Efficiently Implementing Sparsity in Learning

M. Magdon-Ismail

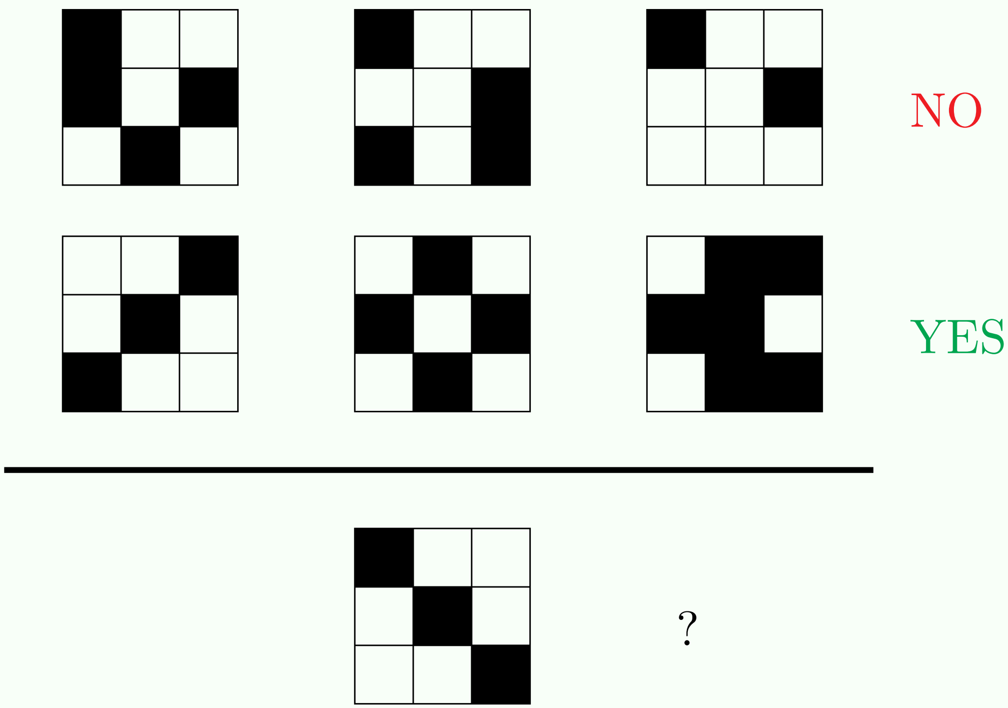
Rensselaer Polytechnic Institute

(Joint Work)

"Waiter! My glass is half empty."

December 9, 2013.

Out-of-Sample is What Counts



- A **pattern** exists
- We **don't know it**
- We **have data** to learn it
- Tested on **new cases**

Data

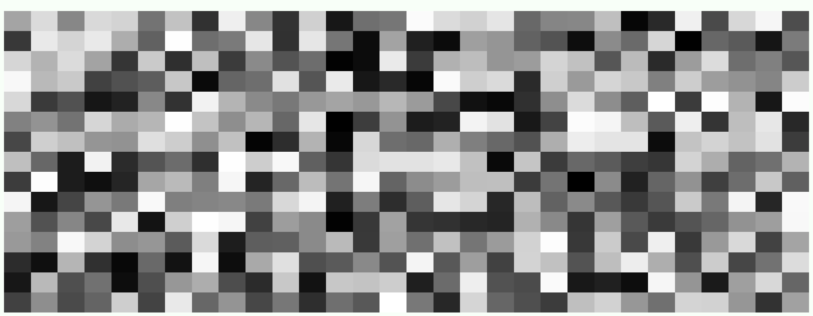
Data Matrix

d dimensions ☹

n data points ☺

$$\begin{array}{c}
\begin{array}{cccccccc} \text{name} & \text{age} & \text{debt} & \text{income} & \cdots & \text{hair} & \text{weight} & \text{sex} \end{array} \\
\left[\begin{array}{cccccccc}
John & 21yrs & -\$10K & \$65K & \cdots & black & 175lbs & M \\
Joe & 74yrs & -\$100K & \$25K & \cdots & blonde & 275lbs & M \\
Jane & 27yrs & -\$20K & \$85K & \cdots & blonde & 135lbs & F \\
\vdots & & & & & & & \\
Jen & 37yrs & -\$400K & \$105K & \cdots & brun & 155lbs & F
\end{array}\right]
\end{array}$$

