Versatility of Singular Value Decomposition (SVD)

January 7, 2015

Assumption : Data = Real Data + Noise

- Each Data Point is a column of the $n \times d$ Data Matrix A.

Assumption : Data = Real Data + Noise

- Each Data Point is a column of the $n \times d$ Data Matrix A.
- $A = \underbrace{B}_{\text{Real Data}} + \underbrace{C}_{\text{Noise}}$.

Assumption : Data = Real Data + Noise

- Each Data Point is a column of the $n \times d$ Data Matrix A.
- $A = \underbrace{B}_{\text{Real Data}} + \underbrace{C}_{\text{Noise}}$.
- rank $(B) \leq k$. $||C||(= \text{Max}_{|u|=1}||Cu|) \quad \leq \quad \Delta$.

Assumption : Data = Real Data + Noise

- Each Data Point is a column of the $n \times d$ Data Matrix A.
- $A = \underbrace{B}_{\text{Real Data}} + \underbrace{C}_{\text{Noise}}$.
- rank $(B) \leq k$. $||C||(= \text{Max}_{|u|=1} ||Cu|) \quad \leq \quad \Delta$.
  - $k << n, d$. Δ small.

Assumption : Data = Real Data + Noise

- Each Data Point is a column of the $n \times d$ Data Matrix A.
- $A = \underbrace{B}_{\text{Real Data}} + \underbrace{C}_{\text{Noise}}$.
- rank $(B) \leq k$. $||C||(= \text{Max}_{|u|=1}||Cu|) \quad \leq \quad \Delta$.
  - $k << n, d$. Δ small.
  - Caution: $||C||_F (= \sqrt{\sum C_{ij}^2})$ need not be smaller than for example $||B||_F$. In words, overall noise can be larger than overall real data.

Assumption : Data = Real Data + Noise

- Each Data Point is a column of the $n \times d$ Data Matrix A.
- $A = \underbrace{B}_{\text{Real Data}} + \underbrace{C}_{\text{Noise}}$.
- rank $(B) \leq k$. $||C||(= \text{Max}_{|u|=1}||Cu|) \quad \leq \quad \Delta$.
  - $k << n, d$. Δ small.
  - Caution: $||C||_F(=\sqrt{\sum C_{ij}^2})$ need not be smaller than for example $||B||_F$. In words, overall noise can be larger than overall real data.
- Given any A, Singular Value Decomposition (SVD) finds B of rank k (or less) for which $||A - B||$ is minimum. Space spanned by columns of B is the best-fit subspace for A in the sense of least sum over all data points of squared distances to subspace.

Assumption : Data = Real Data + Noise

- Each Data Point is a column of the $n \times d$ Data Matrix A.
- $A = \underbrace{B}_{\text{Real Data}} + \underbrace{C}_{\text{Noise}}$.
- rank $(B) \leq k$. $||C||(= \text{Max}_{|u|=1}||Cu|) \quad \leq \quad \Delta$.
  - $k << n, d$. Δ small.
  - Caution: $||C||_F(= \sqrt{\sum C_{ij}^2})$ need not be smaller than for example $||B||_F$. In words, overall noise can be larger than overall real data.
- Given any A, Singular Value Decomposition (SVD) finds B of rank k (or less) for which $||A - B||$ is minimum. Space spanned by columns of B is the best-fit subspace for A in the sense of least sum over all data points of squared distances to subspace.
- A very powerful tool. Decades of theory, algorithms. Here: Example applications.

Example I- Mixture of Spherical Gaussians

- $F(x) = w_1 N(\mu_1, \sigma_1^2) + w_2 N(\mu_2, \sigma_2^2) + \cdots + w_k N(\mu_k, \sigma_k^2)$, in d dimensions.

Example I- Mixture of Spherical Gaussians

- $F(x) = w_1 N(\mu_1, \sigma_1^2) + w_2 N(\mu_2, \sigma_2^2) + \cdots + w_k N(\mu_k, \sigma_k^2)$, in d dimensions.

Example I- Mixture of Spherical Gaussians

- $F(x) = w_1 N(\mu_1, \sigma_1^2) + w_2 N(\mu_2, \sigma_2^2) + \cdots + w_k N(\mu_k, \sigma_k^2)$, in d dimensions.
- **Learning Problem**: Given i.i.d. samples from $F(\cdot)$, find the components (μ_i, σ_i, w_i). Really a Clustering Problem.

Example I- Mixture of Spherical Gaussians

- $F(x) = w_1 N(\mu_1, \sigma_1^2) + w_2 N(\mu_2, \sigma_2^2) + \cdots + w_k N(\mu_k, \sigma_k^2)$, in d dimensions.
- **Learning Problem**: Given i.i.d. samples from $F(\cdot)$, find the components (μ_i, σ_i, w_i). Really a Clustering Problem.
- In 1-dimension, we can solve the learning problem if Means of the component densities are $\Omega(1)$ standard deviations apart.

Example I- Mixture of Spherical Gaussians

- $F(x) = w_1 N(\mu_1, \sigma_1^2) + w_2 N(\mu_2, \sigma_2^2) + \cdots + w_k N(\mu_k, \sigma_k^2)$, in d dimensions.
- **Learning Problem**: Given i.i.d. samples from $F(\cdot)$, find the components (μ_i, σ_i, w_i). Really a Clustering Problem.
- In 1-dimension, we can solve the learning problem if Means of the component densities are $\Omega(1)$ standard deviations apart.
- But in d dimensions: Approximate k means fails. Pair of Sample from different clusters may be closer than a pair from the same !

