Improved Distributed Principal Component Analysis

Maria-Florina Balcan†, Vandana Kanchanapally‡, Yingyu Liang§, and David Woodruff‡

†ninamf@cs.cmu.edu, ‡vvandana@gatech.edu, §yingyul@cs.princeton.edu, ‡dpwoodru@us.ibm.com

Motivation

▶ Principal Component Analysis
Given: data matrix $P$ with zero mean
Traditional Goal: to summarize/visualize the data in low dim, find $t$-dim subspace $V$ to min $\|P - PVV^T\|^2$.

More Important: PCA pre-processes for downstream applications e.g. k-means, regression, kernel methods, ...

Our Goal: find $t$-dim subspace $V$ s.t. for downstream applications, any $\alpha$-approx for $PVV^T$ is $(1 + \epsilon)$-approx for $P$.

▶ Distributed Data Model
PCA is most useful for large-scale data collected distributedly
• $s$ servers linked to a coordinator
• Local datasets $P_1, P_2, \ldots, P_s$; points in dim $d$
• Global dataset $P = \bigcup P_i$

▶ Challenges in Distributed PCA
1. Arbitrary local datasets: no distributional or low rank assumption
2. Relative error guarantees for downstream applications
3. Low communication/computation cost

Distributed PCA Algorithm

▶ Algorithm disPCA
1. Each server: $V_i = U_i \Sigma_i V_i^T$; send $\Sigma_i$ and $V_i$ to coordinator
2. Coordinator: estimate $S = \sum V_i \Sigma_i V_i^T$ (1st $t$ columns);
   factorize $S = VAV^T$ and get $V$.

▶ Equivalent Form of Algorithm disPCA
Let $Y_i = (V_i^T)^T$; then $S = \sum Y_i^T Y_i$.
Factoring $S$ is equivalent to SVD on $Y = [Y_1; Y_2; \ldots; Y_s]$.

More formally:

$$P = \begin{bmatrix} P_1 & \ldots & P_s \end{bmatrix} \text{LocalPCA} \begin{bmatrix} \Sigma_1^T V_1^T \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \ldots \\ Y_s \end{bmatrix} = Y \text{ GlobalPCA } V^T$$

Guarantee of disPCA

▶ Main Theorem (for k-means): Suppose $t = O(k/\epsilon^2)$.
Then any $\alpha$-approx for $P = PVV^T$ is $(1 + \epsilon)$-approx for $P$.

• When $t = O(k^2/\epsilon^2)$, the same holds for $r$-Subspace $k$-Clustering

$$\min_{\ell} \sum_{p \in \ell} \min_{L \in \mathcal{L}} d_\ell(p, L)$$
where $\mathcal{L} = \{L_j\}_{j=1}^k$ is a set of $k$ linear/affine subspaces of dim $r$.

Our Results

▶ disPCA algorithm
• Relative $(1 + \epsilon)$ error for downstream applications
• Holds for arbitrary data and a wide family of problems: includes low rank approx, $k$-means, LDA, NMF, ...
• Low communication cost: $O(skd/\epsilon^2)$ words

▶ Fast disPCA algorithm
• Easy plug-in of randomized techniques
• Morely comprise the quality and communication
• Speedup of orders of magnitude

Review: Classic Non-Distributed PCA Algorithm

1. SVD on Data
2. Factorize Covariance $S = P^T P$

For a matrix $A$, let $A^{(t)}$ denote the first $t$ columns of $A$

Fast Distributed PCA

▶ Techniques: subspace embedding; randomized SVD
▶ Subspace Embedding
Property of embedding matrix $H$:

• With prob $\geq 99/100$, for any $y$

$$\|HPy\| \leq (1 + \epsilon)\|Py\|$$

such that disPCA($HP$) almost equivalent to disPCA($P$)

• HP much smaller than $P$, computed in time $O(nz(P))$;
  particularly suitable for sparse data

• HP can be computed locally

Challenges: boost the constant success prob to high prob

• Solved by our new cross validation style technique

• Dimension and sparsity of HP independent of the success prob

Experimental Results

▶ Performance: cost increase < 5%; ×10 to ×100 speedup
▶ k-Means Clustering: k-means cost/time vs dimension

Analysis

▶ I. Close Projection Property of SVD:
Let $A = US\Sigma W^T$ be its SVD, and $\bar{A} = AWV^T (W^T)^T$.
For any $k$-dim subspace $X$, $AXX^T$ and $\bar{A}XX^T$ are close.

▶ II. Close Projection Property of disPCA:
For any $k$-dim subspace $X$, $\bar{P}XX^T$ and $\bar{P}XX^T$ are close.

$$(P - \bar{P})XX^T \leq \sum_{i=1}^s \|P_i - \bar{P}_i\|XX^T + 2\|\bar{P} - P\|XX^T$$
where each term can be bounded by $P$ or its generalization

▶ III. Cost Decomposition:
Decompose the costs of $\bar{P}$ and $P$ and show that they are close

$$\text{cost}(P) = d^2(P, PXX^T) + d^2(PXX^T, C)$$
$$\text{cost}(\bar{P}) = d^2(\bar{P}, \bar{P}XX^T) + d^2(\bar{P}XX^T, C)$$

difference in each part bounded by

Principal Component Regression: PCR cost/time vs dimension

MNIST

CTShoses