Title

Principal Component Analysis for Big Data

Authors

Fan, Jianqing
Sun, Qiang
Zhou, Wen-Xin
et al.

Publication Date

2022-12-21

DOI

10.1002/9781118445112.stat08122

Peer reviewed

Wiley StatsRef:
Statistics Reference Online

Principal Component Analysis for Big Data

By *Jianqing Fan*[1], *Qiang Sun*[2], *Wen-Xin Zhou*[3], *and Ziwei Zhu*[1]

Keywords: *big data, covariance matrix, dimension reduction, factor analysis, high dimensionality, machine learning, principal components, robust PCA*

Abstract: Big data is transforming our world, revolutionizing operations and analytics everywhere, from financial engineering to biomedical sciences. The complexity of big data often makes dimension reduction techniques necessary before conducting statistical inference. Principal component analysis, commonly referred to as PCA, has become an essential tool for multivariate data analysis and unsupervised dimension reduction, the goal of which is to find a lower dimensional subspace that captures most of the variation in the dataset. This article provides an overview of methodological and theoretical developments of PCA over the past decade, with focus on its applications to big data analytics. We first review the mathematical formulation of PCA and its theoretical development from the view point of perturbation analysis. We then briefly discuss the relationship between PCA and factor analysis as well as its applications to large covariance estimation and multiple testing. PCA also finds important applications in many modern machine learning problems, and in this article, we focus on community detection, ranking, mixture model, and manifold learning.

1 Introduction

Principal component analysis (PCA), first introduced by Karl Pearson[1], is one of the most commonly used techniques for dimension reduction in many disciplines, such as neurosciences, genomics, and finance[2]. We refer the readers to Jolliffe[3] for a recent review. It extracts latent principal factors that preserve most of the variation in the dataset. Let $\boldsymbol{x}$ be a random vector taking values in $\mathbb{R}^d$ with mean zero and covariance matrix $\boldsymbol{\Sigma}$. With this formalism, PCA seeks projection direction vectors, $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_K \in \mathbb{R}^d$, such that

$$\boldsymbol{\nu}_1 \in \operatorname{argmax}_{\|\boldsymbol{\nu}\|_2=1} \boldsymbol{\nu}^\top \boldsymbol{\Sigma} \boldsymbol{\nu}, \quad \boldsymbol{\nu}_2 \in \operatorname{argmax}_{\|\boldsymbol{\nu}\|_2=1, \boldsymbol{\nu} \perp \boldsymbol{\nu}_1} \boldsymbol{\nu}^\top \boldsymbol{\Sigma} \boldsymbol{\nu}, \quad \boldsymbol{\nu}_3 \in \operatorname{argmax}_{\|\boldsymbol{\nu}\|_2=1, \boldsymbol{\nu} \perp \boldsymbol{\nu}_1, \boldsymbol{\nu}_2} \boldsymbol{\nu}^\top \boldsymbol{\Sigma} \boldsymbol{\nu}, \ \ldots$$

In other words, $\{\boldsymbol{\nu}_j\}_{j=1}^K$ are the top K eigenvectors of $\boldsymbol{\Sigma}$. Given $\mathbf{V}_K \equiv (\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_K)$, we can then project the original high-dimensional data onto the low-dimensional space spanned by columns of $\mathbf{V}_K$ to achieve

[1]Princeton University, Princeton, NJ, USA
[2]University of Toronto, Toronto, ON, Canada
[3]University of California, San Diego, CA, USA

DOI: 10.1002/9781118445112.stat08122

the goal of dimensionality reduction. As $\mathbf{V}_K$ captures the most variation in the dataset, these projected data points approximately preserve the geometric properties of the original data, which are amenable to downstream statistical analysis. In real applications, the true covariance matrix $\mathbf{\Sigma}$ is typically unknown; we need to substitute $\mathbf{\Sigma}$ with its empirical version $\hat{\mathbf{\Sigma}}$.

The high complexity of big data, such as massiveness, contamination, nonlinearity, and decentralization, has posed fundamental challenges to statistical inference. PCA, a 100-year-old idea, has been and is still shining as a powerful tool for modern data analytics. Modern developments for PCA focus on attempts that address various challenges created by big data. For example, the massiveness of features in big data has been shown to create many notorious problems, making many conventional inferential procedures ill-posed. Recent study[4] shows that PCA is closely connected to factor analysis. This motivates new methodological developments in multiple testing problems when tens of thousands of possibly dependent statistical tests are evaluated simultaneously[5,6], which partially solve the high-dimensional inference problem. Moreover, big data are often contaminated by outliers or heavy-tailed errors[7,8], motivating the use of robust covariance inputs in the PCA formulation[9,10]. This results in a form of robust PCA. Moreover, machine learning algorithms, such as those in clustering, community detection, ranking, matrix completion, and mixture models, often involve solving a highly nonconvex system. This makes developing practical and efficient computational algorithms a grand challenge. PCA, or spectral method more generally, can often be used to solve a reduction of the highly nonconvex system, without losing much statistical efficiency[11,12]. Manifold sometimes can be used to approximate the nonlinearity structure of a dataset. Surprisingly, PCA finds applications in this setting and achieves a form of nonlinear dimension reduction[13,14].

2 Covariance Matrix Estimation and PCA

We now begin the journey of PCA for big data. Given a small number K, the main goal of PCA is to estimate the K-dimensional principal eigenspace of $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$ such that it captures most of the variation in the dataset. In statistical applications, the true covariance matrix $\mathbf{\Sigma}$ is typically unknown and in practice, PCA is conducted on some estimator $\hat{\mathbf{\Sigma}} = \hat{\mathbf{\Sigma}}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ of $\mathbf{\Sigma}$, where $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are observations from $\boldsymbol{X}$. A significant error in recovering the eigenvalues and eigenvectors of $\mathbf{\Sigma}$ using those of $\hat{\mathbf{\Sigma}}$ would lead to a substantial loss of information contained in the data by PCA projection. As direct applications of matrix perturbation theory (see Section 2.1), bounds on the estimation error $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|$ under some matrix norm $\|\cdot\|$ can be effectively applied to relate the eigenvalues/eigenvectors of $\mathbf{\Sigma}$ and of $\hat{\mathbf{\Sigma}}$ under the spectral gap condition. Therefore, it is important to build a good covariance estimator, say $\hat{\mathbf{\Sigma}}$, with small statistical error in the sense that for any given $\delta > 0$, we are interested in minimizing the value r that satisfies $\mathbb{P}(\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\| > r) \leq \delta$ for some matrix norm $\|\cdot\|$.