SVD to the Rescue

- For a mixture of k spherical Gaussians (with different variances), the best-fit k dimensional subspace (found by SVD) passes through all the k centers. Vempala, Wang.

SVD to the Rescue

- For a mixture of k spherical Gaussians (with different variances), the best-fit k dimensional subspace (found by SVD) passes through all the k centers. Vempala, Wang.
- Beautiful proof: For one spherical Gaussian with non-zero mean, the best fit 1-dim subspace passes through the mean. And any k-dim subspace containing the mean is a best-fit $k-$ dimensional space.

SVD to the Rescue

- For a mixture of k spherical Gaussians (with different variances), the best-fit k dimensional subspace (found by SVD) passes through all the k centers. Vempala, Wang.
- Beautiful proof: For one spherical Gaussian with non-zero mean, the best fit 1-dim subspace passes through the mean. And any k-dim subspace containing the mean is a best-fit $k-$ dimensional space.
- So, now if a $k-$ dimensional space contains all the k means, it is individually the best for each component Gaussian !!

SVD to the Rescue

- For a mixture of k spherical Gaussians (with different variances), the best-fit k dimensional subspace (found by SVD) passes through all the k centers. Vempala, Wang.
- Beautiful proof: For one spherical Gaussian with non-zero mean, the best fit 1-dim subspace passes through the mean. And any k-dim subspace containing the mean is a best-fit $k-$ dimensional space.
- So, now if a $k-$ dimensional space contains all the k means, it is individually the best for each component Gaussian !!
- Simple Observation to finish : Given the $k-$ space containing the means, we need only solve a $k-$ dim problem. Can be done in time exponential only in k

Planted Clique Problem

- Given $G = G(n, 1/2) + S \times S$, ($S$ unknown, $|S| = s$), find S in poly time. Best known: $s \geq \Omega(\sqrt{n})$.

$$A = \left[\begin{array}{ccc|ccccc}
1 & 1 & 1 & \pm 1 & \pm 1 & \pm 1 & \pm 1 & \pm 1 \\
1 & 1 & 1 & \pm 1 & \pm 1 & \pm 1 & \pm 1 & \pm 1 \\
1 & 1 & 1 & \pm 1 & \pm 1 & \pm 1 & \pm 1 & \pm 1 \\
\pm 1 & \pm 1 & \pm 1 & \pm 1 & \pm 1 & \pm 1 & \pm 1 & \pm 1 \\
\pm 1 & \pm 1 & \pm 1 & \pm 1 & \pm 1 & \pm 1 & \pm 1 & \pm 1 \\
\pm 1 & \pm 1 & \pm 1 & \pm 1 & \pm 1 & \pm 1 & \pm 1 & \pm 1 \\
\pm 1 & \pm 1 & \pm 1 & \pm 1 & \pm 1 & \pm 1 & \pm 1 & \pm 1 \\
\pm 1 & \pm 1 & \pm 1 & \pm 1 & \pm 1 & \pm 1 & \pm 1 & \pm 1
\end{array}\right]$$

Planted Clique Problem

- Given $G = G(n, 1/2) + S \times S$, ($S$ unknown, $|S| = s$), find S in poly time. Best known: $s \geq \Omega(\sqrt{n})$.

$$A = \left[\begin{array}{ccc|ccccc} 1 & 1 & 1 & \pm1 & \pm1 & \pm1 & \pm1 & \pm1 \\ 1 & 1 & 1 & \pm1 & \pm1 & \pm1 & \pm1 & \pm1 \\ 1 & 1 & 1 & \pm1 & \pm1 & \pm1 & \pm1 & \pm1 \\ \pm1 & \pm1 & \pm1 & \pm1 & \pm1 & \pm1 & \pm1 & \pm1 \\ \pm1 & \pm1 & \pm1 & \pm1 & \pm1 & \pm1 & \pm1 & \pm1 \\ \pm1 & \pm1 & \pm1 & \pm1 & \pm1 & \pm1 & \pm1 & \pm1 \\ \pm1 & \pm1 & \pm1 & \pm1 & \pm1 & \pm1 & \pm1 & \pm1 \\ \pm1 & \pm1 & \pm1 & \pm1 & \pm1 & \pm1 & \pm1 & \pm1 \end{array}\right]$$

- $||$ Planted Clique $|| = s$. **Random Matrix Theory**: Random ±1 matrix has norm at most $2\sqrt{n}$. So, SVD finds S when $s \geq \sqrt{n}$. Alon, Boppanna-1985.

Planted Clique Problem

- Given $G = G(n, 1/2) + S \times S$, ($S$ unknown, $|S| = s$), find S in poly time. Best known: $s \geq \Omega(\sqrt{n})$.

$$A = \left[\begin{array}{ccc|ccccc} 1 & 1 & 1 & \pm1 & \pm1 & \pm1 & \pm1 & \pm1 \\ 1 & 1 & 1 & \pm1 & \pm1 & \pm1 & \pm1 & \pm1 \\ 1 & 1 & 1 & \pm1 & \pm1 & \pm1 & \pm1 & \pm1 \\ \pm1 & \pm1 & \pm1 & \pm1 & \pm1 & \pm1 & \pm1 & \pm1 \\ \pm1 & \pm1 & \pm1 & \pm1 & \pm1 & \pm1 & \pm1 & \pm1 \\ \pm1 & \pm1 & \pm1 & \pm1 & \pm1 & \pm1 & \pm1 & \pm1 \\ \pm1 & \pm1 & \pm1 & \pm1 & \pm1 & \pm1 & \pm1 & \pm1 \\ \pm1 & \pm1 & \pm1 & \pm1 & \pm1 & \pm1 & \pm1 & \pm1 \end{array}\right]$$

- $||$ Planted Clique $|| = s$. **Random Matrix Theory**: Random ±1 matrix has norm at most $2\sqrt{n}$. So, SVD finds S when $s \geq \sqrt{n}$. Alon, Boppanna-1985.
- Feldman, Grigorescu, Reyzin, Vempala, Xiao (2014): Cannot be beaten by Statistical Learning Algorithms.

