Sampling for Subset Selection and Applications

Amit Deshpande
Microsoft Research India

Sample? why? when?

- Subsampling data when it is formidably large
- Feature selection, dimension reduction
- Randomized algorithms, hedging your bets against the adversary

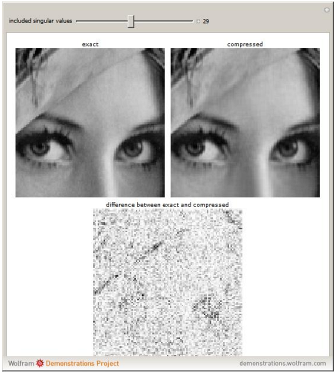
Outline

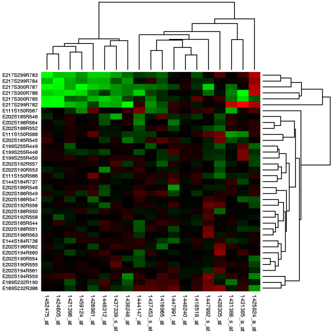
- Low-rank matrix approximation and SVD
- Row/column sampling techniques
- Determinantal Point Processes, rounding Lasserre solutions etc.

1. DPPs for Machine Learning by Kulesza-Taskar (2012), http://arxiv.org/abs/1207.6083
2. Guruswami-Sinop rounding of Lasserre SDPs (2011), http://www.math.ias.edu/~asinop/pubs/qip-gs11.pdf

Data = structure + noise

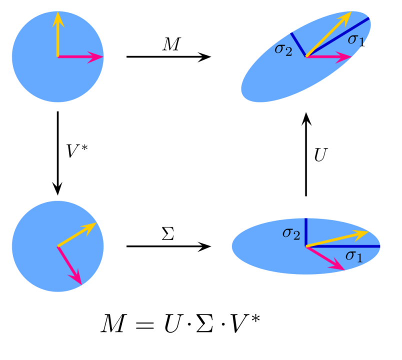
In matrix data, structure is often captured by an underlying low-rank matrix, and can be recovered by SVD.





`http://demonstrations.wolfram.com/ImageCompressionViaTheSingularValueDecomposition/`
`http://upload.wikimedia.org/wikipedia/commons/4/48/Heatmap.png`

Singular vectors and SVD



$$M = \underbrace{\sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \ldots + \sigma_k u_k v_k^T}_{\text{structure}} + \underbrace{\ldots + \sigma_n u_n v_n^T}_{\text{noise}},$$

where $\sigma_1 \geq \ldots \geq \sigma_n \geq 0$ and $\{u_i\}$, $\{v_j\}$ orthonormal.

http://commons.wikimedia.org/wiki/File:Singular-Value-Decomposition.svg

Low-rank matrix approximation

Given $A \in \mathbb{R}^{n \times d}$, find $B \in \mathbb{R}^{n \times d}$ of rank at most k that minimizes

$$\|A - B\|_F^2 = \sum_{ij} (A_{ij} - B_{ij})^2.$$

Low-rank matrix approximation

Given $A \in \mathbb{R}^{n\times d}$, find $B \in \mathbb{R}^{n\times d}$ of rank at most k that minimizes

$$\|A - B\|_F^2 = \sum_{ij}(A_{ij} - B_{ij})^2.$$

- Best rank-k approximation $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$. Geometrically, project each rows of A onto span $(v_1, \dots, v_k)$.
- SVD computation takes time $O\left(\min\{nd^2, n^2 d\}\right)$. Not fast enough for large data streams. Another drawback is that linear combinations of features/objects are not always meaningful. We rather want a subset of features/objects.

Dimension reduction

- Random projection aka Johnson-Lindenstrauss: $R \in \mathbb{R}^{d \times t}$, where $t = O\left(\frac{\log n}{\epsilon^2}\right)$ with i.i.d. $\sqrt{\frac{t}{d}}\ N(0,1)$ entries, followed by SVD of $AR \in \mathbb{R}^{n \times t}$ gives

$$\|A - (AR)_k\|_F^2 \leq \|A - A_k\|_F^2 + \epsilon \|A\|_F^2 , \qquad \text{w.h.p.}$$

Dimension reduction

- Random projection aka Johnson-Lindenstrauss: $R \in \mathbb{R}^{d \times t}$, where $t = O\left(\frac{\log n}{\epsilon^2}\right)$ with i.i.d. $\sqrt{\frac{t}{d}}\, N(0,1)$ entries, followed by SVD of $AR \in \mathbb{R}^{n \times t}$ gives

$$\|A - (AR)_k\|_F^2 \leq \|A - A_k\|_F^2 + \epsilon \|A\|_F^2, \qquad \text{w.h.p.}$$

- Squared-length sampling by Frieze-Kannan-Vempala: Pick $O\left(\frac{k}{\epsilon}\right)$ rows of A with $\Pr(i) \propto \|a_i\|^2$, project all rows onto their span to get $\tilde{A}$, and then compute SVD of $\tilde{A}$, which gives

$$\left\|A - \tilde{A}_k\right\|_F^2 \leq \|A - A_k\|_F^2 + \epsilon \|A\|_F^2, \qquad \text{w.h.p.}$$

w.h.p. here means extra $\log\left(\frac{1}{\delta}\right)$ factor for success probability $1 - \delta$.