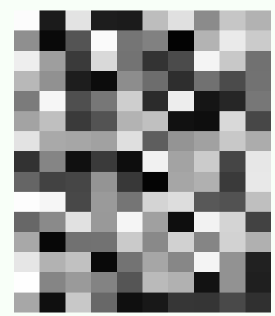
$$X \in \mathbb{R}^{n \times d}$$



Response Matrix

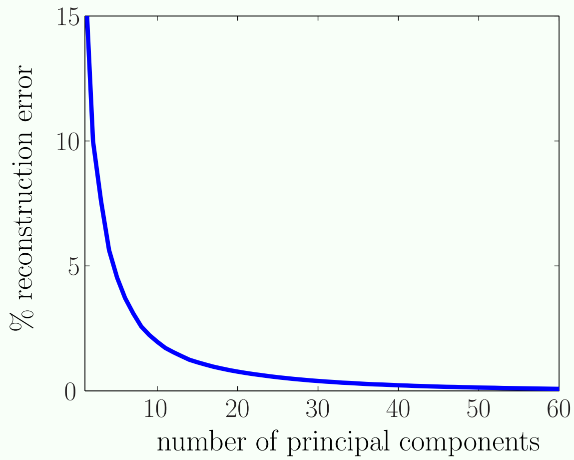
$$\begin{array}{c}
\begin{array}{ccc} \text{credit?} & \text{limit} & \text{risk} \end{array} \\
\left[\begin{array}{ccc}
\checkmark & 2K & high \\
\times & 0 & - \\
\checkmark & 10K & low \\
\vdots & & \\
\checkmark & 15K & high
\end{array}\right]
\end{array}$$

$$Y \in \mathbb{R}^{n \times \omega}$$

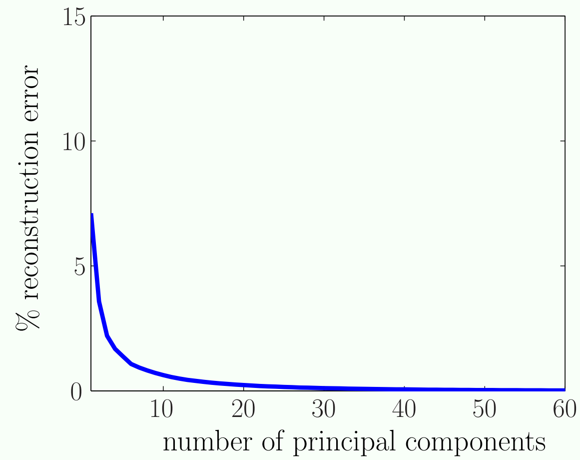


More Beautiful Data

$X \in \mathbb{R}^{231\times174}$



$Y \in \mathbb{R}^{231\times166}$



Throwing Out Unnecessary Features is Good

Sparsity: represent your solution **using only a few features**.

'Sparse' solutions generalize to out-of-sample better – less *overfitting*.

Sparse solutions are easier to interpret – few important features.

Computations are more efficient.

Problem: How to find the few relevant features *quickly*.

