1 Principal Component Analysis

1.1 Motivation

PCA wants to identify a *meaningful* basis to re-express the dataset [1]. PCA assumes that a *meaningful* data representation is one which

1. the features with large variance have meaningful structure and should be preserved
2. the features with small variance are noise and should be discarded
3. correlated features indicate redundancy and should be made uncorrelated

Suppose we have observations $\{x_i\}_{i=1}^N$ where $x_i \in \mathbb{R}^p$ for some random variable x. We want to find linear transformation of x to obtain y. In particular, let $\mathbf{X} \in \mathbb{R}^{N\times p}$ be stacked observations, we want to find a linear map $\mathbf{P} \in \mathbb{R}^{p\times q}$, where columns of $\mathbf{P}$ are orthonormal basis for feature space, i.e. $row(\mathbf{X})$, to re-express data $\mathbf{X}$ to $\mathbf{Y} \in \mathbb{R}^{N\times q}$.

$$\mathbf{Y} = \mathbf{X}\mathbf{P} \tag{1}$$

$\mathbf{Y}$ has a meaningful representation if $cov(\mathbf{Y})$ is a diagonal matrix (decorrelated), and that successive dimension in $\mathbf{Y}$ are rank-ordered according to variance (preserve struture, discard noise). We might want to take the first k principal components that accounts for the majority of variation in data, i.e. $\mathbf{P} \leftarrow \begin{bmatrix}\mathbf{P}_1, \cdots, \mathbf{P}_k\end{bmatrix}$ and compute the projected data according to (1). We can un-project the data to the original feature space by $\widehat{\mathbf{X}} = \mathbf{Y}\mathbf{P}^T$

1.2 Empirical Covariance Matrix

Note that for a random variable x with stacked observations $\mathbf{X} \in \mathbb{R}^{N\times p}$, the empirical covariance for $\mathsf{x}_i, \mathsf{x}_j$ is given by

$$\hat{\sigma}^2(\mathsf{x}_i, \mathsf{x}_j) = \frac{1}{N-1}\sum_i (x_i - \overline{x}_i)(x_j - \overline{x}_j) = \frac{1}{N-1}\left(\mathbf{X}_i - \overline{\mathbf{X}}_i \mathbf{1}_N\right)^T \left(\mathbf{X}_j - \overline{\mathbf{X}}_j \mathbf{1}_N\right)$$

where $\mathbf{X}_i, \mathbf{X}_j$ are i and j-th column of $\mathbf{X}$ and $\overline{\mathbf{X}}_i = \frac{1}{N-1}\sum_j \mathbf{X}_{ji}$. So then,

$$\widehat{\mathrm{Cov}}(\mathsf{x}) = \left[\hat{\sigma}(\mathsf{x}_i, \mathsf{x}_j)\right]_{i,j=1}^p = \frac{1}{N-1}\left(\mathbf{X} - \overline{\mathbf{X}}\mathbf{1}_N\right)^T \left(\mathbf{X} - \overline{\mathbf{X}}\mathbf{1}_N\right)$$

where $\overline{\mathbf{X}}$ is column wise feature average of $\mathbf{X}$. For zero mean observation matrix, the empirical covariance matrix is simply $\frac{1}{N-1}\mathbf{X}^T\mathbf{X}$

1.3 Solving PCA using Eigenvector Decomposition

We first write covariance matrix for $\mathbf{Y}$,

$$\widehat{\mathrm{Cov}}(\mathsf{y}) = \frac{1}{N-1}\mathbf{Y}^T\mathbf{Y} = \frac{1}{N-1}(\mathbf{X}\mathbf{P})^T(\mathbf{X}\mathbf{P}) = \mathbf{P}^T\left(\frac{1}{N-1}\mathbf{X}^T\mathbf{X}\right)\mathbf{P} = \mathbf{P}^T\widehat{\mathrm{Cov}}(\mathsf{x})\mathbf{P}$$

We know that $\widehat{\mathrm{Cov}}(\mathsf{x})$ is a symmetric matrix and therefore can be written as $\widehat{\mathrm{Cov}}(\mathsf{x}) = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$ where $\mathbf{Q} \in \mathbb{R}^{p\times p}$ are eigenvectors of $\widehat{\mathrm{Cov}}(\mathsf{x})$ with corresponding eigenvalues along diagonal entries in $\mathbf{\Lambda}$. Setting projection to be eigenvectors of $\widehat{\mathrm{Cov}}(\mathsf{x})$ diagonalizes $\widehat{\mathrm{Cov}}(\mathsf{y})$,

$$\mathbf{P} \leftarrow \mathbf{Q} \qquad \Rightarrow \qquad \widehat{\mathrm{Cov}}(\mathsf{y}) = \mathbf{P}^T\widehat{\mathrm{Cov}}(\mathsf{x})\mathbf{P} = \mathbf{Q}^T\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T\mathbf{Q} = \mathbf{\Lambda}$$

where $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$. The *principal components* of $\mathbf{X}$ are column vectors of $\mathbf{P}$, i.e. eigenvectors for $\widehat{\mathrm{Cov}}(\mathsf{x})$. y is decorrelated and $\hat{\sigma}^2(\mathsf{y}_i)$ is the variance of x along i-th principal component.

