

Single-Pass PCA of Large High-Dimensional Data

Wenjian Yu, Yu Gu, Jian Li

TNList,
Dept. Computer Science & Tech.,
Tsinghua University, Beijing, China
yu-wj@tsinghua.edu.cn

Shenghua Liu *

Inst. Computing Technology,
Chinese Academy of Sciences,
Beijing, China
liushenghua@ict.ac.cn

Yaohang Li

Dept. Computer Science,
Old Dominion University,
Norfolk, VA 23529, USA
yaohang@cs.odu.edu

Abstract

Principal component analysis (PCA) is a fundamental dimension reduction tool in statistics and machine learning. For large and high-dimensional data, computing the PCA (i.e., the top singular vectors of the data matrix) becomes a challenging task. In this work, a single-pass randomized algorithm is proposed to compute PCA with only one pass over the data. It is suitable for processing extremely large and high-dimensional data stored in slow memory (hard disk) or the data generated in a streaming fashion. Experiments with synthetic and real data validate the algorithm's accuracy, which has orders of magnitude smaller error than an existing single-pass algorithm. For a set of high-dimensional data stored as a 150 GB file, the algorithm is able to compute the first 50 principal components in just 24 minutes on a typical 24-core computer, with less than 1 GB memory cost.

1 Introduction

Many existing machine learning models, no matter supervised or unsupervised, rely on dimension reduction of input data. Even the applications of Deep Neural Networks on natural language processing tasks, prefer to use an embedding of each word in a sentence [Mikolov *et al.*, 2013; Bahdanau *et al.*, 2014], which essentially reduces the data dimensionality. Principal component analysis (PCA) is an efficient and well-structured dimension reduction technique [Halko *et al.*, 2011a; Friedman *et al.*, 2001]. However, how to calculate PCA of large-size (say terabyte) and high-dimensional dense data in a limited-memory computation node is still an open problem. Plus some data are generated in stream, e.g., from internet traffic, and signals from internet of things, we need a kind of pass-efficient algorithm, or even single-pass algorithm, to realize the dimension reduction of input data.

A single-pass algorithm has the benefit of requiring only one pass over the data, and is particularly useful and efficient for streaming data or data stored in slow memory [Halko

et al., 2011b; Ye *et al.*, 2016]. It also allows the computation with small or fixed RAM size [De Stefani *et al.*, 2016]. Although there are provable single-pass truncated SVD algorithms for symmetric positive semi-definite (SPSD) matrices [Wang and Zhang, 2013; Gittens and Mahoney, 2016; Wang *et al.*, 2016], the study for general matrices is not sufficient. [Halko *et al.*, 2011b] proposed a single-pass algorithm for approximately calculating SVD for general matrices, but with a significant cost of accuracy. [Ordonez *et al.*, 2014] developed a PCA algorithm for large-size data, but only applicable to low-dimensional data (less than one thousand in dimension). Frequent-directions (FD) algorithm [Liberty, 2013] was a single-pass and deterministic matrix sketching scheme [Woodruff, 2014], which is useful for matrix multiplication problem [Ye *et al.*, 2016], but the PCA computation.

Randomized matrix computation has gained significant increases in popularity as the data sets are becoming larger and larger. It has been revealed that randomization can be a powerful computational resource for developing algorithms with improved runtime and stability properties [Halko *et al.*, 2011b; Drineas and Mahoney, 2016; Wang, 2015]. Compared with classic algorithms, the randomized algorithm involves the same or fewer floating-point operations (*flops*), and is more efficient for truly large high-dimensional data sets, by exploiting modern computing architectures. An idea of randomization is using random projection to identify the subspace capturing the dominant actions of a matrix \mathbf{A} . With the subspace's orthonormal basis matrix \mathbf{Q} , a so-called QB approximation is obtained: $\mathbf{A} \approx \mathbf{QB}$. This produces a smaller sketch matrix \mathbf{B} , and facilitates the computation of near-optimal decompositions of \mathbf{A} . A simple implementation of this idea and related techniques and theories have been presented in [Halko *et al.*, 2011b]. With the merit of requiring a small constant number of passes over data, this algorithm has been applied to compute PCA of data sets that are too large to be stored in RAM [Halko *et al.*, 2011a]. It has also been employed to speed up the distributed PCA, without compromising the quality of the solution [Liang *et al.*, 2014]. However, this basic `randQB` algorithm still involves several passes instead of a single pass over data, which makes it not efficient enough or infeasible for some situations.

Progress has also been achieved based on the randomized algorithm for QB approximation. In [Mary *et al.*, 2015], the basic `randQB` algorithm [Halko *et al.*, 2011b] was slightly

*Also with CAS Key Laboratory of Network Data Science and Technology.

modified for computing the QR factorization. The main efforts were paid to investigate the algorithm’s performance scaling on shared-memory multi-core CPUs with multiple GPUs, and the comparison with the traditional QR factorization with column pivoting (QRCP). The results demonstrated that the randomized algorithm could be an excellent computational tool for many applications, with growing potential on the emerging parallel computers. In [Martinsson and Voronin, 2016], a randomized blocked algorithm was proposed for computing rank-revealing factorizations in an incremental manner. Although it enables adaptive rank determination, the algorithm needs to access the matrix for a number of times and is not efficient for large-size data.

Therefore, we reconstruct the randomized blocked algorithm [Martinsson and Voronin, 2016] and enforce numerical stability of the algorithm, which results in a single-pass PCA algorithm owning the following advantages.

- *Single-pass*: it involves only one pass over specified large high-dimensional data.
- *Efficiency*: it has $\mathcal{O}(mnk)$ or $\mathcal{O}(mn \log(k))$ time complexity and $\mathcal{O}(k(m+n))$ space complexity for computing k principal components of an $m \times n$ matrix data, and well adapts to parallel computing.
- *Accuracy*: it has theoretical error bounds, and empirically shows much less error than the single-pass algorithm [Halko *et al.*, 2011b], offering good PCA accuracy for matrices with different singular value distributions.

We have examined the effectiveness of the proposed single-pass algorithm for performing PCA on large-size (~ 150 GB) high dimensional data, which cannot be fit in RAM (32 GB). The experimental results show that our single-pass algorithm outperforms the standard SVD and existing competitors, by largely reduced time and memory usage, and accuracy guarantees. For reproducibility, we share the codes of the proposed algorithm and experimental data on <https://github.com/WenjianYu/rSVD-single-pass>.

