Learning Tools for Big Data Analytics

Georgios B. Giannakis

Acknowledgments: Profs. G. Mateos and K. Slavakis

NSF 1343860, 1442686, and MURI-FA9550-10-1-0567
Growing data torrent

$600 to buy a disk drive that can store all of the world’s music

5 billion mobile phones in use in 2010

30 billion pieces of content shared on Facebook every month

40% projected growth in global data generated per year vs. 5% growth in global IT spending

Big Data: Capturing its value

$300 billion
potential annual value to US health care—more than double the total annual health care spending in Spain

€250 billion
potential annual value to Europe’s public sector administration—more than GDP of Greece

$600 billion
potential annual consumer surplus from using personal location data globally

60%
potential increase in retailers’ operating margins possible with big data

Challenges

- Sheer volume of data
  - Decentralized and parallel processing
  - Security and privacy measures

- Modern massive datasets involve many attributes
  - Parsimonious models to ease interpretability and enhance learning performance

- Real-time streaming data
  - Online processing
    - Quick-rough answer vs. slow-accurate answer?

- Outliers and misses
  - Robust imputation approaches
Opportunities

**Theoretical and Statistical Foundations of Big Data Analytics**

- Big tensor data models and factorizations
- High-dimensional statistical SP
- Network data visualization
- Pursuit of low-dimensional structure
- Analysis of multi-relational data
- Common principles across networks
- Resource tradeoffs

**Algorithms and Implementation Platforms to Learn from Massive Datasets**

- Scalable online, decentralized optimization
- Randomized algorithms
- Information processing over graphs
- Convergence and performance guarantees
- Novel architectures for large-scale data analytics
- Robustness to outliers and missing data

Graph SP

5
Roadmap

- Context and motivation

- Critical Big Data tasks
  - Encompassing and parsimonious data modeling
  - Dimensionality reduction
  - Data cleansing, anomaly detection, and inference

- Randomized learning via data sketching

- Conclusions and future research directions
Encompassing model

- **Observed data** $\in \mathbb{R}^{D \times T}$
- **Dictionary** $\in \mathbb{R}^{D \times Q}$
- **Sparse matrix** $\in \mathbb{R}^{Q \times T}$

\[ Y = L + DS + V \]

- **Background (low rank)**
- **Patterns, innovations, (co-)clusters, outliers**
- **Noise**

- Subset $\Omega \subset \{1, \ldots, D\} \times \{1, \ldots, T\}$ of observations and projection operator

\[ [P_\Omega(Y)]_{ij} = \begin{cases} [Y]_{ij}, & \text{if } (i, j) \in \Omega \\ 0, & \text{otherwise} \end{cases} \]

allow for misses

- Large-scale data $D \gg$ and/or $T \gg$
- Any of $\{L, D, S\}$ unknown
Subsumed paradigms

- Structure leveraging criterion

\[
\min \left\{ \frac{1}{2} \| Y \|_F^2 \right\}
\]

Nuclear norm: \( \| L \|_* := \sum_{j=1}^{\text{rank}(L)} \sigma_j(L) \)

\( \{\sigma_j(L)\}_{j=1}^{\text{rank}(L)} \) : singular val. of \( L \)

(With or without misses)

- \( L = 0, D \) known \( \Rightarrow \) Compressive sampling (CS) [Candes-Tao ‘05]
- \( L = 0 \) \( \Rightarrow \) Dictionary learning (DL) [Olshausen-Field ‘97]
- \( L = 0, [D]_{ij} \geq 0, [S]_{ij} \geq 0 \) \( \Rightarrow \) Non-negative matrix factorization (NMF) [Lee-Seung ‘99]
- \( D = I_D \) \( \Rightarrow \) Principal component pursuit (PCP) [Candes etal ‘11]
- \( S = 0, \text{rank}(L) \leq \rho \) \( \Rightarrow \) Principal component analysis (PCA) [Pearson 1901]
PCA formulations

- **Training data** \( \{y_t \in \mathbb{R}^D\}_{t=1}^T \) \( \hat{C}_{yy} := (1/T) \sum_{t=1}^T y_t y_t^\top \)

- **Minimum reconstruction error**
  - **Compression** \( G \in \mathbb{R}^{d \times D} \) \( d \ll D \)
  - **Reconstruction** \( U \in \mathbb{R}^{D \times d} \) \( U^\top U = I_d \)
  \[
  \min_{U,G} \sum_{t=1}^T \|y_t - UGy_t\|_2^2, \quad \text{s.to.} \quad U^\top U = I_d
  \]

