LAWRENCE LIVERMORE NATIONAL LABORATORY

CASIS LLNL

Learning Tools for Big Data Analytics

Georgios B. Giannakis

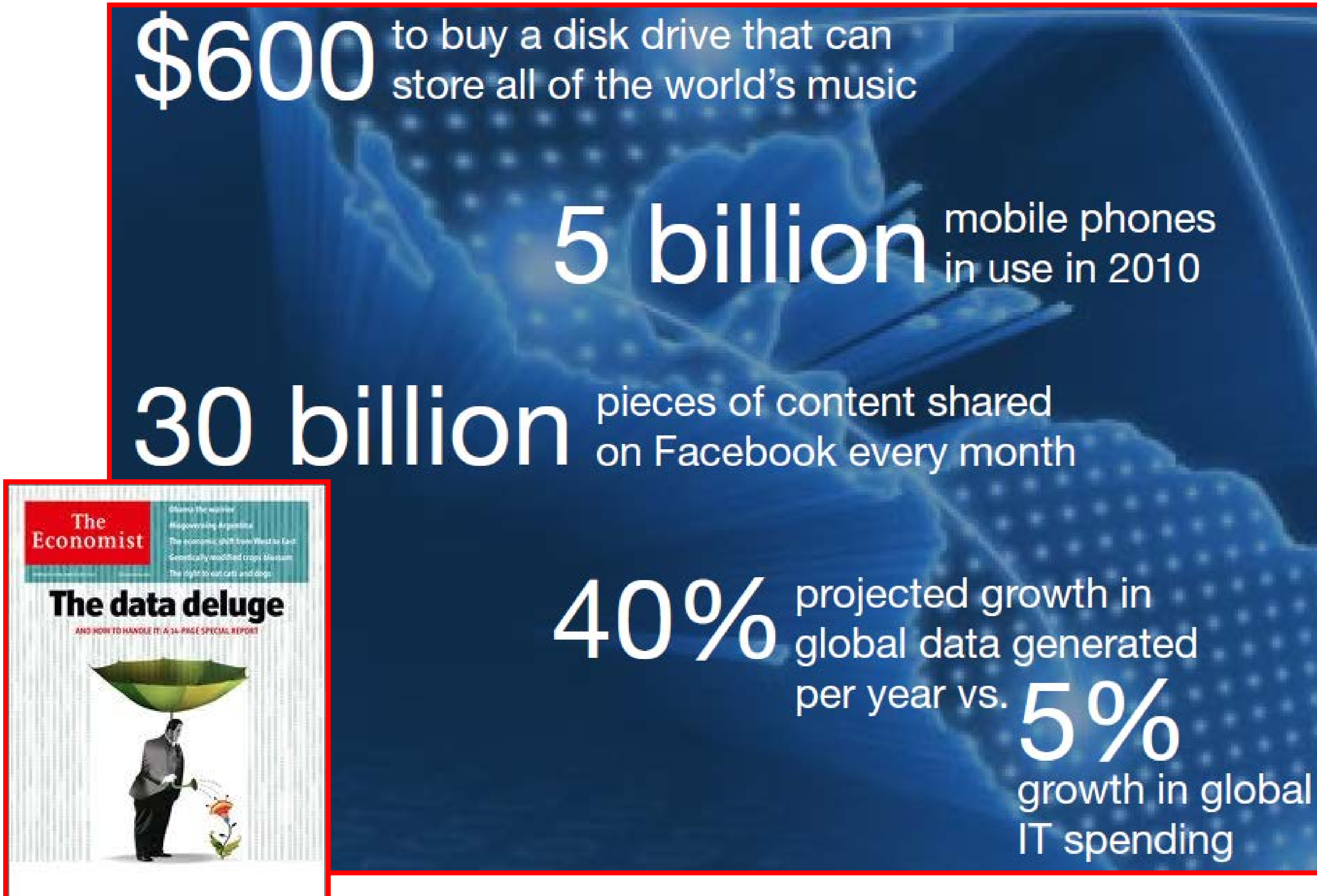
Acknowledgments: Profs. G. Mateos and K. Slavakis
NSF 1343860, 1442686, and MURI-FA9550-10-1-0567

UNIVERSITY OF MINNESOTA
Driven to Discover℠

Center for Advanced Signal and Image Sciences (CASIS)