2 Preliminaries

An orthonormal matrix denotes a matrix whose columns are a set of orthonormal vectors; \mathbf{I} denotes the identity matrix. The Matlab convention is obeyed to specify row/column indices.

2.1 Singular Value Decomposition and PCA

Let \mathbf{A} denote an $m \times n$ matrix. The SVD of \mathbf{A} is

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top, \quad (1)$$

where \mathbf{U} and \mathbf{V} are $m \times \min(m, n)$ and $n \times \min(m, n)$ orthonormal matrices respectively, and $\mathbf{\Sigma}$ is a diagonal matrix. The diagonal entries of $\mathbf{\Sigma}$ are the descending singular values of \mathbf{A} : $\sigma_{11} \geq \sigma_{22} \geq \dots \geq 0$. The columns of matrices \mathbf{U} and \mathbf{V} are the left and right singular vectors, respectively.

Taking the first k , $k < \min(m, n)$, columns of \mathbf{U} and \mathbf{V} respectively, and the first k singular values in $\mathbf{\Sigma}$, we have the truncated (partial) SVD of matrix \mathbf{A} :

$$\mathbf{A}_k = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^\top, \quad (2)$$

where \mathbf{U}_k and \mathbf{V}_k include the first k columns of \mathbf{U} and \mathbf{V} , respectively. $\mathbf{\Sigma}_k$ is the $k \times k$ up-left submatrix of $\mathbf{\Sigma}$. \mathbf{A}_k is

actually the optimal rank- k approximation of \mathbf{A} , in terms of l_2 -norm and Frobenius norm [Eckart and Young, 1936].

The approximation properties of the SVD explain the equivalence between SVD and PCA. Suppose each row of matrix \mathbf{A} is an observed data. The matrix is assumed to be centered, i.e., the mean of each column is equal to zero. Then, the leading right singular vectors $\{\mathbf{v}_i\}$ of \mathbf{A} are the principal components. Particularly, \mathbf{v}_1 is the first principal component.

2.2 The Basic Randomized Algorithm for PCA

The algorithm in [Halko *et al.*, 2011a] is based on the basic randQB algorithm for QB approximation, and described as Algorithm 1. $\mathbf{\Omega}$ is a Gaussian i.i.d. matrix. Replacing it with a *structured* random matrix is also feasible, and can reduce the computational cost for a dense \mathbf{A} from $\mathcal{O}(mnl)$ to $\mathcal{O}(mn \log(l))$ flops [Halko *et al.*, 2011b]. The over-sampling technique which uses $\mathbf{\Omega}$ with more than k columns is employed for better accuracy [Halko *et al.*, 2011b]. Usually, the over-sampling parameter s is a small integer, like 5 or 10. “orth(\mathbf{X})” denotes the orthonormalization of the columns of \mathbf{X} . In practice, it is achieved efficiently by a call to a packaged QR factorization (e.g., `qr(X, 0)` in Matlab), which implements the QR factorization without pivoting.

Algorithm 1 Basic randomized scheme for truncated SVD

Require: $\mathbf{A} \in \mathbb{R}^{m \times n}$, rank k , over-sampling parameter s .

- 1: $l = k + s$;
- 2: $\mathbf{\Omega} = \text{randn}(n, l)$;
- 3: $\mathbf{Q} = \text{orth}(\mathbf{A}\mathbf{\Omega})$;
- 4: $\mathbf{B} = \mathbf{Q}^\top \mathbf{A}$;
- 5: $[\tilde{\mathbf{U}}, \mathbf{S}, \mathbf{V}] = \text{svd}(\mathbf{B})$;
- 6: $\mathbf{U} = \mathbf{Q}\tilde{\mathbf{U}}$;
- 7: $\mathbf{U} = \mathbf{U}(:, 1 : k)$; $\mathbf{V} = \mathbf{V}(:, 1 : k)$; $\mathbf{S} = \mathbf{S}(1 : k, 1 : k)$;
- 8: **return** $\mathbf{U}, \mathbf{S}, \mathbf{V}$.

The first four steps in Algorithm 1 is the basic randQB scheme for building \mathbf{A} ’s QB approximation. This procedure could not produce the optimal low-rank approximation. However, in many applications the optimal approximation is not necessary, and even impossible to obtain due to the high computational complexity of performing SVD. The existing work has revealed that this randomized algorithm often produces a good enough solution. Compared with the classic rank-revealing QR factorization [Golub and Van Loan, 1996] for low-rank approximation, it has less computational cost and can obtain substantial speedup on a parallel computing platform [Martinsson and Voronin, 2016].

The error of the randomized QB approximation could be large for the matrix whose singular values decay slowly [Halko *et al.*, 2011b]. This can be eased by a technique called *power scheme* [Rokhlin *et al.*, 2009]. It is based on the fact that matrix $(\mathbf{A}\mathbf{A}^\top)^P \mathbf{A}$ has exactly the same singular vectors as \mathbf{A} , but its j -th singular value is σ_{jj}^{2P+1} . This largely reduces the relative weight of the tail singular values. Thus, performing the randomized QB procedure on $(\mathbf{A}\mathbf{A}^\top)^P \mathbf{A}$ can achieve more accurate approximation. More theoretical analysis can be found in Sec. 10.4 of [Halko *et al.*, 2011b]. On the other hand, the power scheme increases the number of passes over \mathbf{A} from 2 to $2P+2$ [Halko *et al.*, 2011b].

Notice that the output of the randomized approximation algorithms is a random variable. However, the variation of this random variable is small, which means the output is always very close to the variable's expectation [Halko *et al.*, 2011b].

2.3 An Existing Single-Pass Algorithm

Algorithm 1 involves two passes over matrix \mathbf{A} . In [Halko *et al.*, 2011b], a single-pass algorithm was proposed as a remedy (see Algorithm 2), where only Step 2 need access matrix \mathbf{A} .

Algorithm 2 An existing single-pass algorithm

Require: $\mathbf{A} \in \mathbb{R}^{m \times n}$, rank parameter k .