PCA, K-means, Linear Regression

$$k = 20$$
$$r = 2k$$

PCA

Exact

Approx, fast (relative error)

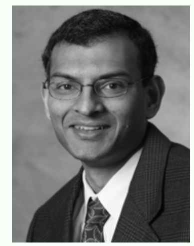
Sparse, approx, fast (relative error)

K-Means

Exact

Sparse, approx, fast (relative error)

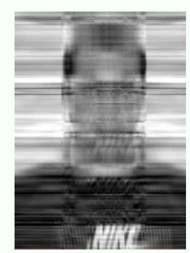
Regression



Exact

$[\ ] =$

top-k PCA regression

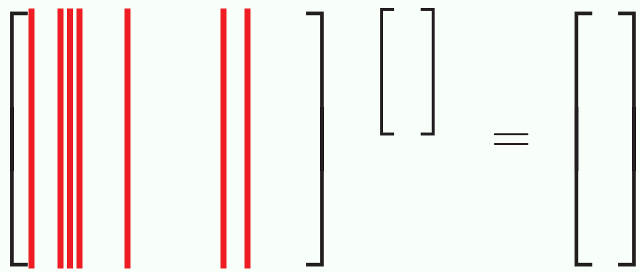


Fast-sparse regression (additive error)

Sparsity

Represent your solution using **only a few ...**

Example: linear regression



$$\mathrm{X}\mathbf{w} = \mathbf{y}$$

$\mathbf{y}$ is an optimal linear combination of **only a few** columns in X.

(sparse regression; regularization ($\|\mathbf{w}\|_0 \leq k$); feature subset selection; ...)

Singular Value Decomposition (SVD)

$$\mathrm{X} = \begin{bmatrix}\mathrm{U}_k & \mathrm{U}_{d-k}\end{bmatrix}\begin{bmatrix}\Sigma_k & 0\\ 0 & \Sigma_{d-k}\end{bmatrix}\begin{bmatrix}\color{red}{\mathrm{V}_k^{\mathrm{T}}}\\ \mathrm{V}_{d-k}^{\mathrm{T}}\end{bmatrix} \qquad O(nd^2)$$

$$\underset{(n\times d)}{\mathrm{U}} \qquad \underset{(d\times d)}{\Sigma} \qquad \underset{(d\times d)}{\mathrm{V}^{\mathrm{T}}}$$

$$\begin{aligned}\mathrm{X}_k &= \mathrm{U}_k\Sigma_k\mathrm{V}_k^{\mathrm{T}}\\ &= \mathrm{X}\mathrm{V}_k\mathrm{V}_k^{\mathrm{T}}\end{aligned}$$

X_k is the best rank-k approximation to X.

Reconstruction of X using **only a few** *deg. of freedom*.

X $\quad$ X_{20} $\quad$ X_{40} $\quad$ X_{60}

V_k is an orthonormal basis for the best k-dimensional subspace of the row space of X.

Fast Approximate SVD

$\mathrm{Z} = \mathrm{XR}$

$\mathrm{Q} = \textsc{qr.factorize}(\mathrm{Z})$

$\hat{\mathrm{V}}_k \leftarrow \textsc{svd}_k(\mathrm{Q}^{\mathrm{T}}\mathrm{X})$

$\mathrm{R} \sim \mathcal{N}(d \times r),\ \mathrm{Z} \in \mathbb{R}^{n \times r}$

Theorem. Let $r = \left\lceil k(1 + \frac{1}{\epsilon}) \right\rceil$ and $\mathrm{E} = \mathrm{X} - \mathrm{X}\hat{\mathrm{V}}_k\hat{\mathrm{V}}_k^{\mathrm{T}}$. Then,

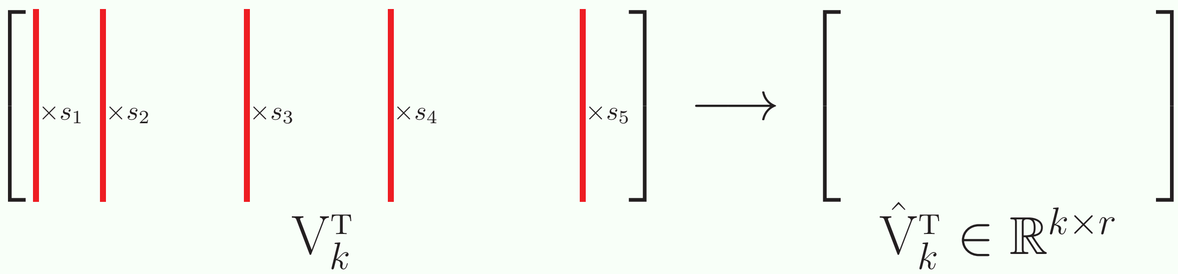
$$\mathbb{E}\left[\|\,\mathrm{E}\,\|\right] \leq (1 + \epsilon)\|\,\mathrm{X} - \mathrm{X}_k\,\|$$

running time is $O(ndk) = o(\textsc{svd})$

[BDM, FOCS 2011]

