complexity is $\tilde{\Theta}(d^2/\epsilon^2)$.

Current state: Many algorithms for Toeplitz covariance estimation, but few formal results on sample complexities/tradeoffs.

Current state: Many algorithms for Toeplitz covariance estimation, but few formal results on sample complexities/tradeoffs.

Our contributions:

OUR CONTRIBUTIONS

Current state: Many algorithms for Toeplitz covariance estimation, but few formal results on sample complexities/tradeoffs.

Our contributions:

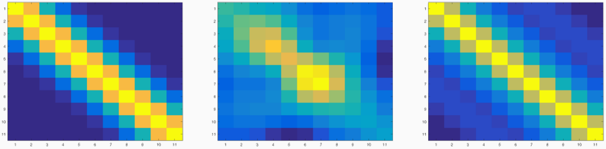
- Non-asymptotic sample complexity bounds by analyzing classic algorithms, including those with sublinear entry sample complexity based on sparse ruler measurements.

Current state: Many algorithms for Toeplitz covariance estimation, but few formal results on sample complexities/tradeoffs.

Our contributions:

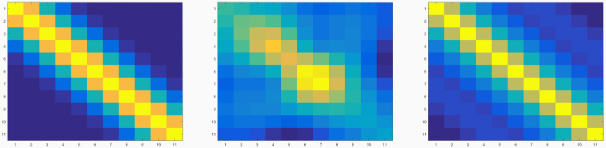
- Non-asymptotic sample complexity bounds by analyzing classic algorithms, including those with sublinear entry sample complexity based on sparse ruler measurements.
- Develop improved algorithms for the case when T is (approximately) low-rank, using techniques from matrix sketching, leverage score-based sampling, and sparse Fourier transform algorithms.

Estimator: $\tilde{T} = \mathrm{avg}\left(\frac{1}{n}\sum x^{(j)}x^{(j)^T}\right)$



True covariance T Empirical covariance $\hat{T}$ Improved estimator $avg(\hat{T})$

Estimator: $\tilde{T} = \operatorname{avg}\left(\frac{1}{n}\sum x^{(j)}{x^{(j)}}^T\right)$



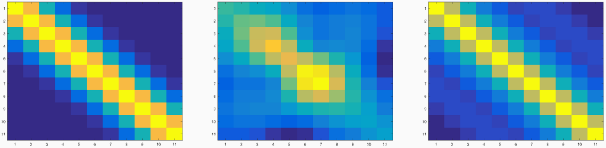
True covariance T — Empirical covariance $\hat{T}$ — Improved estimator $avg(\hat{T})$

- **Vector sample complexity**: $O(\log^2 d/\epsilon^2)$

[2] All assuming Gaussian or sub-Gaussian distribution.

A FIRST RESULT

Estimator: $\tilde{T} = \mathrm{avg}\left(\frac{1}{n}\sum x^{(j)}{x^{(j)}}^T\right)$



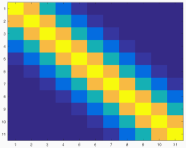
True covariance T Empirical covariance $\hat{T}$ Improved estimator $avg(\hat{T})$

- **Vector sample complexity**: $O(\log^2 d/\epsilon^2)$
- **Entry sample complexity**: d.
- **Total sample complexity**: $O(d \log^2 d/\epsilon^2)$.[2]

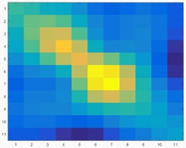
[2] All assuming Gaussian or sub-Gaussian distribution.

A FIRST RESULT

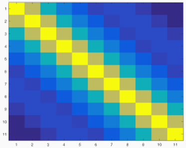
Estimator: $\tilde{T} = \operatorname{avg}\left(\frac{1}{n}\sum x^{(j)}x^{(j)^T}\right)$



True covariance T



Empirical covariance $\hat{T}$



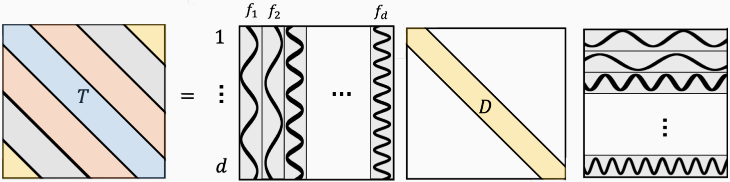
Improved estimator $avg(\hat{T})$

- **Vector sample complexity**: $O(\log^2 d/\epsilon^2)$
- **Entry sample complexity**: d.
- **Total sample complexity**: $O(d\log^2 d/\epsilon^2)$.[2]

Improves over $\tilde{O}(d^2/\epsilon^2)$ for generic covariance matrices.

[2] All assuming Gaussian or sub-Gaussian distribution.