- **Component analysis model** \( y_t = U\psi_t + \varepsilon_t \)
  \[
  \min_{U,\psi_t} \sum_{t=1}^T \|y_t - U\psi_t\|_2^2, \quad \text{s.to.} \quad U^\top U = I_d
  \]

**Solution:** \( \hat{U}_d = d\text{-evecs}(\hat{C}_{yy}) \), \( \hat{G} = \hat{U}_d^\top \), \( \hat{\psi}_t = \hat{U}_d^\top y_t \)
Dual and kernel PCA

\[ \text{SVD: } \underbrace{Y}_{D \times T} = U \Sigma V^\top \]

\[ YY^\top = U \Sigma^2 U^\top \in \mathbb{R}^{D \times D} \quad O(TD^2) \]

\[ Y^\top Y = V \Sigma^2 V^\top \in \mathbb{R}^{T \times T} \quad O(DT^2) \]

Gram matrix

\[ \hat{U}_d = Y \hat{V}_d \hat{\Sigma}_d^{-1} \]

\[ \hat{U}_d y_t = \hat{\Sigma}_d^{-1} \hat{V}_d y_t \]

Inner products

\[ \hat{U}_d \hat{\psi}_t = Y \hat{V}_d \hat{\Sigma}_d^{-1} \hat{\psi}_t \rightarrow \hat{y}_t \]

Q. What if approximating low-dim space not a hyperplane?

A1. Stretch it to become linear: Kernel PCA; e.g., [Scholkopf-Smola’01]
   - maps \( y_t \) to \( \varphi(y_t) \), and leverages dual PCA in high-dim spaces

A2. General (non)linear models; e.g., union of hyperplanes, or, locally linear
   - tangential hyperplanes

Identification of network communities

- Kernel PCA instrumental for partitioning of large graphs (spectral clustering)
  - Relies on graph Laplacian to capture nodal correlations

For $D \gg$ random sketching and validation reduces complexity to $O(d)$

Facebook egonet
744 nodes, 30,023 edges

arXiv collaboration network (General Relativity)
4,158 nodes, 13,422 edges

Local linear embedding

- For each $y_t$ find neighborhood $\{y_{t'}\}_{t' \in \mathcal{N}_t}$, e.g., k-nearest neighbors

- Weight matrix captures local affine relations

\[
\begin{align*}
\min_{W := [w_1, \ldots, w_T] \in \mathbb{R}^{T \times T}} & \sum_{t=1}^{T} \left\| y_t - \sum_{t' \in \mathcal{N}_t} w_{t'} y_{t'} \right\|^2 \\
\text{subject to} & \quad W^T 1_T = 0_T \quad \text{and} \quad \mathcal{N}_t \quad \text{Sparse} \quad \mathcal{N}_t \quad \text{and} \quad W \quad \text{[Elhamifar-Vidal’11]}
\end{align*}
\]

- Identify low-dimensional vectors preserving local geometry [Saul-Roweis’03]

\[
\begin{align*}
\min_{\Psi := [\psi_1, \ldots, \psi_T] \in \mathbb{R}^{d \times T}} & \left\{ \sum_{t=1}^{T} \left\| \psi_t - \sum_{t' = 1}^{T} w_{t'} \psi_{t'} \right\|^2 \right. \\
\text{subject to} & \quad \Psi \Psi^T = I_d \quad \text{and} \quad \Psi 1_T = 0_d \\
\end{align*}
\]

Solution: The rows of $\Psi$ are the $d$ minor, excluding $1_T$, 
\[
evecs\left( \begin{bmatrix} I_T & -W \\ I_T & W^T \end{bmatrix} \right)
\]

Dictionary learning

- Solve for dictionary $D$ and sparse $S$:
  \[
  \min_{D \in \mathcal{D}} \min_{S \in \mathbb{R}^{Q \times T}} \frac{1}{2} \| \mathcal{P}_\Omega (Y - DS) \|_F^2 + \lambda_1 \sum_{t=1}^{T} \| s_t \|_1
  \]

  \[\mathcal{D} := \{ D = [d_1, \ldots, d_Q] : \| d_q \| \leq 1, \forall q \} \quad Q \geq D\]

- \( \ominus \) Alternating minimization; both \( \mathcal{L}(D_k, \cdot) \) and \( \mathcal{L}(\cdot, S_{k+1}) \) are convex

- Under conditions, \( (D_k, S_k)_{k=0}^\infty \) converges to a stationary point of \( \mathcal{L} \) \cite{Tseng’01}

