High Dimensional Data Analysis With Dependency and Under Limited Memory

Yiming Sun

Cornell University

Based on joint work with
Madeleine Udell (Cornell), Sumanta Basu(Cornell)
Yang Guo (UW Madison), Charlene Luo (Columbia)
Joel Tropp (Caltech), Amy Kuceyeski(Cornell University)
Yige Li(Harvard University)

September 11, 2019

Outline

Multivariate Spectral Density Estimation Under Weak Sparsity

Low Rank Tucker Approximation of a Tensor from Streaming Data

A Connection of These Two

*

Motivation

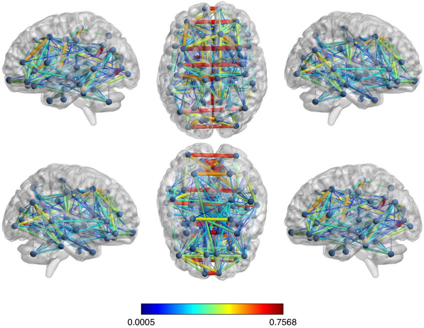


Figure: Interactions Between Regions in Brain

Weakly & strongly stationary time series

Definition (Weak Stationarity)

p-variate time series X is weakly stationary, if $\mathbb{E}X_t = \mathbb{E}X_s$ for any t, s and $\Gamma(\ell) := \mathbb{E}X_t X_{t-\ell}^\top$ only depends on the lag ℓ.

Definition (Strong Stationarity)

p-variate time series X is strongly stationary, if for any sequence $t_1, \cdots, t_n$, $X_{t_1} \cdots X_{t_n}$ has the same distribution of $X_{t_1+\tau} \cdots X_{t_n+\tau}$ for any integer τ,

Gaussian process

Definition (Gaussian Process)

p-variate time series X is Gaussian process if for any sequence $t_1, \cdots, t_n$, $X_{t_1} \cdots X_{t_n}$ are jointly Gaussian distributed.

For Gaussian process, weak stationarity is equivalent to strong stationarity.

Spectral density

Given a weakly stationary p-variate time series X, the spectral density at frequency $\omega \in [-\pi, \pi)$ is defined

$$f(\omega) = \sum_{\ell=-\infty}^{\infty} \Gamma(\ell) e^{-i\omega\ell}$$

where $\Gamma(\ell) = \mathbb{E} X_0 X_{-\ell}^\top$. X_t is independent with X_s, $t \neq s$ iff $f_{rs}(\omega) = 0$ for any ω.

Thresholding estimator under weak sparsity- An example

Suppose that we have n observation of p-variate Gaussian distribution as follows.

$$y_i \overset{i.i.d}{\sim} \mathcal{N}\left(\mu, \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & 0 & \sigma_p^2 \end{bmatrix}\right),$$

$i = 1, \cdots, n$.

An example

The maximum likelihood estimator for μ_j is $\bar{y}_j = \frac{1}{n}\sum_{i=1}^{n} y_{ij}$.
Does not Work Well Under Weak Sparsity

$$\mu \in \left\{ \mu \in \mathbb{R}^p, \sum_{j=1}^{p} |\mu_j|^q \leq c_0(p) \right\}.$$

for some $0 \leq q < 1$ and $c_0(p)$ measures the weak sparsity.

Solution : Thresholding

Suppose $\sigma_i \leq B$, define element-wise thresholding operator

$$T_\lambda(x) = \begin{cases} x & |x| \geq \lambda \\ 0 & \text{else} \end{cases}$$

. Hard thresholding estimator $T_\lambda(\bar{y}_j)$ can be shown asymptotically consistent under weak sparsity where we set

$$\lambda \propto B\sqrt{\frac{\log p}{n}}$$

and assume $\lambda \to 0$.

Two key ingredients for thresholding

Two key ingredients under above example assuming weak sparsity. Cai & Liu 2011

- An element-wise concentration inequality :

$$\mathbb{P}(|\bar{y}_j - \mu_j| \geq \eta) \leq 2\exp(-n\eta^2/2\sigma_j^2).$$

- σ_j are uniformly bounded.