Vandermonde Decomposition: Any Toeplitz $T \in R^{d \times d}$ can be written as $F_S D F_S$ where $F_S \in \mathbb{R}^{d \times d}$ is an 'off-grid' Fourier matrix with frequencies $f_1, \ldots, f_d \in [0, 1]$ and D is a positive diagonal matrix.

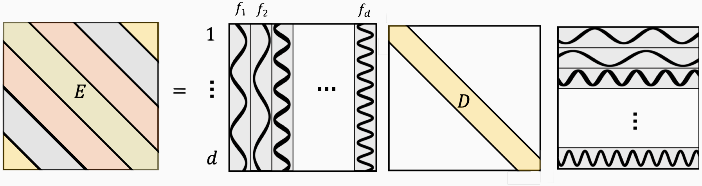


$$F_S(j, k) = \exp\left(-2\pi\sqrt{-1} \cdot j \cdot f_k\right)$$

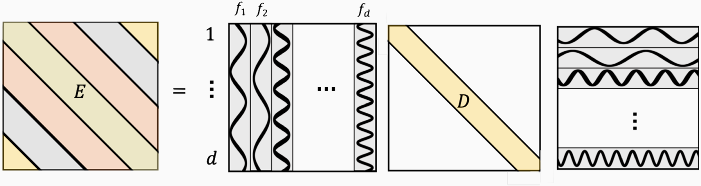
VERY ROUGH PROOF IDEA

Let $\hat{T} = \frac{1}{n} \sum x^{(j)} x^{(j)^T}$. $\quad \tilde{T} = \mathrm{avg}\left(\hat{T}\right)$. $\quad E = T - \tilde{T}$.

Let $\hat{T} = \frac{1}{n} \sum x^{(j)} x^{(j)^T}$. $\tilde{T} = \operatorname{avg}\left(\hat{T}\right)$. $E = T - \tilde{T}$.

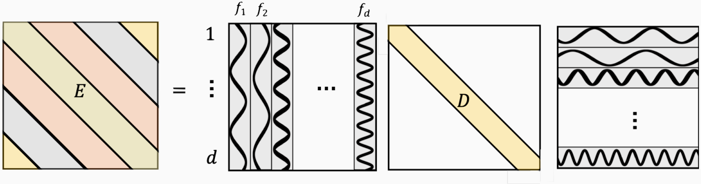


Let $\hat{T} = \frac{1}{n}\sum x^{(j)}{x^{(j)}}^T$. $\quad \tilde{T} = \mathrm{avg}\left(\hat{T}\right)$. $\quad E = T - \tilde{T}$.



- Roughly, to bound $\|E\|_2 = \max_{\|z\|_2=1} |z^T E z|$, it suffices to bound $|f_j^T E f_j|$. Obvious if $f_1, \ldots, f_d$ where eigenvectors of E, but they aren't.

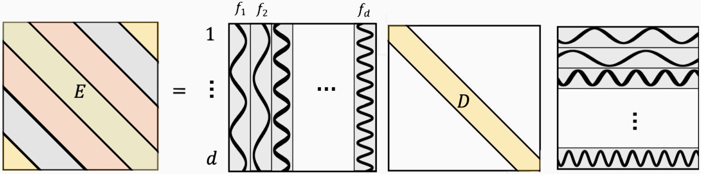
Let $\hat{T} = \frac{1}{n}\sum x^{(j)}x^{(j)T}$. $\qquad \tilde{T} = \text{avg}\left(\hat{T}\right)$. $\qquad E = T - \tilde{T}$.



- Roughly, to bound $\|E\|_2 = \max_{\|z\|_2=1} |z^T E z|$, it suffices to bound $|f_j^T E f_j|$. Obvious if $f_1, \ldots, f_d$ where eigenvectors of E, but they aren't.
- Argue that $|f_j^T(T - \tilde{T})f_j| = |f_j^T(T - \hat{T})f_j| \leq \epsilon\|T\|_2$ for all j using standard matrix concentration (Hanson-Wright inequality) + ϵ-net over frequencies in $[0, 1]$ + union bound.

VERY ROUGH PROOF IDEA

Let $\hat{T} = \frac{1}{n}\sum x^{(j)}x^{(j)T}$. $\tilde{T} = \text{avg}\left(\hat{T}\right)$. $E = T - \tilde{T}$.