- 1: Generate random $n \times k$ matrix Ω and $m \times k$ matrix $\tilde{\Omega}$;
 - 2: Compute $\mathbf{Y} = \mathbf{A}\Omega$ and $\tilde{\mathbf{Y}} = \mathbf{A}^\top \tilde{\Omega}$ in a single pass over \mathbf{A} ;
 - 3: $\mathbf{Q} = \text{orth}(\mathbf{Y})$; $\tilde{\mathbf{Q}} = \text{orth}(\tilde{\mathbf{Y}})$;
 - 4: Solve linear equation $\tilde{\Omega}^\top \mathbf{Q}\mathbf{B} = \tilde{\mathbf{Y}}^\top \tilde{\mathbf{Q}}$ for \mathbf{B} ;
 - 5: $[\tilde{\mathbf{U}}, \mathbf{S}, \tilde{\mathbf{V}}] = \text{svd}(\mathbf{B})$;
 - 6: $\mathbf{U} = \mathbf{Q}\tilde{\mathbf{U}}$; $\mathbf{V} = \tilde{\mathbf{Q}}\tilde{\mathbf{V}}$;
 - 7: **return** $\mathbf{U}, \mathbf{S}, \mathbf{V}$.
-

Step 3 of Algorithm 2 results in matrices \mathbf{Q} and $\tilde{\mathbf{Q}}$ such that $\mathbf{A} \approx \mathbf{Q}\mathbf{Q}^\top \mathbf{A}\tilde{\mathbf{Q}}\tilde{\mathbf{Q}}^\top$. Then, the left problem is how to compute the small matrix $\mathbf{B} = \mathbf{Q}^\top \mathbf{A}\tilde{\mathbf{Q}}$. One can find out:

$$\tilde{\mathbf{Q}}^\top \tilde{\mathbf{Y}} = \tilde{\mathbf{Q}}^\top \mathbf{A}^\top \tilde{\Omega} \approx \tilde{\mathbf{Q}}^\top \mathbf{A}^\top \mathbf{Q}\mathbf{Q}^\top \tilde{\Omega} = \mathbf{B}^\top \mathbf{Q}^\top \tilde{\Omega}. \quad (3)$$

So, \mathbf{B} is approximately computed in Step 4. Due to these approximations, the accuracy of this algorithm or its variants is not good. We will reveal this through experiments.

2.4 The Randomized Blocked Algorithm

The randomized blocked algorithm in [Martinsson and Voronin, 2016] is inspired by a greedy Gram-Schmidt procedure for the orthonormalization step of basic randQB, which constitutes an iterative updating of the QB approximation error. Then, the algorithm is converted to a blocked version to attain high performance of linear algebraic computation (see Fig. 1). It is easy to prove that, if the algorithm is executed in exact arithmetic \mathbf{Q} is orthonormal, $\mathbf{B} = \mathbf{Q}^\top \mathbf{A}$, and after Step (6) \mathbf{A} becomes the approximation error: $\mathbf{A} - \mathbf{Q}\mathbf{B}$.

The algorithm is mathematically equivalent to the basic randQB procedure (the first 4 steps of Algorithm 1), except for the stopping criterion. The re-orthogonalization step (4) is for easing the accumulation of numerical round-off error.

function $[\mathbf{Q}, \mathbf{B}] = \text{randQB}_b(\mathbf{A}, \varepsilon, b)$

- (1) **for** $i = 1, 2, 3, \dots$
- (2) $\Omega_i = \text{randn}(n, b)$;
- (3) $\mathbf{Q}_i = \text{orth}(\mathbf{A}\Omega_i)$;
- (4) $\mathbf{Q}_i = \text{orth}(\mathbf{Q}_i - \sum_{j=1}^{i-1} \mathbf{Q}_j \mathbf{Q}_j^\top \mathbf{Q}_i)$;
- (5) $\mathbf{B}_i = \mathbf{Q}_i^\top \mathbf{A}$;
- (6) $\mathbf{A} = \mathbf{A} - \mathbf{Q}_i \mathbf{B}_i$;
- (7) **if** $\|\mathbf{A}\| < \varepsilon$ **then stop**
- (8) **end for**
- (9) $\mathbf{Q} = [\mathbf{Q}_1 \dots \mathbf{Q}_i]$; $\mathbf{B} = [\mathbf{B}_1^\top \dots \mathbf{B}_i^\top]^\top$.

Figure 1: The blocked randQB algorithm, from [Martinsson and Voronin, 2016].

3 Methodology

3.1 A Pass-Efficient Blocked Algorithm

The blocked randQB procedure in Fig. 1 facilitates adaptive rank determination, but increases the number of passes over the data matrix. For many scenarios of using PCA, the rank parameter k is a known value. Without the evaluation of approximation error, the block randQB procedure can be modified to be a pass-efficient procedure. It is presented as Algorithm 3, with the re-orthogonalization step ignored. For simplicity, we also assume that k is a multiple of b .

Algorithm 3 A pass-efficient blocked algorithm

Require: $\mathbf{A} \in \mathbb{R}^{m \times n}$, rank parameter k , block size b .

- 1: $\mathbf{Q} = []$; $\mathbf{B} = []$;
 - 2: $\Omega = \text{randn}(n, k)$;
 - 3: $\mathbf{G} = \mathbf{A}\Omega$;
 - 4: $\mathbf{H} = \mathbf{A}^\top \mathbf{G}$;
 - 5: **for** $i = 1, 2, \dots, k/b$ **do**
 - 6: $\Omega_i = \Omega(:, (i-1)b + 1 : ib)$;
 - 7: $\mathbf{Y}_i = \mathbf{G}(:, (i-1)b + 1 : ib) - \mathbf{Q}(\mathbf{B}\Omega_i)$;
 - 8: $[\mathbf{Q}_i, \mathbf{R}_i] = \text{qr}(\mathbf{Y}_i)$;
 - 9: $\mathbf{B}_i = \mathbf{R}_i^{-\top} (\mathbf{H}(:, (i-1)b + 1 : ib))^\top - \Omega_i^\top \mathbf{B}^\top \mathbf{B}$;
 - 10: $\mathbf{Q} = [\mathbf{Q}, \mathbf{Q}_i]$; $\mathbf{B} = [\mathbf{B}^\top, \mathbf{B}_i^\top]^\top$;
 - 11: **end for**
-

A major difference between Algorithm 3 and the algorithm in Fig. 1 is the multiplications with \mathbf{A} moved out of the loop. With $\mathbf{G} = \mathbf{A}\Omega$, Steps 7 and 8 in Algorithm 3 perform the same function as Step (3) in the latter. Here, “qr” denotes a standard QR factorization, i.e., $\mathbf{Q}_i \mathbf{R}_i = \mathbf{Y}_i$. During the first iteration, \mathbf{Q} and \mathbf{B} are null matrices and therefore we should drop off the last items in Step 7 and Step 9. The equivalence of the both algorithms is guaranteed with Theorem 1.