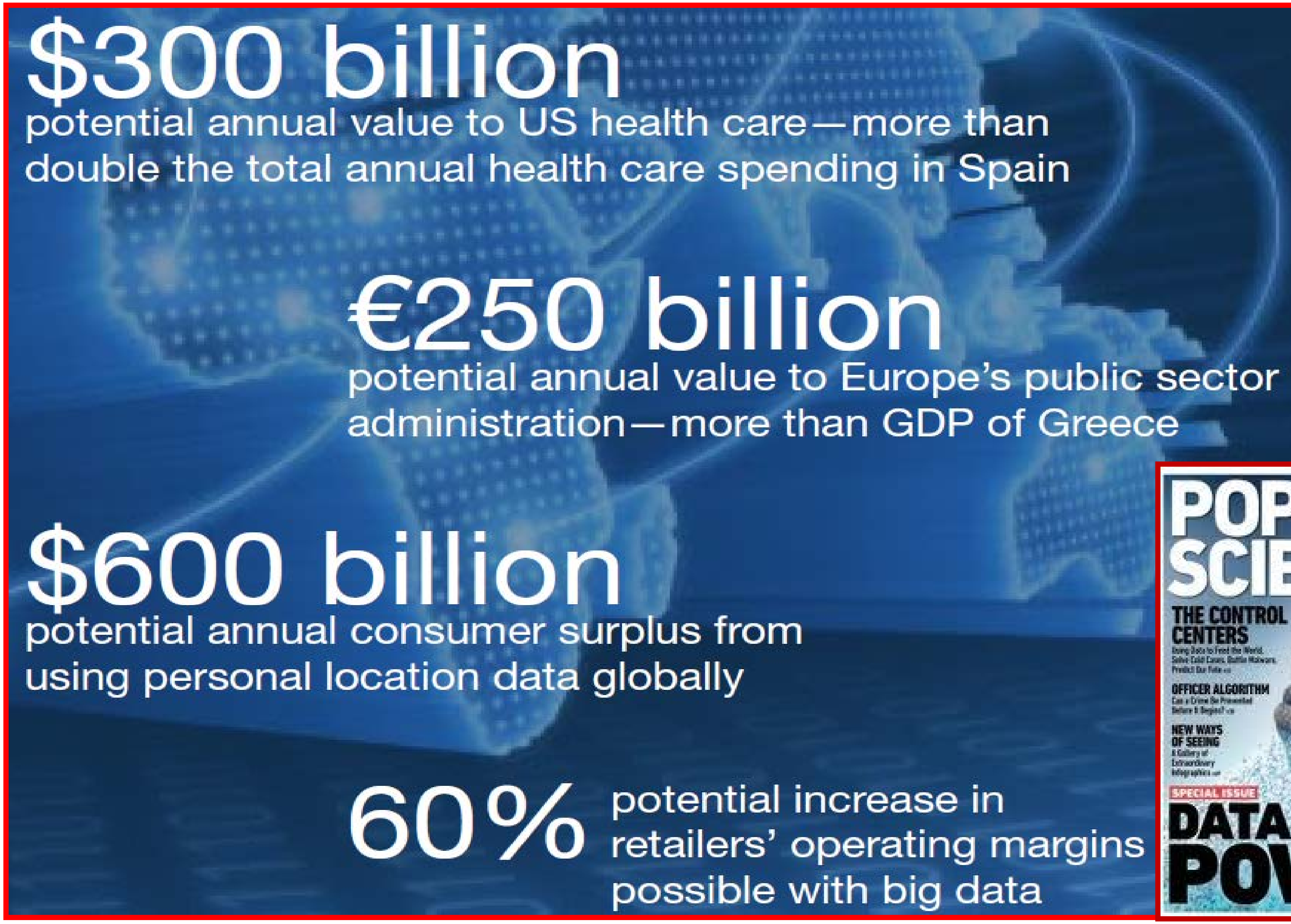
Livermore, May 13, 2015

Growing data torrent



Source: McKinsey Global Institute, "Big Data: The next frontier for innovation, competition, and productivity," May 2011.

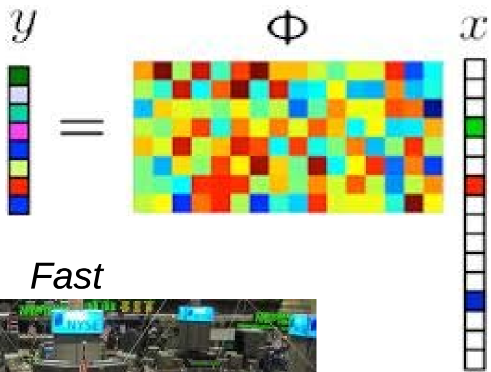
Big Data: Capturing its value



Source: McKinsey Global Institute, "Big Data: The next frontier for innovation, competition, and productivity," May 2011.

Challenges

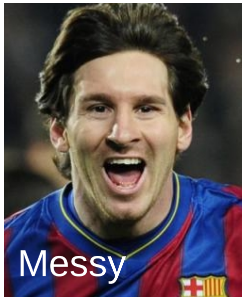
- ❑ Sheer <u>volume</u> of data
  - ➢ Decentralized and parallel processing
  - ➢ Security and privacy measures



- ❑ Modern massive datasets involve many <u>attributes</u>
  - ➢ Parsimonious models to ease interpretability
    and enhance learning performance

Fast

- ❑ <u>Real-time</u> streaming data
  - ➢ Online processing
    Quick-rough answer vs. slow-accurate answer?
- ❑ <u>Outliers</u> and <u>misses</u>
  - ➢ 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
- Resource tradeoffs
- Pursuit of low-dimensional structure
- Analysis of multi-relational data
- Common principles across networks

Algorithms and Implementation Platforms to Learn from Massive Datasets

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

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

Background (low rank)

Patterns, innovations, (co-)clusters, outliers

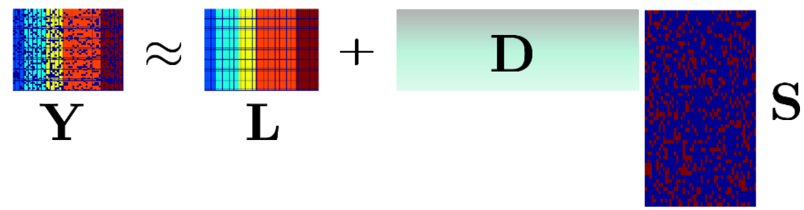
Observed data $\in \mathbb{R}^{D\times T}$

$$\mathbf{Y} = \mathbf{L} + \mathbf{D}\mathbf{S} + \mathbf{V}$$

Noise

Dictionary $\in \mathbb{R}^{D\times Q}$

Sparse matrix $\in \mathbb{R}^{Q\times T}$



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

$$[\mathcal{P}_\Omega(\mathbf{Y})]_{ij} = \begin{cases} [\mathbf{Y}]_{ij}, & \text{if } (i,j)\in\Omega \\ 0, & \text{o.w.} \end{cases}$$

  allow for misses

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

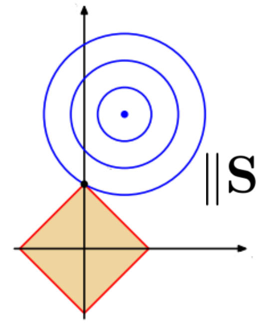
Subsumed paradigms

- Structure leveraging criterion

$$\min_{\{\quad\}} \frac{1}{2}\| \quad \mathbf{Y} \quad \|_{\mathrm{F}}^2$$

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

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



ℓ_1-norm

$\|\mathbf{S}\|_1 := \sum_{q,t} |s_{q,t}|$

(With or without misses)

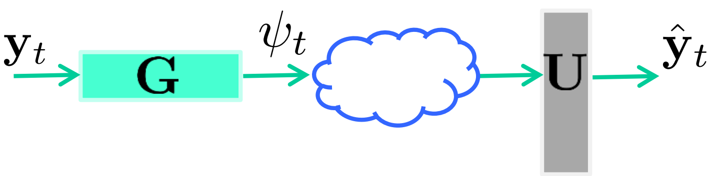
- $\boldsymbol{L} = \mathbf{0}, \boldsymbol{D}$ known $\Rightarrow$ Compressive sampling (CS) [Candes-Tao '05]
- $\boldsymbol{L} = \mathbf{0} \Rightarrow$ Dictionary learning (DL) [Olshausen-Field '97]
- $\boldsymbol{L} = \mathbf{0}, [\boldsymbol{D}]_{ij} \geq 0, [\boldsymbol{S}]_{ij} \geq 0 \Rightarrow$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

PCA formulations

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

- Minimum reconstruction error
  - Compression $\mathbf{G} \in \mathbb{R}^{d\times D}$
  - Reconstruction $\mathbf{U} \in \mathbb{R}^{D\times d}$

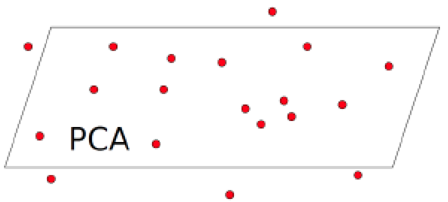
$d \ll D$



$$\min_{\mathbf{U},\mathbf{G}} \sum_{t=1}^T \|\mathbf{y}_t - \mathbf{U}\mathbf{G}\mathbf{y}_t\|_2^2, \quad \text{s.to. } \mathbf{U}^\top\mathbf{U} = \mathbf{I}_d$$