Planted Gaussians: Signal and Noise

- A $n \times n$ matrix and $S \subseteq [n], |S| = k$.

$$A = \left[\begin{array}{ccc|ccccc}
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \mu^+ & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\hline
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & N(0,\sigma^2) & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot
\end{array}\right]$$

Planted Gaussians: Signal and Noise

- A $n \times n$ matrix and $S \subseteq [n], |S| = k$.
- A_{ij} all independent r.v.'s

$$A = \left[\begin{array}{ccc|ccccc}
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \mu^{+} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\hline
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & N(0,\sigma^2) & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot
\end{array}\right]$$

Planted Gaussians: Signal and Noise

- A $n \times n$ matrix and $S \subseteq [n], |S| = k$.
- A_{ij} all independent r.v.'s
- For $i, j \in S$, $\Pr(A_{ij} \geq \mu) \geq 1/2$. (Eg. $N(\mu, \sigma^2)$). Signal $= \mu$.

$$A = \left[\begin{array}{ccc|ccccc}
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \mu^+ & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\hline
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & N(0, \sigma^2) & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot
\end{array}\right]$$

Planted Gaussians: Signal and Noise

- A $n \times n$ matrix and $S \subseteq [n], |S| = k$.
- A_{ij} all independent r.v.'s
- For $i, j \in S$, $\Pr(A_{ij} \geq \mu) \geq 1/2$. (Eg. $N(\mu, \sigma^2)$). Signal $= \mu$.
- For other i, j, A_{ij} is $N(0, \sigma^2)$. Noise $= \sigma$.

$$A = \left[\begin{array}{ccc|ccccc}
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \mu^+ & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\hline
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & N(0,\sigma^2) & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot
\end{array}\right]$$

Planted Gaussians: Signal and Noise

- A $n \times n$ matrix and $S \subseteq [n], |S| = k$.
- A_{ij} all independent r.v.'s
- For $i, j \in S$, $\Pr(A_{ij} \geq \mu) \geq 1/2$. (Eg. $N(\mu, \sigma^2)$). Signal $= \mu$.
- For other i, j, A_{ij} is $N(0, \sigma^2)$. Noise $= \sigma$.
- Given A, μ, σ, find S. [Recall Planted Clique.]

$$A = \left[\begin{array}{ccc|ccccc}
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \mu^+ & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\hline
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & N(0,\sigma^2) & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot
\end{array}\right]$$

Exponential Advantage in SNR by Thresholding

- Brave new step: Threshold entries of A at $\mu \rightarrow$ 0-1 matrix B.

Exponential Advantage in SNR by Thresholding

- Brave new step: Threshold entries of A at $\mu \rightarrow$ 0-1 matrix B.
- $E(B)$:

$$
\left[\begin{array}{ccc|ccccc}
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & (1/2)^{+} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\hline
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \exp(-\mu^2/2\sigma^2) & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot
\end{array}\right]
$$

Exponential Advantage in SNR by Thresholding

- Brave new step: Threshold entries of A at $\mu \rightarrow$ 0-1 matrix B.
- $E(B)$:

$$
\left[\begin{array}{ccc|ccccc}
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & (1/2)^{+} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\hline
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \exp(-\mu^2/2\sigma^2) & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot
\end{array}\right]
$$

Exponential Advantage in SNR by Thresholding

- Brave new step: Threshold entries of A at $\mu \to$ 0-1 matrix B.
- $E(B):$ $\left[\begin{array}{ccc|ccccc} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & (1/2)^{+} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \hline \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \exp(-\mu^2/2\sigma^2) & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{array}\right]$
- Subtract $\exp(-\mu^2/2\sigma^2)$

Exponential Advantage in SNR by Thresholding

- Brave new step: Threshold entries of A at $\mu \rightarrow$ 0-1 matrix B.
- $E(B):$ $\left[\begin{array}{ccc|ccccc} . & . & . & . & . & . & . & . \\ . & (1/2)^{+} & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ \hline . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & \exp(-\mu^2/2\sigma^2) & . & . \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \end{array}\right]$
- Subtract $\exp(-\mu^2/2\sigma^2)$
- $\ldots \rightarrow$ $\left[\begin{array}{c|c} \|\cdot\| \geq k/4 & \\ \hline & \|\cdot\| \leq \quad \sqrt{n}\exp(-c\mu^2/\sigma^2) \\ & \text{Rand. Matrix} \end{array}\right]$

Exponential Advantage in SNR by Thresholding

- Brave new step: Threshold entries of A at $\mu \rightarrow$ 0-1 matrix B.
- $E(B):$
$$\left[\begin{array}{ccc|ccccc}
. & . & . & . & . & . & . & . \\
. & (1/2)^{+} & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . \\
\hline
. & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . \\
. & . & . & . & . & \exp(-\mu^2/2\sigma^2) & . & . \\
. & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & .
\end{array}\right]$$
- Subtract $\exp(-\mu^2/2\sigma^2)$
- $\ldots \rightarrow \left[\begin{array}{c|c}
\|\cdot\| \geq k/4 & \\
\hline
 & \|\cdot\| \leq \sqrt{n}\exp(-c\mu^2/\sigma^2) \\
 & \text{Rand. Matrix}
\end{array}\right]$
- So, SVD finds S provided $\exp(c(\mu/\sigma)^2) > \frac{\sqrt{n}}{k}$.

