Randomization as Regularization in Principal Component Analysis/Regression

Mack Yi
Department of Computer Science
Duke University
Durham, NC 27708
mack.yi@duke.edu

Abstract

Randomized algorithms for dimension reduction are interesting not only for their computational performance, but also for their statistical performance. Following up on recent work showing regularization behavior of randomized Singular Value Decomposition (SVD), we explore the regularization behavior of randomized SVD in Principal Component Analysis (PCA) and Principal Components Regression, and observe on simulated data that randomized PCA displays regularization behavior, especially for noisy data. We also comment on the theoretical questions which arise surrounding this regularization behavior.

1 Introduction

In the regime of large data sets which have both a large number of datapoints as well as a large number of features, traditional machine learning algorithms can become infeasible due to computational constraints. One approach to mitigating this problem has been in the area of dimension reduction. Motivated by the assumption that data can often be succinctly described by considerably fewer features due to colinearity or other relationships between features, some smaller set of features is used in place of the full set of raw features. Principal Component Analysis (PCA) and subsampling algorithms are some common examples of dimension reduction methods. Dimension reduction methods give rise to a large class of algorithms which can generally be viewed as a two step process: In the first step, some dimension reduction algorithm is applied, while in the second step, a traditional machine learning algorithm is applied on the reduced dataset at a small scale, with considerable computational savings.

However, when datasets are of very massive scale, the dimension reduction step itself can be computationally infeasible. For example, computing the exact principal components using a Singular Value Decomposition (SVD) of an nxp matrix requires $O(np * min(n, p))$ floating point operations. Intuitively speaking, it is not necessary to calculate the full SVD, since, if we are reducing to k dimensions, we are interested in only the first k singular values. Various randomized algorithms have been established, often with running time around $O(npk)$ (see [Rokhlin et al.(2008)].

While these algorithms have often been motivated by computational constraints, the statistical quality of the results of the algorithm are the primary interest of this project, specifically with respect to the regularization behavior of randomized algorithms. For datasets with a very large number of features, it can often be the case that $n << p$. This can commonly occur, for example, in genomics data, where the features are alleles. In these cases, some method of regularization becomes vital in order to avoid overfitting.

In general, dimension reduction methods, such as PCA or subsampling methods, will provide some regularization. However, there seems to be evidence that randomized dimension reduction algorithms can provide implicit regularization due to the randomization itself. Understanding of this

regularization behavior could reduce or remove the need for explicit regularization terms, resulting in greater algorithmic simplicity and computational savings, and could also be used in concert with explicit regularization methods.

Previous work on randomized algorithms using randomized SVD has shown that randomized versions of algorithms can provide implicit regularization.[Georgiev & Mukherjee(2012)] observed regularization behavior with supervised regression methods (Sliced Inverse Regression and Localized Sliced Inverse Regression) in the data-poor regime where $n < p$, where the predictive performance of the randomized (using randomized SVD) version of the algorithm improved the predictive performance for both SIR and LSIR. Specifically, the authors observed that LSIR with regularization performed well without adding explicit regularization terms, which are generally required for LSIR to perform well in this regime. Similarly, [Darnell et al.(2015)] observed implicit regularization behavior when using randomized SVD in place of deterministic SVD when fitting a Linear Mixed Model for simulated genomics data.

In light of these previous results on regularization behavior, the goal of this project was to further explore the regularization behavior of algorithms utilizing randomized SVD, as well as attempt to establish a theoretical framework for the guarantees which might exist for the generalization error of algorithms using randomized SVD. Specifically, the use of randomized SVD in Principal Components Regression (PCR) is studied versus unrandomized Principal Components Regression.

This is motivated by the intuition that deterministic PCR may be overfitting variation in the first k principal components, while neglecting variation related to subsequent principal components. While methods such as Partial Least Squares Regression or other supervised methods may in general provide greater statistical accuracy when this is an issue, the question of interest here is whether randomized PCR provides regularization which can limit this overfitting behavior and provide better generalization error.

2 Methods

2.1 Randomized SVD

The core randomized algorithm studied in this project is the Adaptive Randomized Singular Value Decomposition (ARSVD) described by [Darnell et al.(2015)]. We neglect the adaptive choice of t and d^*, choosing to directly specify these parameters instead since our focus is on the regularization behavior of the algorithm. The algorithm is summarized in Algorithm 1.

