ECE598: Representation Learning: Algorithms and Models Fall 2017

Lecture 1: Principal Component Analysis (PCA)

Lecturer: Pramod Viswanath *Scribe: Kaiqing Zhang, Aug. 28, 2017*

In this lecture, we study one ubiquitous technique to analyze and represent high dimensional data in low dimensions, principal component analysis (PCA). We introduce the model of the data, the formulation that leads to the algorithm, and also a probabilistic view of it.

1.1 Introduction

1.1.1 Model and Problem Statement

Consider the following setting

- Data: $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$ with $\mathbf{x}_i \in \mathbb{R}^d, \forall i = 1, \cdots, N$, where usually $d < N$ and d is large;
- Goal: To represent the data in a lower dimension $\{\tilde{\mathbf{x}}_1, \cdots, \tilde{\mathbf{x}}_N\}$ with $\tilde{\mathbf{x}}_i \in \mathbb{R}^l$, where $l < d$;
- Intuition: Consider a special case where $\mathbf{x}_i = \alpha_i \mathbf{a}, \forall i \Longrightarrow \tilde{\mathbf{x}}_i = \alpha_i, \forall i$ with $l = 1$, i.e., all data can be represented by a scale of one vector $\mathbf{a}$.

We introduce one common technique to achieve the goal of dimension reduction: principal component analysis. It is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components (or sometimes, principal modes of variation) [Jol86].

Note: In the intuitive example above, we can generalize the setting to

$$\mathbf{x}_i = \alpha_i \mathbf{a} + \epsilon_i, \forall i = 1, \cdots, N$$

where ϵ_i is some error or noise of the model. The lower dimensional representation of $\tilde{\mathbf{x}}_i = \alpha_i$ is optimal in some sense if ϵ_i follows zero-mean Gaussian distribution. This turns out to be relevant in the probabilistic view of PCA as seen shortly.

1.1.2 Approach

We first summarize the formulation and approach introduced in class. For ease of notation, we concatenate the data in a matrix $\mathbf{X} = \begin{bmatrix}\mathbf{x}_1 & \cdots & \mathbf{x}_N\end{bmatrix}^T \in \mathbb{R}^{N \times d}$. Without loss of generality, we assume the data vectors are all centered, i.e., the average of each column of $\mathbf{X}$ is zero. Then we formulate the following matrix approximation problem subject to a low-rank constraint to find the low-dimension representation of the data. We aim to find the best rank-l approximation $\mathbf{X}_l$ by solving

$$\begin{aligned} \min_{\mathbf{X}_l} \ & \|\mathbf{X} - \mathbf{X}_l\|_F^2 \\ s.t. \ & \text{rank}(\mathbf{X}_l) \leq l \end{aligned} \tag{1.1}$$

where $\|\|\|_F$ is the Frobenius norm of a matrix.

To solve the rank-constrained optimization problem 1.1, we first perform singular value decomposition (SVD) on the matrix $\mathbf{X}$. Based on Fundamental Theorem of Linear Algebra, we know that $\mathbf{X}$ can be decomposed as

$$\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T \tag{1.2}$$

where the columns of $\mathbf{U} \in \mathbb{R}^{N\times N}$ and $\mathbf{V} \in \mathbb{R}^{d\times d}$ are orthonormal [1], and $\mathbf{\Lambda} \in \mathbb{R}^{N\times d}$ is a diagonal matrix with entries $(\lambda_1, \cdots, \lambda_d)$ such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$. The entries λ_i are called singular values of $\mathbf{X}$.

According to Eckart-Young Theorem [EY36], the best rank-l approximation to $\mathbf{X}$ in Frobenius norm $\mathbf{X}_l^*$ has an analytical form

$$\mathbf{X}_l^* = \sum_{i=1}^{l} \lambda_i \cdot \mathbf{u}_i \mathbf{v}_i^T \tag{1.3}$$

where $\mathbf{u}_i$ and $\mathbf{v}_i$ denote the i-th column of $\mathbf{U}$ and $\mathbf{V}$, respectively. Let $\mathbf{\Lambda}_l$ be a diagonal matrix with the first l largest singular values $\lambda_1, \cdots, \lambda_l$ on its diagonal and $\tilde{\mathbf{X}} := \begin{bmatrix}\tilde{\mathbf{x}}_1 & \cdots & \tilde{\mathbf{x}}_N\end{bmatrix}^T \in \mathbb{R}^{N\times l}$ be a concatenation of low-dimensional representation of data $\mathbf{x}_i, \forall i$. Then the representation matrix has the form

$$\tilde{\mathbf{X}} = \mathbf{U}\mathbf{\Lambda}_l. \tag{1.4}$$

Note: The rank-l approximation to $\mathbf{X}$ of the form (1.3) is also optimal with respect to (w.r.t.) the spectral norm of the matrices difference [EY36].

