Singular Value Decomposition and High-Dimensional Data

Genevera I. Allen[1] & Patrick O. Perry[2]

[1]Baylor College of Medicine & Rice University, [2]Harvard University

Keywords: singular value decomposition, dimension reduction, random matrix theory, principal components analysis

Introduction

A data set with n measurements on p variables can be represented by an $n \times p$ data matrix X. In high-dimensional settings where p is large, it is often desirable to work with a low-rank approximation to the data matrix. The most prevalent low-rank approximation is the singular value decomposition (SVD). Given X, an $n \times p$ data matrix, the SVD factorizes X as $X = UDV'$, where $U \in \mathbb{R}^{n \times n}$ and $V \in \mathbb{R}^{p \times p}$ are orthogonal matrices and $D \in \mathbb{R}^{n \times p}$ is zero except on its diagonal with diagonal entries in decreasing order. The best rank K approximation to X, $\hat{X}_K$, in both the Frobenius and operator norms is given by the first K right singular vectors and singular values of the SVD: $\hat{X}_K = \sum_{k=1}^{K} d_k u_k v_k'$. The SVD of X is also closely related to the eigendecomposition of $X'X$. Specifically, if UDV' is an SVD of X, then $V(D'D)V'$ is an eigendecomposition of $X'X$. Thus, the eigenvalues of $X'X$ are the squares of the singular values of X, and the eigenvectors of $X'X$ are the right singular vectors of X. To fully understand the implications of using the SVD in data-processing applications and classical multivariate analysis techniques such as principal components analysis (PCA), one must consider the behavior of the SVD when the elements of X are random.

Random Matrix Theory for the SVD

There are two regimes of interest for random data matrices. In the first regime, the number of samples, n, is large relative to the number of variables, p, and in the second regime the two numbers are comparable. We

call the first regime the "classical" regime and the second regime the "modern" regime. The classical regime is characterized by $n \to \infty$ and p fixed; the modern regime is characterized by $n \to \infty$, $p \to \infty$, and $n/p \to \gamma$, where γ is a fixed scalar in $(0, \infty)$.

One can study the SVD by analyzing the eigendecomposition of $X'X$. The results we summarize consider the columns of X', denoted $x_1, \ldots, x_n$, to be independent and identically distributed (IID) observations and assume that $x_1, x_2, \ldots, x_n$ are independent Normal$(0, \Sigma)$ random variables for some $p \times p$ covariance matrix Σ. Let $\Sigma = \Phi\Lambda\Phi'$ be an eigendecomposition of Σ and assume that the eigenvalues of Σ are ordered as $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. As the eigenvectors are only unique up to a sign change and arbitrary orthogonal rotations, when we state results for the eigendecomposition of $X'X$, we assume that the signs of the columns are chosen such that $\phi_j' v_j > 0$.Also note that results are stated for Gaussian random variables, but many of the results hold for arbitrary distributions with finite fourth moments.

Classical results

In the classical regime, p is fixed and n increases asymptotically. By the strong law of large numbers, $\frac{1}{n}X'X \overset{\text{a.s.}}{\to} \Sigma$, so the eigendecomposition of $\frac{1}{n}X'X$ converges to that of Σ (up to identifiability constraints). Anderson derived the distribution of $X'X$ when p is fixed and n increases asymptotically [3]. When the eigenvalues of Σ are distinct, his result implies the following: after appropriate centering and scaling, D^2 and V converge jointly in distribution and their limits are independent, and for all $1 \leq j \leq p$

$$\frac{d_j^2 - n\lambda_j}{\sqrt{n}} \overset{d}{\to} \text{Normal}\left(0,\, 2\lambda_j^2\right), \qquad \text{and} \qquad \sqrt{n}\left(\Phi' V - I\right) \overset{d}{\to} F,$$

where F is a skew-symmetric matrix with elements above the diagonal independent of each other and distributed as $F_{jk} \sim \text{Normal}\left(0,\, \frac{\lambda_j \lambda_k}{(\lambda_j - \lambda_k)^2}\right)$ for all $1 \leq j < k \leq p$. Anderson's result is more general, and also handles the case when the λ_j are not all unique.

Modern results

In the modern asymptotic regime, the dimension increases with the sample size n, so that the sequence of column dimensions, p_n, replaces the fixed column dimension p, and the sequence of covariance matrices, Σ_n, replaces the fixed covariance matrix Σ. Even though $\frac{1}{n}X'X$ converges elementwise to Σ_n, the

eigendecomposition of the former does not converge to the eigendecomposition of the latter. Even restricting attention to the top K eigenvalue-eigenvector pairs, the sample quantities are inconsistent.