Algorithm 1 Randomized SVD

1: **1. Find an orthonormal basis for the range of X:**
2: a. Set the number of working directions: $l = d_{max} + \Delta$
3: b. Generate random matrix $\Omega \in \mathbb{R}^{nxl}$ with $\Omega_{ij} \overset{iid}{\sim} N(0,1)$
4: c. Construct blocks $F^{(t')} = XX^T F^{(t'-1)}$, $F^{(0)} = \Omega$ for $t' \in 0, ..., t$
5: d. Estimate basis for block using QR decomposition $F = QR$
6: **2. Project data onto the range basis and compute SVD:**
7: a. Project onto the basis: $B = X^T Q \in \mathbb{R}^{rxl}$
8: b. Factorize $B \overset{svd}{=} U\Sigma W^T$
9: c. Take the rank d^* approximation using the first d^* columns of U, the first d^* rows and columns of Σ, and QxW_{d^*}, where W_{d^*} is the first d^* column of W.

The focus here is on the parameter t for the number of power iterations of XX^T, described by [Georgiev & Mukherjee(2012)] as a regularization parameter. This parameter controls the singular value shrinkage, which results in a preference for higher directions of variance in the data as t increases. [Darnell et al.(2015)] observed that as t increases, the randomized singular values (exponentially) rapidly converge to the deterministic singular values. Thus, for relatively small t, the singular values will have already converged, and so any regularization can only be tuned very coarsely when parameterized in this way. However, if regularization is observed, then there would be both statistical as well as computational motivation to reduce the number of power iterations used, since

more regularization can be achieved with fewer power iterations, as it is only in this case that there is significant shrinkage of the randomized singular values versus the deterministic singular values.

2.2 PCR based on randomized SVD

Plugging a randomized SVD into Principal Components Regression is straighforward. The Principal Component Analysis step is replaced by PCA using randomized SVD. The algorithm is given in Algorithm 2. The deterministic version of this algorithm, using an implementation of PCA using a deterministic SVD of the data X, is used for comparison of the regularization behavior of the randomized algorithm.

Algorithm 2 Randomized PCR on data X, Y

1: **1. Dimension Reduction:**
2: a. Take the randomized SVD of $X = U\Sigma V^T$, parameterized by d^* and t, determined either adaptively or given
3: b. Extract the d^* principal components $U\Sigma$
4: **2. Regression:**
5: a. Perform ordinary Least Squares Regression on the principal components and Y to find coefficients β
6: b. Transform the coefficients to the original space $\beta' = \beta V$

3 Results

3.1 Simulated Low Rank Data

Low rank data is simulated using the same method used by [Darnell et al.(2015)]. The data matrix $X \in \mathbb{R}^{nxp}$ is generated as $X = USV^T + E$, where $U^TU = V^TV = I_{d^*}$. The columns of U and V are drawn uniformly at random from the unit sphere and the singular values are generated starting from a baseline value, which is a fraction of the maximum noise singular value, with exponential increments separating consecutive entries. The noise is iid Gaussian.

The signal to noise ratio is controlled by κ, where larger values correspond to greater separation between the signal and the noise. κ is defined as the fraction of the maximum noise singular value used above to determine the baseline value for the singular values of the low-rank signal.

3.2 Regularization of Principal Component Analysis

First, similar to [Darnell et al.(2015)], it is observed that regularization occurs within the SVD step itself, so that the randomized PCA tends to provide a better approximation to the "true" principal components of the data than PCA based on unrandomized SVD.

The randomized SVD differs from the unrandomized SVD more when there are fewer power iterations (small t). In addition, the singular values diverge more in the smaller singular values than the larger ones. This behavior is suggestive of regularization behavior, as it is in the directions of the principal components with smaller singular values that the variance is more affected by noise. By shrinking these singular values, the randomized SVD thus seems to be better fitting the principal components of the signal rather than the noise.

This intuition is supported by simulations on low rank data generated as described in the previous subsection. Data is generated for $n = 500$, $p = 5000$, and $d^* = 20$, which is representative of the typical case where $n < p$ and $d^* << n, p$. The effect of the regularization parameter t is observed by setting it equal to 1, 2, and 3. Since an exponential decrease in relative error between the singular values obtained from the unrandomized and randomized SVD is observed, this small range is generally enough to cover the space where there is interesting variation between the randomized and unrandomized methods. Finally, the effect of different choices of the signal to noise ratio κ is studied.