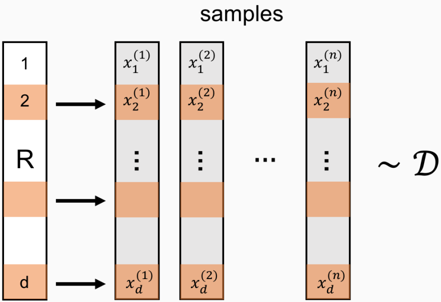
- Roughly, to bound $\|E\|_2 = \max_{\|z\|_2=1} |z^T E z|$, it suffices to bound $|f_j^T E f_j|$. Obvious if $f_1, \ldots, f_d$ where eigenvectors of E, but they aren't.
- Argue that $|f_j^T(T - \tilde{T})f_j| = |f_j^T(T - \hat{T})f_j| \leq \epsilon\|T\|_2$ for all j using standard matrix concentration (Hanson-Wright inequality) + ϵ-net over frequencies in $[0, 1]$ + union bound.

Question: Can $O(\log^2 d)$ samples be improved to $O(\log d)$?

IMPROVING ENTRY SAMPLE COMPLEXITY

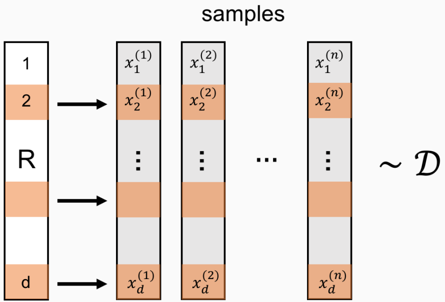
Consider algorithms that sample $x^{(1)}, \ldots, x^{(n)} \sim \mathcal{D}$ and read a fixed subset of entries $R \subseteq [d]$ from each $x^{(j)}$.

Approximate T using $x_R^{(1)}, \ldots, x_R^{(n)} \in \mathbb{R}^{|R|}$.



Entry sample complexity: $|R|$. Total sample complexity: $|R| \cdot n$.

Consider algorithms that sample $x^{(1)}, \ldots, x^{(n)} \sim \mathcal{D}$ and read a fixed subset of entries $R \subseteq [d]$ from each $x^{(j)}$.
Approximate T using $x_R^{(1)}, \ldots, x_R^{(n)} \in \mathbb{R}^{|R|}$.



Entry sample complexity: $|R|$. Total sample complexity: $|R| \cdot n$.

Only get information about $\operatorname{cov}(x_j, x_k)$ for subset of pairs j, k.

How small can R be if T is Toeplitz?

How small can R be if T is Toeplitz? Can take advantage of redundancy.

$$T = \begin{bmatrix} a_0 & a_1 & a_2 & \cdots & a_{d-2} & a_{d-1} \\ a_1 & a_0 & a_1 & \cdots & \cdots & a_{d-2} \\ a_2 & \boxed{a_1} & a_0 & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{d-2} & \cdots & \cdots & \cdots & \cdots & a_1 \\ a_{d-1} & a_{d-2} & \cdots & \cdots & \boxed{a_1} & a_0 \end{bmatrix}$$

How small can R be if T is Toeplitz? Can take advantage of redundancy.

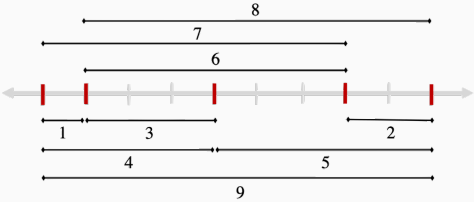
$$T = \begin{bmatrix} a_0 & a_1 & a_2 & \cdots & a_{d-2} & a_{d-1} \\ a_1 & a_0 & a_1 & \cdots & \cdots & a_{d-2} \\ a_2 & a_1 & a_0 & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{d-2} & \cdots & \cdots & \cdots & \cdots & a_1 \\ a_{d-1} & a_{d-2} & \cdots & \cdots & a_1 & a_0 \end{bmatrix}$$

- $a_1 = \mathbb{E}[x_2 \cdot x_3] = \mathbb{E}[x_d \cdot x_{d-1}]$.

Definition (Ruler) A subset $R \subseteq [d]$ is a ruler if for every distance $s \in \{0, \ldots, d-1\}$, there exist $j, k \in R$ with $j - k = s$.

Definition (Ruler) A subset $R \subseteq [d]$ is a ruler if for every distance $s \in \{0, \ldots, d-1\}$, there exist $j, k \in R$ with $j - k = s$.