The Null Case In the "null" case, when $\Sigma_n = I$ and $\frac{n}{p} \to \gamma$, the scaled eigenvalue $\frac{d_1^2}{n}$ converges almost surely to $(1 + \gamma^{-1/2})^2$ [10, 17]. Similar behavior (asymptotic upward bias) holds with a different limit for non-identity Σ_n [5, 28, 32].

Johnstone [13] and Johansson [12] and derived the limiting null distribution (after appropriate centering and scaling) of d_1^2 for real and complex data when $\Sigma_n = I$. This limiting distribution is the so-called Tracy-Widom distribution of order 1, which first appeared as the limit (after appropriate centering and scaling) of the maximum eigenvalue from a certain $n \times n$ symmetric matrix with independent Gaussian entries. [29, 30]. With center $\mu_n = \left(\sqrt{n - \frac{1}{2}} + \sqrt{p_n - \frac{1}{2}}\right)^2$ and scale $\sigma_n = \left(\sqrt{n - \frac{1}{2}} + \sqrt{p_n - \frac{1}{2}}\right)\left(1/\sqrt{n - \frac{1}{2}} + 1/\sqrt{p_n - \frac{1}{2}}\right)^{1/3}$, the quantity $\frac{d_1^2 - \mu_n}{\sigma_n}$ converges in distribution to a random variable with the Tracy-Widom law of order 1. El Karoui [9] and Ma [22] established the rate of convergence.

Patterson, Price, and Reich [23] used these results to test for latent structure in high-dimensional genetic data. Kritchman and Nadler [19, 20] did the same for chemometric and signal processing applications.

The Alternative Case The popular "alternative" case is when the eigenvalues of Σ_n are "spiked," so that the top few eigenvalues are larger than the rest. Specifically, denote the eigenvalues of Σ_n by $\lambda_{n,1}, \ldots, \lambda_{n,p}$. The spiked model is parametrized by fixed values $\lambda_1, \ldots, \lambda_K$, prescribed such that $\lambda_{n,k} = \lambda_k$ for $1 \leq k \leq K$ and $\lambda_{n,k} = 1$ otherwise. Baik, Ben Arous, and Péché [6] discovered a phase transition for complex Gaussian data, whereby the eigenvalue d_k^2 of $X'X$ behaves similarly to the null case whenever the corresponding eigenvalue $\lambda_{n,k}$ lies below the critical threshold $1 + \gamma^{-1/2}$. Baik and Silverstein [7], Paul [24], and Bai and Yao [4] extended this result. For real data, if if $\lambda_{n,k} > 1 + \gamma^{-1/2}$, and $\lambda_1, \ldots, \lambda_K$ are all distinct, then the following identities hold:

1. d_k^2/n converges almost surely to the value $\mu(\lambda_k) = \lambda_k(1 + \frac{1}{\gamma(\lambda_k - 1)})$;
2. $\left(\frac{d_1^2 - n\mu(\lambda_1)}{\sqrt{n}}, \ldots, \frac{d_k^2 - n\mu(\lambda_k)}{\sqrt{n}}\right)$ converges in distribution to a mean-zero multivariate normal random variable with diagonal covariance;
3. $\phi'_{n,k} v_k$ converges almost surely to $\rho(\lambda_k) = \sqrt{\left(1 - \frac{1}{\gamma(\lambda_k - 1)^2}\right) / \left(1 + \frac{1}{\gamma(\lambda_k - 1)}\right)}$;
4. $\left(\sqrt{n}(\phi'_{n,1} v_1 - \rho(\lambda_1)), \ldots, \sqrt{n}(\phi'_{n,k} v_k - \rho(\lambda_k))\right)$ converges in distribution to a mean-zero multivariate normal random variable.

On the other hand, if $\lambda_k \leq 1 + \gamma^{-1/2}$, then

1. d_k^2/n converges almost surely to $(1 + \gamma^{-1/2})^2$;
2. $\phi'_{n,k} v_k$ converges almost surely to 0.

In the case when $\lambda_1, \ldots, \lambda_K$ are not all distinct, similar behavior manifests, but the limiting distributions are no longer Gaussian.

Harding [11] used the phase transition behavior to explain the apparent empirical lack of latent structure in arbitrage pricing data. Perry and Wolfe [26] use these results to estimate the number of latent factors in signal processing applications.

Sparse Principal Components Analysis

Understanding the properties of the SVD in high-dimensional settings, is also important for evaluating the behavior of principal components analysis (PCA), a classical technique for dimension reduction, exploratory analysis, and data visualization. PCA seeks a linear projection of the data that maximizes the sample variance: $\text{Var}(Xv)/||v||_2$. Subsequent PC directions are constrained to be orthogonal to previous directions. It is well known that these PC directions are given by the right singular vectors of X or the eigenvectors of $X'X$. The question remains whether these PC directions can be consistently estimated when the dimension, p, is greater than the number of observations, n.