Exponential Advantage in SNR by Thresholding

- Brave new step: Threshold entries of A at $\mu \rightarrow$ 0-1 matrix B.
- $E(B):\left[\begin{array}{ccc|ccccc} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & (1/2)^{+} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \hline \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \exp(-\mu^2/2\sigma^2) & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{array}\right]$
- Subtract $\exp(-\mu^2/2\sigma^2)$
- $\ldots \rightarrow \left[\begin{array}{c|c} ||\cdot|| \geq k/4 & \\ \hline & ||\cdot|| \leq \sqrt{n}\exp(-c\mu^2/\sigma^2) \\ & \text{Rand. Matrix} \end{array}\right]$
- So, SVD finds S provided $\exp(c(\mu/\sigma)^2) > \frac{\sqrt{n}}{k}$.
- Cf: Ordinary SVD succeeds if $\frac{\mu}{\sigma} > \frac{\sqrt{n}}{k}$.

Thresholding: Second Plus

- Data points $\{A_1, A_2, \ldots, A_j, \ldots\}$ in $\mathbf{R}^d$, d features.

Thresholding: Second Plus

- Data points $\{A_1, A_2, \ldots, A_j, \ldots\}$ in $\mathbf{R}^d$, d features.
- Data points are in 2 “SOFT” clusters: Data point j belongs w_j to cluster 1 and $1 - w_j$ to cluster 2. (More Generally, k clusters)

Thresholding: Second Plus

- Data points $\{A_1, A_2, \ldots, A_j, \ldots\}$ in $\mathbf{R}^d$, d features.
- Data points are in 2 “SOFT” clusters: Data point j belongs w_j to cluster 1 and $1 - w_j$ to cluster 2. (More Generally, k clusters)
- Each cluster has some some dominant features and each data point has a dominant cluster.

Thresholding: Second Plus

- Data points $\{A_1, A_2, \ldots, A_j, \ldots\}$ in $\mathbf{R}^d$, d features.
- Data points are in 2 “SOFT” clusters: Data point j belongs w_j to cluster 1 and $1 - w_j$ to cluster 2. (More Generally, k clusters)
- Each cluster has some some dominant features and each data point has a dominant cluster.
- $A_{ij} \geq \mu$ if feature i is a dominant feature of the dominant topic of data point j.

Thresholding: Second Plus

- Data points $\{A_1, A_2, \ldots, A_j, \ldots\}$ in $\mathbf{R}^d$, d features.
- Data points are in 2 “SOFT” clusters: Data point j belongs w_j to cluster 1 and $1 - w_j$ to cluster 2. (More Generally, k clusters)
- Each cluster has some some dominant features and each data point has a dominant cluster.
- $A_{ij} \geq \mu$ if feature i is a dominant feature of the dominant topic of data point j.
- $A_{ij} \leq \sigma$ otherwise.

Thresholding: Second Plus

- Data points $\{A_1, A_2, \ldots, A_j, \ldots\}$ in $\mathbf{R}^d$, d features.
- Data points are in 2 "SOFT" clusters: Data point j belongs w_j to cluster 1 and $1 - w_j$ to cluster 2. (More Generally, k clusters)
- Each cluster has some some dominant features and each data point has a dominant cluster.
- $A_{ij} \geq \mu$ if feature i is a dominant feature of the dominant topic of data point j.
- $A_{ij} \leq \sigma$ otherwise.
- If variance above μ is larger than gap between μ and σ, a 2-clustering criterion (like 2-means) may split the high weight cluster instead of separating it from the others.

Thresholding: Second Plus

- Data points $\{A_1, A_2, \ldots, A_j, \ldots\}$ in $\mathbf{R}^d$, d features.
- Data points are in 2 “SOFT” clusters: Data point j belongs w_j to cluster 1 and $1 - w_j$ to cluster 2. (More Generally, k clusters)
- Each cluster has some some dominant features and each data point has a dominant cluster.
- $A_{ij} \geq \mu$ if feature i is a dominant feature of the dominant topic of data point j.
- $A_{ij} \leq \sigma$ otherwise.
- If variance above μ is larger than gap between μ and σ, a 2-clustering criterion (like 2-means) may split the high weight cluster instead of separating it from the others.
- Two Differences from Mixtures: Soft, High Variance in dominant features.

Topic Modeling: The Problem

Joint Work with T. Bansal and C. Bhattacharyya

- d features - words in the dictionary. A document is a $d-$ (column) vector.

Topic Modeling: The Problem

Joint Work with T. Bansal and C. Bhattacharyya

- d features - words in the dictionary. A document is a $d-$ (column) vector.
- k topics. Topic l is a $d-$ vector. (Probabilities of words in topic).

Topic Modeling: The Problem

Joint Work with T. Bansal and C. Bhattacharyya

- d features - words in the dictionary. A document is a $d-$ (column) vector.
- k topics. Topic l is a $d-$ vector. (Probabilities of words in topic).
- To generate doc j, generate a random convex combination of topic vectors.