2.1 Perturbation Theory

Considering $\mathbf{\Delta} = \hat{\mathbf{\Sigma}} - \mathbf{\Sigma}$ as a perturbation, it is crucial to understand how the eigenspace of $\hat{\mathbf{\Sigma}} = \mathbf{\Sigma} + \mathbf{\Delta}$ perturbs around that of $\mathbf{\Sigma}$. This problem has been widely studied in the literature[15–19]. Among these results, the $\sin\Theta$ theorems, established by Davis and Kahan[15] and Wedin[16], have become fundamental tools and are commonly used in numerical analysis, statistics, and machine learning. While Davis and Kahan[15] focused on eigenvectors of symmetric matrices, Wedin's $\sin\Theta$ theorem studies the singular vectors for asymmetric matrices and provides a uniform perturbation bound for both the left and right singular spaces in terms of the singular value gap and perturbation level. Over the years, various extensions have been made in different settings. For example, Vu[20], Shabalin and Nobel[21], O'Rourke *et al.*[22], and Wang[23]

Wiley StatsRef: Statistics Reference Online, © 2014–2018 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat08122

considered the rotations of singular vectors after random perturbations; Cai and Zhang[24] established separate perturbation bounds for the left and right singular subspaces (see also Dopico[25] and Stewart[18]). Recently, Fan *et al.*[26] derived new perturbation bounds, measured in the ℓ_∞-norm, for singular vectors (or eigenvectors in the symmetric case), which can be applied to robust covariance estimation in high-dimensional factor models and robust estimation of the false discovery proportion[6] when the sampling distributions are heavy tailed. See Sections 3.2 and 3.3 for details on these statistical applications.

2.2 Robust Covariance Inputs and Robust PCA

Owing to matrix perturbation theory, upper bounds on the spectral norm $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2$ or the elementwise ℓ_∞-norm (also known as the max norm) $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\max}$ can be used to establish corresponding bounds on the ℓ_2 distance and ℓ_∞ distance between the population eigenvectors and their empirical counterparts obtained from $\hat{\mathbf{\Sigma}}$, respectively. Given independent observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from $\boldsymbol{X}$ with $\mathbb{E}\boldsymbol{X} = \mathbf{0}$, the sample covariance matrix, namely $\hat{\mathbf{\Sigma}}_{\text{sam}} := n^{-1}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^\top$, is arguably the most natural choice to estimate $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$ when the dimension d is smaller than the sample size n. The finite sample bound on $\|\hat{\mathbf{\Sigma}}_{\text{sam}} - \mathbf{\Sigma}\|_2$ or $\|\hat{\mathbf{\Sigma}}_{\text{sam}} - \mathbf{\Sigma}\|_{\max}$ has been well studied in the literature[27,28]. If $\boldsymbol{X}$ has a sub-Gaussian distribution in the sense that for all unit vectors $\boldsymbol{v} \in \mathbb{S}^{d-1}$, $\mathbb{E}\exp(\lambda\boldsymbol{v}^\top\boldsymbol{X}) \le \exp(c\lambda^2\boldsymbol{v}^\top\mathbf{\Sigma}\boldsymbol{v})$ for some constant c, then Remark 5.40 in Vershynin[27] implies that for every $t \ge 0$, $\|\hat{\mathbf{\Sigma}}_{\text{sam}} - \mathbf{\Sigma}\|_2 \le \max(\delta, \delta^2)\|\mathbf{\Sigma}\|_2$ with probability at least $1 - 2\mathrm{e}^{-t}$, where $\delta = C_1\sqrt{(d+t)/n}$. Similarly, it can be shown that with probability greater than $1 - 2\mathrm{e}^{-t}$, $\|\hat{\mathbf{\Sigma}}_{\text{sam}} - \mathbf{\Sigma}\|_{\max} \le \max(\eta, \eta^2)$, where $\eta = C_2\sqrt{(\log d + t)/n}$. Here, $C_1, C_2 > 0$ are constants depending only on c.

However, when the distribution is heavy tailed, one cannot expect sub-Gaussian behaviors of the sample covariance in either the spectral or max norm[29]. Therefore, to perform PCA for heavy-tailed data, the sample covariance is a risky choice to begin with. Indeed, alternative robust estimators have been constructed to achieve better finite sample performance. In Ref. 7, the authors constructed an elementwise robustified version of the sample covariance matrix using the adaptive Huber loss minimization[8] and derived a sub-Gaussian-type deviation inequality in the max norm under finite fourth moment condition instead of sub-Gaussianity. On the basis of a novel shrinkage principle, Fan *et al.*[9] and Minsker[10] independently constructed global robustified variants of the sample covariance with sub-Gaussian behavior under the spectral norm as long as the fourth moment of $\boldsymbol{X}$ is finite. A different robust method using the idea of median-of-means was proposed and studied by Minsker[30]. More recently, Giulini[31] studied robust PCA in a more general setting where the data sample is made of independent copies of some random variable ranging in a separable real Hilbert space. Together, these results provide a new perspective on robustness from a nonasymptotic standpoint, and also represent a useful complement to the previous results on robust PCA. For instance, Candés *et al.*[32] focused on a different notion of robustness and showed that it is possible to recover the principal components of a data matrix when the observations are contained in a low-dimensional space but arbitrarily corrupted by additive noise (see also Chandrasekaran *et al.*[33], Zhang and Lerman[34], and the references therein).

3 PCA and Factor Analysis

PCA and factor analysis are two important problems in their respective fields and are seemingly unrelated at first sight. Lately, it is shown in Ref. 4 that the high-dimensional factor model is innately related to PCA, which makes it different from the classical factor model with fixed dimensionality[35] and helps one understand why PCA can be used for the factor analysis when the top eigenvalues are spikes. In addition,

DOI: 10.1002/9781118445112.stat08122

this observation has triggered a series of interesting studies on matrix perturbation theory and robust covariance estimation.

3.1 Factor Model and PCA

Let $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$ be a random vector of outcomes of interest, which may represent financial returns, housing prices, or gene expressions. The impact of dependence among outcomes is currently among the most discussed topics in various statistical problems, such as variable selection, covariance and precision matrix estimation, and multiple testing. For example, financial returns depend on the equity market risks, housing prices depend on the economic health, and gene expressions can be stimulated by cytokines, among others. Because of the presence of common factors, it is unrealistic to assume that many outcomes are uncorrelated. It is thus natural to assume a factor model structure, which relies on the identification of a linear space of random vectors capturing the dependence among the data. To do so, we consider an approximate factor model, which has been frequently used in economic and financial studies[36−38] as well as genomics[39]:

$$X_{ij} = \mu_j + \boldsymbol{b}_j^\top \boldsymbol{f}_i + \varepsilon_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, d,$$
$$\text{or in a matrix form } \boldsymbol{X}_i = \boldsymbol{\mu} + \mathbf{B}\boldsymbol{f}_i + \varepsilon_i \quad \text{with} \quad \mathbf{B} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_d)^\top \tag{1}$$