Theorem 1. *The \mathbf{Q} and \mathbf{B} obtained with Algorithm 3 satisfy: \mathbf{Q} is orthonormal and $\mathbf{B} = \mathbf{Q}^\top \mathbf{A}$.*

Proof. We prove Theorem 1 via induction. For any variable v after the i -th iteration of the loop is executed, we use $v^{(i)}$ to denote its value. Moreover, we assume the random matrix Ω is of full column rank. In the base case, $\mathbf{Q}^{(1)} = \mathbf{Q}_1$ is orthonormal because of Step 8 in Algorithm 3. It also ensures that \mathbf{Q}_i is orthonormal, and $\mathbf{Q}_i \mathbf{R}_i = \mathbf{Y}_i$. So,

$$\begin{aligned} \mathbf{B}^{(1)} &= \mathbf{B}_1 = \mathbf{R}_1^{-\top} \Omega_1^\top \mathbf{A}^\top \mathbf{A} = (\mathbf{A}\Omega_1 \mathbf{R}_1^{-1})^\top \mathbf{A} \\ &= (\mathbf{Y}_1 \mathbf{R}_1^{-1})^\top \mathbf{A} = \left(\mathbf{Q}^{(1)}\right)^\top \mathbf{A}. \end{aligned} \quad (4)$$

Now, suppose the proposition holds for the i -th iteration. We need to prove $\mathbf{Q}^{(i+1)}$ is orthonormal and $\mathbf{B}^{(i+1)} = (\mathbf{Q}^{(i+1)})^\top \mathbf{A}$. We first check the orthogonality of \mathbf{Q}_{i+1} .

$$\begin{aligned} \mathbf{Q}_{i+1}^\top \mathbf{Q}^{(i)} &= \left((\mathbf{A} - \mathbf{Q}^{(i)} \mathbf{B}^{(i)}) \Omega_{i+1} \mathbf{R}_{i+1}^{-1} \right)^\top \mathbf{Q}^{(i)} \\ &= \left((\mathbf{A} - \mathbf{Q}^{(i)} (\mathbf{Q}^{(i)})^\top \mathbf{A}) \Omega_{i+1} \mathbf{R}_{i+1}^{-1} \right)^\top \mathbf{Q}^{(i)} \\ &= ((\mathbf{I} - \mathbf{P}_{\mathbf{Q}^{(i)}}) \mathbf{A} \Omega_{i+1} \mathbf{R}_{i+1}^{-1})^\top \mathbf{Q}^{(i)} \\ &= (\mathbf{A} \Omega_{i+1} \mathbf{R}_{i+1}^{-1})^\top (\mathbf{Q}^{(i)} - \mathbf{P}_{\mathbf{Q}^{(i)}} \mathbf{Q}^{(i)}) = \mathbf{O}. \end{aligned} \quad (5)$$

The last two equalities of (5) is based on the properties of projector matrix $\mathbf{P}_{\mathbf{Q}^{(i)}} \equiv \mathbf{Q}^{(i)} (\mathbf{Q}^{(i)})^\top$. See the Appendix. Eq.

(5) guarantees that $\mathbf{Q}^{(i+1)}$ is an orthonormal matrix. Then,

$$\begin{aligned} \mathbf{B}_{i+1} &= \mathbf{R}_{i+1}^{-\top} \boldsymbol{\Omega}_{i+1}^{\top} \left(\mathbf{A}^{\top} \mathbf{A} - \mathbf{B}^{(i)\top} \mathbf{B}^{(i)} \right) \\ &\stackrel{1}{=} \mathbf{R}_{i+1}^{-\top} \boldsymbol{\Omega}_{i+1}^{\top} \left(\mathbf{A} - \mathbf{Q}^{(i)} \mathbf{B}^{(i)} \right)^{\top} \left(\mathbf{A} - \mathbf{Q}^{(i)} \mathbf{B}^{(i)} \right) \\ &= \left(\mathbf{A} \boldsymbol{\Omega}_{i+1} \mathbf{R}_{i+1}^{-1} - \mathbf{Q}^{(i)} \mathbf{B}^{(i)} \boldsymbol{\Omega}_{i+1} \mathbf{R}_{i+1}^{-1} \right)^{\top} \left(\mathbf{A} - \mathbf{Q}^{(i)} \mathbf{B}^{(i)} \right) \\ &= \mathbf{Q}_{i+1}^{\top} \left(\mathbf{A} - \mathbf{Q}^{(i)} \mathbf{B}^{(i)} \right) \stackrel{2}{=} \mathbf{Q}_{i+1}^{\top} \mathbf{A}, \end{aligned}$$

where equality 1 holds due to $\mathbf{Q}^{(i)}$ is orthonormal and $\mathbf{B}^{(i)} = \left(\mathbf{Q}^{(i)} \right)^{\top} \mathbf{A}$. Equality 2 just follows from (5).

Therefore, \mathbf{Q} is orthonormal and $\mathbf{B} = \mathbf{Q}^{\top} \mathbf{A}$, based on the induction hypothesis and Step 10 in Algorithm 3. \square

As the blocked randQB algorithm is mathematically equivalent to the basic randQB, Algorithm 3 inherits the theoretical error bound (if ignoring the round-off error):

$$\mathbb{E}(\|\mathbf{A} - \mathbf{Q}\mathbf{B}\|_F) \leq \left(1 + \frac{k}{s-1}\right)^{1/2} \left(\sum_{j=k+1}^{\min(m,n)} \sigma_{jj}^2\right)^{1/2}, \quad (6)$$

where \mathbb{E} denotes expectation. We see that the theoretically minimal error is only magnified by a factor of $(1 + \frac{k}{s-1})^{1/2}$. If measuring the error with l_2 -norm, we have a similar error bound formula (see Theorem 10.6 of [Halko *et al.*, 2011b]).

3.2 The Version with Re-Orthogonalization

Due to the accumulation of round-off errors, the orthonormality among the columns in $\{\mathbf{Q}_1, \mathbf{Q}_2, \dots\}$ may lose. This affects the correctness of some statements in Algorithm 3, and increases the error of its output. To fix this problem, we explicitly reproject \mathbf{Q}_i away from the span of the previously computed basis vectors, just as what is done in [Martinsson and Voronin, 2016]. Then, the formula for matrix \mathbf{B}_i is revised to incorporate the modified \mathbf{Q}_i .