Shortcomings for hard thresholding

- σ_j may variate much
- B will appear in the thresholding value making convergence rate slow

Solution: adaptive thresholding

Simply estimate σ_j, say with sample standard deviation:

$$\hat{\sigma}_j = \sqrt{1/(n-1)\sum_{i=1}^{n}(y_{ij} - \bar{y}_j)^2}$$

and replace B: $\lambda_j \propto \hat{\sigma}_j\sqrt{\frac{\log p}{n}}$. Now we can relax constraint in upper bound for σ_j and upper bound will not appear in rate of convergence.

An Similar Example: Covariance Matrix

$$y_i \overset{i.i.d}{\sim} \mathcal{N}(0, \Sigma_{p \times p})$$

Goal: Estimate Σ assuming weak sparsity, $\|\Sigma\|_1 \leq c_0(p)$. Bickel & Levina 2008

A similar example: covariance matrix

Estimate the expectation of a vector of length p^2:

$$[(y_1 y_1^\top)_{rs}, 1 \le r, s \le p].$$

MLE: $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n y_i y_i^\top$. But we need to perform thresholding. Remember two ingredients:

- $\mathbb{P}(|\hat{\Sigma}_{rs} - \Sigma_{rs}| \ge \eta) \le c_1 \exp(-c_2 n \eta^2)$
- $\mathbf{var}((y_1 y_1^\top)_{rs}) = \Sigma_{rr}\Sigma_{ss} + \Sigma_{rs}^2 \le 2\max_{r=1}^p \Sigma_{rr}^2$

Thus Bickel & Levina 2008 presents an assumption $\max_{r=1}^p \Sigma_{rr}$ is bounded.

A similar example: covariance matrix

hard thresholding: $\lambda_{rs} \propto (\max_{r=1}^{p} \Sigma_{rr})\sqrt{\frac{\log p}{n}}$

adaptive thresholding: $\lambda_{rs} \propto \sqrt{\widehat{\mathbf{var}(y_1 y_1^\top)}_{rs}}\sqrt{\frac{\log p}{n}}$ where

$$\widehat{\mathbf{var}(y_1 y_1^\top)}_{rs} = \frac{1}{n-1}\sum_{i=1}^{n}\left[(y_i y_i^\top)_{rs} - \frac{1}{n}\sum_{i=1}^{n}(y_i y_i^\top)_{rs}\right]^2$$

,

Outline

Multivariate Spectral Density Estimation Under Weak Sparsity

Low Rank Tucker Approximation of a Tensor from Streaming Data

A Connection of These Two

*

Motivation

We listed three scenarios for Motivation Borrowed from Professor Udell's Recent Talk

Big data, small laptop

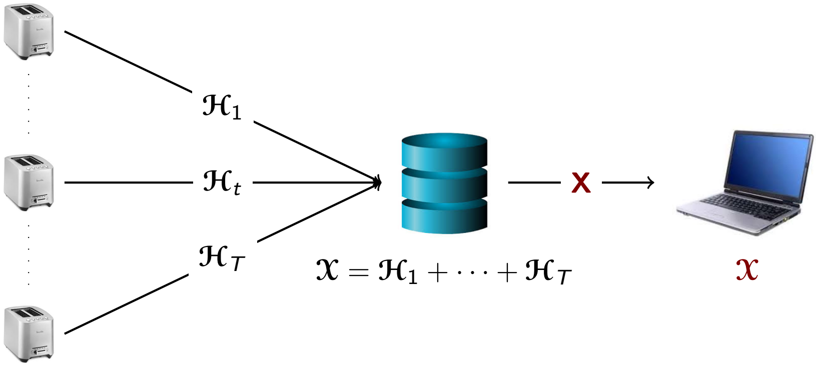
$\xrightarrow{\mathcal{H}_T}$

$\xrightarrow{\mathcal{X}}$

$$\mathcal{X} = \mathcal{H}_1 + \cdots + \mathcal{H}_T$$

$$\mathcal{X}$$

Distributed data



Streaming data

$\mathcal{H}_t$

$$\mathcal{X}^{(t)} = \mathcal{H}_1 + \cdots + \mathcal{H}_t$$

Notation

tensor to compress:

- tensor $\mathcal{X} \in \mathbf{R}^{I_1 \times \cdots \times I_N}$ with N modes
- sometimes assume $I_1 = \cdots = I_N = I$ for simplicity

indexing:

- $[N] = 1, \ldots, N$
- $I_{(-n)} = I_1 \times \cdots \times I_{n-1} \times I_{n+1} \times I_N$

tensor operations:

- mode n product: for $\mathcal{A} \in \mathbf{R}^{k \times I_n}$,
  $\mathcal{X} \times_n \mathbf{A} \in \mathbf{R}^{I_1 \times \cdots \times I_{n-1} \times k \times I_{n+1} \times \cdots \times I_N}$
- unfolding $\mathbf{X}^{(n)} \in \mathbf{R}^{I_n \times I_{(-n)}}$ stacks mode-n fibers of $\mathcal{X}$ as columns of matrix

Review of Our Tool: Linear Sketch

A linear random projection can be represented as a random matrix $\mathbf{\Omega} \in \mathbf{R}^{d \times k}$, operating on a vector $\mathbf{x} \in \mathbf{R}^d$ or a matrix $\mathbf{X} \in \mathbf{R}^{m \times d}$ to reduce the dimension:

$$\begin{aligned} \mathbf{x} \in \mathbf{R}^n &\rightarrow \mathbf{\Omega}^\top \mathbf{x} \in \mathbf{R}^k \\ \mathbf{X} \in \mathbf{R}^{m \times d} &\rightarrow \mathbf{X}\mathbf{\Omega} \in \mathbf{R}^{m \times k}. \end{aligned} \tag{1}$$

Properties Preserved after Projection

Lemma (Arriaga & Vempala 2006)

Let $\mathbf{x} \in \mathbf{R}^d$, assume that the entries in $\mathbf{\Omega} \in \mathbf{R}^{d \times k}$ are sampled independently from $\mathcal{N}(0, 1)$. Then

$$\mathbf{Prob}\left((1-\epsilon)\|\mathbf{x}\|^2 \le \left\|\frac{1}{\sqrt{k}}\mathbf{\Omega}^\top \mathbf{x}\right\| \le (1+\epsilon)\|\mathbf{x}\|^2\right) \le 1 - 2e^{-(\epsilon^2-\epsilon^3)k/4}. \tag{2}$$

Lemma (Halko, Martinsson & Tropp 2011)

Let $\mathbf{X} \in \mathbf{R}^{m \times d}$, assume that the entries in $\mathbf{\Omega} \in \mathbf{R}^{d \times (k+p)}$ are sampled independently from $\mathcal{N}(0, 1)$. Then let $\mathbf{Q}$ be the orthonormal matrix from QR factorization $\mathbf{X\Omega} = \mathbf{QR}$, then

$$\|\mathbf{X} - \mathbf{QQ}^\top \mathbf{X}\|_F \le \left(1 + \frac{k}{p-1}\right)^{1/2} \left(\sum_{j>k} \sigma_j^2\right)^{1/2}. \tag{3}$$

Tucker factorization

rank $\mathbf{r} = (r_1, \ldots, r_N)$ **Tucker factorization** of $\mathcal{X} \in \mathbf{R}^{I_1 \times \cdots \times I_N}$:

$$\mathcal{X} \;=\; \mathcal{G} \times_1 \mathbf{U}_1 \cdots \times_N \mathbf{U}_N =: [\![\mathcal{G}; \mathbf{U}_1, \ldots, \mathbf{U}_N]\!]$$

where

- $\mathcal{G} \in \mathbf{R}^{r_1 \times \cdots \times r_N}$ is the **core matrix**
- $\mathbf{U}_n \in \mathbf{R}^{I_n \times r_n}$ is the **factor matrix** for each mode $n \in [N]$

(sometimes assume $r_1 = \cdots = r_N = r$ for simplicity)

Tucker is useful for compression: when N is small,

- Tucker stores $O(rNI)$ numbers for rank r^3 approximation
- CP stores $O(rNI)$ numbers for rank r approximation