The results on simulated data suggest that the unrandomized SVD consistently overestimates the singular values relative to the singular values which we use to generate the signal component of

Table 1: Percent relative error of singular values

Parameters	RSVD vs SVD	RSVD vs TRUE	SVD vs TRUE
$\kappa = 0.5, t = 1$	18.137	8.981	21.457
$\kappa = 0.5, t = 2$	7.551	10.899	21.457
$\kappa = 0.5, t = 3$	3.578	16.179	21.457
$\kappa = 1.0, t = 1$	16.010	10.831	22.222
$\kappa = 1.0, t = 2$	6.946	11.950	22.222
$\kappa = 1.0, t = 3$	4.136	15.960	22.222
$\kappa = 2.0, t = 1$	12.515	8.865	5.522
$\kappa = 2.0, t = 2$	4.782	5.000	5.522
$\kappa = 2.0, t = 3$	0.249	5.237	5.522
$\kappa = 3.0, t = 1$	14.034	10.981	4.230
$\kappa = 3.0, t = 2$	2.121	4.091	4.230
$\kappa = 3.0, t = 3$	0.013	4.216	4.230

the data, and the randomized SVD is able to outperform the unrandomized version by shrinking the singular values, favoring the shrinkage of the smaller singular values, as can be observed in Figure 1. It seems that this shrinkage is not always of an appropriate degree, for example, in the $\kappa = 3$ and $t = 1$ case, the shrinkage appears to be too extreme for the smaller singular values, and the error is actually significantly larger for the randomized SVD compared to the unrandomized SVD. However, choosing $t = 2$ instead appears to provide better behavior, after which the randomized SVD rapidly converges to the unrandomized SVD.

Thus, here appears to be a balance between inaccuracy of the randomized SVD and regularization behavior when adjusting the regularization parameter t. The results of the simulation suggest that there may be some point which is "optimal" in finding the true signal singular values, often this at $t = 2$ for these particular cases. The adaptive estimation of t described by [Darnell et al.(2015)] could be seen as working to find this point.

In addition, it seems that this regularization behavior is stronger for smaller κ, i.e. when there is a low signal to noise ratio. This can be observed qualitatively in Figure 1 or more explicitly in Table 1. In the cases where the signal to noise ratio is quite high, we observe that the randomized SVD converges very rapidly to the unrandomized SVD, and that the error is lower in all cases. On the other hand, for small κ, the randomized SVD has greater error relative to the unrandomized SVD, but provides a significantly better estimate of the "true" singular values of the signal. In fact, for $\kappa = 0.5$ and $\kappa = 1$, there is actually a trade-off between error relative to the unrandomized SVD and error relative to the "true" singular values.

3.3 Principal Components Regression

The regularization behavior in the Principal Component Analysis step suggests that the randomized SVD could also provide regularization in the PCR process as a whole. Given that the assumption when applying PCR is generally that the dependence between X and Y is captured by the principal components. Thus, better estimation of the "true" principal components might provide better generalization error when the X and Y are related in a way matching the assumptions of PCR, especially when Y depends strongly on the principal components with smaller singular values, for which the unrandomized PCR is significantly more accurate on based on the PCA simulations. Regularization behavior in this area was not further studied or simulated, but the observations on randomized PCA suggest that there should be some kind of regularization behavior here.