- Component analysis model $\mathbf{y}_t = \mathbf{U}\psi_t + \boldsymbol{\varepsilon}_t$

$$\min_{\mathbf{U},\psi_t} \sum_{t=1}^T \|\mathbf{y}_t - \mathbf{U}\psi_t\|_2^2, \quad \text{s.to. } \mathbf{U}^\top\mathbf{U} = \mathbf{I}_d$$



Solution: $\hat{\mathbf{U}}_d = d\text{-evecs}(\hat{\mathbf{C}}_{yy}),\ \hat{\mathbf{G}} = \hat{\mathbf{U}}_d^\top,\ \hat{\psi}_t = \hat{\mathbf{U}}_d^\top\mathbf{y}_t$

Dual and kernel PCA

- SVD: $\underbrace{\mathbf{Y}}_{D\times T} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$

$T \gg D$: $\mathbf{Y}\mathbf{Y}^\top = \mathbf{U}\boldsymbol{\Sigma}^2\mathbf{U}^\top \in \mathbb{R}^{D\times D} \quad \mathcal{O}(TD^2)$

$D \gg T$: $\boxed{\mathbf{Y}^\top\mathbf{Y}} = \mathbf{V}\boldsymbol{\Sigma}^2\mathbf{V}^\top \in \mathbb{R}^{T\times T} \quad \mathcal{O}(DT^2)$

Gram matrix

$$\hat{\mathbf{U}}_d = \mathbf{Y}\hat{\mathbf{V}}_d\hat{\boldsymbol{\Sigma}}_d^{-1}$$

$$\xrightarrow{\mathbf{y}_t} \hat{\mathbf{U}}_d^\top\mathbf{y}_t = \hat{\boldsymbol{\Sigma}}_d^{-1}\hat{\mathbf{V}}_d^\top\boxed{\mathbf{Y}^\top\mathbf{y}_t} \xrightarrow{\hat{\boldsymbol{\psi}}_t} \quad \xrightarrow{\hat{\boldsymbol{\psi}}_t} \hat{\mathbf{U}}_d\hat{\boldsymbol{\psi}}_t = \mathbf{Y}\hat{\mathbf{V}}_d\hat{\boldsymbol{\Sigma}}_d^{-1}\hat{\boldsymbol{\psi}}_t \rightarrow \hat{\mathbf{y}}_t$$

Inner products

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

A1. Stretch it to become linear: Kernel PCA; e.g., [Scholkopf-Smola'01]

- maps $\mathbf{y}_t$ to $\varphi(\mathbf{y}_t)$, and leverages dual PCA in high-dim spaces

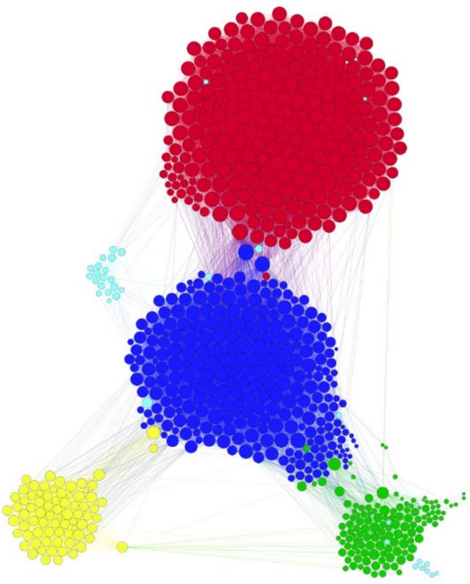
A2. General (non)linear models; e.g., union of hyperplanes, or, locally linear

- tangential hyperplanes

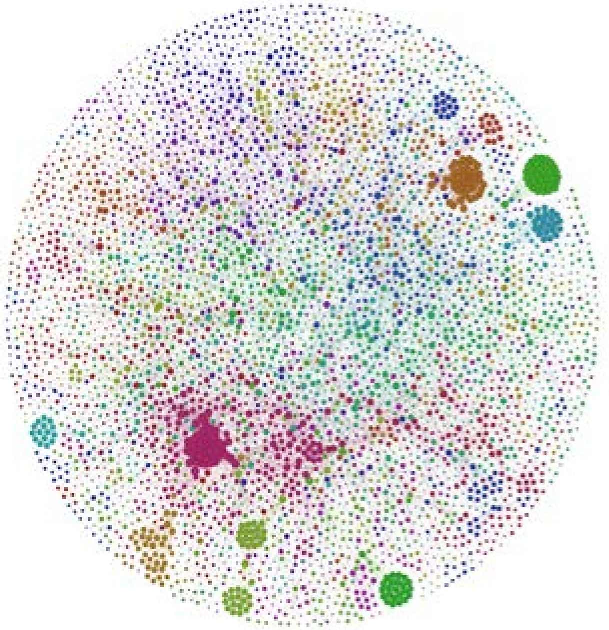
B. Schölkopf and A. J. Smola, "*Learning with Kernels*," Cambridge, MIT Press, 2001

Identification of network communities

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



Facebook egonet
744 nodes, 30,023 edges



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

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

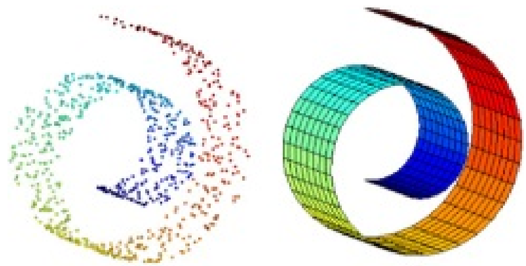
P. A. Traganitis, K. Slavakis, and G. B. Giannakis, "Spectral clustering of large-scale communities via random sketching and validation," *Proc. Conf. on Info. Science and Systems,* Baltimore, Maryland, March 18-20, 2015.

Local linear embedding

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

- Weight matrix captures local affine relations

$$\min_{\substack{\mathbf{W}:=[\mathbf{w}_1,\ldots,\mathbf{w}_T]\in\mathbb{R}^{T\times T}\\ \mathbf{W}^\top\mathbf{1}_T=\mathbf{0}_T}} \sum_{t=1}^{T}\left\|\mathbf{y}_t - \sum_{t'\in\mathcal{N}_t} w_{t't}\mathbf{y}_{t'}\right\|^2$$


(a) (b)

Sparse $\mathcal{N}_t$ and $\mathbf{W}$ [Elhamifar-Vidal‘11]

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

$$\min_{\substack{\boldsymbol{\Psi}:=[\boldsymbol{\psi}_1,\ldots,\boldsymbol{\psi}_T]\in\mathbb{R}^{d\times T}\\ \boldsymbol{\Psi}\boldsymbol{\Psi}^\top=\mathbf{I}_d\\ \boldsymbol{\Psi}\mathbf{1}_T=\mathbf{0}_d}} \left\{\sum_{t=1}^{T}\left\|\boldsymbol{\psi}_t - \sum_{t'=1}^{T} w_{t't}\boldsymbol{\psi}_{t'}\right\|^2 = \mathrm{trace}\Big[\boldsymbol{\Psi}(\mathbf{I}_T-\mathbf{W})(\mathbf{I}_T-\mathbf{W})^\top\boldsymbol{\Psi}^\top\Big]\right\}$$

Solution: The rows of $\boldsymbol{\Psi}$ are the d minor, excluding $\mathbf{1}_T$, $\mathrm{evecs}\big[(\mathbf{I}_T-\mathbf{W})(\mathbf{I}_T-\mathbf{W})^\top\big]$

L. K. Saul and S. T. Roweis, “Think globally, fit locally: Unsupervised learning of low dimensional manifolds,” *J. Machine Learning Research,* vol. 4, pp. 119-155, 2003.

