Topic Modeling, Threshold SVD

October 27, 2014

Simple Setting: Signal and Noise

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

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

Simple Setting: Signal and Noise

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

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

Simple Setting: Signal and Noise

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

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

Simple Setting: Signal and Noise

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

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

Simple Setting: Signal and Noise

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

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

Condition on Signal-to-Noise ratio

- SNR $= \frac{\mu}{\sigma}$.

Condition on Signal-to-Noise ratio

- SNR $= \frac{\mu}{\sigma}$.
- Standard Planted Clique (PC) problem is like having SNR = $O(1)$.

Condition on Signal-to-Noise ratio

- SNR $= \frac{\mu}{\sigma}$.
- Standard Planted Clique (PC) problem is like having SNR = $O(1)$.
- Known Results for PC : If SNR$\geq \frac{\sqrt{n}}{k}$, then we can find S.

Condition on Signal-to-Noise ratio

- SNR $= \frac{\mu}{\sigma}$.
- Standard Planted Clique (PC) problem is like having SNR = $O(1)$.
- Known Results for PC : If SNR$\geq \frac{\sqrt{n}}{k}$, then we can find S.
- We have not been able to beat this lower bound requirement on SNR for PC. In fact, Feldman, Grigorescu, Reyzin, Vempala, Xiao have shown: Cannot be beaten by Statistical Learning Algorithms.

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.
- $E(B)$:

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

Exponential Advantage in SNR by Thresholding

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

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

Exponential Advantage in SNR by Thresholding

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

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

- Subtract $\exp(-\mu^2/2\sigma^2)$

Exponential Advantage in SNR by Thresholding

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

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

- Subtract $\exp(-\mu^2/2\sigma^2)$
- $\ldots \to$

$$\left[\begin{array}{c|c} ||\cdot|| \geq k/4 & \\ \hline & ||\cdot|| \leq \sqrt{n}\exp(-c\mu^2/\sigma^2) \\ & \text{Rand. Matrix} \end{array}\right]$$

Exponential Advantage in SNR by Thresholding

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

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

- Subtract $\exp(-\mu^2/2\sigma^2)$
- $\ldots \to$

$$\left[\begin{array}{c|c} ||\cdot|| \geq k/4 & \\ \hline & ||\cdot|| \leq \sqrt{n}\exp(-c\mu^2/\sigma^2) \\ & \text{Rand. Matrix} \end{array}\right]$$

- 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 $\mu \to$ 0-1 matrix B.
- $E(B)$:

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

- Subtract $\exp(-\mu^2/2\sigma^2)$
- $\ldots \to$

$$\left[\begin{array}{c|c} ||\cdot|| \geq k/4 & \\ \hline & ||\cdot|| \leq \sqrt{n}\exp(-c\mu^2/\sigma^2) \\ & \text{Rand. Matrix} \end{array}\right]$$

- 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.

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.
- 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 $\equiv$ 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.
- 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.

Topic Modeling = Soft Clustering

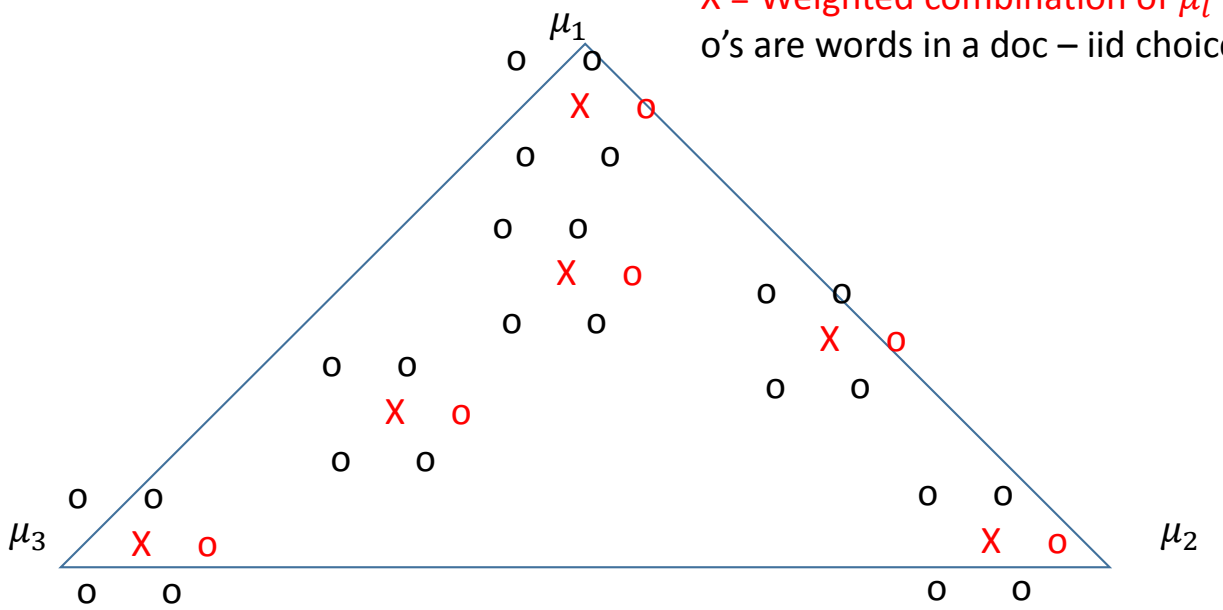
Given doc's (means of o's), find μ_l.

