A Randomized Algorithm for PCA

MATH285K - Spring 2010 *Presenter: Ryan Compton*

Principle component analysis is an essential tool for the study of population genetics. Genetic datasets are typically very large and therefore difficult to study with classical techniques. In this note we review a new randomized algorithm for accurate and efficient PCA [2].

1 Low rank approximation of A

For a given data matrix $A \in \mathbb{R}^{m \times n}$ we seek a rank k approximation matrix B. With respect to the spectral norm

$$\min_{\text{rank}(B)=k} ||A - B|| = \sigma_{k+1}(A) \tag{1}$$

where $\sigma_{k+1}(A)$ is the kth singular value of A. This can be seen by forming an SVD of A and throwing out all components past the kth singular value.

Given an oversampling parameter p the first step in our method is to construct a matrix Q such that

$$||A - QQ^*A|| \approx \min_{\text{rank}(B)=k} ||A - B||. \tag{2}$$

Q orthogonalizes the range of the principle components of A. A minimizing Q exists for $p = 0$ however taking p small will allow us to produce a computationally efficient method.

For intuition consider the case where the rank of A is exactly k. Select k random vectors ω_i and form the products $y_i = A\omega_i$. The set y_i will be linearly independent and thus span the range of A. To compute our Q we need only orthogonalize the y_i.

Now suppose that $A = B + E$ where B has rank k and form the products y_i again. In this case the y_i likely do not span the range of A due to the perturbation E. However, it turns out that sampling only a few more (eg 12) directions ω_i will very likely capture the full range of A.

We execute the following algorithm for the computation of Q:

- Draw a random matrix $\Omega \in \mathbb{R}^{n \times (k+p)}$.

- Form $Y = A\Omega$.
- QR decompose Y and discard R.

The main theoretical result is:

$$\mathbb{E}||A - QQ^*A|| \leq \left(1 + \frac{4 * \sqrt{k+p}}{p-1}\sqrt{\min(m,n)}\right)\sigma_{k+1}(A). \tag{3}$$

Proof Sketch.

Apply the triangle inequality many times in order to split the error into a part that involves optimizing over a space of dimension k and a separate high dimensional part.

Let $\Omega \in \mathbb{R}^{n\times(k+12)}$, $W \in \mathbb{R}^{(k+12)\times n}$ and $Z \in \mathbb{R}^{k\times(k+12)}$

$$||A - QQ^*A|| \leq 2||A - A\Omega W|| + 2||A\Omega - QZ||||W||. \tag{4}$$

we want to choose W and Z to show

$$||A - QQ^*A|| \leq C\sigma_{k+1}(A). \tag{5}$$

The algorithm forms Q's columns from singular vectors corresponding to the $k+p$ greatest singular values of $A\Omega$. This lets us choose Z such that

$$||A - QQ^*A|| \leq \sigma_{k+1}(A\Omega) \leq ||\Omega||\sigma_{k+1}(A) \tag{6}$$

where we understand the second inequality by recalling that we are working with the spectral norm in this note.

The existence of a $(k+p)\times n$ matrix W such that $||A - A\Omega W|| \leq C\sigma_{k+1}(A)$ is tedious and shown in the appendix of [4] using results from [1].□

A few notes about the result:

- A few iterations of the power method in our computation of Y can improve the accuracy of our method.
- We expect the bound in (3) to involve a factor of $\sigma_{k+1}(A)$ as $\sigma_{k+1}(A)$ is the theoretical best bound we can find.
- Notice that increasing p greatly improves accuracy.

2 SVD in reduced subspace.

Now that we have identified a subspace that captures most of the action of A we cheaply compute an SVD in the reduced space. Form

$$B = Q^* A \tag{7}$$

SVD B

$$B = \hat{U}\Sigma V^T \tag{8}$$

replace

$$U = Q\hat{U} \tag{9}$$

then

$$A \approx U\Sigma V^T. \tag{10}$$

We make the following observations about the method

- This is an "out-of-core" technique requiring only 2 passes over A (one to form Y, another to form B). When A too large to fit in RAM we may load A in stages to form the products $A\Omega$ and Q^*A saving only Y and B in memory.
- The major bottleneck of the method is the formation of the products $A\Omega$ and Q^*A. This is readily accelerated on parallel architectures.
- It turns out that replacing Ω with an undersampled DFT matrix provides similar error bounds. This allows one to asymptotically speed up the computation of $A\Omega$ by an exponential factor by using FFT libraries to compute the product [3].

References

[1] Zizhong Chen, Jack, J. Dongarra, and (gmn P. Condition numbers of Gaussian random matrices, 2004.

[2] Edo Liberty, Franco Woolfe, Per-Gunnar Martinsson, Vladimir Rokhlin, and Mark Tygert. Randomized algorithms for the low-rank approximation of matrices. *Proceedings of the National Academy of Sciences of the United States of America*, 104(51):20167–72, December 2007.

[3] Computational Mathematics. FINDING STRUCTURE WITH RANDOMNESS : STOCHASTIC ALGORITHMS FOR CONSTRUCTING APPROXIMATE MATRIX DECOMPOSITIONS N . HALKO , P . G . MARTINSSON , AND J . A . TROPP Technical Report No . 2009-05 September 2009. *Techniques*, 2009.

[4] Vladimir Rokhlin, Arthur Szlam, and Mark Tygert. A RANDOMIZED ALGORITHM FOR. *Computer*, 0811203:1–26.