Here, X_{ij} is the response for the jth feature of the ith observation $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{id})^\top$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^\top$ is the intercept, $\boldsymbol{b}_j$ is a vector of factor loadings, $\boldsymbol{f}_i$ is a K-dimensional vector of common factors, and $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{id})^\top$ is the error term, which is typically called the idiosyncratic component, uncorrelated with or independent of $\boldsymbol{f}_i$. We emphasize that, in model (1), only X_{ij}s are observable. Intuitively, the unobserved common factors can only be inferred reliably when $d \to \infty$. Under model (1), $\boldsymbol{\Sigma} = \mathrm{cov}(\boldsymbol{X})$ is given by

$$\boldsymbol{\Sigma} = \mathbf{B}\mathrm{cov}(\boldsymbol{f}_i)\mathbf{B}^\top + \boldsymbol{\Sigma}_\varepsilon, \quad \boldsymbol{\Sigma}_\varepsilon = (\sigma_{\varepsilon,jk})_{1\le j,k\le d} = \mathrm{cov}(\varepsilon_i) \tag{2}$$

The literature on approximate factor models typically assumes that the first K eigenvalues of $\mathbf{B}\mathrm{cov}(\boldsymbol{f}_i)\mathbf{B}^\top$ diverge at rate $O(d)$, whereas all the eigenvalues of $\boldsymbol{\Sigma}_\varepsilon$ are bounded as $d \to \infty$. This assumption holds naturally when the factors are pervasive in the sense that a nonnegligible fraction of factor loadings should be nonvanishing. The decomposition (Equation 2) is then asymptotically identifiable as $d \to \infty$.

Now we elucidate why PCA can be used for the factor analysis in the presence of spiked eigenvalues. Note that the linear space spanned by the first K principal components of $\mathbf{B}\mathrm{cov}(\boldsymbol{f}_i)\mathbf{B}^\top$ coincides with that spanned by the columns of $\mathbf{B}$ when $\mathrm{cov}(\boldsymbol{f}_i)$ is nondegenerate. Therefore, we can assume without loss of generality that the columns of $\mathbf{B}$ are orthogonal, and $\mathrm{cov}(\boldsymbol{f}_i)$ is the identity matrix[40,41]. Let $\overline{\boldsymbol{b}}_1, \ldots, \overline{\boldsymbol{b}}_K$ be the columns of $\mathbf{B}$, ordered such that $\|\overline{\boldsymbol{b}}_1\|_2 \ge \cdots \ge \|\overline{\boldsymbol{b}}_K\|_2$. Then, $\{\overline{\boldsymbol{\xi}}_j := \overline{\boldsymbol{b}}_j/\|\overline{\boldsymbol{b}}_j\|_2\}_{j=1}^K$ are the top K eigenvectors of $\mathbf{B}\mathbf{B}^\top$ with eigenvalues $\{\overline{\lambda}_j := \|\overline{\boldsymbol{b}}_j\|_2\}_{j=1}^K$. Let $\{\lambda_j\}_{j=1}^d$ be the eigenvalues of $\boldsymbol{\Sigma}$ in a nonincreasing order with the corresponding eigenvectors being $\{\boldsymbol{\xi}_j\}_{j=1}^p$. From the *pervasiveness* assumption that the eigenvalues of $d^{-1}\mathbf{B}^\top\mathbf{B}$ are distinct and bounded away from zero and infinity, it follows that $\{\overline{\lambda}_j\}_{j=1}^K$ grow at rate $O(d)$. In addition, by Weyl's theorem and the $\sin\Theta$ theorem of Davis and Kahan[15],

$$\max_{1\le j\le K} |\lambda_j - \overline{\lambda}_j| \bigvee \max_{j>K} |\lambda_j| \le \|\boldsymbol{\Sigma}_\varepsilon\|_2 \quad \text{and} \quad \max_{1\le j\le K} \|\boldsymbol{\xi}_j - \overline{\boldsymbol{\xi}}_j\|_2 = O(d^{-1}\|\boldsymbol{\Sigma}_\varepsilon\|_2)$$

These results state that PCA and factor analysis are approximately the same if $\|\boldsymbol{\Sigma}_\varepsilon\|_2 = o(d)$. This is assured through a sparsity condition on $\boldsymbol{\Sigma}_\varepsilon$, which is usually measured through $m_d = \max_{1\le j\le d} \sum_{k=1}^d |\sigma_{\varepsilon,jk}|^q$ for some $q \in [0, 1]$. The intuition is that, after subtracting out the common factors, many pairs of the cross-sectional units become weakly correlated. This generalized notion of sparsity was used by Bickel and

Wiley StatsRef: Statistics Reference Online, © 2014–2018 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat08122

Levina[42], under which it holds that $\|\mathbf{\Sigma}_\varepsilon\|_2 \leq \max_{1\leq j\leq d}\sum_{k=1}^{d}|\sigma_{\varepsilon,jk}|^q(\sigma_{\varepsilon,jj}\sigma_{\varepsilon,kk})^{(1-q)/2} = O(m_d)$ if the variances $\sigma_{\varepsilon,jj}$s are uniformly bounded. Therefore, in the approximate sparse setting where $m_d = o(d)$, the pervasiveness assumption implies that the principal components $\{\lambda_j\}_{j=1}^{K}$ and the remaining components $\{\lambda_j\}_{j=K+1}^{d}$ are well separated, and the first K principal components are approximately the same as the standardized columns of the factor loading matrix. In this setting, PCA serves as a valid approximation to factor analysis only if $d \to \infty$.

3.2 Application to Large Covariance Estimation

Covariance structure plays a particularly important role in high-dimensional data analysis. Apart from PCA, a large collection of fundamental statistical methods, such as linear and quadratic discriminant analysis, clustering analysis, and regression analysis, require the knowledge of the covariance structure or certain aspects thereof. Realizing the importance of estimating large covariance matrices and the challenges that are brought by the high dimensionality, since the seminal work of Bickel and Levina[42] and Fan *et al.*[43], various regularization techniques have been proposed to estimate $\mathbf{\Sigma}$ consistently. See Cai *et al.*[44] and Fan *et al.*[45] for two comprehensive surveys on this topic. One commonly assumed low-dimensional structure is that the covariance matrix is sparse, namely many entries are zero or nearly so. In many applications, however, the sparsity assumption directly on $\mathbf{\Sigma}$ is not appropriate. A natural extension is conditional sparsity[43]. Given the common factors, the outcomes are weakly correlated. In other words, in decomposition (Equation 2) we assume that $\mathbf{\Sigma}_\varepsilon$ is approximately sparse as in Bickel and Levina[42]. Motivated by the innate connection between PCA and factor analysis in high dimensions, Fan *et al.*[4] proposed a principal orthogonal complement thresholding method (POET) to estimate $\mathbf{\Sigma}$ in model (1): (i) run the singular value decomposition on the sample covariance matrix of $\boldsymbol{X}$; (ii) keep the covariance matrix that is formed by the first K principal components; and (iii) apply the adaptive thresholding procedure[46,47] to the remaining components of the sample covariance matrix.