Joint DL-LLE paradigm

\[
\begin{align*}
\min_{D_y, S, S^\top 1_Q = 1_T} & \quad \frac{1}{2} \| \mathcal{P}_{\Omega} (Y - D_y S) \|_F^2 + \frac{\lambda_y}{2} \| D_y - D_y W \|_F^2 + \sum_{t=1}^{T} \lambda_{st} \| s_t \|_1 + \sum_{q=1}^{Q} \lambda_{wq} \| w_q \|_1 \\
\text{DL fit} & \quad \text{LLE fit} & \quad \text{Sparsity regularization}
\end{align*}
\]

- Dictionary morphs data to a smooth basis; reduces noise and complexity

Original image  Degraded image  Restored image

50% missing entries

From low-rank matrices to tensors

- Data cube $\mathbf{X} \in \mathbb{R}^{M \times N \times T}$, e.g., sub-sampled MRI frames $\mathbf{Y}_t^\Omega \approx \mathcal{F}_\Omega (\mathbf{X}_t)$

- PARAFAC decomposition per slab $t$ [Harshman '70]

$$\mathbf{X}_t = \sum_{r=1}^{R} \gamma_{t,r} \mathbf{a}_r \mathbf{b}_r^\top = \mathbf{A} \text{diag}(\gamma_t) \mathbf{B}^\top$$

- Tensor subspace comprises $R$ rank-one matrices $\{\mathbf{a}_r \mathbf{b}_r^\top\}_{r=1}^{R}$

Goal: Given streaming $\mathbf{Y}_t^\Omega \approx \mathcal{F}_\Omega (\mathbf{A} \text{diag}(\gamma_t) \mathbf{B}^\top)$, learn the subspace matrices $(\mathbf{A}, \mathbf{B})$ recursively, and impute possible misses of $\mathbf{Y}_t$

Online tensor subspace learning

- Image domain low tensor rank \( Y_t^\Omega \approx F_\Omega (A\text{diag}(\gamma_t)B^\top) \)

\[
(\hat{A}_t, \hat{B}_t) = \arg\min_{A,B} \frac{1}{t} \sum_{\tau=1}^{t} \min_{\gamma_\tau} \left\{ \|Y_\tau^\Omega - F_\Omega (A\text{diag}(\gamma_\tau)B^\top)\|_F^2 + \frac{\lambda}{2} \|\gamma_\tau\|_2^2 \right\} \\
+ \frac{\lambda}{2t} (\|A\|_F^2 + \|B\|_F^2)
\]

- Tikhonov regularization promotes low rank

Proposition [Bazerque-GG ‘13]: With \( [\sigma]_r = \|a_r\|_{\|b_r\|,\|c_r\|} \)

\[
\|\sigma\|_{2/3}^{2/3} := \arg\min_{\{A,B,C\}} (\|A\|_F^2 + \|B\|_F^2 + \|C\|_F^2)
\]

- Stochastic alternating minimization; parallelizable across bases

- Real-time reconstruction (FFT per iteration) \( \hat{X}_t = \hat{A}_t\text{diag}(\hat{\gamma}_t)\hat{B}_t^\top \)

Dynamic cardiac MRI test

- *in vivo* dataset: 256 k-space 200x256 frames

Ground-truth frame

Sampling trajectory

- Potential for accelerating MRI at high spatio-temporal resolution

- Low-rank $\mathcal{F}_{\Omega_t}(X_t)$ plus $\mathcal{F}_{\Omega_t}(DS_t)$ can also capture motion effects

Roadmap

- Context and motivation
- Critical Big Data tasks
- Randomized learning via data sketching
  - Johnson-Lindenstrauss lemma
  - Randomized linear regression
  - Randomized clustering
- Conclusions and future research directions
Randomized linear algebra

- **Basic tools**: Random sampling and random projections

- **Attractive features**
  - Reduced dimensionality to lower complexity with Big Data
  - Rigorous error analysis at reduced dimension

**Ordinary least-squares (LS)**

Given \( y \in \mathbb{R}^D, \ X \in \mathbb{R}^{D \times p} \)

\[
\theta_{LS} := \arg \min_{\theta \in \mathbb{R}^p} \| y - X\theta \|_2^2
\]

If \( \text{rank}(X) = p \) \( \Rightarrow \quad \theta_{LS} = (X^TX)^{-1}X^Ty \)

- SVD incurs complexity \( \mathcal{O}(Dp^2) \).