Note: Since the data are centered, the matrix $\mathbf{X}^T\mathbf{X}$ is also recognized as the empirical sample covariance matrix of the data. Hence, as in other formulations [Jol86, Smi02], the PCA can also be conducted by performing eigenvalue decomposition over $\mathbf{X}^T\mathbf{X}$ as

$$\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^T$$

where the columns of $\mathbf{V}$ are now the right eigenvectors of $\mathbf{X}^T\mathbf{X}$ and the matrix of low-dimensional data $\tilde{\mathbf{X}}$ becomes

$$\tilde{\mathbf{X}} = \mathbf{X}\mathbf{V}_l$$

where $\mathbf{V}_l$ consists of the first l columns of the matrix $\mathbf{V}$. This eigen-decomposition-based approach is essentially the same as the SVD-based approach we introduced.

1.1.3 Another Interpretation

Another way to think about PCA is to transform the data to a new coordinate system such that the projected data with the largest variance lies on the first coordinate (the first component), the second largest variance on the second coordinate, etc. This can be understood as in [Jol86] that the components with small variances can be omitted without losing too much information of the data.

[1] A set of vectors are orthonormal if each is of length one and they are pairwise orthogonal.

To this end, first note that since the data has zero-mean, $\mathbb{E}[(\mathbf{x}^T\mathbf{v})^2]$ becomes the population variance of the random variable $\mathbf{x}$ projected on the direction $\mathbf{v}$ provided $\|\mathbf{v}\| = 1$. Accordingly, its sample variance becomes $\|\mathbf{X}\mathbf{v}\|_2$ with the data samples $\{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ [Jol86]. Therefore, to find the axes of the coordinate system, the transform is performed in the following recursive way

$$\begin{aligned} \mathbf{v}_1 &= \arg\max_{\|\mathbf{v}\|_2=1} \|\mathbf{X}\mathbf{v}\|_2 \\ \mathbf{v}_2 &= \arg\max_{\|\mathbf{v}\|_2=1, \mathbf{v}\perp\mathbf{v}_1} \|\mathbf{X}\mathbf{v}\|_2 \\ &\cdots \\ \mathbf{v}_l &= \arg\max_{\|\mathbf{v}\|_2=1, \mathbf{v}\perp\mathbf{v}_1,\cdots,\mathbf{v}_{l-1}} \|\mathbf{X}\mathbf{v}\|_2 \end{aligned} \tag{1.5}$$

where the vectors, i.e., the axes $\mathbf{v}_1, \cdots, \mathbf{v}_l$ are exactly the first l columns of the orthonormal matrix $\mathbf{V}$ as in (1.2).

Note: This interpretation from a population point of view will be important in understanding the canonical correlation analysis (CCA) in the next lecture.

1.2 More on SVD and Applications to PCA

As a significant step in performing PCA, singular value decomposition is useful in many tasks, including the CCA discussed later. Hence we elaborate more on SVD based on the theoretical introduction in Chapter 3 of [BHK16].

1.2.1 Relation to Eigen-decomposition

The SVD satisfies some analogous relationship as eigen-decomposition. The SVD is defined for all matrices, while eigen-decomposition is only defined for square matrices. Moreover, to ensure the eigenvectors to be orthogonal, more conditions are required such as the symmetry of the matrix $\mathbf{X}$. While the right (left) singular vectors of $\mathbf{X}$ are intrinsically orthonormal with no assumptions on $\mathbf{X}$. In fact, for symmetric and positive semidefinite (PSD) matrices, the singular values and eigenvalues are identical, so are the singular vectors and eigenvectors.