The sketch

approximate factor matrices and core:

- **Factor sketch** ($\mathbf{k}$). For each $n \in [N]$,
  fix random DRM $\boldsymbol{\Omega}_n \in \mathbb{R}^{I_{(-n)} \times k_n}$ and compute the sketch

$$\mathbf{V}_n = \mathbf{X}^{(n)} \boldsymbol{\Omega}_n \quad \in \mathbf{R}^{I_n \times k_n}.$$

- **Core sketch** ($\mathbf{s}$). For each $n \in [N]$,
  fix random DRM $\boldsymbol{\Phi}_n \in \mathbf{R}^{I_n \times s_n}$. Compute the sketch

$$\mathcal{H} = \mathcal{X} \times_1 \boldsymbol{\Phi}_1^\top \cdots \times_N \boldsymbol{\Phi}_N^\top \quad \in \mathbf{R}^{s_1 \times \cdots \times s_N}.$$

- *Rule of thumb.* Pick $\mathbf{k}$ as big as you can afford, pick $\mathbf{s} = 2\mathbf{k}$.
- define $(\mathcal{H}, \mathbf{V}_1, \ldots, \mathbf{V}_N) = \textsc{Sketch}\left(\mathcal{X}; \{\boldsymbol{\Phi}_n, \boldsymbol{\Omega}_n\}_{n \in [N]}\right)$

Low memory DRMs

factor sketch DRMs are big! Same size of the tensor

- $I_{(-n)} \times k_n$ for each $n \in [N]$
- Solution: Generate random matrix $\mathbf{A}_n \in \mathbf{R}^{I_n \times k}$ [Sun, Guo, Luo, Tropp & Udell 2019]

$$\boldsymbol{\Omega} := (\mathbf{A}_1 \odot \cdots \odot \mathbf{A}_N)$$

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} A_{11}\mathbf{B} & \cdots & A_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ A_{m1}\mathbf{B} & \cdots & A_{mn}\mathbf{B} \end{bmatrix}.$$

We let $\mathbf{X} \odot \mathbf{Y}$ denotes the *Khatri-Rao product*, $\mathbf{A} \in \mathbb{R}^{I \times K}, \mathbf{B} \in \mathbb{R}^{J \times K}$, i.e. the "matching column-wise" Kronecker product. The resulting matrix of size $(IJ) \times K$ is given by:

$$\mathbf{A} \odot \mathbf{B} = [\mathbf{A}_{(1,.)} \otimes \mathbf{B}_{(1,.)}, \ldots, \mathbf{A}_{(K,.)} \otimes \mathbf{B}_{(K,.)}]. \tag{4}$$

Two pass algorithm

Algorithm Two Pass Sketch and Low Rank Recovery

Given: tensor $\mathcal{X}$, DRMs $\{\mathbf{\Phi}_n, \mathbf{\Omega}_n\}_{n\in[N]}$ with parameters $\mathbf{k}$ and $\mathbf{s} \geq \mathbf{k}$

1. *Sketch.* $(\mathcal{H}, \mathbf{V}_1, \ldots, \mathbf{V}_N) = \textsc{Sketch}\left(\mathcal{X}; \{\mathbf{\Phi}_n, \mathbf{\Omega}_n\}_{n\in[N]}\right)$
2. *Recover factor matrices.* For $n \in [N]$,
$$(\mathbf{Q}_n, \sim) \leftarrow \mathrm{QR}(\mathbf{V}_\mathrm{n})$$
3. *Recover core.*
$$\mathcal{W} \leftarrow \mathcal{X} \times_1 \mathbf{Q}_1 \cdots \times_N \mathbf{Q}_N$$

Return: Tucker approximation $\tilde{\mathcal{X}} = [\![\mathcal{W}; \mathbf{Q}_1, \ldots, \mathbf{Q}_N]\!]$ with rank $\leq \mathbf{k}$

accesses $\mathcal{X}$ twice: 1) to sketch 2) to recover core

Intuition: one pass core recovery

- we want to know $\mathcal{W}$:
  compression of $\mathcal{X}$ using factor range approximations $\mathbf{Q}_n$
- we observe $\mathcal{H}$:
  compression of $\mathcal{X}$ using random projections $\mathbf{\Phi}_n$

how to approximate $\mathcal{W}$?