As discussed in Section 2.2, in the presence of heavy-tailed data, the use of the sample covariance matrix is controversial and therefore is in question to begin with. Tailored to elliptical factor models, Fan *et al.*[48] proposed to use the marginal Kendall's tau to estimate $\mathbf{\Sigma}$ and its top K eigenvalues, and separately, use the spatial Kendall's tau to construct estimators for the corresponding leading eigenvectors. In more general settings where no shape constraints or parametric assumptions are imposed (normality, symmetry, elliptically symmetry, etc.), robust alternatives, such as the U-type covariance estimator and adaptive Huber covariance estimator considered in Ref. 6, are preferred to be taken as the initial estimators in the POET procedure.

3.3 Application to Factor-Adjusted Multiple Testing

An iconic example of model (1) is the factor pricing model in financial economics, where X_{ij} is the excess return of fund/asset j at time i and $\boldsymbol{f}_i$s are the systematic risk factors related to some specific linear pricing model, such as the capital asset pricing model[49], and the Fama-French three-factor model[50]. Although the key implication from the multifactor pricing theory is that the intercept μ_j should be zero, known as the "mean-variance efficiency" pricing, for any asset j, an important question is whether such a pricing theory can be validated by empirical data. In fact, a very small proportion of μ_js might be nonzero according to the Berk and Green equilibrium[51,52]. It is practically important to identify those positive μ_js, which amounts to conducting the following d hypothesis testing problems simultaneously:

$$H_{0j} : \mu_j = 0 \text{ versus } H_{1j} : \mu_j \neq 0, \quad j = 1, \ldots, d \tag{3}$$

DOI: 10.1002/9781118445112.stat08122

In the presence of common factors, $X_1, \ldots, X_p$ are highly correlated, and therefore directly applying classical false discovery rate (FDR) controlling procedures[53,54] can lead to inaccurate false discovery control and spurious outcomes. To improve the efficiency and to incorporate the strong dependence information, various factor-adjusted procedures have been proposed[5,55,56]. These works assume that both the factor and idiosyncratic noise follow multivariate normal distributions. However, the Gaussian assumption on the sampling distribution is often unrealistic in many applications, especially in genomics and finance. Also, as noted in Ref. 57, only non-Gaussian latent variables are detectable. To address the two challenges that are brought by strong dependence and heavy tailedness, Fan *et al.*[6] proposed a factor-adjusted robust multiple testing (FARM Test) procedure with control of the false discovery proportion. Starting with a robust initial estimate of the covariance matrix, the FARM-Test method uses its first several principal components to estimate the factor loading matrix; next, using these estimated loading vectors, runs adaptive Huber regression to estimate the realized factors; robust test statistics are then constructed by subtracting out the realized common factors from the robust mean estimators; and finally, the data-driven critical value is computed so that the estimated false discovery proportion is controlled at the prespecified level. The first step is motivated by the approximate equivalence between PCA and factor analysis as discussed in Section 3.2.

4 Applications to Statistical Machine Learning

PCA finds applications in many statistical machine learning problems. In this section, we focus on its applications to four major problems in the machine learning literature: clustering and community detection, ranking, mixture model, and manifold learning.

4.1 Clustering and Community Detection

Clustering and community detection is an important task in data analysis[58]. We focus on the stochastic block model (SBM), which is widely regarded as a canonical model for this purpose. We refer the readers to Ref. 12 for a review of recent developments on SBM. We start with the definition of SBM. Let n be the number of vertices, k the number of communities, $\mathbf{p} = (p_1, \ldots, p_k)^\top$ the probability vector, and $\mathbf{W}$ the $k \times k$ symmetric transition probability matrix with entries in $[0, 1]$. We say a pair $(\boldsymbol{X}, G)$ is sampled from SBM$(n, \mathbf{p}, \mathbf{W})$ if $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ is an n-dimensional random vector with independent and identically distributed components distributed under $\mathbf{p}$ and G is an n-vertex simple graph where vertices i and j are connected with probability W_{X_i,X_j}, independently of other pairs of vertices. The ith community set is defined by $\Omega_i := \{v : X_v = i, v = 1, \ldots, n\}$. Roughly speaking, an SBM is a random graph with planted clusters: the cluster sizes follow a multinomial distribution with probability vector $\boldsymbol{p}$ and the probability that a member in the i cluster and a member in the jth group get connected is $W_{i,j}$.

To fix idea, we consider the SBM with two communities, where the inner-cluster probability is a and the across-cluster probability is b with $b < a$. Further assume for simplicity that the two clusters have exactly size $n/2$ and index the first cluster with the first $n/2$ vertices. Given an observed graph with the adjacency matrix $\boldsymbol{A}$ (indicating whether or not two vertices are connected), the expected value $\mathbb{E}\mathbf{A}$ of this graph has four blocks given by

$$\mathbb{E}\mathbf{A} = \begin{bmatrix} a \cdot \mathbf{I}_{n/2} & b \cdot \mathbf{I}_{n/2} \\ b \cdot \mathbf{I}_{n/2} & a \cdot \mathbf{I}_{n/2} \end{bmatrix}$$

Wiley StatsRef: Statistics Reference Online, © 2014–2018 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat08122

It can be verified that the above matrix has three eigenvalues $a + b$, $a - b$, and 0, where 0 has multiplicity $n - 2$, and the eigenvectors associated with the two largest eigenvectors are

$$\begin{bmatrix} \mathbf{1}_{n/2} \\ \mathbf{1}_{n/2} \end{bmatrix} \text{ and } \begin{bmatrix} \mathbf{1}_{n/2} \\ -\mathbf{1}_{n/2} \end{bmatrix}$$

respectively. In other words, if one were to work with the expected adjacency matrix, communities could be simply recovered by taking the eigenvector associated with the second largest eigenvalue. In practice, we only observe a single shot of the SBM graph $\mathbf{A}$, which can be viewed as a perturbation of the expected adjacency matrix:

$$\mathbf{A} = \mathbb{E}\mathbf{A} + \mathbf{Z}$$

where $\mathbf{Z} = \mathbf{A} - \mathbb{E}\mathbf{A}$ is the perturbation. Therefore, the spectral method to achieve community detection can be formulated as solving the following optimization problem:

$$\max_{\|\boldsymbol{x}\|_2^2 = n,\ \boldsymbol{x}^\top \mathbf{1}_n = 0} \boldsymbol{x}^\top \mathbf{A} \boldsymbol{x}, \quad \text{where } \mathbf{1}_n = (1, \ldots, 1)^\top \in \mathbb{R}^n$$

which is equivalent to finding the second eigenvector, as the first eigenvector is $\mathbf{1}_n / \sqrt{n}$.