Dictionary learning

- Solve for dictionary $\boldsymbol{D}$ and sparse $\boldsymbol{S}$: $\min_{\substack{\mathbf{D}\in\mathfrak{D} \\ \mathbf{S}\in\mathbb{R}^{Q\times T}}} \underbrace{\frac{1}{2}\|\mathcal{P}_\Omega(\mathbf{Y}-\mathbf{DS})\|_F^2 + \lambda_1 \sum_{t=1}^{T}\|\mathbf{s}_t\|_1}_{=:\mathcal{L}(\mathbf{D},\mathbf{S})}$

$$\mathfrak{D} := \{\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_Q] : \|\mathbf{d}_q\| \leq 1, \forall q\} \quad Q \geq D$$

$$\begin{cases} \mathbf{S}_{k+1} \in \arg \min_{\mathbf{S}\in\mathbb{R}^{Q\times T}} \overbrace{\frac{1}{2}\|\mathbf{Y}-\mathbf{D}_k\mathbf{S}\|_F^2 + \lambda_1\|\mathbf{S}\|_1}^{\mathcal{L}(\mathbf{D}_k,\mathbf{S})} \\ \boldsymbol{D}_{k+1} \in \arg \min_{\mathbf{D}\in\mathfrak{D}} \|\mathbf{Y}-\mathbf{D}\mathbf{S}_{k+1}\|_F^2 = \arg \min_{\mathbf{D}\in\mathfrak{D}} \mathcal{L}(\mathbf{D},\mathbf{S}_{k+1}) \end{cases}$$

(Lasso task; sparse coding)

(Constrained LS task)

- Alternating minimization; both $\mathcal{L}(\boldsymbol{D}_k,\cdot)$ and $\mathcal{L}(\cdot,\boldsymbol{S}_{k+1})$ are convex

- Under conditions, $(\boldsymbol{D}_k,\boldsymbol{S}_k)_{k=0}^{\infty}$ converges to a stationary point of $\mathcal{L}$ [Tseng'01]

B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?" *Vis. Res.*, vol. 37, no. 23, pp. 3311–3325, 1997.

Joint DL-LLE paradigm

$$\min_{\substack{\mathbf{D}_y,\ \mathbf{S},\ \mathbf{S}^\top \mathbf{1}_Q = \mathbf{1}_T \\ \mathbf{W},\ \mathbf{W}^\top \mathbf{1}_Q = \mathbf{1}_Q \\ \mathrm{diag}(\mathbf{W}) = \mathbf{0}}} \overbrace{\frac{1}{2}\|\mathcal{P}_\Omega(\mathbf{Y} - \mathbf{D}_y\mathbf{S})\|_{\mathrm{F}}^2}^{\text{DL fit}} + \overbrace{\frac{\lambda_y}{2}\|\mathbf{D}_y - \mathbf{D}_y\mathbf{W}\|_{\mathrm{F}}^2}^{\text{LLE fit}} + \overbrace{\sum_{t=1}^{T}\lambda_{st}\|\mathbf{s}_t\|_1 + \sum_{q=1}^{Q}\lambda_{wq}\|\mathbf{w}_q\|_1}^{\text{Sparsity regularization}}$$

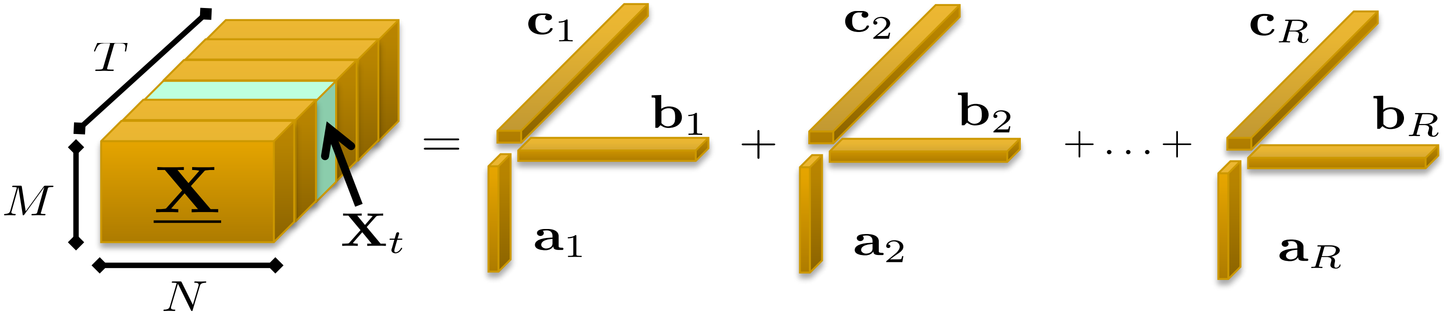
- Dictionary morphs data to a smooth basis; reduces noise and complexity

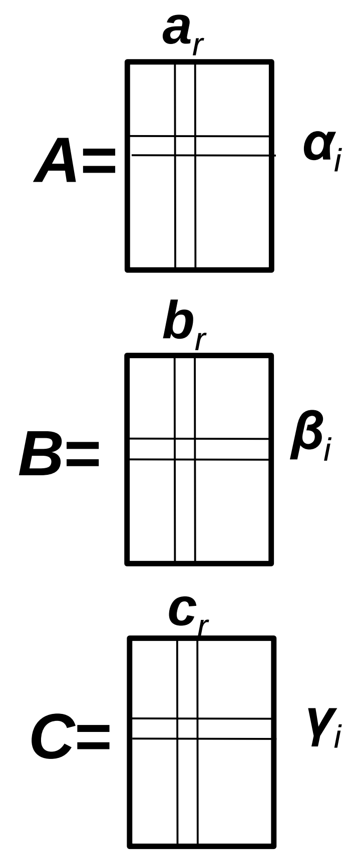
- Inpainting by local-affine-geometry-preserving DL for 50% misses; PSNR $\approx$ 32dB
- DL-LLE offers data-driven non-linear embedding for robust (de-)compression

K. Slavakis, G. B. Giannakis, and G. Leus, "Robust sparse embedding and reconstruction via dictionary learning," *Proc. of Conf. on Info. Science and Systems, JHU,* Baltimore, March 2013.

From low-rank matrices to tensors

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





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

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

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

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

J. A. Bazerque, G. Mateos, and G. B. Giannakis, "Rank regularization and Bayesian inference for tensor completion and xtrapolation," *IEEE Trans. on Signal Processing*, vol. 61, no. 22, pp. 5689-5703, November 2013.