V_k and Sparsity

Important "dimensions" of $\mathrm{V}_k^{\mathrm{T}}$ are important for X



The sampled r columns are "good" if

$$\mathrm{I} = \mathrm{V}_k^{\mathrm{T}}\mathrm{V}_k \approx \hat{\mathrm{V}}_k^{\mathrm{T}}\hat{\mathrm{V}}_k.$$

Sampling schemes: Largest norm (Jollife, 1972);
Randomized norm sampling (Rudelson, 1999; RudelsonVershynin, 2007);
Greedy (Batson et al, 2009; **BDM, 2011**).

Sparse PCA – Algorithm

1: Choose a few columns C of X; $\mathrm{C} \in \mathbb{R}^{n\times r}$.
2: Find the best rank-k approximation of X in the span of C, $\mathrm{X}_{\mathrm{C},k}$.
3: Compute the SVD$_k$ of

$$\mathrm{X}_{\mathrm{C},k} = \mathrm{U}_{\mathrm{C},k}\Sigma_{\mathrm{C},k}\mathrm{V}_{\mathrm{C},k}^{\mathrm{T}}.$$

4:

$$\mathrm{Z} = \mathrm{X}\mathrm{V}_{\mathrm{C},k}.$$

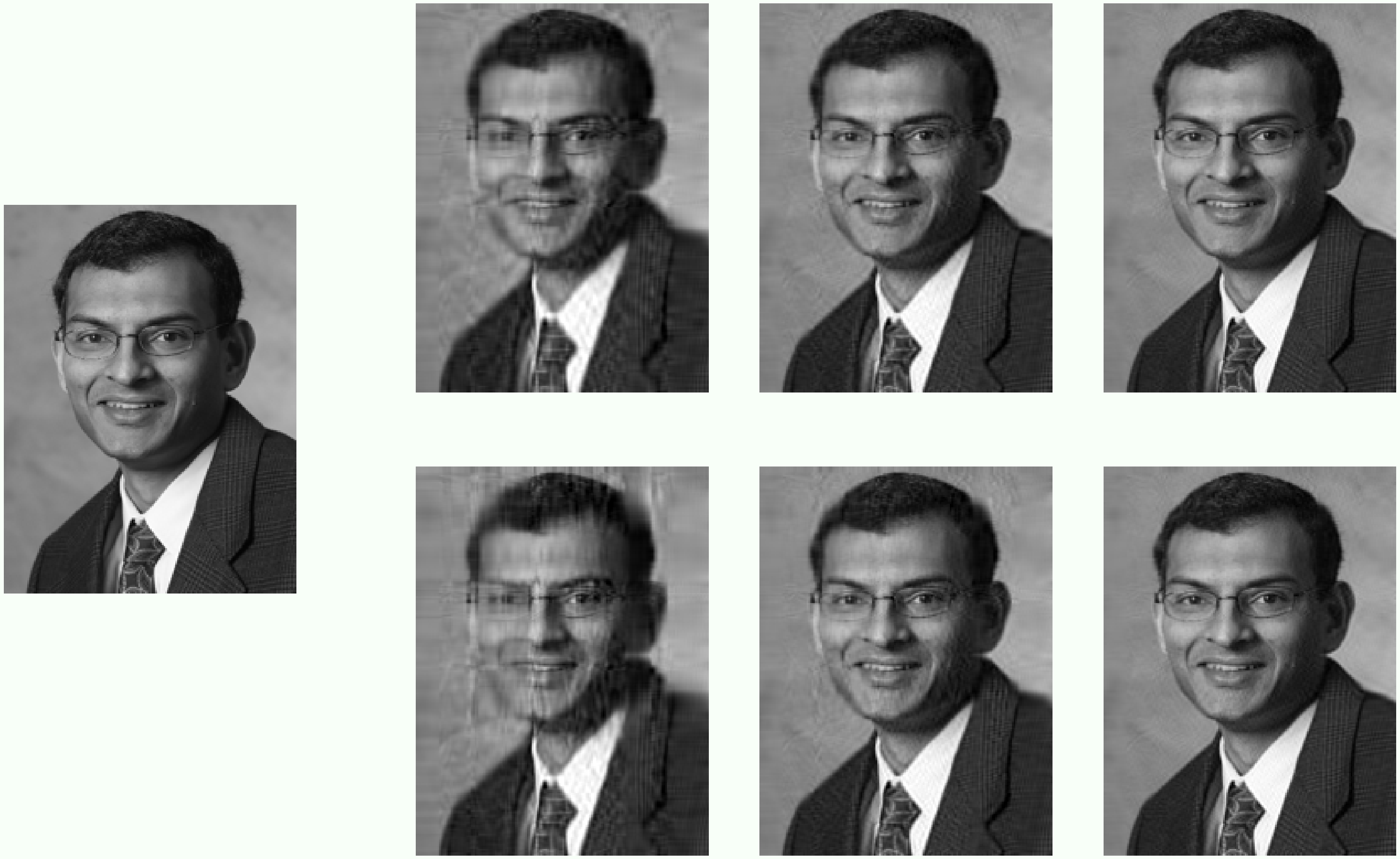
Each feature in Z is a mixture of **only the few** original r feature dimensions in C.