However, as pointed by Abbe *et al.*[59], it has been an open problem whether the above spectral method is optimal for achieving the exact recovery property until very recently. We say that an algorithm achieves the exact recovery property if the community memberships are recovered exactly with probability going to 1. Abbe *et al.*[60] bridged this gap by providing a sharp entrywise eigenvector analysis for random matrices with low expected rank. Other than applications in SBM, the result can also be applied in other statistical machine learning problems, such as factor analysis and matrix completion, which we will not discuss in detail.

4.2 Ranking

Suppose that we have a large collection of n items and we are given partially revealed comparisons between pairs of items. For example, player A defeats player B; video A is preferred to video B when both are recommended. The goal is to identify the K items that receive the highest ranks based on these pair comparisons. This problem, which is called the top-K rank aggregation, has wide applications including web search[61], recommendation systems[62], and sports competition[63]. One of the most widely used parametric models is the Bradley–Terry–Luce (BTL) model[64]. In the BTL model, we have

$$\mathbb{P}(\text{item } j \text{ is preferred over item } i) = \frac{\omega_j^*}{\omega_i^* + \omega_j^*}$$

where ω_i^* is the preference score of item i or the ability of the ith person. The task then boils down to finding the K items with the highest preference scores.

To see how PCA can be applied to this problem, we introduce one spectral method called "rank centrality" proposed by Negahban *et al.*[65]. Consider each item as a node in a graph, and construct a random walk on this graph where at each time, the random walk is possible to go from vertex i to vertex j if items i and j were ever compared; and if so, the likelihood of going from i to j depends on how often i lost to j. That is, the random walk is more likely to move to a neighbor who has more "wins." The frequency this walk visits a particular node in the long run, or equivalently the stationary distribution, is the score of the corresponding item. Specifically, define the edge set $\mathcal{E} := \{(i, j) : \text{item } i \text{ and } j \text{ were compared}\}$ and consider

DOI: 10.1002/9781118445112.stat08122

the transition matrix $\mathbf{P}^* \in \mathbb{R}^{n\times n}$ such that

$$P^*_{i,j} = \begin{cases} \dfrac{1}{d}\dfrac{\omega^*_j}{\omega^*_i + \omega^*_j} & \text{if } (i,j) \in \mathcal{E} \\ 1 - \dfrac{1}{d}\displaystyle\sum_{k:(i,k)\in\mathcal{E}} \dfrac{\omega^*_k}{\omega^*_i + \omega^*_k} & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

where d is a sufficiently large constant that makes every row nonnegative (a proper probability distribution). We can verify that the normalized score vector

$$\boldsymbol{\pi}^* := \frac{1}{\sum_{i=1}^n \omega^*_i}[\omega^*_1, \ldots, \omega^*_n]^\top$$

is the stationary distribution of the Markov chain induced by $\mathbf{P}^*$, as $\mathbf{P}^*$ and $\boldsymbol{\pi}^*$ are in detailed balance, namely, $\pi^*_i P^*_{i,j} = \pi^*_j P^*_{j,i}$. Hence, by definition, $\boldsymbol{\pi}^*$ is the top left singular vector of $\mathbf{P}^*$: $\boldsymbol{\pi}^*\mathbf{P}^* = \boldsymbol{\pi}^*$. This motivates the algorithm of "rank centrality" that uses the top left singular vector of the empirical transition probability matrix $\hat{\mathbf{P}}$ as an estimate of the preference score.

As for the statistical guarantee, Negahban *et al.*[65] showed that $\Omega(n \log n)$ pairs are needed for consistency of estimating $\boldsymbol{\omega}^*$ in terms of the Euclidean norm. This is also the sample size needed for the Erdös-Rényi comparison graph to get connected, which is the minimum condition that makes the identification of top K items possible. However, this does not lead to accurate identification of the top K items. A recent work Chen *et al.*[66] showed that the same sample complexity ensures exact top-K identification, thus matching the minimax lower bound established by Chen and Suh[67] before. This was accomplished via optimal control of the entrywise error of the score estimates.

4.3 Mixture Model

PCA can also be applied to learning latent variable models or mixture regression models, which are important models for investigating heterogeneous data. To illustrate the idea, we consider a mixture of k Gaussian distributions with spherical covariances. Let $w_i \in (0, 1)$ be the probability of choosing component $i \in \{1, \ldots, k\}$, $\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k\} \subseteq \mathbb{R}^d$ be the component mean vectors, and $\sigma^2\mathbf{I}$ be the common covariance matrix. An observation in this model is given by

$$\boldsymbol{x} = \boldsymbol{\mu}_h + \boldsymbol{z} \tag{4}$$

where h is a discrete random variable such that $\mathbb{P}(h = j) = w_j$ for $j = 1, \ldots, k$ and $\boldsymbol{z} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$. In other words, $\boldsymbol{x} \sim w_1\mathcal{N}(\boldsymbol{\mu}_1, \sigma^2\mathbf{I}) + \cdots + w_k\mathcal{N}(\boldsymbol{\mu}_k, \sigma^2\mathbf{I})$ follows the mixture of the Gaussian distribution. We remark here that this is different from the topic model as every realization of $\boldsymbol{x}$ corresponds to a different realization of h. The parameters of interest are the component mean vectors, $\boldsymbol{\mu}_j$s.

Let $\mathbf{M} = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^\top) - \sigma^2\mathbf{I}$. Then, it is easy to see Ref. 68

$$\mathbf{M} = \sum_{j=1}^k w_j\ \boldsymbol{\mu}_j\boldsymbol{\mu}_j^\top$$

This indicates that running PCA on the population level would recover the k-dimensional linear subspace spanned by the mean component vectors. It cannot fully identify all $\boldsymbol{\mu}_j$s owing to identifiability issues in general, but does help reduce the dimension of the parameter space, enabling the possibility of random

Wiley StatsRef: Statistics Reference Online, © 2014–2018 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat08122

initializations for a more delicate method in a second stage, such as higher order tensor decomposition or likelihood-based approaches. See Anandkumar *et al.*[11] for a comprehensive review on tensor methods for latent variable models, which includes the above-discussed setting as a special case.

It is possible to extend the tensor decomposition method to study the finite mixture linear regression problems. One can replace the component mean vector in model (4) by $\langle \boldsymbol{x}, \boldsymbol{\beta}_h \rangle$, where h is again a random variable indicating the label of submodels. We refer to Yi *et al.*[69] and the references therein for more discussions on such generalizations. We point out that there are still many open problems in this direction. For example, the global landscape of mixture models is not well understood in general.