Topic Modeling: The Problem

Joint Work with T. Bansal and C. Bhattacharyya

- d features - words in the dictionary. A document is a $d-$ (column) vector.
- k topics. Topic l is a $d-$ vector. (Probabilities of words in topic).
- To generate doc j, generate a random convex combination of topic vectors.
- Generate words of doc. j in i.i.d. trials , each from the multinomial with prob.s = Convex Combination. ***DRAW PICTURE ON BOARD WITH SPORTS, POLITICS, WEATHER***

Topic Modeling: The Problem

Joint Work with T. Bansal and C. Bhattacharyya

- d features - words in the dictionary. A document is a $d-$ (column) vector.
- k topics. Topic l is a $d-$ vector. (Probabilities of words in topic).
- To generate doc j, generate a random convex combination of topic vectors.
- Generate words of doc. j in i.i.d. trials , each from the multinomial with prob.s = Convex Combination. ***DRAW PICTURE ON BOARD WITH SPORTS, POLITICS, WEATHER***
- **The Topic Modeling Problem** Given only A, find an approximation to all topic vectors so that **the l_1 error** in each topic vector is at most ε. l_1 error crucial. (l_2 misses small words.)

Topic Modeling: The Problem

Joint Work with T. Bansal and C. Bhattacharyya

- d features - words in the dictionary. A document is a $d-$ (column) vector.
- k topics. Topic l is a $d-$ vector. (Probabilities of words in topic).
- To generate doc j, generate a random convex combination of topic vectors.
- Generate words of doc. j in i.i.d. trials , each from the multinomial with prob.s = Convex Combination. ***DRAW PICTURE ON BOARD WITH SPORTS, POLITICS, WEATHER***
- **The Topic Modeling Problem** Given only A, find an approximation to all topic vectors so that **the l_1 error** in each topic vector is at most ε. l_1 error crucial. (l_2 misses small words.)
- Generally NP-hard.

Topic Modeling is Soft Clustering

- Topic Vectors $\equiv$ Cluster Centers

Topic Modeling is Soft Clustering

- Topic Vectors $\equiv$ Cluster Centers
- Each data point (doc) belongs to a weighted combination of clusters. Generated from a distribution (happens to be multinomial) with expectation = weighted combination.

Topic Modeling is Soft Clustering

- Topic Vectors ≡ Cluster Centers
- Each data point (doc) belongs to a weighted combination of clusters. Generated from a distribution (happens to be multinomial) with expectation = weighted combination.
- Even if we manage to solve the clustering problem somehow, it is not true that cluster centers are averages of documents. Big Distinction from Learning Mixtures which is hard clusetring.

Geometry

Topic Modeling = Soft Clustering

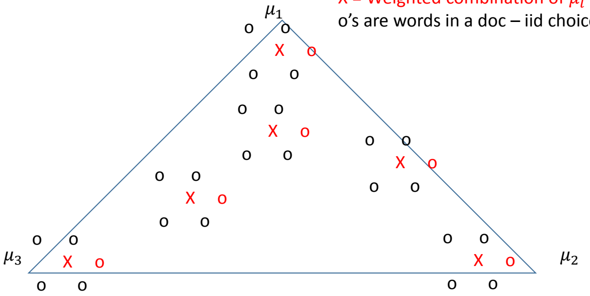
Given doc's (means of o's), find μ_l.

Helps to find nearly pure docs (X near corner)

$\mu_l = l$ th topic vector

X = Weighted combination of μ_l

o's are words in a doc – iid choices with mean X



Prior Results and Assumptions

- Under Pure Topics and Primary Words ($1-\varepsilon$ of words are primary) Assumptions, SVD solves it. Papadimitriou, Raghavan, Tamaki, Vempala.

- **Our Goals**

Prior Results and Assumptions

- Under Pure Topics and Primary Words ($1-\varepsilon$ of words are primary) Assumptions, SVD solves it. Papadimitriou, Raghavan, Tamaki, Vempala.
- Belief: SVD cannot do the non-pure topic case.
- **Our Goals**

Prior Results and Assumptions

- Under Pure Topics and Primary Words ($1-\varepsilon$ of words are primary) Assumptions, SVD solves it. Papadimitriou, Raghavan, Tamaki, Vempala.
- Belief: SVD cannot do the non-pure topic case.
- LDA : Most used model. Blei, Ng, Jordan. Multiple topics per doc.
- **Our Goals**

Prior Results and Assumptions

- Under Pure Topics and Primary Words ($1-\varepsilon$ of words are primary) Assumptions, SVD solves it. Papadimitriou, Raghavan, Tamaki, Vempala.
- Belief: SVD cannot do the non-pure topic case.
- LDA : Most used model. Blei, Ng, Jordan. Multiple topics per doc.
- Anandkumar, Foster, Hsu, Kakade, Liu do Topic Modeling under LDA, to l_2 error using clever tensor methods. Parameters.
- **Our Goals**

Prior Results and Assumptions

- Under Pure Topics and Primary Words ($1-\varepsilon$ of words are primary) Assumptions, SVD solves it. Papadimitriou, Raghavan, Tamaki, Vempala.
- Belief: SVD cannot do the non-pure topic case.
- LDA : Most used model. Blei, Ng, Jordan. Multiple topics per doc.
- Anandkumar, Foster, Hsu, Kakade, Liu do Topic Modeling under LDA, to l_2 error using clever tensor methods. Parameters.
- Arora, Ge, Moitra Assume Anchor Word + Other parameters : Each topic has one word (a) occurring only in that topic (b) with high frequency. First Provable Algorithm.
- **Our Goals**

Prior Results and Assumptions

- Under Pure Topics and Primary Words ($1-\varepsilon$ of words are primary) Assumptions, SVD solves it. Papadimitriou, Raghavan, Tamaki, Vempala.
- Belief: SVD cannot do the non-pure topic case.
- LDA : Most used model. Blei, Ng, Jordan. Multiple topics per doc.
- Anandkumar, Foster, Hsu, Kakade, Liu do Topic Modeling under LDA, to l_2 error using clever tensor methods. Parameters.
- Arora, Ge, Moitra Assume Anchor Word + Other parameters : Each topic has one word (a) occurring only in that topic (b) with high frequency. First Provable Algorithm.
- **Our Goals**
  - Intuitive, empirically verified assumptions.