The re-orthogonalization step corresponds to:

$$\tilde{\mathbf{Q}}_i \tilde{\mathbf{R}}_i = \mathbf{Q}_i - \mathbf{Q}\mathbf{Q}^{\top} \mathbf{Q}_i, \quad (7)$$

where $\tilde{\mathbf{Q}}_i \neq \mathbf{Q}_i$ and $\tilde{\mathbf{R}}_i \neq \mathbf{I}$ due to round-off error. And, $\tilde{\mathbf{Q}}_i$ is better orthogonal to the previously generated $\{\mathbf{Q}_1, \mathbf{Q}_2, \dots, \mathbf{Q}_{i-1}\}$ than \mathbf{Q}_i . Since $\mathbf{Q}_i \mathbf{R}_i = \mathbf{Y}_i$,

$$\tilde{\mathbf{Q}}_i = (\mathbf{I} - \mathbf{Q}\mathbf{Q}^{\top}) \mathbf{Y}_i \mathbf{R}_i^{-1} \tilde{\mathbf{R}}_i^{-1}, \quad (8)$$

$$\begin{aligned} \tilde{\mathbf{B}}_i &= \tilde{\mathbf{Q}}_i^{\top} \mathbf{A} = (\tilde{\mathbf{R}}_i \mathbf{R}_i)^{-\top} \mathbf{Y}_i^{\top} (\mathbf{I} - \mathbf{Q}\mathbf{Q}^{\top}) \mathbf{A} \\ &= (\tilde{\mathbf{R}}_i \mathbf{R}_i)^{-\top} (\mathbf{H}_i^{\top} - \mathbf{Y}_i^{\top} \mathbf{Q}\mathbf{B} - \boldsymbol{\Omega}_i^{\top} \mathbf{B}^{\top} \mathbf{B}), \end{aligned} \quad (9)$$

where \mathbf{H}_i denotes $\mathbf{H}(:, (i-1)b+1 : ib)$. The last equality utilizes that $\mathbf{B} = \mathbf{Q}^{\top} \mathbf{A}$, although this may not hold after a large number of iterations due to numerical round-off error.

Based on (7) and (9), the version with re-orthogonalization can be obtained by replacing Step 9 with the following steps:

$$\begin{aligned} 9: & [\mathbf{Q}_i, \tilde{\mathbf{R}}_i] = \text{qr}(\mathbf{Q}_i - \mathbf{Q}(\mathbf{Q}^{\top} \mathbf{Q}_i)); \\ 9': & \mathbf{R}_i = \tilde{\mathbf{R}}_i \mathbf{R}_i; \\ 9'': & \mathbf{B}_i = \mathbf{R}_i^{-\top} (\mathbf{H}(:, (i-1)b+1 : ib)^{\top} - \mathbf{Y}_i^{\top} \mathbf{Q}\mathbf{B} - \boldsymbol{\Omega}_i^{\top} \mathbf{B}^{\top} \mathbf{B}); \end{aligned}$$

Here, \mathbf{Q}_i and \mathbf{B}_i are overwritten to stand for $\tilde{\mathbf{Q}}_i$ and $\tilde{\mathbf{B}}_i$.

3.3 The Single-Pass Algorithm for PCA

An important feature of Algorithm 3 is that Steps 3 and 4 can be executed with only one pass over matrix \mathbf{A} . Suppose \mathbf{a}_i

and \mathbf{g}_i denote the i -th rows of matrix \mathbf{A} and \mathbf{G} , respectively.

$$\mathbf{H} = [\mathbf{a}_1^{\top}, \mathbf{a}_2^{\top}, \dots, \mathbf{a}_m^{\top}] \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \\ \vdots \\ \mathbf{g}_m \end{bmatrix} = \sum_{i=1}^m \mathbf{a}_i^{\top} \mathbf{g}_i. \quad (10)$$

So, with the i -th row of \mathbf{A} , we can calculate the i -th row of \mathbf{G} with Step 3, and then the i -th item in the summation for calculating \mathbf{H} as (10). Combined with the over-sampling, the single-pass algorithm for computing PCA is as Algorithm 4.

Algorithm 4 A single-pass algorithm for computing PCA

Require: $\mathbf{A} \in \mathbb{R}^{m \times n}$, rank parameter k , block size b .

- 1: $\mathbf{Q} = []$; $\mathbf{B} = []$;
- 2: Choose $l = tb$, which is slightly larger than k ;
- 3: $\boldsymbol{\Omega} = \text{randn}(n, l)$; $\mathbf{G} = []$; Set \mathbf{H} to an $n \times l$ zero matrix;
- 4: **while** \mathbf{A} is not completely read through **do**
- 5: Read next few rows of \mathbf{A} into RAM, denoted by \mathbf{a} ;
- 6: $\mathbf{g} = \mathbf{a}\boldsymbol{\Omega}$; $\mathbf{G} = [\mathbf{G}; \mathbf{g}]$;
- 7: $\mathbf{H} = \mathbf{H} + \mathbf{a}^{\top} \mathbf{g}$;
- 8: **end while**
- 9: **for** $i = 1, 2, \dots, t$ **do**
- 10: $\boldsymbol{\Omega}_i = \boldsymbol{\Omega}(:, (i-1)b+1 : ib)$;
- 11: $\mathbf{Y}_i = \mathbf{G}(:, (i-1)b+1 : ib) - \mathbf{Q}(\mathbf{B}\boldsymbol{\Omega}_i)$;
- 12: $[\mathbf{Q}_i, \tilde{\mathbf{R}}_i] = \text{qr}(\mathbf{Y}_i)$;
- 13: $[\mathbf{Q}_i, \tilde{\mathbf{R}}_i] = \text{qr}(\mathbf{Q}_i - \mathbf{Q}(\mathbf{Q}^{\top} \mathbf{Q}_i))$;
- 14: $\mathbf{R}_i = \tilde{\mathbf{R}}_i \mathbf{R}_i$;
- 15: $\mathbf{B}_i = \mathbf{R}_i^{-\top} (\mathbf{H}(:, (i-1)b+1 : ib)^{\top} - \mathbf{Y}_i^{\top} \mathbf{Q}\mathbf{B} - \boldsymbol{\Omega}_i^{\top} \mathbf{B}^{\top} \mathbf{B})$;
- 16: $\mathbf{Q} = [\mathbf{Q}, \mathbf{Q}_i]$; $\mathbf{B} = [\mathbf{B}^{\top}, \mathbf{B}_i^{\top}]^{\top}$;
- 17: **end for**
- 18: $[\tilde{\mathbf{U}}, \mathbf{S}, \mathbf{V}] = \text{svd}(\mathbf{B})$;
- 19: $\mathbf{U} = \mathbf{Q}\tilde{\mathbf{U}}$;
- 20: $\mathbf{U} = \mathbf{U}(:, 1 : k)$; $\mathbf{V} = \mathbf{V}(:, 1 : k)$; $\mathbf{S} = \mathbf{S}(1 : k, 1 : k)$;
- 21: **return** $\mathbf{U}, \mathbf{S}, \mathbf{V}$.