1.4 Singular Value Decomposition

The singular value decomposition of an arbitrary matrix $\mathbf{X} \in \mathbb{R}^{N \times p}$ is

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

where left singular vectors $\mathbf{U} \in \mathbb{R}^{N \times N}$ is orthogonal, singular values $\mathbf{\Sigma} \in \mathbb{R}^{N \times p}$ is diagonal, right singular vectors $\mathbf{V} \in \mathbb{R}^{p \times p}$ is orthogonal. If $\mathbf{X}^T\mathbf{X}$ is has rank r, then column vectors of $\mathbf{V}$ are eigenvectors with eigenvalues $\{\lambda_i\}_{i=1}^r$ (assuming descending ordering) for symmetric matrix $\mathbf{X}^T\mathbf{X}$, i.e. $(\mathbf{X}^T\mathbf{X})\mathbf{v}_i = \lambda_i \mathbf{v}_i$. Entries along the diagonals of $\mathbf{\Sigma}$ are singular values $\sigma_i = \sqrt{\lambda_i}$. The column vectors of $\mathbf{U}$ are given by $\mathbf{u}_i = \frac{1}{\sigma_i}\mathbf{X}\mathbf{v}_i$. We can show that column vectors of $\mathbf{U}$ are unit orthonormal vectors. Grouping linear relationships $\mathbf{X}\mathbf{v}_i = \sigma_i \mathbf{u}_i$ yield $\mathbf{X}\mathbf{V} = \mathbf{\Sigma}\mathbf{U}$. Note, $\mathbf{V}$ acts similarly to the projection matrix $\mathbf{P}$.

1.5 Solving PCA using SVD

From previous, the principal components of $\mathbf{X}$ are the eigenvectors of $\widehat{\text{Cov}}(\mathsf{x})$. Let $\mathbf{W} = \frac{1}{\sqrt{N-1}}\mathbf{X}$, then right singular vectors of $\mathbf{W}$ are the principal components desired

$$\mathbf{W}^T\mathbf{W} = \left(\frac{1}{\sqrt{N-1}}\mathbf{X}\right)^T \left(\frac{1}{\sqrt{N-1}}\mathbf{X}\right) = \frac{1}{N-1}\mathbf{X}^T\mathbf{X} = \widehat{\text{Cov}(\mathsf{x})}$$

and that $\hat{\sigma}^2(y_i) = \mathbf{\Sigma}_{ii}^2$. In particular, the projected data is

$$\mathbf{Y} = \mathbf{X}\mathbf{P} = (\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T)\mathbf{V} = \mathbf{U}\mathbf{\Sigma}$$

1.6 Limitations

PCA works well with Gaussian observations, in particular the transformed data is guaranteed to be independent. If x is jointly Gaussian, then any linear function of x is also jointly Gaussian. Suppose $\mathsf{x} \sim \mathcal{N}(\boldsymbol{\mu}_\mathsf{x}, \mathbf{\Sigma}_\mathsf{x})$ and $\mathsf{b} \sim \mathcal{N}(\boldsymbol{\mu}_\mathsf{b}, \mathbf{\Sigma}_\mathsf{b})$, then

$$A\mathsf{x} + \mathsf{b} \sim \mathcal{N}(A\boldsymbol{\mu}_\mathsf{x} + \boldsymbol{\mu}_\mathsf{b}, A\mathbf{\Sigma}_\mathsf{x}A^T + \mathbf{\Sigma}_\mathsf{b})$$

i.e. the transformed data y are jointly Gaussian. Any two variables $\mathsf{y}_i, \mathsf{y}_j$ are un-correlated (by diagonal $\widehat{\text{Cov}}(\mathsf{y})$) and therefore independent (by y jointly Gaussian). For non jointly Gaussian x, we can not assume independence in y. In other words, PCA is not able to reveal non-linear relationships between features.

2 Eigenfaces for Recognition

Eigenfaces project a set of faces to the *face space*, spanned by a set of orthonormal *eigenfaces*, which best encode variation amongst faces [2, 3, 4]. In practice this means doing SVD on the set of zero mean faces $\mathbf{X}$, pick first M right singular vectors $\mathbf{V} \in \mathbb{R}^{p \times M}$ associated with largest singular values. columns of $\mathbf{V}$ are called *eigenfaces* and *col*$(\mathbf{V})$ is the *face space*. We can project a new image $\mathbf{x} \in \mathbb{R}^{1 \times p}$ to the face space, $\mathbf{y} = (\mathbf{x} - \overline{\mathbf{X}})\mathbf{V}$ and classify faces to class $k = \arg\min_{k \in 1:N} \|\mathbf{y} - (\mathbf{X}\mathbf{V})_k\|$.

References

[1] Jonathon Shlens. "A Tutorial on Principal Component Analysis". In: *arXiv:1404.1100 [cs, stat]* (Apr. 3, 2014). arXiv: 1404.1100. URL: http://arxiv.org/abs/1404.1100 (visited on 01/31/2020).

[2] Alex P Pentland. "Face Recognition Using Eigenfaces". In: (1991), p. 6.

[3] Matthew Turk and Alex Pentland. "Eigenfaces for Recognition". In: *Journal of Cognitive Neuroscience* 3.1 (Jan. 1991), pp. 71–86. ISSN: 0898-929X, 1530-8898. DOI: 10.1162/jocn.1991.3.1.71. URL: http://www.mitpressjournals.org/doi/10.1162/jocn.1991.3.1.71 (visited on 02/04/2020).

[4] Jun Zhang, Yong Yan, and M. Lades. "Face recognition: eigenface, elastic matching, and neural nets". In: *Proceedings of the IEEE* 85.9 (Sept. 1997), pp. 1423–1435. ISSN: 1558-2256. DOI: 10.1109/5.628712.