Online tensor subspace learning

- Image domain low tensor rank $\mathbf{Y}_t^{\Omega} \approx \mathcal{F}_{\Omega_t}(\mathbf{A}\mathrm{diag}(\boldsymbol{\gamma}_t)\mathbf{B}^{\top})$

$$(\hat{\mathbf{A}}_t, \hat{\mathbf{B}}_t) = \arg\min_{\mathbf{A},\mathbf{B}} \frac{1}{t}\sum_{\tau=1}^{t} \min_{\boldsymbol{\gamma}_\tau} \left\{ \|\mathbf{Y}_\tau^{\Omega} - \mathcal{F}_{\Omega_\tau}(\mathbf{A}\mathrm{diag}(\boldsymbol{\gamma}_\tau)\mathbf{B}^{\top})\|_{\mathrm{F}}^2 + \frac{\lambda}{2}\|\boldsymbol{\gamma}_\tau\|^2 \right\} + \frac{\lambda}{2t}(\|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2)$$

- Tikhonov regularization promotes low rank

Proposition [Bazerque-GG '13]: With $[\boldsymbol{\sigma}]_r = \|\mathbf{a}_r\|\|\mathbf{b}_r\|\|\mathbf{c}_r\|$

$$\|\boldsymbol{\sigma}\|_{2/3}^{2/3} := \arg\min_{\{\mathbf{A},\mathbf{B},\mathbf{C}\}} \left(\|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2 + \|\mathbf{C}\|_F^2\right)$$

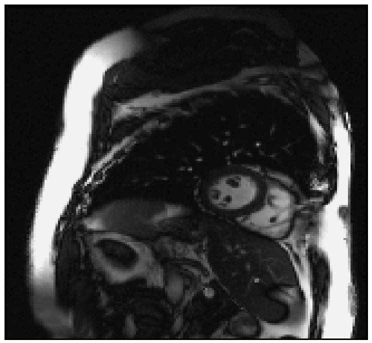
- Stochastic alternating minimization; parallelizable across bases

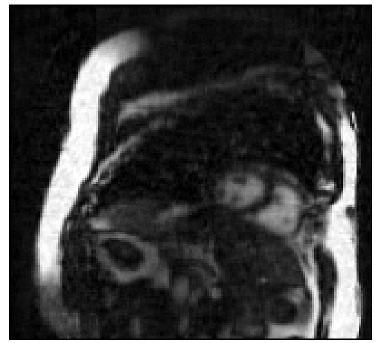
- Real-time reconstruction (FFT per iteration) $\hat{\mathbf{X}}_t = \hat{\mathbf{A}}_t \mathrm{diag}(\hat{\gamma}_t)\hat{\mathbf{B}}_t^{\top}$

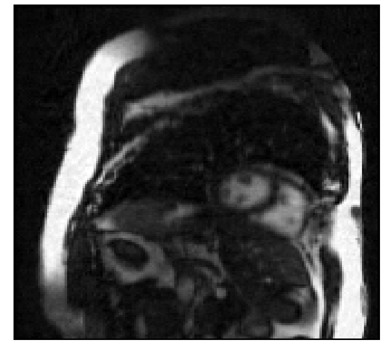
M. Mardani, G. Mateos, and G. B. Giannakis, "Subspace learning and imputation for streaming big data matrices and tensors," *IEEE Trans. on Signal Processing,* vol. 63, pp. 2663 - 2677, May 2015.

Dynamic cardiac MRI test

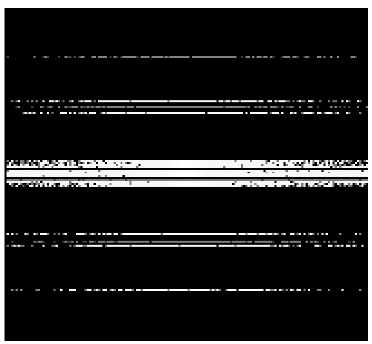
- *in vivo* dataset: 256 k-space 200x256 frames

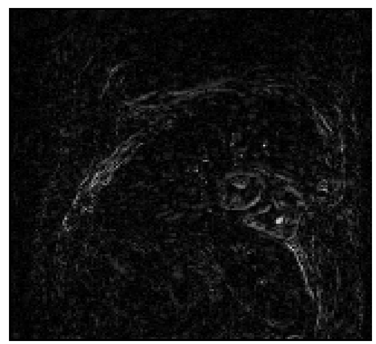


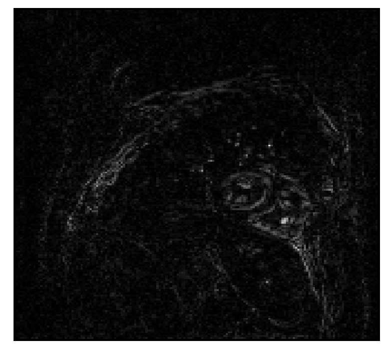




Ground-truth frame







Sampling trajectory

R=100, 90% misses

R=150, 75% misses

- Potential for accelerating MRI at high spatio-temporal resolution
- Low-rank $\mathcal{F}_{\Omega_t}(\mathbf{X}_t)$ plus $\mathcal{F}_{\Omega_t}(\mathbf{DS}_t)$ can also capture motion effects

M. Mardani and G. B. Giannakis, "Accelerating dynamic MRI via tensor subspace learning,“ *Proc. of ISMRM 23rd Annual Meeting and Exhibition*, Toronto, Canada, May 30 - June 5, 2015.

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 $\mathbf{y} \in \mathbb{R}^D, \ \mathbf{X} \in \mathbb{R}^{D\times p}$

$$\boldsymbol{\theta}_{\mathrm{LS}} := \arg \min_{\boldsymbol{\theta}\in\mathbb{R}^p} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2$$

If $\mathrm{rank}(\mathbf{X}) = p \Rightarrow \boldsymbol{\theta}_{\mathrm{LS}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}$

- SVD incurs complexity $\mathcal{O}(Dp^2)$. **Q:** What if $D \gg p$?

M. W. Mahoney, Randomized Algorithms for Matrices and Data, *Foundations and Trends In Machine Learning*, vol. 3, no. 2, pp. 123-224, Nov. 2011.

Randomized LS for linear regression

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

$$\check{\boldsymbol{\theta}}_{\text{LS}} = \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^p} \| \overbrace{\underbrace{\boldsymbol{\Gamma}_d \mathbf{S}_d}_{\mathbf{R}_2} \underbrace{\mathbf{H}_D \boldsymbol{\Delta}_D}_{\mathbf{R}_1}}^{\mathbf{R}} (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) \|_2^2$$

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

$$\mathbf{H}_D = \frac{1}{\sqrt{D}} \begin{bmatrix} \mathbf{H}_{D/2} & \mathbf{H}_{D/2} \\ \mathbf{H}_{D/2} & -\mathbf{H}_{D/2} \end{bmatrix}, \quad \mathbf{H}_2 := \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

  - subsets of data obtained by uniform sampling/scaling via $\mathbf{S}_d, \boldsymbol{\Gamma}_d$ 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 $o(Dp^2)$

N. Ailon and B. Chazelle, "The fast Johnson-Lindenstrauss transform and approximate nearest neighbors," *SIAM Journal on Computing*, 39(1):302–322, 2009.