$$
\begin{aligned}
\mathcal{X} &\approx \mathcal{X} \times_1 \mathbf{Q}_1\mathbf{Q}_1^\top \times \cdots \times_N \mathbf{Q}_N\mathbf{Q}_N^\top \\
&= \left(\mathcal{X} \times_1 \mathbf{Q}_1^\top \times_N \cdots \times \mathbf{Q}_N^\top\right) \times_1 \mathbf{Q}_1 \cdots \times_N \mathbf{Q}_N \\
&= \mathcal{W} \times_1 \mathbf{Q}_1 \cdots \times_N \mathbf{Q}_N \\
\underbrace{\mathcal{X} \times_1 \mathbf{\Phi}_1^\top \cdots \times_N \mathbf{\Phi}_N^\top}_{\mathcal{H}} &\approx \mathcal{W} \times_1 \mathbf{\Phi}_1^\top\mathbf{Q}_1 \times \cdots \times_N \mathbf{\Phi}_N^\top\mathbf{Q}_N
\end{aligned}
$$

we can solve for $\mathcal{W}$: $\mathbf{s} > \mathbf{k}$, so each $\mathbf{\Phi}_n^\top\mathbf{Q}_n$ has a left inverse (whp):

$$\mathcal{W} \approx \mathcal{H} \times_1 (\mathbf{\Phi}_1^\top\mathbf{Q}_1)^\dagger \times \cdots \times_N (\mathbf{\Phi}_N^\top\mathbf{Q}_N)^\dagger$$

One pass algorithm

Algorithm One Pass Sketch and Low Rank Recovery

Given: tensor $\mathcal{X}$, rank $\mathbf{r} = (r_1, \ldots, r_N)$, DRMs $\{\mathbf{\Phi}_n, \mathbf{\Omega}_n\}_{n \in [N]}$

- *Sketch.* $(\mathcal{H}, \mathbf{V}_1, \ldots, \mathbf{V}_N) = \textsc{Sketch}\left(\mathcal{X}; \{\mathbf{\Phi}_n, \mathbf{\Omega}_n\}_{n \in [N]}\right)$
- *Recover factor matrices.* For $n \in [N]$,

$$(\mathbf{Q}_n, \sim) \leftarrow \mathrm{QR}(\mathbf{V}_\mathrm{n})$$

- *Recover core.*

$$\mathcal{W} \leftarrow \mathcal{H} \times_1 (\mathbf{\Phi}_1^\top \mathbf{Q}_1)^\dagger \times \cdots \times_N (\mathbf{\Phi}_N^\top \mathbf{Q}_N)^\dagger$$

Return: Tucker approximation $\hat{\mathcal{X}} = [\![\mathcal{W}; \mathbf{Q}_1, \ldots, \mathbf{Q}_N]\!]$

---

accesses $\mathcal{X}$ only once, to sketch

Source: [Sun et al. 2019]

Fixed rank approximation

to truncate reconstruction to rank $\mathbf{r}$, truncate core:

Lemma

For a tensor $\mathcal{W} \in \mathbb{R}^{k_1 \times \cdots \times k_N}$, orthogonal matrices $\mathbf{Q}_n \in \mathbb{R}^{k_n \times r_n}$,

$$[\![\mathcal{W} \times_1 \mathbf{Q}_1 \cdots \times_N \mathbf{Q}_N]\!]_{\mathbf{r}} = [\![\mathcal{W}]\!]_{\mathbf{r}} \times_1 \mathbf{Q}_1 \cdots \times_N \mathbf{Q}_N,$$

where $[\![\cdot]\!]$ denotes the best rank $\mathbf{r}$ Tucker approximation.