4.4 Manifold Learning

Some of the most popular methods in manifold learning are also based on spectral methods. For example, the classical multidimensional scaling (MDS) is used as a numerical tool for manifold learning, which is frequently used in psychometrics as a means of visualizing the level of similarity (or dissimilarity) of individual cases in a dataset[13]. It is also known as principal coordinate analysis (PCoA), emphasizing the fact that the classical MDS takes the dissimilarity matrix as an input and outputs a coordinate matrix by assigning each object a coordinate. The classical MDS uses the fact that the coordinate matrix $\mathbf{U}$ can be derived by eigenvalue decomposition from $\mathbf{B} = \mathbf{U}\mathbf{U}^\top$, and the matrix $\mathbf{B}$ can be computed from proximity matrix $\mathbf{D}$ by using double centering:

1. Set up the matrix of squared dissimilarities $\mathbf{D}^2 = [d_{ij}^2]$.
2. Apply the double centering $\mathbf{B} = -\frac{1}{2}\mathbf{J}\mathbf{D}^2\mathbf{J}$ using the centering matrix $\mathbf{J} = \mathbf{I} - n^{-1}11^\top$, where n is the number of objects.
3. Extract the m largest positive eigenvalues $\lambda_1, \ldots, \lambda_m$ of $\mathbf{B}$ and the corresponding eigenvectors $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_m$.
4. Let $\mathbf{U} = \mathbf{V}_m \boldsymbol{\Lambda}_m^{1/2}$, where $\mathbf{V}_m = (\boldsymbol{\nu}_1, \ldots \boldsymbol{\nu}_m)$ and $\boldsymbol{\Lambda}_m = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$.

In the above algorithm, the dimension m can be chosen using criteria based on the eigenvalue ratios such as those in the factor analysis[70,71]. The classical MDS is a linear dimension reduction method that uses the Euclidean distances between objects as the dissimilarity measures. It has many extensions by designing different input matrices based on dataset and objectives. These extensions include the metric MDS, the nonmetric MDS, and the generalized MDS, largely extending the scope of MDS, especially to the nonlinear setting. We refer readers to Borg and Groenen[72] for various forms of MDS and corresponding algorithms. Theoretical analysis in the literature of MDS concentrates on the case whether the objects from a higher dimensional space can be embedded into a lower dimensional space, Euclidean or non-Euclidean[73]. However, we are not aware of any statistical results measuring the performance of MDS under randomness, such as perturbation analysis when the objects are sampled from a probabilistic model.

Other nonlinear methods for manifold learning include ISOMAP[74], local linear embedding (LLE)[14], and diffusion maps (DM)[75], to name a few. Most of these procedures rely on PCA or local PCA. We illustrate this using the LLE, which is designed to be simple and intuitive, and can be computed efficiently. It mainly contains two steps: the first step is to determine the nearest neighbors for each data point and catch the local geometric structure of the dataset through finding the barycenter coordinate for those neighboring points; the second step is, by viewing the barycenter coordinates as the "weights" for the neighboring points, to evaluate the eigenvectors corresponding to the first several largest eigenvalues of the associated "affinity matrix" to locally embed the data to a lower dimensional Euclidean space. Surprisingly, despite its popularity in the manifold learning literature, Wu and Wu[76] provided the asymptotic analysis of LLE until very recently.

DOI: 10.1002/9781118445112.stat08122

5 Discussion

PCA is a powerful tool for data analytics[3]. Entering the era of big data, it is also finding many applications in modern machine learning problems. In this article, we focus on clustering and community detection[58], ranking, mixture model, and manifold learning. Other applications, such as matrix completion[77], phase synchronization[78], image segmentation[79], and functional data analysis[80], are not discussed here owing to space limitations. Motivated by the fact that data are often collected and stored in distant places, many distributed algorithms for PCA have been proposed[81,82] and shown to provide strong statistical guarantees[83].

Another important application of PCA is augmented principal component regression, which is an extension of the classical principal regression[84−86]. The basic idea is to assume that latent factors that impact on a large fraction of covariates also impact on the response variable. Therefore, we use PCA to extract latent factors from the covariates and then regress the response variable on these latent factors along with any augmented variables, resulting in an augmented factor models. An example of this is the multi-index prediction based on estimated latent factors in Ref. 87. A related topic to this is the supervised principal component regression[88].

Recently, a number of interesting theoretical results on the empirical principal components under weak signals have been developed[89−94], which are closely related to the rapid advances in random matrix theory. Interested readers are referred to the literature above for details.

References

[1] Pearson, K. (1901) On lines and planes of closest fit to systems of points in space. *Lond. Edinb. Dubl. Phil. Mag. J. Sci.* **2**, 559–572.

[2] Izenman, A.J. (2008) *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*, Springer, New York.

[3] Jolliffe, I. (2014) Principal component analysis, *Wiley StatsRef: Statistics Reference Online*.

[4] Fan, J., Liao, Y., and Mincheva, M. (2013) Large covariance estimation by thresholding principal orthogonal complements. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **75**, 1–44.

[5] Fan, J., Han, X., and Gu, W. (2012) Estimating false discovery proportion under arbitrary covariance dependence. *J. Amer. Statist. Assoc.* **107**, 1019–1048.

[6] Fan, J., Ke, Y., Sun, Q., and Zhou, W.-X. (2017) *FARM-Test: Factor-Adjusted Robust Multiple Testing with False Discovery Control*, https://arxiv.org/abs/1711.05386 (accessed 29 March 2018).

[7] Fan, J., Li, Q., and Wang, Y. (2017) Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **79**, 247–265.

[8] Sun, Q., Zhou, W.-X., and Fan, J. (2017) *Adaptive Huber Regression: Optimality and Phase Transition*, https://arxiv.org/abs/1706.06991 (accessed 29 March 2018).

[9] Fan, J., Wang, W., and Zhu, Z. (2016) *A Shrinkage Principle for Heavy-Tailed Data: High-Dimensional Robust Low-Rank Matrix Recovery*, https://arxiv.org/abs/1603.08315 (accessed 29 March 2018).

[10] Minsker, S. (2016) *Sub-Gaussian Estimators of the Mean of a Random Matrix with Heavy-Tailed Entries*, https://arxiv.org/abs/1605.07129 (accessed 29 March 2018).

[11] Anandkumar, A., Ge, R., Hsu, D., et al. (2014) Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.* **15**, 2773–2832.

[12] Abbe, E. (2017) *Community Detection and Stochastic Block Models: Recent Developments*, https://arxiv.org/abs/1703.10146 (accessed 29 March 2018).

[13] Torgerson, W.S. (1952) Multidimensional scaling: I. theory and method. *Psychometrika* **17**, 401–419.

[14] Roweis, S.T. and Saul, L.K. (2000) Nonlinear dimensionality reduction by locally linear embedding. *Science* **290**, 2323–2326.

[15] Davis, C. and Kahan, W.M. (1970) The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.* **7**, 1–46.