In the algorithm, the while loop corresponds to Steps 3 and 4 in Algorithm 3, but involves only one pass over \mathbf{A} . In every step, small matrices in size $m \times l$ or $n \times l$ (noting $l \ll \min(m, n)$) are used. So, the memory cost of this algorithm is small, which can be bounded by that for storing $(m+2n)l$ floating numbers. The computational cost of this algorithm is the same as Algorithm 3 and 1 [Halko *et al.*, 2011a], i.e., $\mathcal{O}(mnk)$ or $\mathcal{O}(mn \log(k))$ flops. The theoretical error bounds of $\mathbf{A} - \mathbf{Q}\mathbf{B}$ also apply to $\mathbf{A} - \mathbf{U}\mathbf{S}\mathbf{V}^{\top}$ in Algorithm 4, as the latter hardly induces new error.

This single-pass algorithm requests that the data matrix \mathbf{A} is stored in a row-major format. Otherwise, we can apply the algorithm to \mathbf{A}^{\top} instead. In case there is a request for higher accuracy, the power scheme can be applied with like $P=1$, which is equivalent to replacing \mathbf{A} with $\mathbf{A}\mathbf{A}^{\top}\mathbf{A}$ in the algorithm. It requests one additional pass over \mathbf{A} , even the orthonormalization is enforced [Martinsson and Voronin, 2016; Voronin and Martinsson, 2016]. Nevertheless, the single-pass algorithm works well in many applications.

4 Experiments

All experiments are carried out on a Linux server with two 12-core Intel Xeon E5-2630 CPUs (2.30 GHz), and 32 GB

RAM. The algorithms are implemented in C with OpenMP derivatives for multi-thread computing, and MKL libraries [Int, 2016]. QR factorization and other basic linear algebra operations are realized through LAPACK routines which are automatically executed in parallel on the multi-core CPUs.

We first validate the accuracy of the proposed single-pass algorithm. Then, large test cases stored on hard disk in IEEE single-precision float format are used to validate the algorithm’s efficiency. In all experiments, the block size $b = 10$.

4.1 Accuracy Validation

We consider test matrices owning the following singular spectrums with different decaying behavior, where σ_{ii} denotes the i -th singular value (i.e., a diagonal element of matrix Σ).

- Type 1:
$$\sigma_{ii} = \begin{cases} 10^{-4(i-1)/19}, & i = 1, 2, \dots, 20, \\ 10^{-4/(i-20)^{1/10}}, & i = 21, 22, \dots, \min(m, n). \end{cases}$$
- Type 2: $\sigma_{ii} = i^{-2}, i = 1, 2, \dots$
- Type 3: $\sigma_{ii} = i^{-3}, i = 1, 2, \dots$
- Type 4: $\sigma_{ii} = e^{-i/7}, i = 1, 2, \dots$
- Type 5: $\sigma_{ii} = 10^{-i/10}, i = 1, 2, \dots$

Type 1 is from [Halko *et al.*, 2011a], and Type 3 and Type 5 are from [Mary *et al.*, 2015]. These singular value distributions are shown in Fig. 2. It reveals that the singular values of Type 1 and Type 2 matrices decay asymptotically slowly, although they attenuate fast at the start. The singular values of Type 4 and Type 5 matrices decay asymptotically faster.

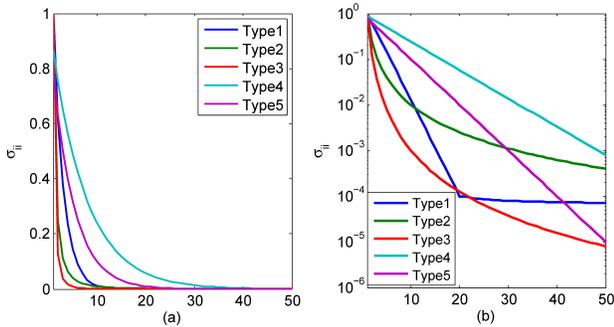


Figure 2: Different decay behavior of the singular values of the test matrices. (a) Normal plot, (b) Semi-logarithmic plot.

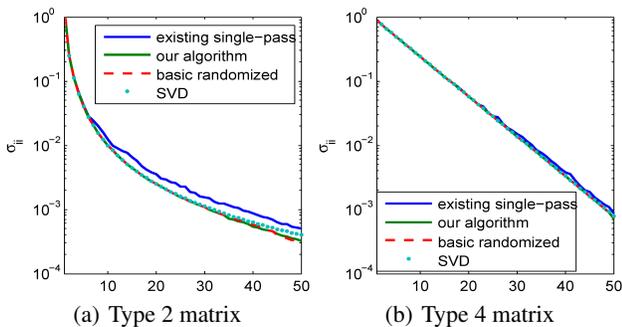


Figure 3: The computed singular values for a slow-decay and a fast-decay matrix, showing the accuracy of our algorithm.

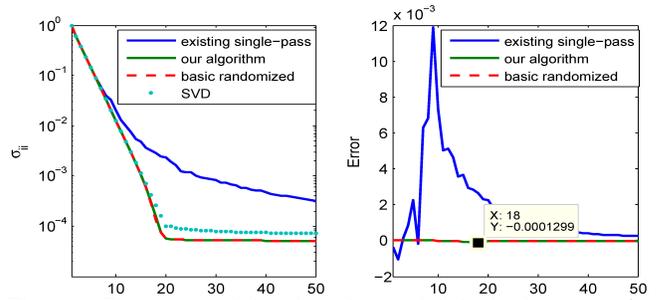


Figure 4: The computed singular values and their absolute errors for a very slow-decay matrix (Type 1), showing the advantage of our algorithm over Algorithm 2.