PCA is Inconsistent in High-Dimensions

One can study the behavior of the PC directions by considering the spiked covariance model introduced in the previous section. Under the conditions previously stated, $\phi'_{n,k} v_k$ converges almost surely to $\rho(\lambda_k) = \sqrt{\left(1 - \frac{1}{\gamma(\lambda_k - 1)^2}\right) / \left(1 + \frac{1}{\gamma(\lambda_k - 1)}\right)}$ when $\lambda_{n,k} > 1 + \gamma^{-1/2}$ and converges almost surely to zero otherwise. These results imply that the PC direction vectors, v_k, are consistent (that is $\phi'_{n,k} v_k \to 1$) if and only if $\frac{p}{n} \to 0$. Thus, in high-dimensional settings when $p >> n$, classical PCA is inconsistent. Johnstone and Lu [14] first proved this result for the first PC direction vector and Paul [24] expanded this later to the multi-component PCA model.

Jung and Marron [18] prove similar results by considering a model where the sample size, n, is fixed, the dimension, $p \to \infty$, and the eigenvectors, $\lambda_{p,k}$ grow with p such that $\lambda_{p,k} = p^{\alpha_k}$. They show that under certain conditions on the eigenvalues of the spiked covariance model:

1. (Consistency) If $\alpha_k > 1$, $\angle(\phi_{p,k}, v_k) \xrightarrow{P} 0$.
2. (Strong Inconsistency) If $\alpha_k < 1$, $\angle(\phi_{p,k}, v_k) \xrightarrow{P} \frac{\pi}{2}$.

Here, $\angle(\phi_{p,k}, v_k)$ denotes the angle between the true eigenvector, $\phi_{p,k}$ and the estimated PC direction v_k. In other words, this result states that if the magnitude of the eigenvalues associated with the PC directions of interest do not grow with the dimension, the estimated PC directions are orthogonal to the true eigenvectors and are essentially random.

Sparse PCA

Given that the PC direction vectors are inconsistent in high-dimensional settings, many have proposed to find PC directions using only a subset of the variables, a method termed sparse PCA. This method seeks linear projections that maximize the sample variance such that these projection vectors have a limited number of non-zero elements. In other words, one seeks a direction vector v that maximizes $\text{Var}(Xv)/v'v$ subject to $||v||_0 \leq t$, where $||\cdot||_0$ is the ℓ_0-norm, summing the number of non-zero elements. Jolliffe, Trendafilov and Uddin [16] first proposed to estimate sparse PC directions by relaxing the ℓ_0-norm to an ℓ_1-norm, placing this penalty on the PC directions to encourage sparsity. Using similar penalization approaches, many have proposed alternative formulations to achieve sparse PCA by employing the elastic net [33], semi-definite relaxations [8], and regression-based extensions of the power method [27]. Others have proposed to select variables in the PC directions in a two step approach: first, finding a good subset of variables via thresholding, and then applying classical PCA to these selected variables [15].

Several sparse PCA methods have been shown to be consistent in high-dimensional settings where classical PCA is inconsistent. Johnstone and Lu [15] propose a simple filtering method, thresholding variables based upon the dimensions and noise level of the latent variable model. They show that under certain conditions, the angle between the first sparse PC direction obtained in this manner and the true factor converges almost surely to zero. Paul and Johnstone [25] show convergence in the ℓ_2-norm between PC directions estimated by augmented sparse PCA, an extension of the simple thresholding method, and the population eigenvectors of

a spiked covariance model. Amini and Wainwright [2] also consider a spiked covariance model and show that under certain conditions depending on the data dimensions and number of true latent factors, the two step thresholding and semi-definite programming [8] sparse PCA methods have consistent support for estimating the true non-zero variables in the first eigenvector.

More recently, several have proposed to encourage sparsity in both the PC directions as well as the sample principal components, forming a penalized SVD or sparse matrix factorization [1, 21, 31] of the following form: $\hat{X} = \sum_{k=1}^{K} d_k u_k v_k^T$ where $||u_k||_0 \leq t_k$ and $||v_k||_0 \leq s_k$. These penalized SVDs have been used to reduce the dimensions of high-dimensional two-way data.

References

[1] G. I. Allen, L. Grosenick, and J. Taylor. A generalized least squares matrix decomposition. Rice University Technical Report No. TR2011-03, 2011.