Wiley StatsRef: Statistics Reference Online, © 2014–2018 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat08122

[16] Wedin, P.-A.A. (1972) Perturbation bounds in connection with singular value decomposition. *BIT Numer. Math.* **12**, 99–111.

[17] Stewart, G.W. (1991) Perturbation theory for the singular value decomposition, in *SVD and Signal Processing II: Algorithms, Analysis and Applications*, (ed. R.J. Vaccaro), Elsevier Science, Amsterdam, The Netherlands pp. 99–109.

[18] Stewart, M. (2006) Perturbation of the SVD in the presence of small singular values. *Linear. Algebra. Appl.* **419**, 53–77.

[19] Yu, Y., Wang, T., and Samworth, R.J. (2015) A useful variate of the Davis–Kahan theorem for statisticians. *Biometrika* **102**, 315–323.

[20] Vu, V. (2011) Singular vectors under random perturbation. *Random Struct. Alg.* **39**, 526–538.

[21] Shabalin, A.A. and Nobel, A.B. (2013) Reconstruction of a low-rank matrix in the presence of Gaussian noise. *J. Multivar. Anal.* **118**, 67–76.

[22] O'Rourke, S., Vu, V., and Wang, K. (2018) Random perturbation of low rank matrices: Improving classical bounds. *Linear. Algebra. Appl.* **540**, 26–59.

[23] Wang, R. (2015) Singular vector perturbation under Gaussian noise. *SIAM J. Matrix Anal. Appl.* **36**, 158–177.

[24] Cai, T.T. and Zhang, A. (2018) Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *Ann. Statis.* **46**, 60–89.

[25] Dopico, F.M. (2000) A note on $\sin\Theta$ theorems for singular subspace variations. *BIT Numer. Math.* **40**, 395–403.

[26] Fan, J., Wang, W., and Zhong, Y. (2016) *An ℓ_∞ Eigenvector Perturbation Bound and its Application to Robust Covariance Estimation*, https://arxiv.org/abs/1603.03516 (accessed 29 March 2018).

[27] Vershynin, R. (2012) Introduction to the non-asymptotic analysis of random matrices, in *Compressed Sensing*, Cambridge University Press, Cambridge, pp. 210–268.

[28] Tropp, J.A. (2012) User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.* **12**, 389–434.

[29] Catoni, O. (2012) Challenging the empirical mean and empirical variance: a deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.* **48**, 1148–1185.

[30] Minsker, S. (2015) Geometric median and robust estimation in Banach spaces. *Bernoulli* **21**, 2308–2335.

[31] Giulini, I. (2017) Robust PCA and pairs of projections in a Hilbert space. *Electron. J. Stat.* **11**, 3903–3926.

[32] Candés, E.J., Li, X., Ma, Y., and Wright, J. (2011) Robust principal component analysis? *J. ACM* **58**, 11.

[33] Chandrasekaran, V., Sanghavi, S., Parrilo, P.A., and Willsky, A.S. (2011) Rank-sparsity incoherence for matrix decomposition. *SIAM J. Optim.* **21**, 572–596.

[34] Zhang, T. and Lerman, G. (2014) A novel M-estimator for robust PCA. *J. Mach. Learn. Res.* **15**, 749–808.

[35] Lawley, D. and Maxwell, A. (1971) *Factor Analysis as a Statistical Method*, 2nd edn, Butterworth, London.

[36] Fama, E. and French, K. (1992) The cross-section of expected stock returns. *J. Finance* **47**, 427–465.

[37] Chamberlain, G. and Rothschild, M. (1983) Arbitrage, factor structure, and mean variance analysis on large asset markets. *Econometrika* **51**, 1281–1304.

[38] Bai, J. and Ng, S. (2002) Determining the number of factors in approximate factor models. *Econometrica* **70**, 191–221.

[39] Leek, J. and Storey, J. (2008) A general framework for multiple testing dependence. *Proc. Natn. Acad. Sci. USA* **105**, 18718–18723.

[40] Bai, J. and Li, K. (2012) Statistical analysis of factor models of high dimension. *Ann. Statist.* **40**, 436–465.

[41] Bai, J. and Ng, S. (2013) Principal components estimation and identification of static factors. *J. Econom.* **176**, 18–29.

[42] Bickel, P.J. and Levina, E. (2008) Covariance regularization by thresholding. *Ann. Statist.* **36**, 2577–2604.

[43] Fan, J., Fan, Y., and Lv, J. (2008) Large dimensional covariance matrix estimation using a factor model. *J. Econom.* **147**, 186–197.

[44] Cai, T.T., Ren, Z., and Zhou, H.H. (2016) Estimating structured high-dimensional covariance and precision matrices: optimal rates and adaptive estimation. *Electron. J. Stat.* **10**, 1–59.

[45] Fan, J., Liao, Y., and Liu, H. (2016) An overview of the estimation of large covariance and precision matrices. *Econom. J.* **19**, C1–C32.

[46] Rothman, A.J., Levina, E., and Zhu, J. (2009) Generalized thresholding of large covariance matrices. *J. Amer. Statist. Assoc.* **104**, 177–186.

[47] Cai, T.T. and Liu, W. (2011) Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.* **106**, 672–684.

[48] Fan, J., Liu, H., and Wang, W. (2017+) Large covariance estimation through elliptical factor models. *Ann. Statist.* (in press).

[49] Sharpe, W.F. (1964) Capital asset prices: a theory of market equilibrium under conditions of risk. *J. Finance* **19**, 425–442.

[50] Fama, E.F. and French, K.R. (1993) Common risk factors in the returns on stocks and bonds. *J. Financ. Econ.* **33**, 3–56.

[51] Berk, J.B. and Green, R.C. (2004) Mutual fund flows and performance in rational markets. *J. Polit. Econ.* **112**, 1269–1295.

[52] Fan, J., Liao, Y., and Yao, J. (2015) Power enhancement in high-dimensional cross-sectional tests. *Econometrica* **83**, 1497–1541.

[53] Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **57**, 289–300.