E.g., for $d = 10$, $R = \{1, 2, 5, 8, 10\}$ is a ruler.



$$T = \begin{bmatrix} a_0 & a_1 & a_2 & \cdots & a_{d-2} & a_{d-1} \\ a_1 & a_0 & a_1 & \cdots & \cdots & a_{d-2} \\ a_2 & a_1 & a_0 & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{d-2} & \cdots & \cdots & \cdots & \cdots & a_1 \\ a_{d-1} & a_{d-2} & \cdots & \cdots & a_1 & a_0 \end{bmatrix}$$

- If R is a ruler, for each $s \in \{0, \ldots, d-1\}$, there is at least one $k, \ell \in R$ with $|k - \ell| = s$ and thus with covariance
$$\mathbb{E}[x_k^{(j)} \cdot x_\ell^{(j)}] = a_s.$$

$$T = \begin{bmatrix} a_0 & a_1 & \boxed{a_2} & \cdots & a_{d-2} & \boxed{a_{d-1}} \\ a_1 & a_0 & a_1 & \cdots & \cdots & a_{d-2} \\ a_2 & \boxed{a_1} & \boxed{a_0} & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{d-2} & \cdots & \cdots & \cdots & \cdots & a_1 \\ a_{d-1} & \boxed{a_{d-2}} & \cdots & \cdots & a_1 & a_0 \end{bmatrix}$$

- If R is a ruler, for each $s \in \{0, \ldots, d-1\}$, there is at least one $k, \ell \in R$ with $|k - \ell| = s$ and thus with covariance
$$\mathbb{E}[x_k^{(j)} \cdot x_\ell^{(j)}] = a_s.$$
- Get at least one independent sample of a_s from every $x_R^{(j)}$.

$$T = \begin{bmatrix} a_0 & a_1 & a_2 & \cdots & a_{d-2} & a_{d-1} \\ a_1 & a_0 & a_1 & \cdots & \cdots & a_{d-2} \\ a_2 & a_1 & a_0 & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{d-2} & \cdots & \cdots & \cdots & \cdots & a_1 \\ a_{d-1} & a_{d-2} & \cdots & \cdots & a_1 & a_0 \end{bmatrix}$$

- If R is a ruler, for each $s \in \{0, \ldots, d-1\}$, there is at least one $k, \ell \in R$ with $|k - \ell| = s$ and thus with covariance

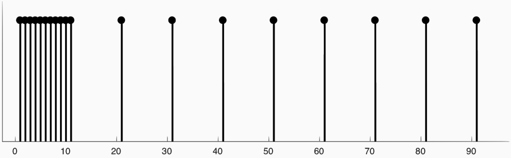
$$\mathbb{E}[x_k^{(j)} \cdot x_\ell^{(j)}] = a_s.$$

- Get at least one independent sample of a_s from every $x_R^{(j)}$.
- With enough samples from $\mathcal{D}$, can estimate each a_s to high accuracy, and thus get an estimate for T.

Claim: For any d there exists a sparse ruler R with $|R| = 2\sqrt{d}$

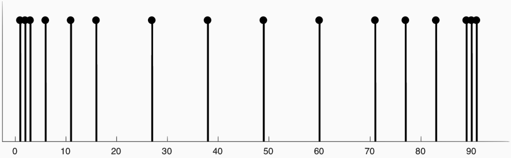
Claim: For any d there exists a sparse ruler R with $|R| = 2\sqrt{d}$

- Suffices to take $R = [1, 2, \ldots, \sqrt{d}] \cup [2\sqrt{d}, 3\sqrt{d}, \ldots, d]$.



Claim: For any d there exists a sparse ruler R with $|R| = 2\sqrt{d}$

- Suffices to take $R = [1, 2, \ldots, \sqrt{d}] \cup [2\sqrt{d}, 3\sqrt{d}, \ldots, d]$.



- Best possible leading constant is between $\sqrt{2 + \frac{4}{3\pi}}$ and $\sqrt{8/3}$ (Erdös, Gal, Leech, ‘48, ‘56)

How many vector samples do we need? What do we pay for the optimal entry sample complexity of sparse rulers?

How many vector samples do we need? What do we pay for the optimal entry sample complexity of sparse rulers?

We prove:

- Upper bound: $\tilde{O}(d)$ vector samples.
- Lower bound: $O(d)$ vector samples.

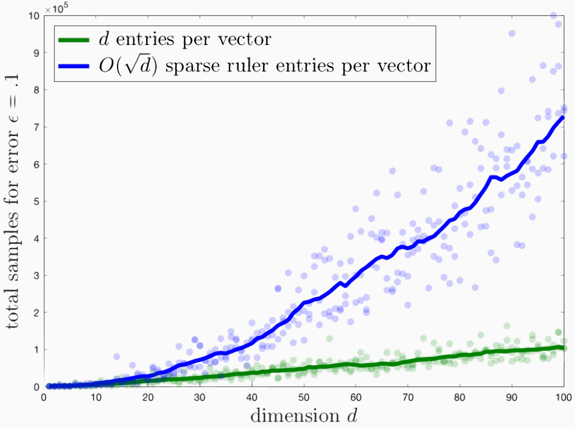
How many vector samples do we need? What do we pay for the optimal entry sample complexity of sparse rulers?