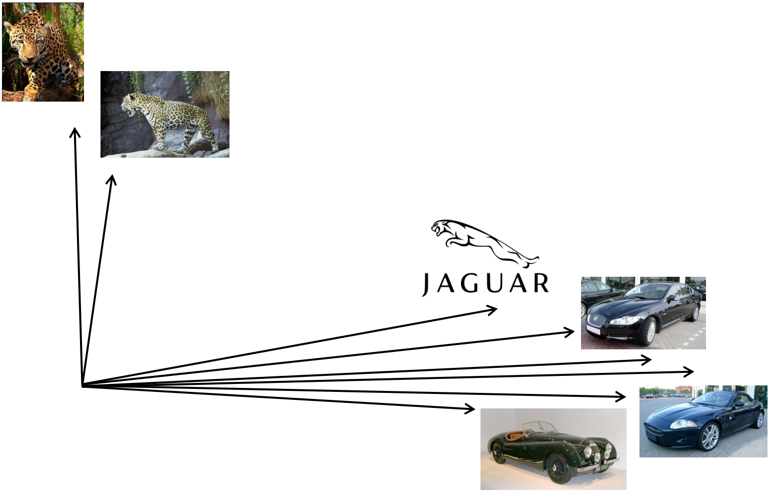
Adaptive sampling and volume sampling

We can pick $O\left(\frac{k}{\epsilon}\right)$ rows of A, in time $\tilde{O}\left(nd\frac{k}{\epsilon}\right)$, such that projecting onto their span followed by SVD gives

$$\left\|A - \tilde{A}_k\right\|_F^2 \leq (1+\epsilon)\left\|A - A_k\right\|_F^2, \qquad \text{w.h.p.}$$

- D-Rademacher-Vempala-Wang and D-Vempala (2006), D-Rademacher (2010), Guruswami-Sinop (2012)
- Drineas-Mahoney-Muthukrishnan (2006), Boutsidis-Drineas-Magdon Ismail (2011), using leverage scores and Batson-Spielman-Srivastava sparsification technique
- Sarlos (2007), Dasgupta-Kumar-Sarlos (2010), Clarkson-Woodruff (2012), no row/column subset selection but much faster algorithms using sparse subspace embeddings

Adaptive sampling



Volume sampling

Probability distribution over all k-subsets of $[n]$, where

$$\text{probability of picking } S \propto \operatorname{vol}(P_S)^2,$$

where P_S is a parallelepiped formed by the rows $\{a_i \; : \; i \in S\}$.

Volume sampling

Probability distribution over all k-subsets of $[n]$, where

$$\text{probability of picking } S \propto \operatorname{vol}(P_S)^2,$$

where P_S is a parallelepiped formed by the rows $\{a_i : i \in S\}$.

- $k = 1$ gives squared-length sampling
- Can we sample from this distribution efficiently?
  Yes, in $O(knd^2)$ time. In fact, $(1+\epsilon)$-approximate sampling in $\tilde{O}\left(nd\frac{k^2}{\epsilon^2}\right)$ time, using generalization of JL lemma to volumes by Magen-Zouzias (2008).

Why can we do volume sampling efficiently?

- Interesting identity using coeffs. of characteristic polynomial

$$\sum_{|S|=k} \mathrm{vol}\left(P_S\right)^2 = \sum_{i_1<\ldots<i_k} \sigma_{i_1}^2 \sigma_{i_2}^2 \cdots \sigma_{i_k}^2 = \left| c_{n-k}(AA^T) \right|.$$

Easy cases $\sum_i \|a_i\|^2 = \sum_i \sigma_i^2$ and $\mathrm{vol}\left(P_{[n]}\right)^2 = \sigma_1^2 \cdots \sigma_n^2$.

- Nice expression for marginals

$$\begin{aligned} \Pr\left(i \in S\right) &\propto \sum_{|S|=k \text{ and } i\in S} \mathrm{vol}\left(P_S\right)^2 \\ &= \|a_i\|^2 \sum_{|T|=k-1} \mathrm{vol}\left(P'_T\right)^2, \end{aligned}$$

where parallelepiped P'_T is formed by projections of a_j, for $j \in T$, orthogonal to a_i.