Prior Results and Assumptions

- Under Pure Topics and Primary Words ($1 - \varepsilon$ of words are primary) Assumptions, SVD solves it. Papadimitriou, Raghavan, Tamaki, Vempala.
- Belief: SVD cannot do the non-pure topic case.
- LDA : Most used model. Blei, Ng, Jordan. Multiple topics per doc.
- Anandkumar, Foster, Hsu, Kakade, Liu do Topic Modeling under LDA, to l_2 error using clever tensor methods. Parameters.
- Arora, Ge, Moitra Assume Anchor Word + Other parameters : Each topic has one word (a) occurring only in that topic (b) with high frequency. First Provable Algorithm.
- **Our Goals**
  - Intuitive, empirically verified assumptions.
  - Natural, provable Algorithm.

Our Assumptions

- Intuitive to Topic Modeling, not numerical parameters like condition number.

Our Assumptions

- Intuitive to Topic Modeling, not numerical parameters like condition number.
- **Catchwords**: Each topic has a set of words: (a) each occurs more frequently in the topic than others and (b) together, they have high frequency.

Our Assumptions

- Intuitive to Topic Modeling, not numerical parameters like condition number.
- **Catchwords**: Each topic has a set of words: (a) each occurs more frequently in the topic than others and (b) together, they have high frequency.
- **Dominant Topics** Each Document has a dominant topic which has weight (in that doc) of at least some α, whereas, non-dominant topics have weight at most some β.

Our Assumptions

- Intuitive to Topic Modeling, not numerical parameters like condition number.
- **Catchwords**: Each topic has a set of words: (a) each occurs more frequently in the topic than others and (b) together, they have high frequency.
- **Dominant Topics** Each Document has a dominant topic which has weight (in that doc) of at least some α, whereas, non-dominant topics have weight at most some β.
- **Nearly Pure Documents** Each topic has a (small) fraction of documents which are $1-\delta$ pure for that topic.

Our Assumptions

- Intuitive to Topic Modeling, not numerical parameters like condition number.
- **Catchwords**: Each topic has a set of words: (a) each occurs more frequently in the topic than others and (b) together, they have high frequency.
- **Dominant Topics** Each Document has a dominant topic which has weight (in that doc) of at least some α, whereas, non-dominant topics have weight at most some β.
- **Nearly Pure Documents** Each topic has a (small) fraction of documents which are $1-\delta$ pure for that topic.
- **No Local Min.**: For every word, the plot of number of documents versus number of occurrences of word (conditioned on dominant topic) has no local min. [Zipf's law Or Unimodal.]

The Algorothm - Threshold SVD (TSVD)

- $s =$ No. of docs. For this talk, probability that each topic is dominant is $1/k$.

The Algorothm - Threshold SVD (TSVD)

- $s =$ No. of docs. For this talk, probability that each topic is dominant is $1/k$.
- **Threshold** Compute the threshold for each word i: First "Gap": Maxζ : $A_{ij} \geq \zeta$ for $\geq (s/2k)\, j's$ and $A_{ij} = \zeta$ for $\leq \varepsilon s\, j's$.

The Algorothm - Threshold SVD (TSVD)

- $s =$ No. of docs. For this talk, probability that each topic is dominant is $1/k$.
- **Threshold** Compute the threshold for each word i: First “Gap”: Maxζ : $A_{ij} \geq \zeta$ for $\geq (s/2k)\, j's$ and $A_{ij} = \zeta$ for $\leq \varepsilon s\, j's$.
- **SVD** Use SVD on thresholded matrix to get starting centers for $k-$means algorithm.

The Algorothm - Threshold SVD (TSVD)

- $s =$ No. of docs. For this talk, probability that each topic is dominant is $1/k$.
- **Threshold** Compute the threshold for each word i: First "Gap": Maxζ : $A_{ij} \geq \zeta$ for $\geq (s/2k)\, j's$ and $A_{ij} = \zeta$ for $\leq \varepsilon s\, j's$.
- **SVD** Use SVD on thresholded matrix to get starting centers for $k-$means algorithm.
- **$k-$means** Run $k-$means. Will show: This identifies dominant topic.

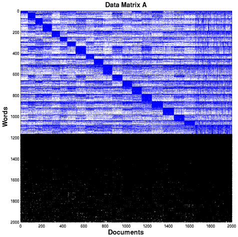
The Algorothm - Threshold SVD (TSVD)

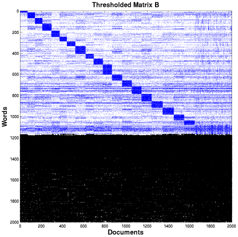
- $s =$ No. of docs. For this talk, probability that each topic is dominant is $1/k$.
- **Threshold** Compute the threshold for each word i: First "Gap": Maxζ : $A_{ij} \geq \zeta$ for $\geq (s/2k)\, j's$ and $A_{ij} = \zeta$ for $\leq \varepsilon s\, j's$.
- **SVD** Use SVD on thresholded matrix to get starting centers for $k-$means algorithm.
- **$k-$means** Run $k-$means. Will show: This identifies dominant topic.
- **Identify Catchwords** Find the set of high frequency words in each cluster. Will show: Set of Catchwords for topic.