We prove:

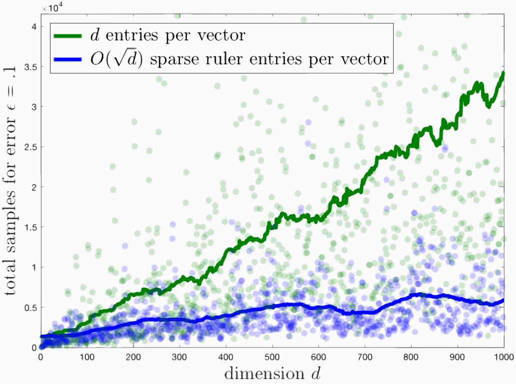
- Upper bound: $\tilde{O}(d)$ vector samples.
- Lower bound: $O(d)$ vector samples.

Recall that $O(\log^2 d)$ samples were possible when reading all entries of each sample.

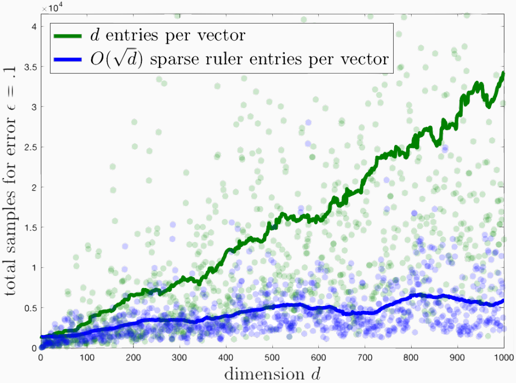
Total sample complexity is $O(\sqrt{d}) \cdot \tilde{O}(d) = \tilde{O}(d^{3/2})$ for sparse ruler vs. $d \cdot \tilde{O}(1) = \tilde{O}(d)$ for full sample estimation.



NOT WHATS OBSERVED IN PRACTICE...



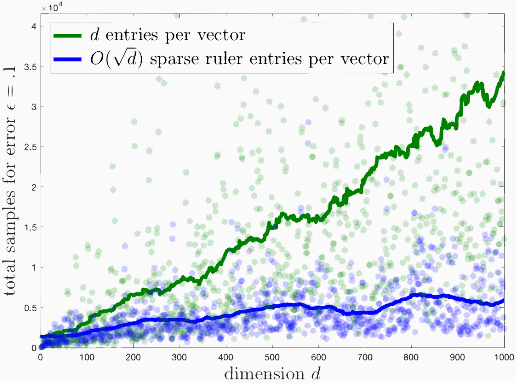
NOT WHATS OBSERVED IN PRACTICE...

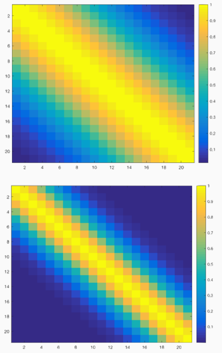


- Total sample complexity appears to be $\tilde{O}(\sqrt{d})$ for sparse rulers vs. $\tilde{O}(d)$ for full samples.

NOT WHATS OBSERVED IN PRACTICE...

Sparse rulers give much better total sample complexity when T is (approximately) low-rank.





- Total sample complexity appears to be $\tilde{O}(\sqrt{d})$ for sparse rulers vs. $\tilde{O}(d)$ for full samples.

How many vector samples do we need when T is (approximately) rank k and samples are collected with a $O(\sqrt{d})$-sparse ruler?

How many vector samples do we need when T is (approximately) rank k and samples are collected with a $O(\sqrt{d})$-sparse ruler?

We prove:

- Upper bound: $\tilde{O}\left(k^2\right)$ vector samples.
- Lower bound: $O\left(k\right)$ vector samples.

How many vector samples do we need when T is (approximately) rank k and samples are collected with a $O(\sqrt{d})$-sparse ruler?

We prove:

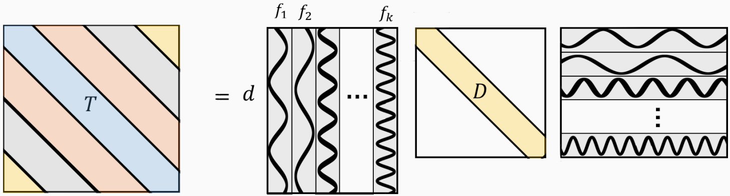
- Upper bound: $\tilde{O}\left(k^2\right)$ vector samples.
- Lower bound: $O\left(k\right)$ vector samples.

Take-away: Sublinear total sample complexity $\tilde{O}(k^2\sqrt{d})$ is possible when T is low-rank.