Helps to find nearly pure docs (X near corner)

μ_l = center of cluster l

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 provably solves Topic Modeling. Papadimitriou, Raghavan, Tamaki, Vempala.

Prior Results and Assumptions

- Under Pure Topics and Primary Words ($1 - \varepsilon$ of words are primary) Assumptions, SVD provably solves Topic Modeling. Papadimitriou, Raghavan, Tamaki, Vempala.
- Long Standing Question/Belief: SVD cannot do the non-pure topic case.

Prior Results and Assumptions

- Under Pure Topics and Primary Words ($1 - \varepsilon$ of words are primary) Assumptions, SVD provably solves Topic Modeling. Papadimitriou, Raghavan, Tamaki, Vempala.
- Long Standing Question/Belief: SVD cannot do the non-pure topic case.
- LDA : Most popular non-pure model. Blei, Ng, Jordan. Multiple topics per doc are allowed. Topic weights (in a doc.) are (essentially) uncorrelated. Correlations: Blei, Lafferty

Prior Results and Assumptions

- Under Pure Topics and Primary Words ($1 - \varepsilon$ of words are primary) Assumptions, SVD provably solves Topic Modeling. Papadimitriou, Raghavan, Tamaki, Vempala.
- Long Standing Question/Belief: SVD cannot do the non-pure topic case.
- LDA : Most popular non-pure model. Blei, Ng, Jordan. Multiple topics per doc are allowed. Topic weights (in a doc.) are (essentially) uncorrelated. Correlations: Blei, Lafferty
- Anandkumar, Foster, Hsu, Kakade, Liu do Topic Modeling under LDA, to l_2 error using tensor methods. Parameters.

Prior Results and Assumptions

- Under Pure Topics and Primary Words ($1 - \varepsilon$ of words are primary) Assumptions, SVD provably solves Topic Modeling. Papadimitriou, Raghavan, Tamaki, Vempala.
- Long Standing Question/Belief: SVD cannot do the non-pure topic case.
- LDA : Most popular non-pure model. Blei, Ng, Jordan. Multiple topics per doc are allowed. Topic weights (in a doc.) are (essentially) uncorrelated. Correlations: Blei, Lafferty
- Anandkumar, Foster, Hsu, Kakade, Liu do Topic Modeling under LDA, to l_2 error using 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. Provable algorithm: Do Topic Modeling with l_1 error per word. First provable algorithm.

Prior Results and Assumptions

- Under Pure Topics and Primary Words ($1 - \varepsilon$ of words are primary) Assumptions, SVD provably solves Topic Modeling. Papadimitriou, Raghavan, Tamaki, Vempala.
- Long Standing Question/Belief: SVD cannot do the non-pure topic case.
- LDA : Most popular non-pure model. Blei, Ng, Jordan. Multiple topics per doc are allowed. Topic weights (in a doc.) are (essentially) uncorrelated. Correlations: Blei, Lafferty
- Anandkumar, Foster, Hsu, Kakade, Liu do Topic Modeling under LDA, to l_2 error using 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. Provable algorithm: Do Topic Modeling with l_1 error per word. First provable algorithm.
  - Our Aim: Intuitive, empirically verified assumptions , Natural 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.