The Algorothm - Threshold SVD (TSVD)

- $s =$ No. of docs. For this talk, probability that each topic is dominant is $1/k$.
- **Threshold** Compute the threshold for each word i: First "Gap": Maxζ : $A_{ij} \geq \zeta$ for $\geq (s/2k)\, j's$ and $A_{ij} = \zeta$ for $\leq \varepsilon s\, j's$.
- **SVD** Use SVD on thresholded matrix to get starting centers for $k-$means algorithm.
- ***k*−means** Run $k-$means. Will show: This identifies dominant topic.
- **Identify Catchwords** Find the set of high frequency words in each cluster. Will show: Set of Catchwords for topic.
- **Identify Pure Docs** Find the set of documents with highest total number of occurrences of set of catchwords. Show: Nearly Pure Docs. Their average $\approx$ topic vector.

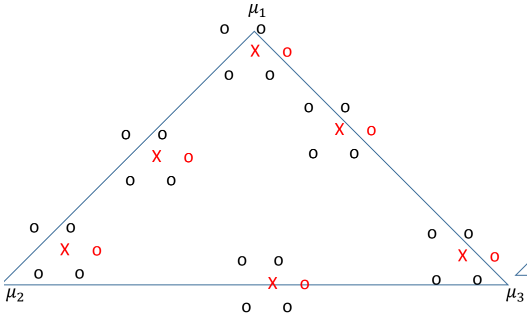
The advantage of Thresholding

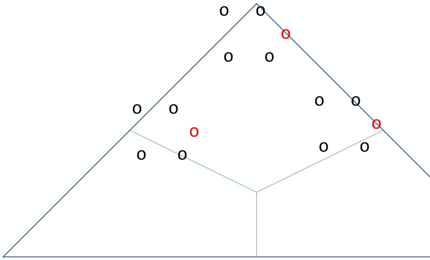




Diagonal blue blocks are Catchwords for each topic.
Black: Non-Catchwords.

Thresh+SVD+k-means → Dominant Topics





Properties of Thresholding

- Using no local min., show: No threshold splits any dominant topic in the “middle”. So, threshlded matrix is a “block” matrix for catchwords. But for non-catchwords, can be high on several topics.

Properties of Thresholding

- Using no local min., show: No threshold splits any dominant topic in the “middle”. So, threshlded matrix is a “block” matrix for catchwords. But for non-catchwords, can be high on several topics.
- PICTURE ON THE BOARD OF A BLOCK MATRIX.

Properties of Thresholding

- Using no local min., show: No threshold splits any dominant topic in the “middle”. So, threshlded matrix is a “block” matrix for catchwords. But for non-catchwords, can be high on several topics.
- PICTURE ON THE BOARD OF A BLOCK MATRIX.
- Done ? No. Need inter-cluster separation $\geq$ intra-cluster spread (variance inside cluster).

Properties of Thresholding

- Using no local min., show: No threshold splits any dominant topic in the "middle". So, threshlded matrix is a "block" matrix for catchwords. But for non-catchwords, can be high on several topics.
- PICTURE ON THE BOARD OF A BLOCK MATRIX.
- Done ? No. Need inter-cluster separation $\geq$ intra-cluster spread (variance inside cluster).
- Catchwords provide sufficient inter-cluster separation.

Properties of Thresholding

- Using no local min., show: No threshold splits any dominant topic in the “middle”. So, threshlded matrix is a “block” matrix for catchwords. But for non-catchwords, can be high on several topics.
- PICTURE ON THE BOARD OF A BLOCK MATRIX.
- Done ? No. Need inter-cluster separation $\geq$ intra-cluster spread (variance inside cluster).
- Catchwords provide sufficient inter-cluster separation.
- Inside-cluster variance bounded with machinery from Random Matrix Theory. Beware: Only columns are independent. Rows are not.

Properties of Thresholding

- Using no local min., show: No threshold splits any dominant topic in the "middle". So, threshlded matrix is a "block" matrix for catchwords. But for non-catchwords, can be high on several topics.
- PICTURE ON THE BOARD OF A BLOCK MATRIX.
- Done ? No. Need inter-cluster separation $\geq$ intra-cluster spread (variance inside cluster).
- Catchwords provide sufficient inter-cluster separation.
- Inside-cluster variance bounded with machinery from Random Matrix Theory. Beware: Only columns are independent. Rows are not.
- Appeal to a result on $k-$means (Kumar, K.: If inter-cluster separation $\geq$ inside-cluster directional stan. dev, then SVD followed by $k-$means clusters.

Getting Topic Vectors

PICTURE OF SIMPLEX with columns of M as extreme points and cluster of doc.s with each dominant topic.
Taking average of docs in T_l no good.

Emprircal Results: Datasets

- **NIPS**: 1,500 NIPS full papers
- **NYT**: Random subset of 30,000 documents from the New York Times dataset
- **Pubmed**: Random subset of 30,000 documents from the Pubmed abstracts dataset
- **20NG**: 13,389 documents from 20NewsGroup dataset

Empirical Results: Assumptions

Corpus	Documents	K	Fraction of Documents		
			$\alpha = 0.4$	$\alpha = 0.8$	$\alpha = 0.9$
NIPS	1,500	50	56.6%	10.7%	4.8%
NYT	30,000	50	63.7%	20.9%	12.7%
Pubmed	30,000	50	62.2%	20.3%	10.7%
20NG	13,389	20	74.1%	54.4%	44.3%

Table: Fraction of documents satisfying dominant topic assumption.