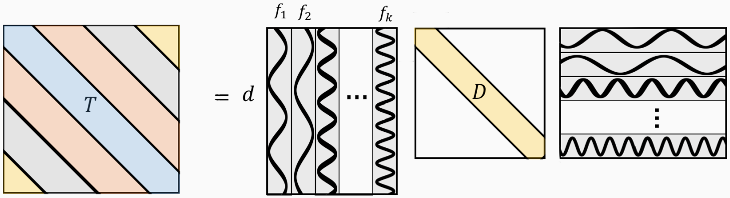
Question: Can we reduce the dependence on d even more?

Remainder of the talk: Sketch an entirely different approach to low-rank Toeplitz covariance estimation using sparse Fourier transform methods.

Low-rank Vandermonde Decomposition: Any rank-k Toeplitz $T \in R^{d \times d}$ can be written as $F_S D F_S$ where $F_S \in \mathbb{R}^{d \times k}$ is an 'off-grid' Fourier transform matrix with frequencies $f_1, \ldots, f_k$ and D is a positive diagonal matrix.



Low-rank Vandermonde Decomposition: Any rank-k Toeplitz $T \in R^{d\times d}$ can be written as $F_S D F_S$ where $F_S \in \mathbb{R}^{d\times k}$ is an 'off-grid' Fourier transform matrix with frequencies $f_1, \ldots, f_k$ and D is a positive diagonal matrix.

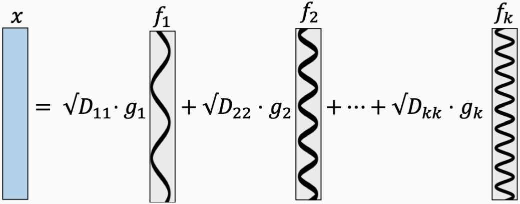


- Any sample $x \sim \mathcal{N}(0, T)$ can be written as $T^{1/2}g = F_S D^{1/2} g$ for $g \sim \mathcal{N}(0, I)$.

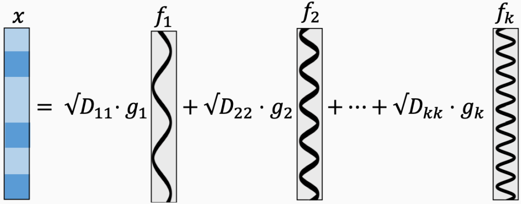
$x \sim \mathcal{N}(0, T) = F_S D^{1/2} g$ is a Fourier sparse function.

SAMPLE SECOVERY VIA SPARSE FOURIER FRANSFORM

$x \sim \mathcal{N}(0, T) = F_S D^{1/2} g$ is a Fourier sparse function.



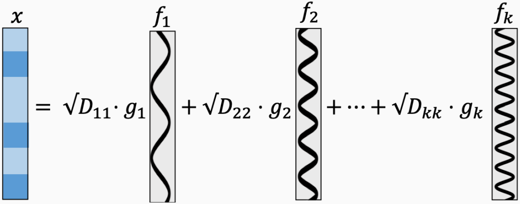
$x \sim \mathcal{N}(0, T) = F_S D^{1/2} g$ is a Fourier sparse function.



- Can recover exactly e.g. via Prony's sparse Fourier transform method by reading any $2k$ entries.

SAMPLE SECOVERY VIA SPARSE FOURIER FRANSFORM

$x \sim \mathcal{N}(0, T) = F_S D^{1/2} g$ is a Fourier sparse function.



- Can recover exactly e.g. via Prony's sparse Fourier transform method by reading any $2k$ entries.
- Take $n = O(\log^2 d/\varepsilon^2)$ samples, recover each in full by reading $2k$ entries, and then apply our earlier result for full ruler $R = [d]$. Total sample complexity: $\tilde{O}(k/\varepsilon^2)$.

What about when T is close to, but not exactly rank-k?

What about when T is close to, but not exactly rank-k?

- Prony's method totally fails in this case.

What about when T is close to, but not exactly rank-k?

- Prony's method totally fails in this case.

Step 1: Prove that when T is close to low-rank, there is are k frequencies that approximately span each $x^{(j)} \sim \mathcal{N}(0, T)$.

What about when T is close to, but not exactly rank-k?

- Prony's method totally fails in this case.

Step 1: Prove that when T is close to low-rank, there is are k frequencies that approximately span each $x^{(j)} \sim \mathcal{N}(0, T)$.

- Not as easy as it sounds.

What about when T is close to, but not exactly rank-k?

- Prony's method totally fails in this case.

Step 1: Prove that when T is close to low-rank, there is are k frequencies that approximately span each $x^{(j)} \sim \mathcal{N}(0, T)$.

- Not as easy as it sounds.

Step 2: Use a robust sparse Fourier transform method to recover $x^{(1)}, \ldots, x^{(n)}$ and estimate T from these samples.