For each type, we construct a 3000×3000 matrix through multiplying Σ with randomly drawn orthogonal matrices \mathbf{U} and \mathbf{V} . We compute the first 50 singular values and singular vectors for each matrix with the basic randomized Algorithm 1, the existing single-pass algorithm (Algorithm 2) and our Algorithm 4, and compare the results with the accurate values obtained by SVD. The over-sampling parameter is set to 10 (i.e., $l = 60$). Fig. 3 shows the computed singular values of two matrices, which demonstrates the single-pass algorithm in [Halko *et al.*, 2011b] produces much larger error, and the results of Algorithms 1 and 4 are indistinguishable. It also reveals that the algorithms produce better results for matrices with asymptotically faster decay of singular values. This is a common property of the randomized algorithms based on QB approximation [Halko *et al.*, 2011b; Mary *et al.*, 2015; Martinsson and Voronin, 2016]. So, we will focus on the accuracy for the matrices with slow decay of singular values.

For the Type 1 matrix, the accuracy of the randomized algorithms all decreases; the existing single-pass algorithm [Halko *et al.*, 2011b] produces considerably large error (up to 1.2×10^{-2}), as shown in Fig. 4. While using the proposed Algorithm 4, we can reduce the maximum error to 1.3×10^{-4} (**~92X smaller**). Its accuracy looks acceptable. Fig. 5(a) shows the first principal components (i.e., \mathbf{v}_1) computed by SVD and our algorithm respectively, which look indistinguishable (only 2.8×10^{-5} difference in l_∞ -norm). For other principal components, we calculate the correlation coefficient between the results obtained with the both methods. As shown in Fig. 5(b), the correlation coefficients are close to 1. The largest deviation occurs for the 10th principal component, with value 0.9993. For other matrices with faster decay of singular val-

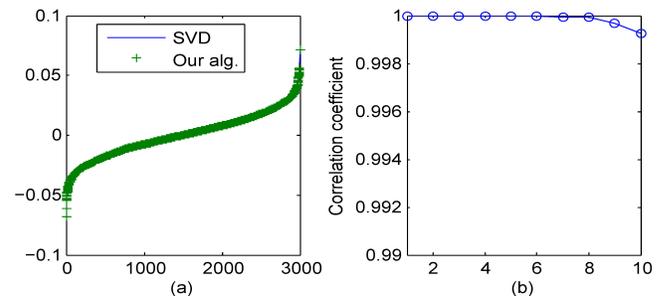


Figure 5: The accuracy of our algorithm on principal components (with comparison to the results from SVD). (a) The numeric values of the first principal component (\mathbf{v}_1). (b) The correlation coefficients for the first 10 principal components.

ues (Types 2~5), the randomized algorithm exhibits better accuracy and outputs more accurate principal components.

4.2 Runtime Comparison

Following [Halko *et al.*, 2011a], we construct several large data using the unitary discrete cosine transform (command “`dct`” in Matlab). They are $200,000 \times 200,000$ matrices following the singular value distributions given in last subsection. Each matrix is stored as a 149 GB file on hard disk. We use `fread` function to read the file and run Algorithm 4 for computing PCA. Each time we read l rows of matrix, to avoid extra memory cost. Once they are loaded into RAM, the data are converted to the IEEE double-precision format. Algorithm 1 and Algorithm 2 are also tested for comparison.

Some results for the matrices with slow-decay singular values are listed in Table 1. l in Algorithm 4 is set to 20 or 30. t_{read} and t_{PCA} mean the total time (in seconds) for reading the data and the total runtime of the algorithm (including t_{read}), respectively. “`max_err`” is the maximum error of the computed singular values. From the table we see that the time for reading data dominates the total runtime, and the proposed algorithm is about **2X faster** than the basic randomized algorithm used in [Halko *et al.*, 2011a] while keeping same accuracy. If comparing Algorithm 2 and ours, we see that the former may be slightly faster but produces much larger error.

To improve the accuracy, the power scheme with $P = 1$ could be applied, which corresponds to one more pass over the data. For Algorithm 4, we just run the while loop once again with Ω replaced by \mathbf{H} after “`orth`” operation. In our experiments, this two-pass algorithm has similar runtime as Algorithm 1, but dramatically reduces “`max_err`” to 4.6×10^{-7} and 3×10^{-6} for the Type 1 and Type 2 matrices, respectively.

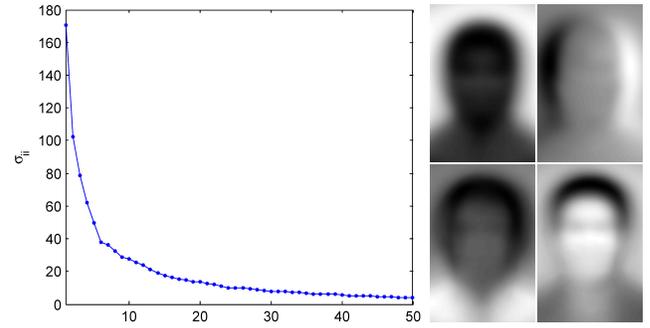
In the experiments, the memory cost of Algorithm 4 ranges from **402 MB** to **490 MB**. In contrast, the standard SVD (including “`svds`” in Matlab for truncated SVD) requests much larger memory than the available physical RAM, and therefore does not work. To have a taste of how fast the proposed randomized algorithm runs, we test a $10,000 \times 10,000$ matrix. Performing a complete SVD and “`svds`” for the first 50 principal components take 226 and 219 seconds, respectively, while the proposed algorithm costs only 0.69 seconds.

4.3 Real Data

We apply the single-pass algorithm with $k=50$ to the matrix representing the images of faces from the FERET database [Phillips *et al.*, 2000]. As in [Halko *et al.*, 2011a], we add two duplicates for each image into the data. For each duplicate, the value of a random choice of 10% of the pixels is set to random numbers uniformly chosen from $0, 1, \dots, 255$.

Table 1: The results for several $200,000 \times 200,000$ data, which demonstrate the efficiency of our Algorithm 4. (unit of time: second)

Matrix	k	Algorithm 1		Algorithm 2		Algorithm 4				
		t_{read}	t_{PCA}	t_{read}	t_{PCA}	t_{read}	t_{PCA}			
Type1	16	2390	2607	1.7e-3	1186	1404	2.2e-2	1206	1426	1.8e-3
Type1	20	2420	2616	9e-4	1198	1380	1.6e-1	1217	1413	1.2e-3
Type1	24	2401	2593	1e-3	1216	1400	1.5e-1	1216	1414	1.2e-3
Type2	12	2553	2764	5e-4	1267	1477	3e-2	1276	1490	5e-4
Type3	24	2587	2777	1e-5	1312	1500	1.7e-3	1310	1502	2e-5



(a) Computed singular values (b) Four eigenfaces
Figure 6: The computational results for the FERET matrix.