Corpus	K	Mean per topic frequency of CW	% Topics with CW
NIPS	50	0.05	95%
NYT	50	0.11	100%
Pubmed	50	0.05	90%
20NG	20	0.06	100%

Table: CatchWords (CW) assumption with $\rho = 1.1$, $\varepsilon = 0.25$

Empirical Results: Semi-synthetic Data

- Generate semi-synthetic corpora from LDA model trained by MCMC, to ensure that the synthetic corpora retain the characteristics of real data
- Gibbs sampling is run for 1000 iterations on all the four datasets and the final word-topic distribution is used to generate varying number (s) of synthetic documents with document-topic distribution drawn from a symmetric Dirichlet with hyper-parameter 0.01
- Note that the synthetic data is *not* guaranteed to satisfy dominant topic assumption for every document, on average about 80% documents satisfy the assumption

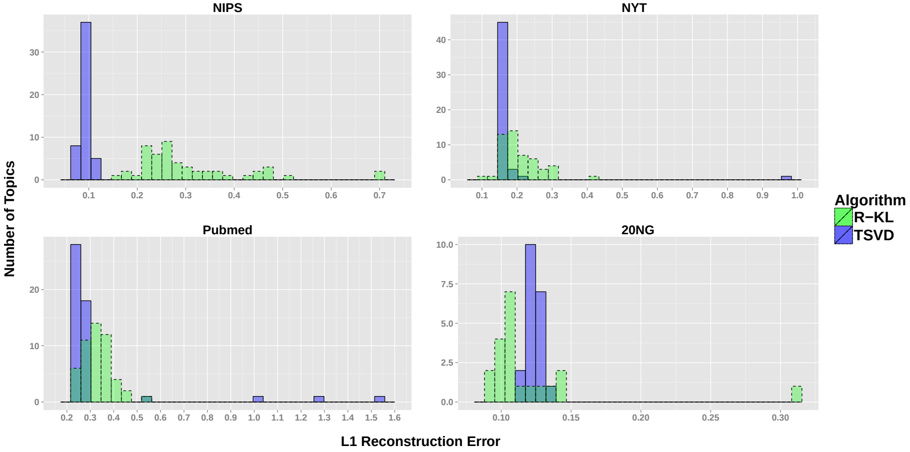
Empirical Results: L1 Recnstruction Error

And percent improvement over Recover-KL. Total average improvement over R-KL is **20%**

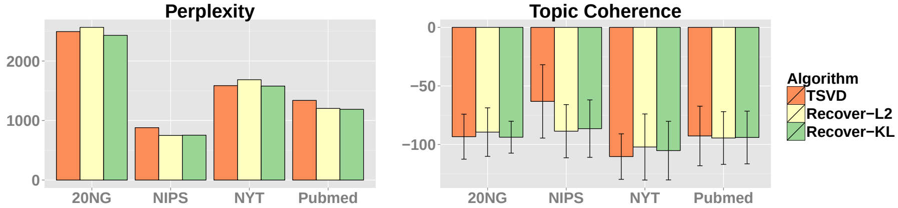
Corpus	Documents	Tensor	R-L2	R-KL	TSVD	% Improvement
NIPS	40,000	0.298	0.342	0.308	**0.094**	**68.5%**
	60,000	0.296	0.346	0.311	**0.089**	**69.9%**
	80,000	0.285	0.335	0.303	**0.087**	**69.4%**
	100,000	0.280	0.344	0.306	**0.086**	**69.3%**
	150,000	0.320	0.336	0.302	**0.084**	**72.2%**
	200,000	0.322	0.335	0.301	**0.113**	**62.5%**
Pubmed	40,000	0.379	0.388	0.332	**0.326**	**1.8%**
	60,000	0.317	0.372	0.328	**0.287**	**9.5%**
	80,000	0.321	0.358	0.320	**0.276**	**13.8%**
	100,000	0.304	0.350	0.315	**0.276**	**9.2%**
	150,000	0.355	0.344	0.313	**0.239**	**23.6%**
	200,000	0.322	0.334	0.309	**0.225**	**27.3%**
20NG	40,000	0.174	0.126	**0.120**	0.124	-3.3%
	60,000	0.207	0.114	0.110	**0.106**	**3.6%**
	80,000	0.203	0.110	0.108	**0.095**	**12.0%**
	100,000	0.151	0.103	0.102	**0.087**	**14.7%**
	200,000	0.162	0.096	0.097	**0.072**	**25.8%**
NYT	40,000	0.316	0.214	0.208	**0.174**	**16.3%**
	60,000	0.330	0.205	0.200	**0.156**	**22.0%**
	80,000	0.330	0.198	0.196	**0.168**	**14.3%**
	100,000	0.353	0.198	0.196	**0.163**	**16.8%**

Empirical Results: L1 Recnstruction Eror

Histogram of L1 error across topics for 40k synthetic documents. On majority of the topic ($> 90\%$) the recovery error for TSVD is significantly smaller than Recover-KL.



Empirical Results: Perplexity & Topic Coherence



Top 5 words of some topics on the real NYT dataset. Catchwords, anchor highlighted. “zzz”- identifier placed by NYT dataset.

TSVD	Recover-KL	Gibbs
cup minutes **add tablespoon oil**	cup minutes **tablespoon** add oil	cup minutes add tablespoon oil
team season coach zzz_ram game	game team season play **zzz_ram**	team season game coach zzz_nfl
patient doctor drug cancer study	patient drug doctor percent found	patient doctor drug medical cancer
zzz_john_mccain zzz_mccain zzz_bush zzz_george_bush campaign	**zzz_john_mccain** zzz_george_bush campaign republican voter	zzz_john_mccain zzz_george_bush campaign zzz_bush zzz_mccain
house room building wall floor	room show look home house	room look water house hand
film movie actor character zzz_oscar	film show movie music book	film movie character play director
zzz_god christian religious **zzz_jesus** church	**pope** church book jewish religious	religious church jewish jew zzz_god