[2] A.A. Amini and M.J. Wainwright. High-dimensional analysis of semidefinite relaxations for sparse principal components. *The Annals of Statistics*, 37(5B):2877–2921, 2009.

[3] T. W. Anderson. Asymptotic theory for principal component analysis. *Ann. Math. Statist.*, 34:122–148, 1963.

[4] Z. Bai and J.-f. Yao. Central limit theorems for eigenvalues in a spiked population model. *Annales de l'Institut Henri Poincaré - Probabilités et Statistiques*, 44:447–474, 2008.

[5] Z.D. Bai, J.W. Silverstein, and Y.Q. Yin. A note on the largest eigenvalue of a large dimensional sample covariance matrix. *J. Multivariate Anal.*, 26(2):166–168, 1988.

[6] J. Baik, G. Ben Arous, and S. Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.*, 33(5):1643–1697, 2005.

[7] J. Baik and J.W. Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *J. Multivariate Anal.*, 97(6):1382–1408, 2006.

[8] A. D'Aspremont, L. El Ghaoui, M.I. Jordan, and G.R.G. Lanckriet. A direct formulation for sparse pca using semidefinite programming. *SIAM review*, 49(3):434, 2007.

[9] N. El Karoui. A rate of convergence result for the largest eigenvalue of complex white Wishart matrices. *Ann. Probab.*, 34(6):2077–2117, 2006.

[10] S. Geman. A limit theorem for the norm of random matrices. *Ann. Probab.*, 8(2):252–261, 1980.

[11] M. C. Harding. Explaining the single factor bias of arbitrage pricing models in finite samples. *Economics Letters*, 99(1):85–88, 2008.

[12] K. Johansson. Shape fluctuations and random matrices. *Comm. Math. Phys.*, 209(2):437–476, 2000.

[13] I.M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.*, 29(2):295–327, 2001.

[14] I.M. Johnstone and A.Y. Lu. Sparse Principal Components Analysis. 2004. Unpublished Manuscript.

[15] I.M. Johnstone and A.Y. Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 2009.

[16] I.T. Jolliffe, N.T. Trendafilov, and M. Uddin. A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics*, 12(3):531–547, 2003.

[17] D. Jonsson. On the largest eigenvalue of a sample covariance matrix. In P.R. Krishnaiah, editor, *Multivariate Analysis VI*. North-Holland, 1983.

[18] S. Jung and JS Marron. PCA consistency in high dimension, low sample size context. *The Annals of Statistics*, 37(6B):4104–4130, 2009.

[19] S. Kritchman and B. Nadler. Determining the number of components in a factor model from limited noisy data. *Chemometrics and Intelligent Laboratory Systems*, 94(1):19–32, 2008.

[20] S. Kritchman and B. Nadler. Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory. *Signal Processing, IEEE Transactions on*, 57(10):3930–3941, 2009.

[21] M. Lee, H. Shen, J.Z. Huang, and JS Marron. Biclustering via sparse singular value decomposition. *Biometrics*, 2010.

[22] Z. Ma. Accuracy of the Tracy-Widom limit for the largest eigenvalue in white Wishart matrices. arXiv:0810.1329v1 [math.ST], 2008.

[23] N. Patterson, A. L. Price, and D. Reich. Population structure and eigenanalysis. *PLoS genetics*, 2(12):e190, 2006.

[24] D. Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Stat. Sinica*, 17(4):1617–1642, 2007.

[25] D. Paul and I. Johnstone. Augmented sparse principal component analysis for highdimensional data. Technical report, U.C. Davis, 2007.

[26] P. O. Perry and P. J. Wolfe. Minimax rank estimation for subspace tracking. *Selected Topics in Signal Processing, IEEE Journal of*, 4(3):504–513, 2010.

[27] H. Shen and J.Z. Huang. Sparse principal component analysis via regularized low rank matrix approximation. *Journal of multivariate analysis*, 99(6):1015–1034, 2008.

[28] J.W. Silverstein. On the largest eigenvalue of a large dimensional sample covariance matrix. Unpublished manuscript, 1984.

[29] C.A. Tracy and H. Widom. Level-spacing distributions and the Airy kernel. *Comm. Math. Phys.*, 159(1):151–174, 1994.

[30] C.A. Tracy and H. Widom. On orthogonal and symplectic matrix ensembles. *Comm. Math. Phys.*, 177(3):727–754, 1996.

[31] D. M. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009.

[32] Y.Q. Yin, Z.D. Bai, and P.R. Krishnaiah. On the limit of the largest eigenvalue of the large dimensional sample covariance matrix. *Probab. Theor. Relat. Field.*, 78(4):509–521, 1988.

[33] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006.