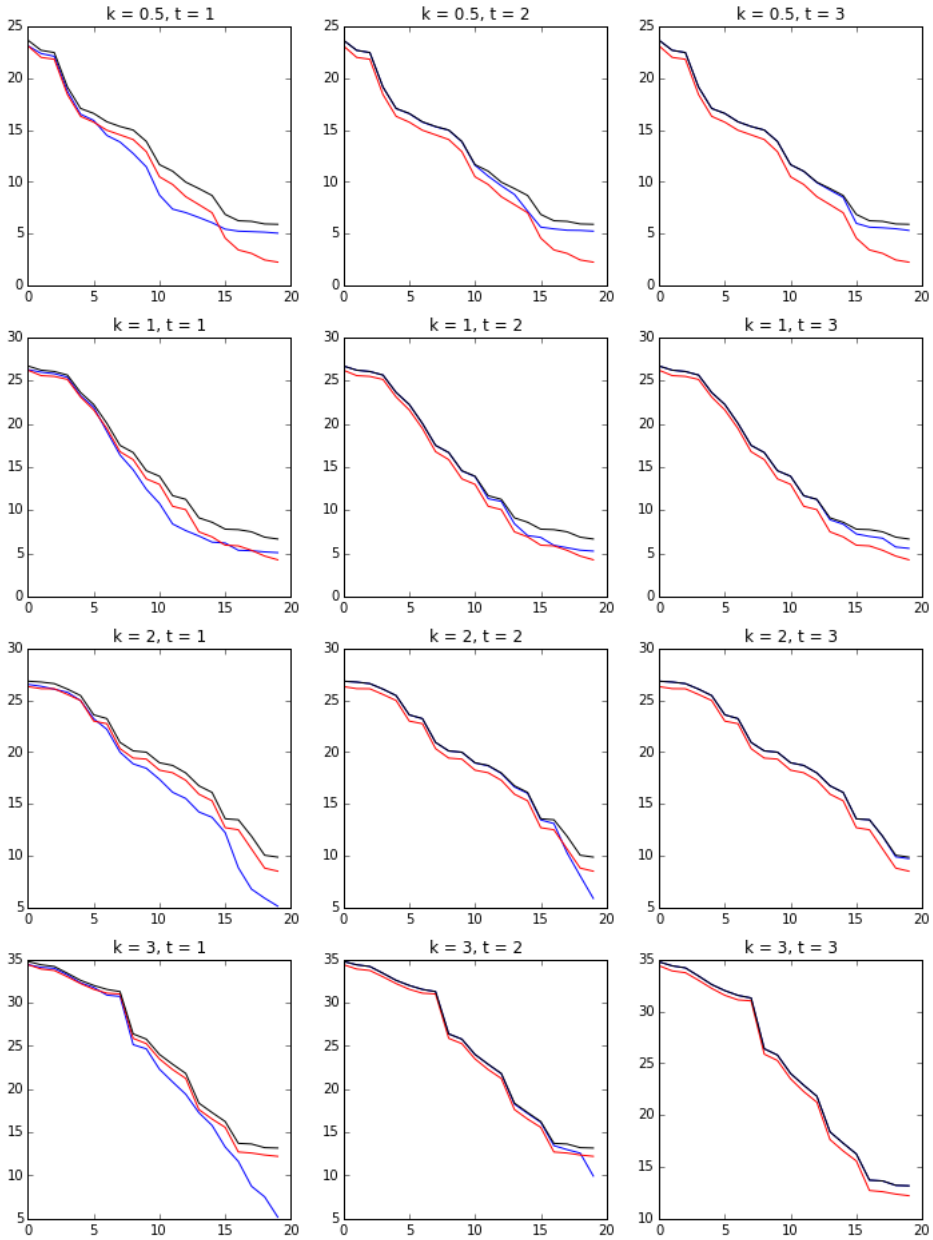

Figure 1: Singular values of the signal (red), unrandomized SVD (black), and randomized SVD(blue) for the simulated low rank data. The x axis indexes the singular values from largest to smallest up to d^*, while the y axis is the value of each singular value. The plots vary κ through 0.5, 1, 2, and 3, and t through 1, 2, and 3, with all other parameters held constant as described for the simulation.

4 Conclusions

The simulations of Principal Component Analysis corroborate the intuition that the randomization SVD algorithm is providing regularization in the estimation of the singular values, especially in the case where there is a relatively low signal to noise ratio. This result lines up well with previous observations when using the same algorithm.

However, the generalization bound guarantees which might exist for this randomized SVD process remain poorly understood. Study by [Rudi et al.(2015)] on Nystrm Subsampling suggests that there exists a relationship between the method of randomization (different sampling schemes, e.g. uniform or based on statistical leverage scores) and the regularization behavior of the algorithm. It seems fairly clear that some regularization behavior is occurring, but more theoretical study will be required in this area to gain a more precise understanding of this behavior.

In addition, the effect of this regularization of the SVD in machine learning algorithms remains unclear. While the regularization behavior within, for example, PCA seems relatively straightforward to interpret, the way in which that behavior might translate as a component of a machine learning algorithm such as Principal Components Regression is less obvious, and was not simulated in depth here.

References

[Darnell et al.(2015)] Darnell, G., Georgiev, S., Mukherjee, S., & Engelhardt, B. E 2015, arXiv:1504.03183

[Georgiev & Mukherjee(2012)] Georgiev, S., & Mukherjee, S. 2012, arXiv:1211.1642

[Rokhlin et al.(2008)] Rokhlin, V., Szlam, A., & Tygert, M. 2008, arXiv:0809.2274

[Rudi et al.(2015)] Rudi, A., Camoriano, R., & Rosasco, L. 2015, arXiv:1507.04717