Any eigenvalue λ and corresponding eigenvector $\mathbf{v}$ of a square matrix $\mathbf{A}$ satisfy $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$. Similarly, for SVD we have

$$\mathbf{X}\mathbf{v}_i = \lambda_i \mathbf{u}_i \quad \text{and} \quad \mathbf{X}^T\mathbf{u}_i = \lambda_i \mathbf{v}_i.$$

Note: Since $\mathbf{X}^T\mathbf{X}\mathbf{v}_i = \lambda_i^2\mathbf{v}_i$, this shows that λ_i^2 and $\mathbf{v}_i$ are the eigenvalue and eigenvector of $\mathbf{X}^T\mathbf{X}$, respectively.

1.2.2 Best-fit Subspaces

As shown in [BHK16, Chapter 3], the SVD actually finds the best-fitting l-dimensional space of the N data points. It is best in the sense that it minimizes the sum of squares of the perpendicular distances of the data to the subspace. Interestingly, this is equivalent to maximizing the sum of squares of the lengths of the projections of the points onto this subspace, which is due to the Pythagorean Theorem [BHK16]. In fact, the length of the projection of $\mathbf{x}_i$ onto any vector $\mathbf{v}$ of unit length is $\|\mathbf{x}_i^T \mathbf{v}\|$. Hence the best fit line is the one maximizes $\|\mathbf{X}\mathbf{v}\|_2^2$, or equivalently $\|\mathbf{X}\mathbf{v}\|_2$, which is identical to the variance-maximization-based formulation shown in Section 1.1.3. Therefore, the recursive procedure (1.5) happens to be the procedure of generating the right singular vectors. The Theorem 3.1 in [BHK16] verifies that this *greedy* algorithm works well.

Theorem 1.1 (Theorem 3.1 in [BHK16]). *Let $\mathbf{v}_1, \cdots, \mathbf{v}_d$ be the singular vectors of the matrix $\mathbf{X} \in \mathbb{R}^{N\times d}$ generated following (1.5), then for $1 \leq l \leq d$, the subspace spanned by $\mathbf{v}_1, \cdots, \mathbf{v}_l$ is the best-fit l-dimensional subspace for $\mathbf{X}$.*

1.2.3 Power Method for SVD

There have been substantial developments on how to compute the SVD in numerical analysis, see more detailed discussions in [SB13]. Here we present a polynomial-time method, *power method*, which is simple to implement but serves as the conceptual basis for many advanced methods. Recall that $\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^T$, making the k-th power of $\mathbf{X}^T\mathbf{X}$ become

$$(\mathbf{X}^T\mathbf{X})^k = \mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^T \cdots \mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^T = \mathbf{V}\mathbf{\Lambda}^2 k\mathbf{V}^T = \sum_{i=1}^{d} \lambda_i^{2k}\mathbf{v}_i\mathbf{v}_i^T.$$

Assume $\lambda_1 > \lambda_2$, then the first term in the summation dominates, i.e., $(\mathbf{X}^T\mathbf{X})^k \to \lambda_1^{2k}\mathbf{v}_1\mathbf{v}_1^T$. Hence a good estimate of $\mathbf{v}_1$ can be computed by the first column of $(\mathbf{X}^T\mathbf{X})^k$ and normalize it to a unit vector. Some improvements on speeding up the method, handling very large and sparse matrices, and handling the tie for the case $\lambda_1 = \lambda_2$, are also introduced in [BHK16, Chapter 3].

1.3 A Probabilistic View of PCA

Most of the classical interpretations of PCA is lack of a probabilistic model for the observed data, which prevents the comparison of PCA with other probabilistic techniques and its applications to Bayesian settings. A probabilistic view of PCA (PPCA) was thus first proposed by [TB99].

1.3.1 Probability Model

PCA is closely related to another latent variable analysis technique, *factor analysis* [TB99]. The setting of the model is

- $\mathbf{t} \in \mathbb{R}^l \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ represents the latent variable of low dimensions $l < d$
- Data $\mathbf{x}$ is viewed as observations of $\mathbf{t}$
- $\mathbf{W} \in \mathbb{R}^{d\times l}$ relates two sets of variables $\mathbf{x}$ and $\mathbf{t}$
- $\boldsymbol{\epsilon}$ represents the observation noise following isotropic [2] Gaussian distribution $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$.

[2] Isotropic model means the noise at each coordinate is independent and has identical variance σ^2.