What about when T is close to, but not exactly rank-k?

- Prony's method totally fails in this case.

Step 1: Prove that when T is close to low-rank, there is are k frequencies that approximately span each $x^{(j)} \sim \mathcal{N}(0, T)$.

- Not as easy as it sounds.

Step 2: Use a robust sparse Fourier transform method to recover $x^{(1)}, \ldots, x^{(n)}$ and estimate T from these samples.

- Well studied in TCS, but almost exclusively in the case when $f_1, \ldots, f_k$ are 'on grid' frequencies.

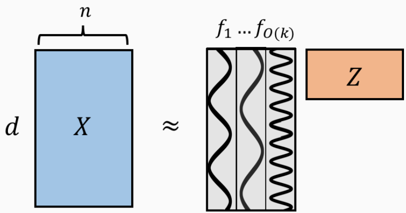
Step 1: Prove that when T is close to low-rank, there is are k frequencies that approximately span each $x^{(j)} \sim \mathcal{N}(0, T)$.

Step 1: Prove that when T is close to low-rank, there is are k frequencies that approximately span each $x^{(j)} \sim \mathcal{N}(0, T)$.

- Use several tools from Randomized Numerical Linear Algebra: Specifically a column subset selection result (see e.g., Guruswami, Sinop '12) + a projection-cost preservation bound (Cohen, Elder, Musco, Musco, Persu, '15).

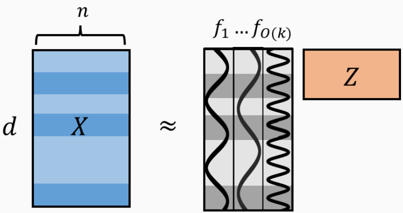
Step 2: Suffices to solve multiple regression problems of the form:

$$min_Y \|X - F_M Y\|_F^2.$$



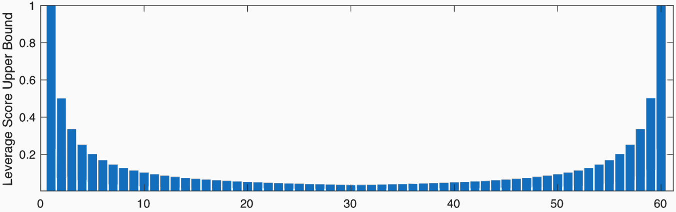
Step 2: Suffices to solve multiple regression problems of the form:

$$min_Y \|X - F_M Y\|_F^2.$$

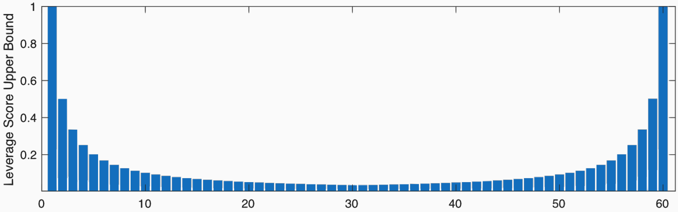


- Suffices to sample $\tilde{O}(k)$ rows by the leverage scores of F_M and solve the regression problem just considering these rows.
- This corresponds to only looking at $\tilde{O}(k)$ entries in each sample $x^{(j)}$ from $\mathcal{D}$!

Extend bounds of [Chen Kane Price Song '16] to give explicit function upper bounding the leverage scores of any F_M:

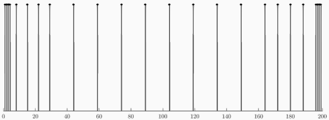


Extend bounds of [Chen Kane Price Song '16] to give explicit function upper bounding the leverage scores of any F_M:

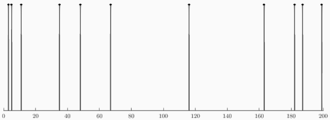


Note the resemblance to the distribution of marks in an optimal sparse ruler!

1. Sample poly(k/ε) indices $R \subset [d]$ according to the sparse Fourier leverage distribution (random 'ultra-sparse' ruler)

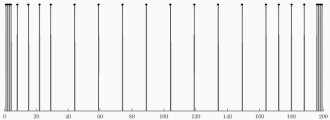


Deterministic sparse ruler pattern.

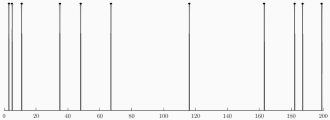


Randomly generated pattern.

1. Sample poly(k/ε) indices $R \subset [d]$ according to the sparse Fourier leverage distribution (random 'ultra-sparse' ruler)



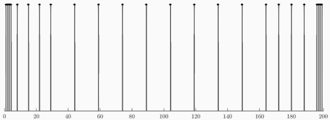
Deterministic sparse ruler pattern.