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 $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_T] \in \mathbb{R}^{D \times T}$ there exists a mapping $f : \mathbb{R}^D \to \mathbb{R}^d$ s.t.

$$(1-\epsilon)\|\mathbf{y}_{t_1} - \mathbf{y}_{t_2}\|^2 \leq \|f(\mathbf{y}_{t_1}) - f(\mathbf{y}_{t_2})\|^2 \leq (1+\epsilon)\|\mathbf{y}_{t_1} - \mathbf{y}_{t_2}\|^2 \quad (\star)$$

Almost preserves pairwise distances!

- If $f(\mathbf{y}) := d^{-1/2}\mathbf{R}\mathbf{y}$ with i.i.d. $\mathcal{N}(0,1)$ entries of $\mathbf{R}$ 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 $f(\mathbf{y}) := d^{-1/2}\mathbf{R}\mathbf{y}$ with i.i.d. uniform over {+1,-1} entries of $\mathbf{R}$ and reduced dimension as in JL lemma, then $(\star)$ holds w.h.p. [Achlioptas'01]

W. B. Johnson and J. Lindenstrauss, "Extensions of Lipschitz maps into a Hilbert space," *Contemp. Math,* vol. 26, pp. 189–206, 1984.

Performance of randomized LS

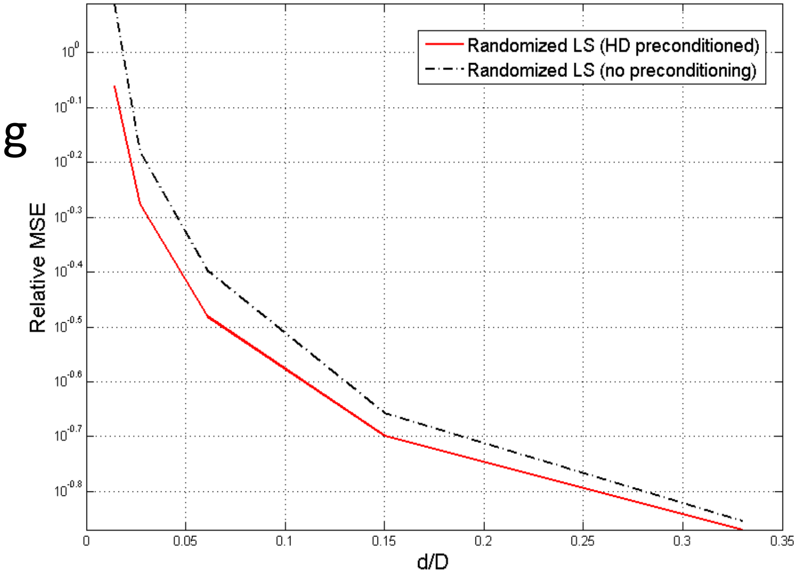
Theorem For any $\epsilon > 0$, if $d = \mathcal{O}(p \log p/\epsilon^2)$, then w.h.p.

$$\|\mathbf{y} - \mathbf{X}\check{\boldsymbol{\theta}}_{\text{LS}}\|_2 \leq (1+\epsilon)\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}_{\text{LS}}\|_2$$

$$\|\boldsymbol{\theta}_{\text{LS}} - \check{\boldsymbol{\theta}}_{\text{LS}}\|_2 \leq \sqrt{\epsilon}\,\kappa(\mathbf{X})\sqrt{\gamma^{-2}-1}\,\|\boldsymbol{\theta}_{\text{LS}}\|_2$$

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

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



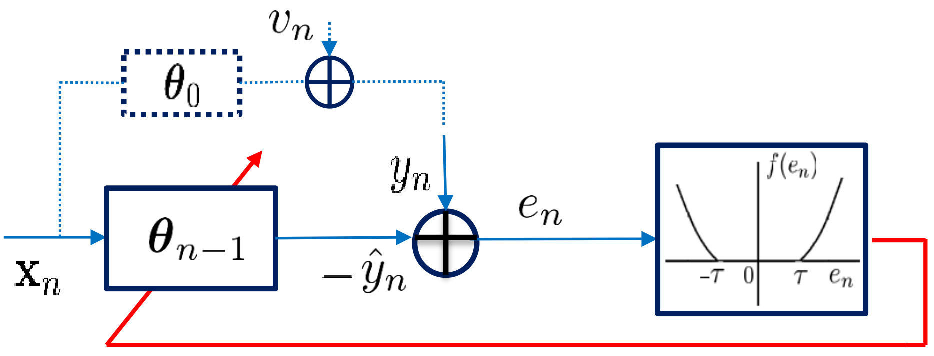
M. W. Mahoney, Randomized Algorithms for Matrices and Data, *Foundations and Trends In Machine Learning*, vol. 3, no. 2, pp. 123-224, Nov. 2011.

Online censoring for large-scale regression

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

- Adaptive censoring (AC) rule

$$c_n = \begin{cases} 1, & \frac{|y_n - \mathbf{x}_n^T \boldsymbol{\theta}_{n-1}|}{\sigma} \leq \tau \\ 0, & \text{otherwise.} \end{cases}$$



- Criterion reveals "causal" support vectors (SVs)

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

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

D. K. Berberidis, G. Wang, G. B. Giannakis, and V. Kekatos, "Adaptive Estimation from Big Data via Censored Stochastic Approximation," *Proc. of Asilomar Conf.*, Pacific Grove, CA, Nov. 2014.

Censoring algorithms and performance

- AC least mean-squares (LMS)

$$\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_{n-1} + \mu(1 - c_n)\mathbf{x}_n(y_n - \mathbf{x}_n^T\hat{\boldsymbol{\theta}}_{n-1})$$

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

$$\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_{n-1} + (1 - c_n)\frac{1}{n}\hat{\mathbf{C}}_n\mathbf{x}_n(y_n - \mathbf{x}_n^T\hat{\boldsymbol{\theta}}_{n-1})$$

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

Proposition: **AC-RLS** $\frac{1}{n}\mathrm{tr}\left(\mathbf{R}_{\mathrm{x}}^{-1}\right)\sigma^2 \leq \mathbf{E}\left[\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_2^2\right] \leq \frac{1}{n}\frac{\mathrm{tr}\left(\mathbf{R}_{\mathrm{x}}^{-1}\right)\sigma^2}{2Q(\tau)} \quad \forall n \geq k$

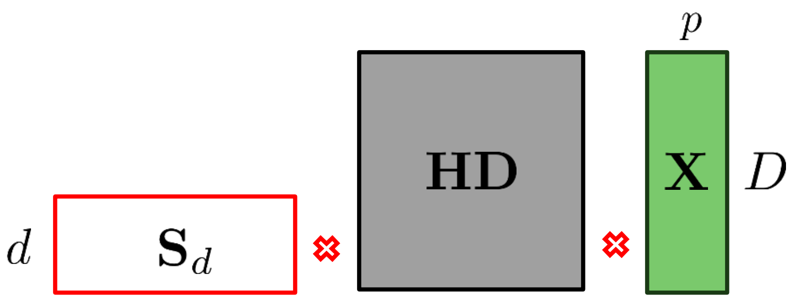
AC-LMS $\mathbb{E}\left[\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_2^2\right] \leq \frac{\exp(4L^2/\alpha^2)}{n^2}\left(\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0\|_2^2 + \frac{\Delta}{L^2}\right) + 8\frac{\Delta}{\alpha^2}\frac{\log n}{n}$

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

D. K. Berberidis, and G. B. Giannakis, "Online Censoring for Large-Scale Regressions," *IEEE Trans. on SP,* 2015 (submitted); also in *Proc. of ICASSP*, Brisbane, Australia, April 2015.

