Augmented Leverage Score Sampling

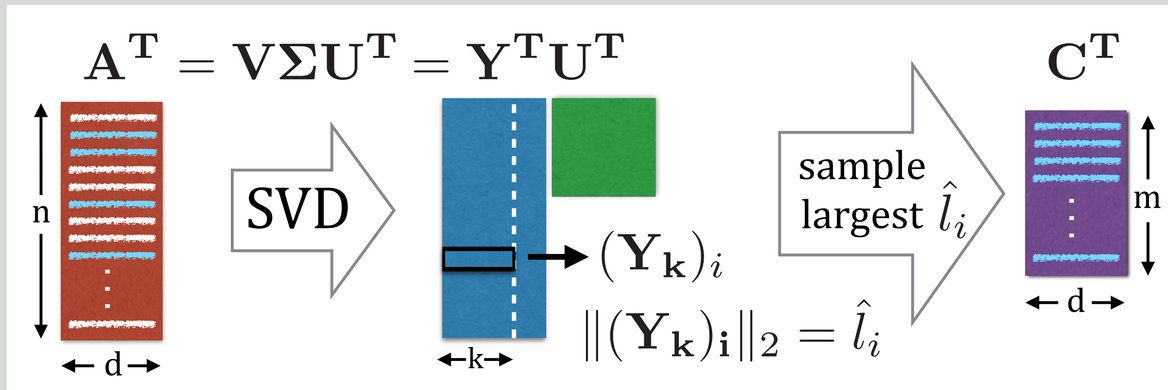
Daniel J. Perry and Ross T. Whitaker
SCI Institute, University of Utah, USA

Objective for Column Subset Selection Problem (CSSP)

Let $\mathbf{A} \in \mathbb{R}^{d \times n}$ and let $m < n$ be a sampling parameter. Find m columns for A - denoted as $\mathbf{C} \in \mathbb{R}^{d \times m}$ that minimize

$$\|\mathbf{A} - \mathbf{C}\mathbf{C}^{\dagger}\mathbf{A}\|_{\eta},$$

for $\eta \in \{F, 2\}$, and where $\mathbf{C}^{\dagger}$ denotes the Moore-Penrose pseudo-inverse.



Algorithm: Augmented Leverage Score Sampling

Input $\mathbf{A} \in \mathbb{R}^{d \times n}$, k, θ
Compute $\mathbf{Y_k} = \mathbf{V_k}\hat{\mathbf{\Sigma}}_\mathbf{k} \in \mathbb{R}^{n \times k}$
Compute $\hat{l}_i^{(k)} = \|[\mathbf{Y_k}]_{i,:}\|_2^2 \forall i = 1, 2, \ldots, n$
Let $\hat{l}_i^{(k)}$'s be sorted, $\hat{l}_1^{(k)} \geq \ldots \geq \ldots \geq \hat{l}_n^{(k)}$
Find index $m \in \{1, \ldots, n\}$ such that:

$$m = \arg\min_m \left(\sum_{i=1}^{m} \hat{l}_i^{(k)} > \theta \right).$$

If $m < k$, set $m = k$.
Output $\mathbf{S} \in \mathbb{R}^{n \times m}$, s.t. $\mathbf{AS}$ has the top m columns of $\mathbf{A}$.

Theorem: Frobenius and spectral error bound

Let $\theta = k \cdot \hat{\sigma}_1^2(\mathbf{\Sigma_k}) - \epsilon$ for some $\epsilon \in (0, 1)$, and let $\mathbf{S} \in \mathbb{R}^{n \times m}$ be the sampling matrix from the augmented leverage sampling algorithm, then, for $\mathbf{C} = \mathbf{AS}$ and $\zeta = \{2, F\}$

$$\|\mathbf{A} - \mathbf{C}\mathbf{C}^{\dagger}\mathbf{A}\|_\zeta^2 < \frac{\hat{\sigma}_1^2(\mathbf{\Sigma_k})}{1 - \epsilon} \cdot \|\mathbf{A} - \mathbf{A_k}\|_\zeta^2,$$

where $\hat{\sigma}_1 = \sigma_1 / \sigma_k$. We can rewrite the bound as

$$\|\mathbf{A} - \mathbf{C}\mathbf{C}^{\dagger}\mathbf{A}\|_\zeta^2 < (1 + 2\epsilon) \cdot \hat{\sigma}_1^2(\mathbf{\Sigma_k}) \cdot \|\mathbf{A} - \mathbf{A_k}\|_\zeta^2$$

if $\epsilon < \frac{1}{2}$.