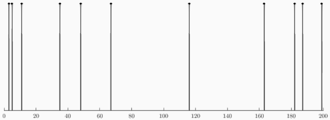
Randomly generated pattern.

2. Solve an exponential number of regression problems to recover $\tilde{X} \approx X$.

1. Sample poly(k/ε) indices $R \subset [d]$ according to the sparse Fourier leverage distribution (random 'ultra-sparse' ruler)



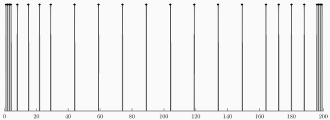
Deterministic sparse ruler pattern.



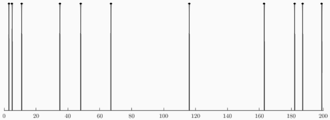
Randomly generated pattern.

2. Solve an exponential number of regression problems to recover $\tilde{X} \approx X$.
3. Return $\tilde{T} = avg(\tilde{X}\tilde{X}^T)$.

1. Sample poly(k/ε) indices $R \subset [d]$ according to the sparse Fourier leverage distribution (random 'ultra-sparse' ruler)



Deterministic sparse ruler pattern.



Randomly generated pattern.

2. Solve an exponential number of regression problems to recover $\tilde{X} \approx X$.
3. Return $\tilde{T} = avg(\tilde{X}\tilde{X}^T)$.

Vector, entry, total sample complexity: $O(\text{poly}(k \log d/\epsilon))$.

Bound: $\|T - \tilde{T}\|_2 \leq \varepsilon\|T\|_2 + f(T - T_k)$

OPEN QUESTIONS AND FUTURE DIRECTIONS

Concrete.

Concrete.

- Runtime efficiency.

Concrete.

- Runtime efficiency.
    - Can hopefully avoid exponential time net approach using off-grid sparse FFT of [Chen Kane Price Song '16.]
    - Convex optimization-based approaches and 'off-grid' RIP?
    - Matrix sparse Fourier transform $X \approx F_M \cdot Z$. Connections to MUSIC, ESPRIT, etc.

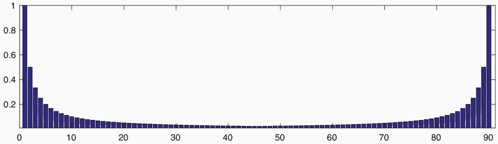
Concrete.

- Runtime efficiency.
  - Can hopefully avoid exponential time net approach using off-grid sparse FFT of [Chen Kane Price Song '16.]
  - Convex optimization-based approaches and 'off-grid' RIP?
  - Matrix sparse Fourier transform $X \approx F_M \cdot Z$. Connections to MUSIC, ESPRIT, etc.
- Improve sample complexity.

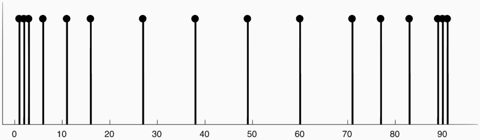
Concrete.

- Runtime efficiency.
  - Can hopefully avoid exponential time net approach using off-grid sparse FFT of [Chen Kane Price Song '16.]
  - Convex optimization-based approaches and 'off-grid' RIP?
  - Matrix sparse Fourier transform $X \approx F_M \cdot Z$. Connections to MUSIC, ESPRIT, etc.
- Improve sample complexity.
  - We give entry sample complexity of $\tilde{O}(k^2)$ but likely can be improved. Possibly to $\tilde{O}(\sqrt{k})$. Work in progress.

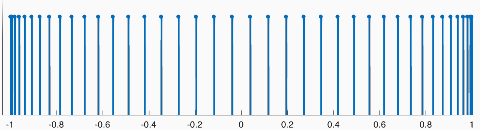
CONNECTIONS BETWEEN SAMPLING SCHEMES



Fourier Sparse Leverage Scores

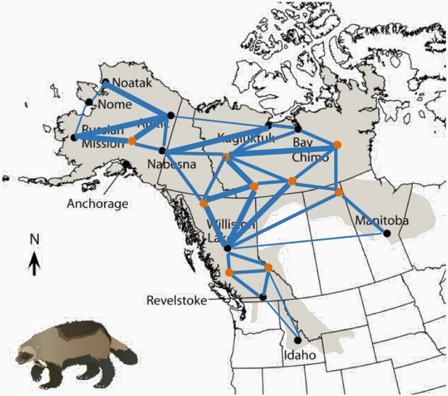


Optimal Sparse Ruler for d=91



Degree 40 Chebyshev Nodes

Not much known for more complicated spatial structure...



Example: Spatially structured genetic covariance in ecology.

thanks! questions?