**Q**: What if \( D \gg p \) ?

Randomized LS for linear regression

- LS estimate using (pre-conditioned) random projection matrix $R_{d \times D}$

$$\tilde{\theta}_{LS} = \arg \min_{\theta \in \mathbb{R}^p} \left\| \Gamma_d S_d H_D \Delta_D (y - X\theta) \right\|_2^2$$

- Random diagonal w/ $[\Delta_D]_{ii} \in \{1, -1\} \sim \text{Ber}(1/2)$ and Hadamard matrix

$$H_D = \frac{1}{\sqrt{D}} \begin{bmatrix} H_{D/2} & H_{D/2} \\ H_{D/2} & -H_{D/2} \end{bmatrix}, \quad H_2 := \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

- Subsets of data obtained by uniform sampling/scaling via $S_d, \Gamma_d$ yield LS estimates of "comparable quality"

- Select reduced dimension $d = \mathcal{O}(p \log p \cdot \log D + \epsilon^{-1} D \log p)$

- Complexity reduced from $\mathcal{O}(Dp^2)$ to $o(Dp^2)$

Johnson-Lindenstrauss lemma

- The “workhorse” for proofs involving random projections

**JL lemma:** If $0 < \epsilon < 1$, integer $T$, and reduced dimension satisfies

$$d \geq 4(\epsilon^2/2 - \epsilon^3/3)^{-1} \ln T$$

then for any $Y = [y_1, \ldots, y_T] \in \mathbb{R}^{D \times T}$ there exists a mapping $f : \mathbb{R}^D \rightarrow \mathbb{R}^d$ s.t.

$$(1 - \epsilon)\|y_{t1} - y_{t2}\|^2 \leq \|f(y_{t1}) - f(y_{t2})\|^2 \leq (1 + \epsilon)\|y_{t1} - y_{t2}\|^2 \quad (\star)$$

Almost preserves pairwise distances!

- If $f(y) := d^{-1/2} R y$ with i.i.d. $\mathcal{N}(0, 1)$ entries of $R$ and reduced dimension

$$d \geq 4(\epsilon^2/2 - \epsilon^3/3)^{-1} \ln T + \mathcal{O}(\log \log T),$$

then $(\star)$ holds w.h.p. [Indyk-Motwani'98]

- If $f(y) := d^{-1/2} R y$ with i.i.d. uniform over $\{+1,-1\}$ entries of $R$ and reduced dimension as in JL lemma, then $(\star)$ holds w.h.p. [Achlioptas’01]
Performance of randomized LS

**Theorem**

For any $\epsilon > 0$, if $d = \mathcal{O}(p \log p / \epsilon^2)$, then w.h.p.

$$
\|y - X\hat{\theta}_{LS}\|_2 \leq (1 + \epsilon)\|y - X\theta_{LS}\|_2
$$

$$
\|\theta_{LS} - \hat{\theta}_{LS}\|_2 \leq \sqrt{\epsilon} \kappa(X) \sqrt{\gamma^{-2} - 1} \|\theta_{LS}\|_2
$$

$\kappa(X)$ condition number of $X$; and $\gamma = \|\hat{y}\|_2 / \|y\|_2$

- Uniform sampling vs Hadamard preconditioning
  - $D = 10,000$ and $p = 50$
  - Performance depends on $X$

---

Online censoring for large-scale regression

- **Key idea**: Sequentially test and update RLS estimates only for informative data

- **Adaptive censoring (AC) rule**

\[
c_n = \begin{cases} 
1, & \frac{|y_n - x_n^T \theta_{n-1}|}{\sigma} \leq \tau \\
0, & \text{otherwise.}
\end{cases}
\]

- **Criterion reveals “causal” support vectors (SVs)**

\[
f_n(\theta) = f(e_n) := \begin{cases} 
\frac{e_n^2}{2} - \frac{\tau^2 \sigma^2}{2}, & |e_n| > \tau \sigma \\
0, & |e_n| \leq \tau \sigma
\end{cases}
\]

- **Threshold controls avg. data reduction**: \( \tau \approx Q^{-1}\left(\frac{1}{2}(1 - \frac{d}{D})\right), \quad D \gg p \)

Censoring algorithms and performance

- AC least mean-squares (LMS)

\[ \hat{\theta}_n = \hat{\theta}_{n-1} + \mu (1 - c_n) x_n (y_n - x^T_n \hat{\theta}_{n-1}) \]