Deterministic row/column subset selection

- Volume sampling can be derandomized using the method of conditional expectations.
- Adaptive sampling part only uses pairwise independence, so can also be derandomized.
- Combining these almost matches the *deterministic* row/column subset selection of Boutsidis-Drineas-Magdon Ismail (2011) that used Batson-Spielman-Srivastava sparsification technique instead.
- Provides efficient rank-revealing RRQR decomposition improving upon Gu-Eisenstat (1996).

From volume sampling to DPPs

- Volume sampling is a special case of *Determinantal Point Processes* arising in quantum physics and random matrix theory. DPPs capture many interesting distributions including random spanning trees, non-intersecting random walks, eigenvalues of random matrices etc.

- Distribution over *all* subsets of $[n]$ such that for a random subset R, $\Pr(S \subseteq R) = \det(M_{S,S})$, where $0 \preccurlyeq M \preccurlyeq I$.

- Ben Hough-Krishnapur-Peres-Virág (2006)
`http://front.math.ucdavis.edu/math.PR/0503110`

ML and big data applications of subset selection

- Determinantal point processes for machine learning, Kulesza-Taskar, Foundations and Trends in ML, NOW Publishers, December 2012. http://arxiv.org/pdf/1207.6083v4.pdf
- Sampling methods for the Nyström method, Kumar-Mohri-Talwalkar, JMLR'12.
  adaptive sampling to speed up kernel algorithms for image segmentation, manifold learning
- Spectral methods in machine learning and new strategies for very large datasets, Belabbas-Wolfe, PNAS'09.
  heuristic Metropolis algorithm for volume sampling
- CUR matrix decompositions for improved data analysis, Drineas-Mahoney, PNAS'09.
  row/column sampling on gene expression data

k-means++ clustering

- k-means clustering: Given points $a_1, a_2, \ldots, a_n \in \mathbb{R}^d$, find k centers $c_1, \ldots, c_k \in \mathbb{R}^d$ that minimize sum of squared distances of all points to their nearest centers, respectively.
- Lloyd's iterative method starts with k initial centers, computes the corresponding clusters, then reassigns c_i's as their means, and iterates. Converges only to a local minimum and does not have good theoretical guarantees.
- k-means++ by Arthur-Vassilvitskii (2007) is initialization via *adaptive sampling*, and gives $O(\log k)$ approximation in expectation.
- Aggarwal-D-Kannan (2009) k-means++ actually gives $O(1)$ approximation using $2k$ centers, w.h.p.

Guruswami-Sinop rounding of Lasserre SDPs

Lasserre SDP for sparsest cut problem produces vectors $x_S(f)$ for *small* subsets S of vertices and $f \in \{0,1\}^{|S|}$, and adds constraints to the usual SDP.

$$\begin{aligned}
\text{minimize} \quad & \sum_{ij \in E} \left\| x_{\{i\}}(1) - x_{\{j\}}(1) \right\|_2^2, \\
\text{subject to} \quad & \sum_{i<j} \left\| x_{\{i\}}(1) - x_{\{j\}}(1) \right\|_2^2 = 1, \\
& \|x_\emptyset\|_2^2 > 0, \quad \text{and} \\
& x_S(f) \text{ satisfy Lasserre conditions for } |S| \le r.
\end{aligned}$$

Can we round $x_{\{i\}}(1)$'s using the extra information about $x_S(f)$'s?

Guruswami-Sinop rounding of Lasserre SDPs

- To round sparsest cut SDP, suffices to give a *good* ℓ_2^2-to-ℓ_1 embedding of $x_{\{i\}}(1)$'s.

- Guruswami-Sinop give such embedding as $y_i = \left(\langle x_S(f), x_{\{i\}}(1)\rangle\right)_{f \in \{0,1\}^{|S|}}$, and show that

  $$\left\|x_{\{i\}}(1) - x_{\{j\}(1)}\right\|_2^2 \geq \|y_i - y_j\|_1 \geq \left\|\Pi_S\left(x_{\{i\}}(1) - x_{\{j\}}(1)\right)\right\|_2^2,$$

  where Π_S is orthogonal projection onto the span of $\{x_{\{i\}}(1) \ : \ i \in S\}$.

- And use row/column subset selection to pick S and obtain *good* approximation guarantees. For details, see `http://arxiv.org/abs/1104.4746` and `http://arxiv.org/abs/1112.4109`.

Summary

- Adaptive/volume sampling as generalizations of squared-length sampling
- Determinantal Point Processes (DPPs)
- Applications to clustering, machine learning, optimization

Thank you. Any questions?