$$\|\,\mathrm{X} - \mathrm{X}\mathrm{V}_{\mathrm{C},k}\mathrm{V}_{\mathrm{C},k}^{\mathrm{T}}\,\| \leq \|\,\mathrm{X} - \mathrm{X}_{\mathrm{C},k}\mathrm{V}_{\mathrm{C},k}\mathrm{V}_{\mathrm{C},k}^{\mathrm{T}}\,\| = \|\,\mathrm{X} - \mathrm{X}_{\mathrm{C},k}\,\| \leq \left(1 + O(\tfrac{2k}{r})\right)\|\,\mathrm{X} - \mathrm{X}_k\,\|.$$

[BDM, FOCS 2011]

Sparse PCA

$k = 20$ $k = 40$ $k = 60$



Dense PCA

Sparse PCA, $r = 2k$

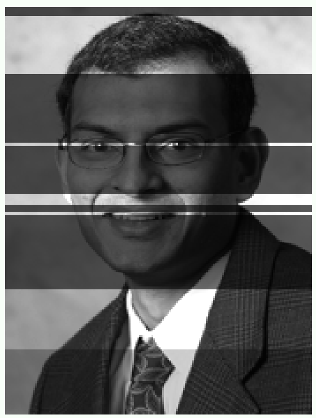
Theorem. One can construct, in $o(\textsc{svd})$, k features that are r-sparse, $r = O(k)$, that are as good as exact dense top-k PCA-features.