- AC recursive least-squares (RLS) at complexity \( \mathcal{O}(dp^2) \)

\[ \hat{\theta}_n = \hat{\theta}_{n-1} + (1 - c_n) \frac{1}{n} \hat{C}_n x_n (y_n - x^T_n \hat{\theta}_{n-1}) \]

\[ \hat{C}_n = \frac{n}{n-1} \left[ \hat{C}_{n-1} - (1 - c_n) \hat{C}_{n-1} x_n x^T_n \hat{C}_{n-1} \left( n - 1 + x^T_n \hat{C}_{n-1} x_n \right)^{-1} \right] \]

Proposition: AC-RLS

\[ \frac{1}{n} \text{tr} \left( R_x^{-1} \right) \sigma^2 \leq \mathbb{E} \left[ \| \hat{\theta}_n - \theta_0 \|_2^2 \right] \leq \frac{1}{n} \text{tr} \left( R_x^{-1} \right) \sigma^2 \forall n \geq k \]

AC-LMS

\[ \mathbb{E} \left[ \| \hat{\theta}_n - \theta_0 \|_2^2 \right] \leq \frac{\exp(4L^2/\alpha^2)}{n^2} \left( \| \theta_1 - \theta_0 \|_2^2 + \frac{\Delta}{L^2} \right) + 8 \frac{\Delta}{\alpha^2} \frac{\log n}{n} \]

- AC Kalman Filtering and Smoothing for “tracking with a budget”

Censoring vis-a-vis random projections

- Random projections for linear regression [Mahoney ‘11]
  - **Data-agnostic** reduction decoupled from LS solution

\[ \hat{\theta}_d = \arg \min_{\theta} \| S_d \, \text{HD}(y - X\theta) \|_2^2 \]

- Adaptive censoring (AC-RLS)
  - **Data-driven** measurement selection
  - Suitable also for streaming data
  - Minimal memory requirements
Performance comparison

- **Synthetic:** $D=10,000$, $p=300$ (50 MC runs); **Real data:** $\theta_0$, $\sigma$ estimated from full set

- Highly non-uniform data

- AC-RLS outperforms alternatives at comparable complexity

- Robustness to uniform data (all rows of $X$ equally “important”)

![Graph showing performance comparison]
Big data clustering

- Given $\{y_t\}_{t=1}^T$ with $\text{dim}(y_t) = D \gg$ assign them to clusters
- **Key idea:** Reduce dimensionality via random projections
- **Desiderata:** Preserve the pairwise data distances in lower dimensions

$$Y := [y_1, \ldots, y_T]$$

**Feature extraction**
- Construct $d \ll D$ “combined” features (e.g., via $RY$)
- Apply K-means to $d$-space

**Feature selection**
- Select $d \ll D$ of input features (rows of $Y$)
- Apply K-means to $d$-space
Random sketching and validation (SkeVa)

- Randomly select $d \ll D$ "informative" dimensions

- Algorithm
  
  For $r = 1, ..., R_{\text{max}}$
  
  - Sketch $d \ll D$ dimensions: $\mathbf{X} \rightarrow \tilde{\mathbf{X}}(r) \in \mathbb{R}^{d \times N}$
  
  - Run k-means on $\tilde{\mathbf{X}}(r) \rightarrow \{\tilde{C}_k(r)\}_{k=1}^{K}$, $\{\tilde{c}_k(r)\}_{k=1}^{K}$
  
  - Re-sketch $d' \leq D - d$ dimensions $\rightarrow \tilde{\mathbf{X}}(r') \in \mathbb{R}^{d' \times N}$
  
  - Augment centroids $\tilde{c}_k(r) := [\tilde{c}_k(r)^\top, \tilde{c}_k(r')^\top]^\top$ $\forall k$, $\tilde{c}_k(r') = \frac{1}{|\tilde{C}_k(r)|} \sum x_n(r) \in \tilde{C}_k(r) \tilde{x}(r')$
  
  - Validate using consensus set $S(r) = \{x_n | \tilde{x}_n(r) \in \tilde{C}_{k_1}(r), x_n(r) \in \tilde{C}_{k_2}(r), \text{ and } k_1 = k_2\}$
  
  $r^* = \arg\max_r f(S(r))$

- Similar approaches possible for $N \gg D$

- Sequential and kernel variants available

---

RP versus SkeVA comparisons

KDDb dataset (subset)
\(D = 2,990,384, \; T = 10,000, \; K = 2\)

RP: [Boutsidis et al. '13]
SkeVa: Sketch and validate

Closing comments

- **Big Data modeling and tasks**
  - Dimensionality reduction
  - Succinct representations
  - Vectors, matrices, and tensors

- **Learning algorithms**
  - Data sketching via random projections
  - Streaming, parallel, decentralized

- **Implementation platforms**
  - Scalable computing platforms
  - Analytics in the cloud

---