$\implies$ compute fixed rank approximation using, e.g., HOOI on (small) core approximation $\mathcal{W}$

Tail Energy

For each unfolding $\mathbf{X}^{(n)}$, define its ρ*th tail energy* as

$$(\tau_\rho^{(n)})^2 := \sum_{k>\rho}^{\min(I_n, I_{(-n)})} \sigma_k^2(\mathbf{X}^{(n)}),$$

where $\sigma_k(\mathbf{X}^{(n)})$ is the kth largest singular value of $\mathbf{X}^{(n)}$.

Guarantees for two pass

Theorem ([Sun, Guo, Tropp & Udell 2018])

Sketch the tensor $\mathcal{X}$ using a Tucker sketch with parameters $\mathbf{k}$ using DRMs with i.i.d. Gaussian $\mathcal{N}(0,1)$ entries. Then the approximation $\hat{\mathcal{X}}_2$ computed with the two pass method satisfies

$$\mathbb{E}\|\mathcal{X} - \hat{\mathcal{X}}_2\|_F^2 \leq \min_{1 \leq \rho_n < k_n - 1} \sum_{n=1}^{N} \left(1 + \frac{\rho_n}{k_n - \rho_n - 1}\right) (\tau_{\rho_n}^{(n)})^2.$$

Guarantees for one pass

Theorem ([Sun et al. 2018])

Sketch $\mathfrak{X}$ with Gaussian DRMs of parameters $\mathbf{k}$, $\mathbf{s} \geq 2\mathbf{k} + 1$. Form a rank $\mathbf{r}$ Tucker approximation $\hat{\mathfrak{X}}$ using the one pass algorithm. Then

$$\mathbb{E}\|\mathfrak{X} - \hat{\mathfrak{X}}\|_F^2 \leq (1 + \Delta) \min_{1 \leq \rho_n < k_n - 1} \sum_{n=1}^{N} \left(1 + \frac{\rho_n}{k_n - \rho_n - 1}\right) (\tau_{\rho_n}^{(n)})^2$$

where $\Delta = \max_{n=1}^{N} k_n/(s_n - k_n - 1)$

Comparison to other methods in pseudo optimality

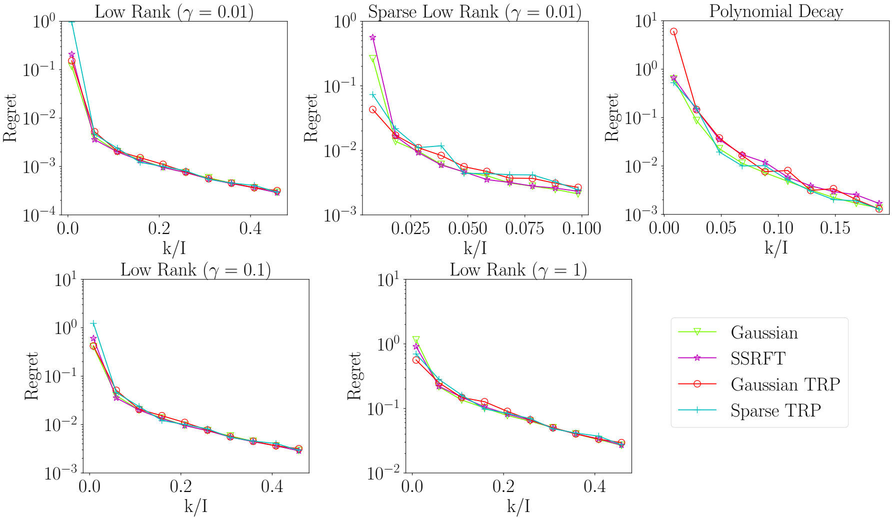
- HOSVD and ST-HOSVD is pseudo optimal with factor N:

$$\|\mathcal{X} - [\![\mathcal{X}]\!]_{\mathbf{ST-r}}\|_F \le \sqrt{\sum_{n=1}^{N} (\tau_{r_n}^{(n)})^2} \le \sqrt{N}\|\mathcal{X} - [\![\mathcal{X}]\!]_{\mathbf{r}}\|_F, \tag{5}$$

- Set $\mathbf{k} = 2\mathbf{r} + 1$ and $\mathbf{s} = 2\mathbf{k} + 1$, and use truncated QR factorization to get $\mathbf{Q} \in \mathbf{R}^{I_n \times r_n}$ from factor sketch.