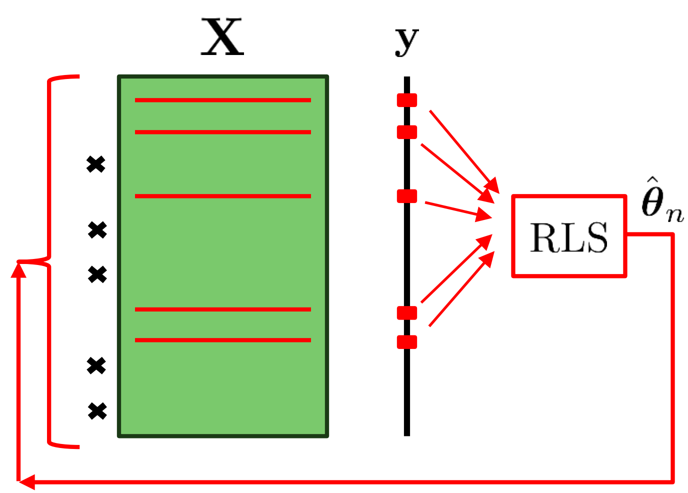
Censoring vis-a-vis random projections

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



$$\hat{\boldsymbol{\theta}}_d = \arg\min_{\boldsymbol{\theta}} \|\mathbf{S}_d \mathbf{HD}(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})\|_2^2$$

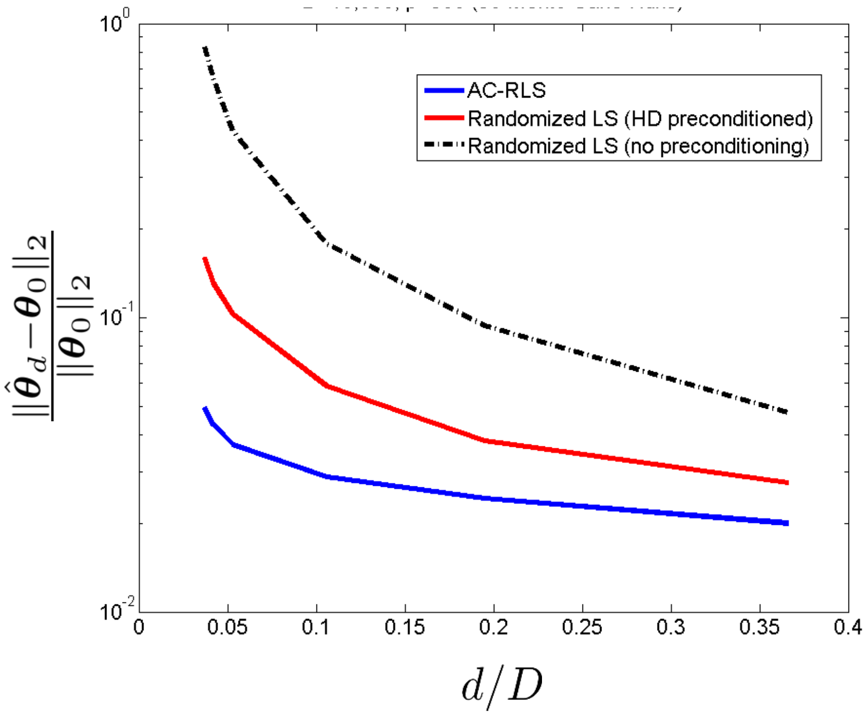
- Adaptive censoring (AC-RLS)
  - Data-driven measurement selection
  - Suitable also for streaming data
  - Minimal memory requirements

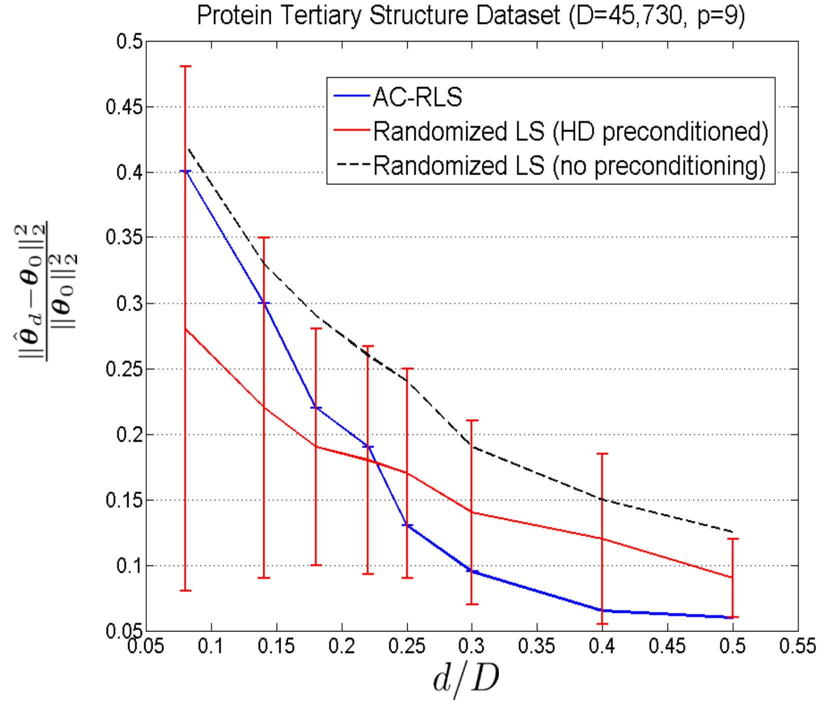


Performance comparison

- **Synthetic**: *D=10,000, p=300* (50 MC runs); **Real data**: $\boldsymbol{\theta}_0, \sigma$ estimated from full set

Highly non-uniform data





- AC-RLS outperforms alternatives at comparable complexity
- Robustness to uniform data (all rows of $\boldsymbol{X}$ equally "important")