DOI: 10.1002/9781118445112.stat08122

[54] Storey, J.D., Taylor, J.E., and Siegmund, D. (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66**, 187–205.
[55] Barras, L., Scaillet, O., and Wermers, R. (2010) False discoveries in mutual fund performance: measuring luck in estimated alphas. *J. Finance* **65**, 179–216.
[56] Lan, W. and Du, L. (2017+) A factor-adjusted multiple testing procedure with application to mutual fund selection. *J. Bus. Econom. Statist.*, https://arXiv:1407.5515 (accessed 12 April 2018).
[57] Perry, P.O. and Owen, A.B. (2010) A rotation test to verify latent structure. *J. Mach. Learn. Res.* **11**, 603–624.
[58] Lowrimore, G. and Manton, K.G. (2016) Cluster analysis: Overview, *Wiley StatsRef: Statistics Reference Online*, 1–19.
[59] Abbe, E., Bandeira, A.S., and Hall, G. (2016) Exact recovery in the stochastic block model. *IEEE Trans. Inf. Theory* **62**, 471–487.
[60] Abbe, E., Fan, J., Wang, K., and Zhong, Y. (2017) *Entrywise Eigenvector Analysis of Random Matrices with Low Expected Rank*, https://arxiv.org/pdf/1709.09565 (accessed 29 March 2018).
[61] Dwork, C., Kumar, R., Naor, M., and Sivarkumar, D. (2001) *Rank Aggregation Methods for the Web*. Proceedings of the 10th International Conference on World Wide Web, ACM, New York, USA, pp. 613–622.
[62] Baltrunas, L., Makcinskas, T., and Ricci, F. (2010) *Group Recommendations with Rank Aggregation and Collaborative Filtering*. Proceedings of the 4th ACM conference on Recommender systems, pp. 119–126.
[63] Massey, K. (1997) Statistical models applied to the rating of sports teams. Master's thesis. Bluefield College, Bluefield, VA.
[64] Bradley, R.A. and Terry, M.E. (1952) Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* **39**, 324–345.
[65] Negahban, S., Oh, S., and Shah, D. (2017) Rank centrality: ranking from pairwise comparisons. *Oper. Res.* **65**, 266–287.
[66] Chen, Y., Fan, J., Ma, C., and Wang, K. (2017) *Spectral Method and Regularized MLE are Both Optimal for Top-K Ranking*, arXiv:1707.09971.
[67] Chen, Y. and Suh, C. (2015) *Spectral MLE: Top-K Rank Aggregation from Pairwise Comparisons*. Proceedings of the 32nd International Conference on Machine Learning, vol. 37, JMLR.org, pp. 371–380.
[68] Hsu, D. and Kakade, S.M. (2013) *Learning Mixtures of Spherical Gaussians: Moment Methods and Spectral Decompositions*. Proceedings of the 2013 ACM Conference on Innovations in Theoretical Computer Science, ACM, New York, pp. 11–19.
[69] Yi, X., Caramanis, C., and Sanghavi, S. (2016) *Solving a Mixture of Many Random Linear Equations by Tensor Decomposition and Alternating Minimization*, https://arxiv.org/abs/1608.05749 (accessed 29 March 2018) .
[70] Ahn, S.C. and Horenstein, A.R. (2013) Eigenvalue ratio test for the number of factors. *Econometrica* **81**, 1203–1227.
[71] Lam, C. and Yao, Q. (2012) Factor modeling for high-dimensional time series: Inference for the number of factors. *Ann. Statist.* **40**, 694–726.
[72] Borg, I. and Groenen, P.J. (2005) *Modern Multidimensional Scaling: Theory and Applications*, Springer Science & Business Media, New York.
[73] de Leeuw, J. and Heiser, W. (1982) Theory of multidimensional scaling, in *Handbook of Statistics*, vol. 2, (eds P.R. Krishnaiah and L.N. Kanal), Elsevier, 285–316.
[74] Tenenbaum, J.B., de Silva, V., and Langford, J.C. (2000) A global geometric framework for nonlinear dimensionality reduction. *Science* **290**, 2319–2323.
[75] Coifman, R.R. and Lafon, S. (2006) Diffusion maps. *Appl. Comput. Harmon. Anal.* **21**, 5–30.
[76] Wu, H.-T. and Wu, N. (2017) *Think Globally, Fit Locally under the Manifold Setup: Asymptotic Analysis of Locally Linear Embedding*, https://arxiv.org/abs/1703.04058.
[77] Keshavan, R., Montanari, A., and Oh, S. (2010) Matrix completion from a few entries. *IEEE Trans. Inf. Theory* **56**, 2980–2998.
[78] Zhong, Y. and Boumal, N. (2018) Near-optimal bounds for phase synchronization. *SIAM J. Optim.* **28**(2), 989–1016.
[79] Dambreville, S., Rathi, Y., and Tannen, A. (2006) Shape-based approach to robust image segmentation using kernel PCA. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* **1**, 977–984.
[80] Ramsay, J.O. (2016) Functional data analysis – Theory, *Wiley StatsRef: Statistics Reference Online*, 1–13.
[81] Qu, Y., Ostrouchov, G., Samatova, N., and Geist, A. (2002) Principal component analysis for dimension reduction in massive distributed data sets, in (eds B. Thuraisingham and D. Cook) *Proc. IEEE Int. Conf. Data Mining (ICDM)*, IEEE.
[82] Feldman, D., Schmidt, M., and Sohler, C. (2013) *Turning Big Data Into Tiny Data: Constant-Size Coresets for k-Means, PCA and Projective Clustering*. Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, Society for Industrial and Applied Mathematics, Philadelphia, PA, pp. 1434–1453.
[83] Fan, J., Wang, D., Wang, K., and Zhu, Z. (2018+) Distributed estimation of principal eigenspaces. *Ann. Statis.* (in press).
[84] Kendall, M. (1957) *A Course in Multivariate Analysis*, Griffin, London.
[85] Hotelling, H. (1957) The relations of the newer multivariate statistical methods to factor analysis. *Br. J. Math. Stat. Psychol.* **10**, 69–79.
[86] Stock, J.H. and Watson, M.W. (2002) Forecasting using principal components from a large number of predictors. *J. Amer. Statist. Assoc.* **97**, 1167–1179.
[87] Fan, J., Xue, L., and Yao, J. (2017) Sufficient forecasting using factor models. *J. Econom.* **201**, 292–306.

DOI: 10.1002/9781118445112.stat08122

[88] Bair, E., Hastie, T., Paul, D., and Tibshirani, R. (2006) Prediction by supervised principal components. *J. Amer. Statist. Assoc.* **101**, 119–137.
[89] Johnstone, I.M. (2001) On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29**, 295–327.
[90] Hoyle, D. and Rattray, M. (2003) *Limiting Form of the Sample Covariance Eigenspectrum in PCA and Kernel PCA*. Proceedings of the 16th International Conference on Neural Information Processing Systems, MIT Press, pp. 1181–1188.
[91] Baik, J., Ben Arous, G., and Péché, S. (2005) Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.* **33**, 1643–1697.
[92] Paul, D. (2007) Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17**, 1617–1642.
[93] Koltchinskii, V. and Lounici, K. (2017) Normal approximation and concentration of spectral projectors of sample covariance. *Ann. Statist.* **45**, 121–157.
[94] Wang, W. and Fan, J. (2017) Asymptotics of empirical eigen-structure for ultra-high dimensional spiked covariance model. *Ann. Statist.* **45**, 1342–1374.

DOI: 10.1002/9781118445112.stat08122