Formally we have

$$\mathbf{x} = \mathbf{W}\mathbf{t} + \boldsymbol{\epsilon}. \tag{1.6}$$

Hence the probability distribution of the observations $\mathbf{x}$ is

$$\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$$

where $\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$. Hence the corresponding log-likelihood $\mathcal{L}$ becomes

$$\mathcal{L} = -\frac{N}{2}\{d \cdot \ln(2\pi) + \ln|\mathbf{C}| + \mathrm{tr}(\mathbf{C}^{-1}\mathbf{S})\} \tag{1.7}$$

where $|\cdot|$ is the determinant of a matrix and $\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^T$ is the sample covariance.

1.3.2 Maximum Likelihood Estimation

Maximum likelihood estimators (MLE) are used to find the best model parameters, i.e., $\mathbf{W}$ and σ^2. As shown in [TB99], the MLE has closed-form solution as

$$\hat{\mathbf{W}} = \mathbf{U}_l(\boldsymbol{\Lambda}_l - \sigma^2\mathbf{I})^{1/2}\mathbf{R} \;\text{ and }\; \hat{\sigma}^2 = \frac{1}{d-l}\sum_{i=l+1}^{d}\lambda_i, \tag{1.8}$$

where $\mathbf{R}$ is an arbitrary $l \times l$ orthogonal rotation matrix, $\mathbf{U}_l$ is the first l columns of $\mathbf{U}$, and $\boldsymbol{\Lambda}_l$ is the diagonal matrix with the first l largest singular values on its diagonal.

Note: Interestingly, due to the non-convexity of $\mathcal{L}$ w.r.t. $\mathbf{W}$, Setting the gradient of $\mathcal{L}$ will only lead to stationary points, where some of them are represented by minor singular values (not the first l principal components). These stationary points are actually saddlepoints on the likelihood surface [TB99].

1.3.3 Connection to PCA

Unlike the standard PCA, PPCA treats the dimensionality reduction in terms of the conditional distribution of the latent variable over observations $\mathbf{t}|\mathbf{x}$, which has the conditional mean of the latent variable as

$$\mathbb{E}[\mathbf{t}_i|\mathbf{x}_i] = \mathbf{M}^{-1}\hat{\mathbf{W}}^T\mathbf{x}_i, \forall i = 1, \cdots, N \tag{1.9}$$

with $\mathbf{M} = \hat{\mathbf{W}}^T\hat{\mathbf{W}} + \sigma^2\mathbf{I}$. Interestingly, as $\sigma^2 \to 0$, $\mathbf{M}^{-1} \to (\hat{\mathbf{W}}^T\hat{\mathbf{W}})^{-1}$, meaning that (1.9) performs an orthogonal projection of $\mathbf{x}_i$ to obtain $\mathbf{t}_i$. This reduces to the standard PCA.

Note: In general for the cases $\sigma^2 > 0$, the reconstruction is not an orthogonal projection of $\mathbf{x}_i$ and is therefore not optimal in the squared reconstruction error sense. Although it is still optimal in the MLE sense.

1.3.4 EM-based Approach and Applications of PPCA

Besides the closed-form solution as (1.8), the MLE can also be obtained iteratively by the expectation-maximization (EM) algorithm. EM is a computationally efficient algorithm for probabilistic inference [DLR77]. Thanks to the probabilistic re-formulation of PCA, this EM algorithm can be applied to handle the case even with *incomplete* or *missing* data. See more applications of PCA that are benefited from this probabilistic point of view in Section 4 of [TB99].

1.4 Summary

In this lecture, we introduce one of the most commonly used dimensionality reduction algorithm, PCA. We present an approach to perform PCA based on the low-rank matrix approximation. Another interpretations and complementary materials about SVD are also provided. In addition, we also introduce a probabilistic point of view about PCA.

Bibliography

[BHK16] Avrim Blum, John Hopcroft, and Ravindran Kannan. Foundations of data science. *Vorabversion eines Lehrbuchs*, 2016.

[DLR77] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[EY36] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[Jol86] Ian T Jolliffe. Principal component analysis and factor analysis. In *Principal component analysis*, pages 115–128. Springer, 1986.

[SB13] Josef Stoer and Roland Bulirsch. *Introduction to numerical analysis*, volume 12. Springer Science & Business Media, 2013.

[Smi02] Lindsay I Smith. A tutorial on principal components analysis. *Cornell University, USA*, 51(52):65, 2002.

[TB99] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.