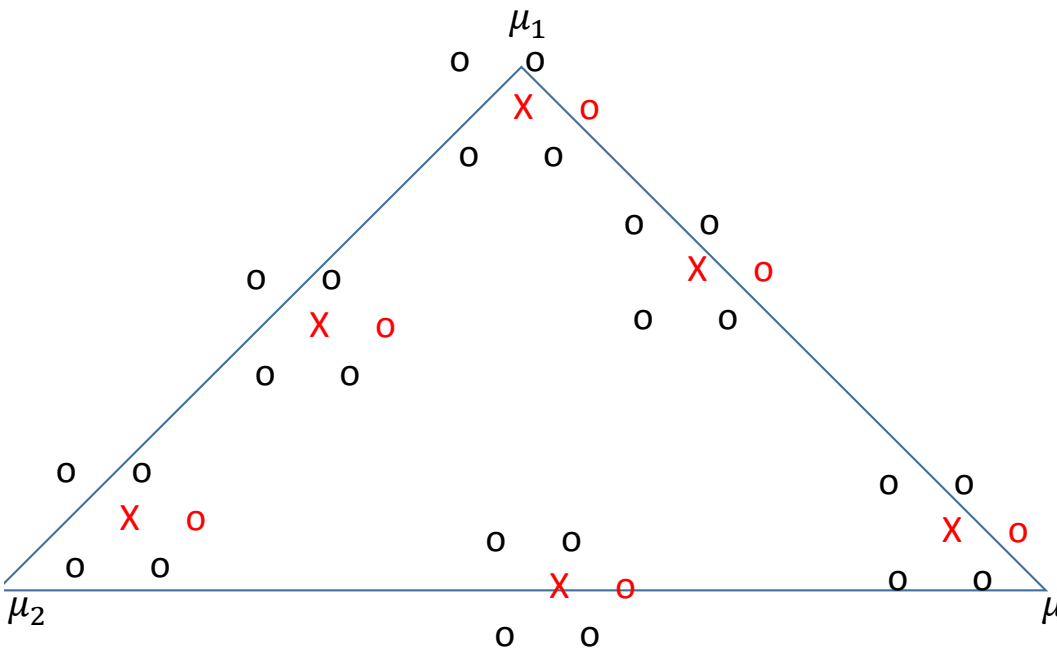
The Algorothm - Threshold SVD (TSVD)

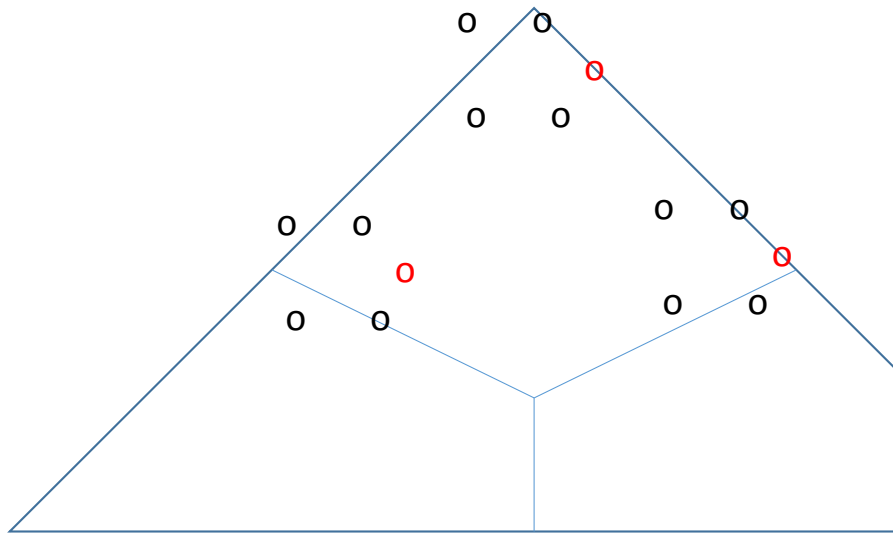
- $s =$ No. of docs. For this talk, probability that each topic is dominant is $1/k$.
- **Threshold** Compute the threshold for each word i: First "Gap": Maxζ : $A_{ij} \geq \zeta$ for $\geq (s/2k)\ j's$ and $A_{ij} = \zeta$ for $\leq \varepsilon s\ j's$.
- **SVD** Use SVD on thresholded matrix to get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

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.

Experimental Results