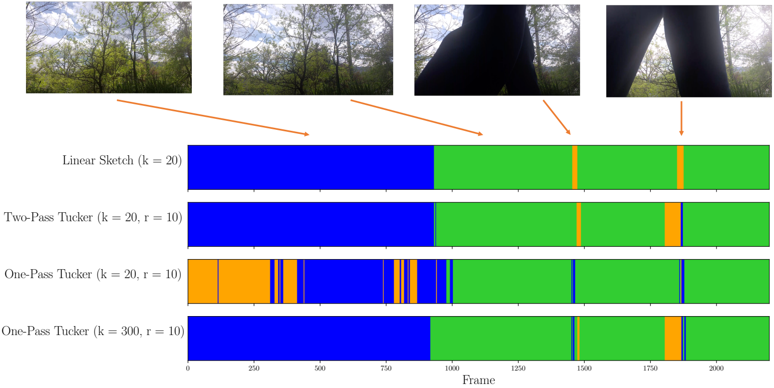
$$\|\mathcal{X} - \hat{\mathcal{X}}_2|_F \le \sqrt{\sum_{n=1}^{N} (\tau_{r_n}^{(n)})^2} \le \sqrt{2N}\|\mathcal{X} - [\![\mathcal{X}]\!]_{\mathbf{r}}\|_F. \tag{6}$$

Different DRMs perform similarly



Comments: Synthetic data, $I = 600$ and $\mathbf{r} = (5, 5, 5)$. $k/I = .4 \implies 20\times$ compression.

Video scene classification



Comments: Video data $2200 \times 1080 \times 1980$. Classify scenes using k-means on: 1) linear sketch along the time dimension $k = 20$ (Row 1); 2) The Tucker factor along the time dimension, computed via our two pass (Row 2) and one pass (Row 3) sketching algorithm $(r, k, s) = (10, 20, 41)$. 3) The Tucker factor along the time dimension, computed via our one pass (Row 4) sketching algorithm $(r, k, s) = (10, 300, 601)$.

Property of tensor random projection

Preserve Pair-wise Distance: Fix $\mathbf{x} \in \mathbf{R}^{\prod_{n=1}^{N} d_n}$. Generate random matrix

$$\mathbf{\Omega} = (\mathbf{A}_1 \odot \cdots \odot \mathbf{A}_N)$$

Theorem

Fix $\mathbf{x} \in \mathbb{R}^{\prod_{n=1}^{N} d_n}$. Form a TRP and TRP_T of order N with range k composed of independent matrices with independent columns whose entries are mean zero, variance one, and within each column every pair of elements has covariance zero. Then

$$\mathbb{E}\|\text{TRP}(\mathbf{x})\|^2 = \|\mathbf{x}\|^2 \tag{7}$$

Property of tensor random projection

AA

Outline

Multivariate Spectral Density Estimation Under Weak Sparsity

Low Rank Tucker Approximation of a Tensor from Streaming Data

A Connection of These Two

*

Outline

Multivariate Spectral Density Estimation Under Weak Sparsity

Low Rank Tucker Approximation of a Tensor from Streaming Data

A Connection of These Two

*

References

Arriaga, R. I. & Vempala, S. (2006). An algorithmic theory of learning: Robust concepts and random projection. *Machine Learning*, *63*(2), 161–182.

Bickel, P. J. & Levina, E. (2008). Covariance regularization by thresholding. *The Annals of Statistics*, 2577–2604.

Cai, T. & Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, *106*(494), 672–684.

Halko, N., Martinsson, P.-G., & Tropp, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, *53*(2), 217–288.

Sun, Y., Guo, Y., Luo, C., Tropp, J. A., & Udell, M. (2019). Low rank tucker approximation of a tensor from streaming data. *In preparation*.

Sun, Y., Guo, Y., Tropp, J. A., & Udell, M. (2018). Tensor random projection for low memory dimension reduction. In *NeurIPS Workshop on Relational Representation Learning*.