Previous work

- *Deterministic leverage score sampling* bound due to (Papailiopoulos, et al., KDD 2014),

$$\|\mathbf{A} - \mathbf{C}\mathbf{C}^{\dagger}\mathbf{A}\|_\zeta^2 < (1 + 2\epsilon) \cdot \|\mathbf{A} - \mathbf{A_k}\|_\zeta^2.$$

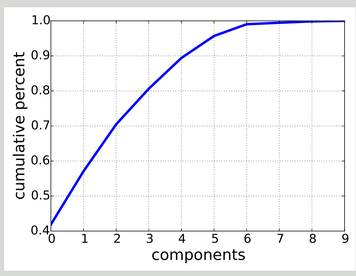
  for $\epsilon \in (0, 0.5)$.
- *Greedy column subset selection* tight bound due to (Altschuler, et al. ICML 2016)

$$\|\mathbf{C}\mathbf{C}^{\dagger}\mathbf{A}\|_F \geq (1 + \epsilon)\|\mathbf{D}\mathbf{D}^{\dagger}\mathbf{A}\|_F$$

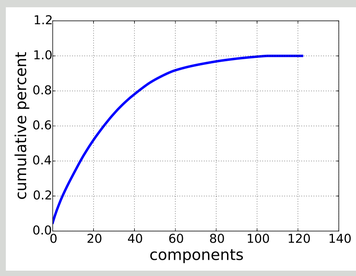
  where $\mathbf{D}$ is the optimal subset of size m from A.

Datasets

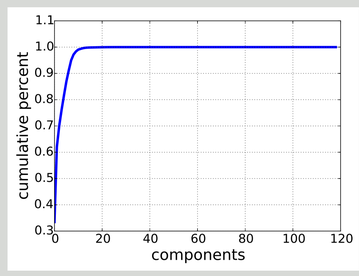
- Methods were compared on real and synthetic datasets, including three real datasets below from the UCI machine learning repository.