Clustering: K-Means

Full, slow

Fast, sparse

3 clusters



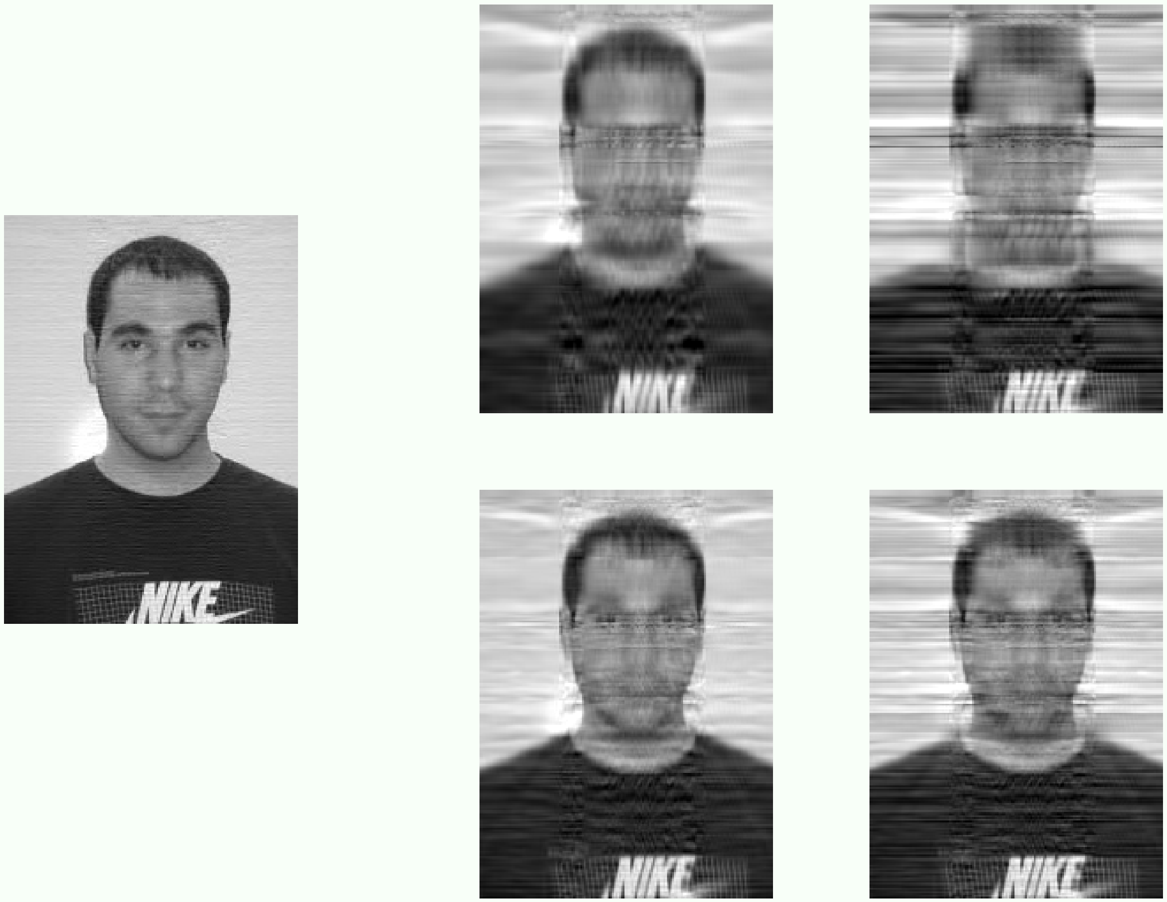
4 Clusters

Theorem. There is a subset of features of size $O(\#clusters)$ which produces nearly the optimal partition (within a constant factor). One can quickly produce features with a log-approximation factor.

[BDM,2013]

Fast Regression using Few Important Features

PCA, slow, dense — Sparse, fast



$k = 20$

$k = 40$

Theorem. Can find $O(k)$ pure features which performs as well top-k PCA-regression (additive error controlled by $\| \mathrm{X} - \mathrm{X}_k \|_F / \sigma_k$).

[BDM,2013]

The Proofs

All the algorithms use the sparsifier of V_k^T in [BDM,FOCS2011].

1. Choose columns of V_k^T to preserve its singular values.
2. Ensure that the selected columns preserve the structural properties of the objective with respect to the columns of X that are sampled.
3. Use dual set sparsification algorithms to accomplish (2).

THANKS!

- **Data compression (PCA):**
  quick and reveals few important features

- **Unsupervised clustering:**
  quick and reveals few important features

- **Supervised Regression:**
  quick and reveals few important features

Few features: easy to interpret; better generalizers; faster computations.