This forms a $102,042 \times 393,216$ matrix, whose rows consist of images. Before processing, we normalize the matrix by subtracting from each row its mean, and then dividing it by its Euclidean norm. With the proposed algorithm, it takes 1453 seconds (**~ 24 minutes**) to process all 150 GB of this data stored on disk. The computed singular values are plotted in Fig. 6. We have also checked the computed “eigenfaces”, which well match those presented in [Halko *et al.*, 2011a].

5 Conclusions

An algorithm for single-pass PCA of large and high-dimensional data is proposed. It involves only one pass over the data, and keeps the comparable accuracy to the existing randomized algorithms. Experiments demonstrate the algorithm’s effectiveness for computing the principal components of large-size (~ 150 GB) data with high dimension that cannot be fit in memory, in terms of runtime and memory usage.

6 Appendix: Orthogonal Projector Matrix

An orthogonal projector matrix corresponds to a linear transformation which converts any vector to its orthogonal projection on a subspace. It is uniquely determined by the subspace, e.g., $\text{range}(\mathbf{X})$ corresponding to a projector matrix denoted by \mathbf{P}_X . Based on the theory of linear least squares, if \mathbf{X} has full column rank [Golub and Van Loan, 1996],

$$\mathbf{P}_X = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T. \quad (11)$$

It is simplified to $\mathbf{P}_X = \mathbf{X}\mathbf{X}^T$, if \mathbf{X} is an orthonormal matrix. It is easy to see the following properties of a projector.

Lemma 1. For a real-valued matrix \mathbf{X} with full column rank,

- \mathbf{P}_X is a symmetric matrix, and $\mathbf{P}_X^2 = \mathbf{P}_X$.
- $\mathbf{I} - \mathbf{P}_X$ is the orthogonal projector determined by the orthogonal complement of $\text{range}(\mathbf{X})$.
- $\mathbf{P}_X \mathbf{X} - \mathbf{X} = \mathbf{O}$, where \mathbf{O} is the zero matrix.

Acknowledgments

This work was supported by NSFC under grant No. 61422402, and in part by Beijing NSF No. 4172059 and NSF No. CCF-1066471.

Portions of the research in this paper use the FERET database of facial images collected under the FERET program, sponsored by the DOD Counterdrug Technology Development Program Office.

References

- [Bahdanau *et al.*, 2014] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- [De Stefani *et al.*, 2016] Lorenzo De Stefani, Alessandro Epasto, Matteo Riondato, and Eli Upfal. TRIEST: Counting local and global triangles in fully-dynamic streams with fixed memory size. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.
- [Drineas and Mahoney, 2016] Petros Drineas and Michael W Mahoney. RandNLA: Randomized numerical linear algebra. *Communications of the ACM*, 59(6):80–90, 2016.
- [Eckart and Young, 1936] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.
- [Friedman *et al.*, 2001] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The Elements of Statistical Learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.
- [Gittens and Mahoney, 2016] Alex Gittens and Michael W Mahoney. Revisiting the nyström method for improved large-scale machine learning. *J. Mach. Learn. Res.*, 17:1–65, 2016.
- [Golub and Van Loan, 1996] Gene H Golub and Charles F Van Loan. *Matrix Computations*. Johns Hopkins University Press, 1996.
- [Halko *et al.*, 2011a] Nathan Halko, Per-Gunnar Martinsson, Yoel Shkolnisky, and Mark Tygert. An algorithm for the principal component analysis of large data sets. *SIAM Journal on Scientific Computing*, 33(5):2580–2594, 2011.
- [Halko *et al.*, 2011b] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.
- [Int, 2016] Intel Parallel Studio XE Cluster Edition for Linux. <https://software.intel.com/en-us/intel-parallel-studio-xe>, 2016.
- [Liang *et al.*, 2014] Yingyu Liang, Maria-Florina F Balcan, Vandana Kanchanapally, and David Woodruff. Improved distributed principal component analysis. In *Advances in Neural Information Processing Systems*, pages 3113–3121, 2014.
- [Liberty, 2013] Edo Liberty. Simple and deterministic matrix sketching. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 581–588. ACM, 2013.
- [Martinsson and Voronin, 2016] P.-G. Martinsson and S. Voronin. A randomized blocked algorithm for efficiently computing rank-revealing factorizations of matrices. *SIAM Journal on Scientific Computing*, 38(5):S485–S507, 2016.
- [Mary *et al.*, 2015] Théo Mary, Ichitaro Yamazaki, Jakub Kurzak, Piotr Luszczek, Stanimire Tomov, and Jack Dongarra. Performance of random sampling for computing low-rank approximations of a dense matrix on GPUs. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC ’15, pages 60:1–60:11, New York, NY, USA, 2015. ACM.
- [Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.
- [Ordonez *et al.*, 2014] Carlos Ordonez, Naveen Mohanam, and Carlos Garcia-Alvarado. PCA for large data sets with parallel data summarization. *Distrib. Parallel Databases*, 32(3):377–403, September 2014.
- [Phillips *et al.*, 2000] P Jonathon Phillips, Hyeonjoon Moon, Syed A Rizvi, and Patrick J Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000.
- [Rokhlin *et al.*, 2009] Vladimir Rokhlin, Arthur Szlam, and Mark Tygert. A randomized algorithm for principal component analysis. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1100–1124, 2009.
- [Voronin and Martinsson, 2016] S. Voronin and P.-G. Martinsson. RSVDPACK: Subroutines for computing partial singular value decompositions via randomized sampling on single core, multi core, and GPU architectures. *arXiv preprint, arXiv:1502.05366v3*, 2016.
- [Wang and Zhang, 2013] Shusen Wang and Zhihua Zhang. Improving cur matrix decomposition and the nyström approximation via adaptive sampling. *Journal of Machine Learning Research*, 14(1):2729–2769, 2013.
- [Wang *et al.*, 2016] Shusen Wang, Zhihua Zhang, and Tong Zhang. Towards more efficient spsd matrix approximation and cur matrix decomposition. *Journal of Machine Learning Research*, 17(210):1–49, 2016.
- [Wang, 2015] Shusen Wang. A practical guide to randomized matrix computations with MATLAB implementations. *arXiv preprint arXiv:1505.07570*, 2015.
- [Woodruff, 2014] David P Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.
- [Ye *et al.*, 2016] Qiaomin Ye, Luo Luo, and Zhihua Zhang. Frequent direction algorithms for approximate matrix multiplication with applications in CCA. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI-16)*, pages 2301–2307, 2016.