DIABETES
$\mathbf{442 \times 9}$

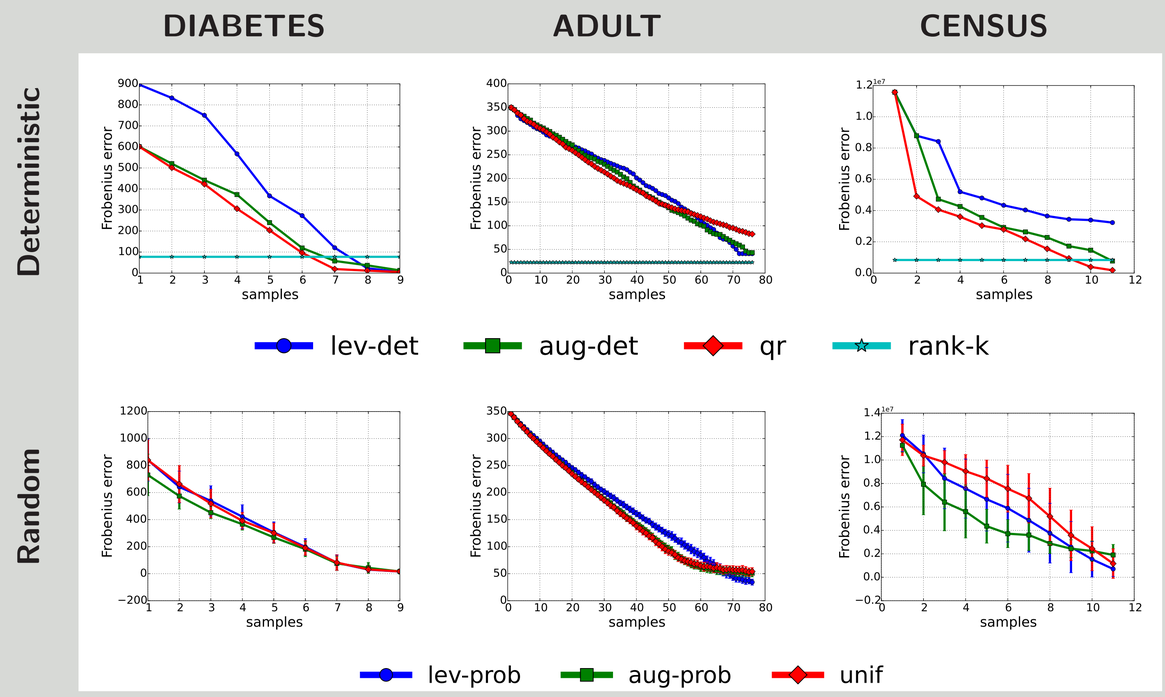


ADULT
$\mathbf{16281 \times 123}$

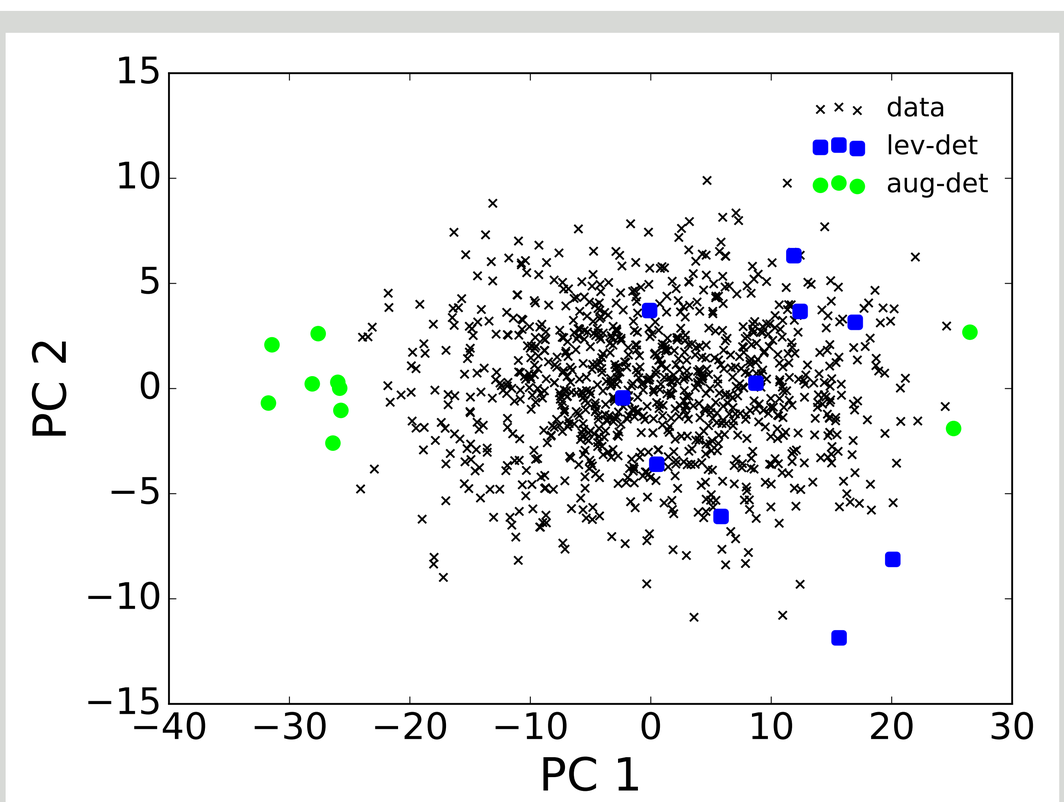


CENSUS
$\mathbf{2273 \times 119}$

Results

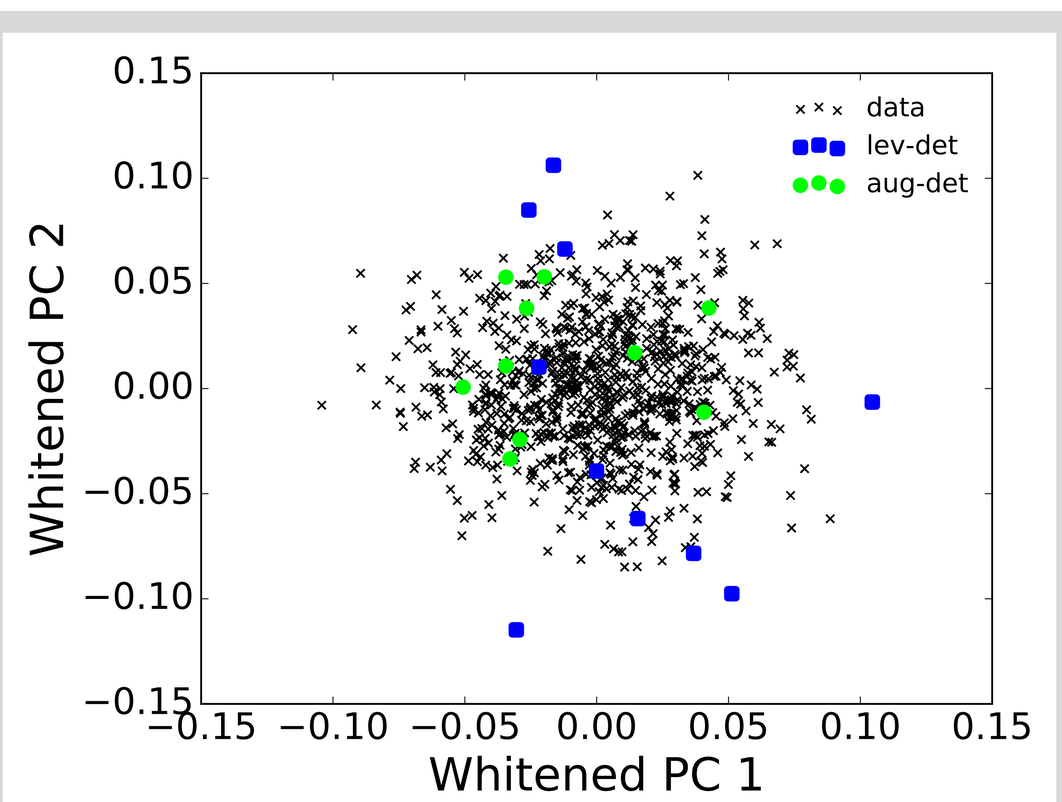


Augmented leverage score space



The data points shown are $\mathbf{P}_{\text{aug-lev}} = \mathbf{\Sigma}_2\mathbf{V}_2^{\mathsf{T}} = \mathbf{U}_2^{\mathsf{T}}\mathbf{A}$ (PCA projection)

Leverage score space



The data points shown are $\mathbf{P}_{\text{lev}} = \mathbf{V}_2^{\mathsf{T}} = \mathbf{\Sigma}^{\dagger}\mathbf{U}_2^{\mathsf{T}}\mathbf{A}$ (Whitened PCA projection)