Big data clustering

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

$$\mathbf{Y} := [\mathbf{y}_1, \ldots, \mathbf{y}_T]$$

Feature extraction

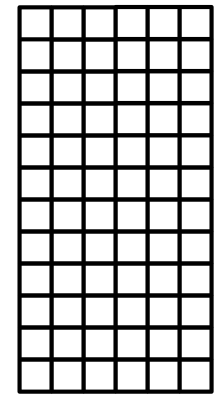
- Construct $d \ll D$ "combined" features (e.g., via $\mathbf{RY}$)
- Apply *K*-means to d-space

Feature selection

- Select $d \ll D$ of input features (rows of $\mathbf{Y}$)
- Apply *K*-means to d-space

Random sketching and validation (SkeVa)

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

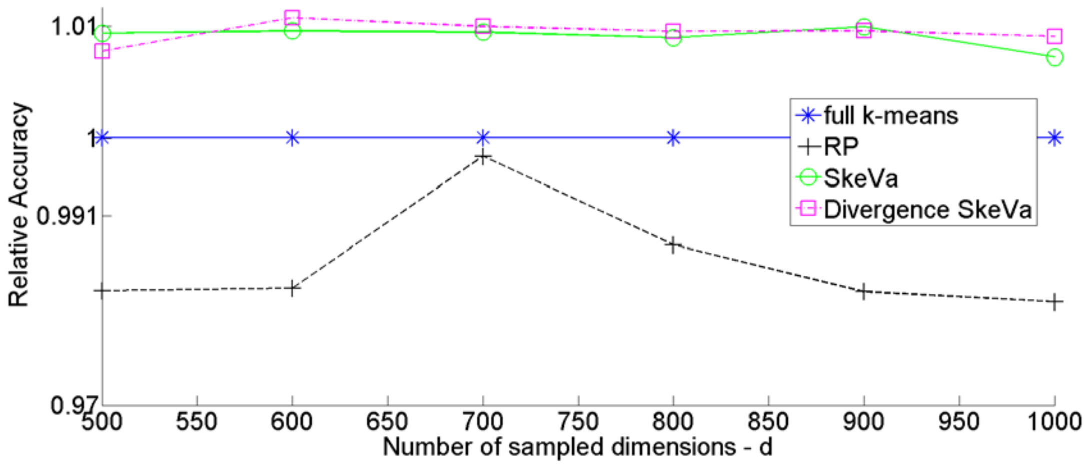


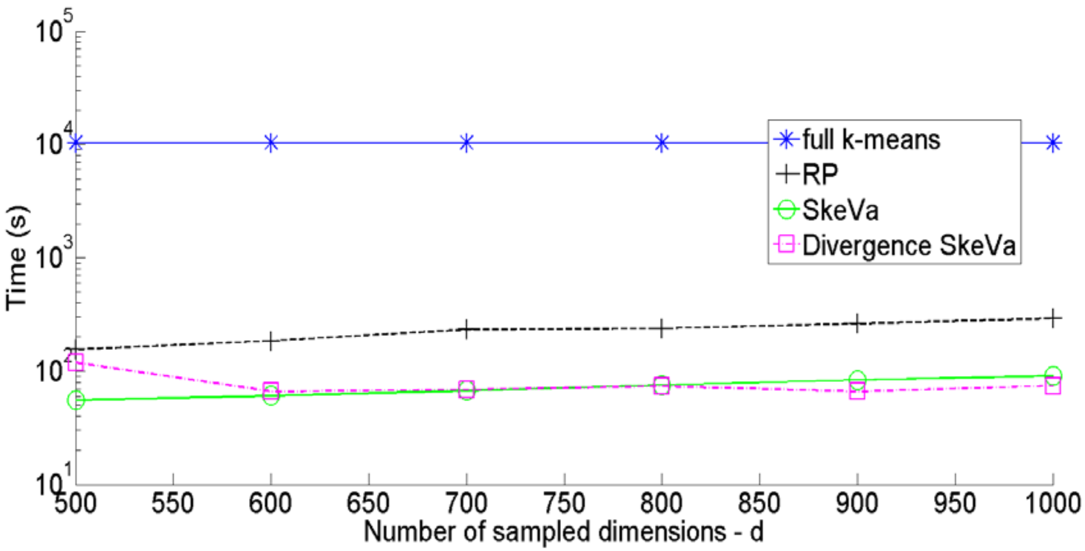
- Algorithm   For $r = 1, \ldots, R_{\max}$
  - Sketch $d \ll D$ dimensions: $\mathbf{X} \rightarrow \check{\mathbf{X}}^{(r)} \in \mathbb{R}^{d \times N}$
  - Run k-means on $\check{\mathbf{X}}^{(r)} \rightarrow \{\check{\mathcal{C}}_k^{(r)}\}_{k=1}^{K}, \{\check{\boldsymbol{c}}_k^{(r)}\}_{k=1}^{K}$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

- Similar approaches possible for $N \gg$
- Sequential and kernel variants available

P. A. Traganitis, K. Slavakis, and G. B. Giannakis, "Clustering High-Dimensional Data via Random Sampling and Consensus," *Proc. of GlobalSIP*, Atlanta, GA, December 2014.

RP versus SkeVA comparisons





KDDb dataset (subset)

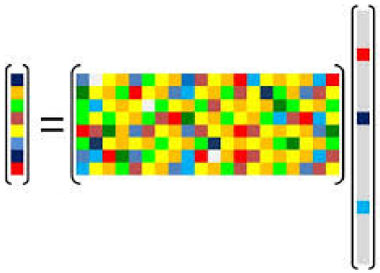
***D* = 2,990,384, *T* = 10,000,** *K* = 2

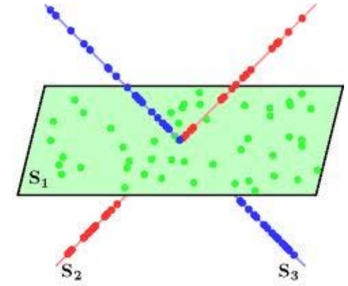
RP: [Boutsidis etal '13]

SkeVa: Sketch and validate

P. A. Traganitis, K. Slavakis, and G. B. Giannakis, "Sketch and Validate for Big Data Clustering," *IEEE Journal on Special Topics in Signal Processing*, June 2015.

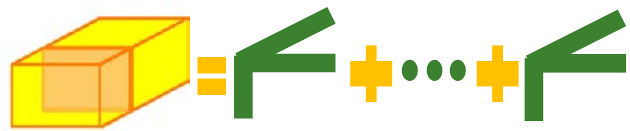
Closing comments





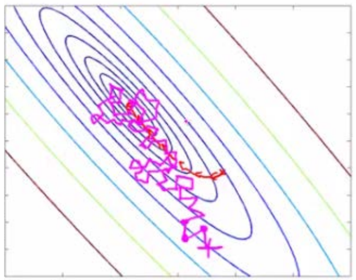
Big Data modeling and tasks

- 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

K. Slavakis, G. B. Giannakis, and G. Mateos, "Modeling and optimization for Big Data analytics," *IEEE Signal Processing Magazine*, vol. 31, no. 5, pp. 18-31, Sep. 2014.