# NOTES ON PRINCIPAL COMPONENT ANALYSIS

IORDAN GANEV

## 1. Introduction

Principal component analysis is a technique for finding a new ordered basis (or partial basis) of the predictor space in such a way that most of the variability in the data can be captured in fewer dimensions. In these expository notes, we first provide a formulation of principal component analysis from the point of view of finding directions that maximize variability. We then explain the relationship between principal component analysis and the singular value decomposition.

## 2. Formulation

Let $X \in \mathbb{R}^{n \times p}$ be the matrix of data. The number of columns is the number of predictors (i.e., the dimension of the predictor space), while the number of rows is the number of samples. For simplicity, we shift the data so that the columns have mean zero. We also assume that $n \geq p$, so there are more samples than predictor dimensions. Let $x_i$ be the $i$-th row of $X$, for $i = 1, \ldots, n$.

**Definition 2.1.** For a unit vector $w \in \mathbb{R}^p$, the *variability of $X$ in the direction of $w$* is defined as $\sum_{i=1}^{n} (x_i \cdot w)^2$.

How to interpret this equation? The projection of $x_i$ onto the span of $w$ is a vector with norm given by the dot product $x_i \cdot w$. Thus, if we if we project all the samples onto the span of $w$, we obtain a one-dimension collection of sample data $\{x_1 \cdot w, x_2 \cdot w, \ldots, x_n \ldots w\}$. The assumption that the mean of the columns is zero implies that the mean of these values is also zero. Hence, the sample variance is proportional to:

$$\text{Sample variance of } X \text{ in Span}(w) \quad \propto \quad \sum_{i=1}^{n} (x_i \cdot w)^2$$

For our purposes, we the averaging factor of $1/(n-1)$ is irrelevant, so we use the term *variability* for the sum $\sum_{i=1}^{n} (x_i \cdot w)^2$.

**Definition 2.2.** The *first principal component of $X$* is a unit vector $w_1^{(X)} \in \mathbb{R}^p$ that maximizes the variance of $X$ in the direction of $w$. That is:

$$w_1^{(X)} \;=\; \arg\max_{|w|=1} \sum_{i=1}^{n} (x_i \cdot w)^2$$

For $k = 2, \ldots, p$, a *k-th principal component* $w_k^{(X)}$ of $X$ is defined recursively as according first principal component of the matrix:

$$X - X \sum_{j=1}^{k-1} w_j^{(X)} \left( w_j^{(X)} \right)^T$$

Note that principal components are not be unique; if $w_k^{(X)}$ is a $k$-th principal component of $X$, then $-w_k^{(X)}$ is also a $k$-th principal component. Generically, however, the principal components will be uniquely defined up to sign. As we discuss below, this is the case when the singular values of $X$ are all distinct.

## 3. Relation to the Singular Value Decomposition

We now relate the principal components to the singular value decomposition. Let $X = U\Sigma W^T$ be the singular value decomposition of $X$, so that $U \in \mathbb{R}^{n \times n}$ and $W \in \mathbb{R}^{p \times p}$ are orthogonal matrices, and $\Sigma$ is a diagonal matrix with non-negative diagonal entries, known as the singular values of $X$, ordered in non-increasing order $\sigma_1 \geq \sigma_1 \geq \cdots \geq \sigma_p$. (Note that we have assumed $n \geq p$.)

**Proposition 3.1.** *For $k = 1, \ldots, p$, the k-th column of $W$ is a k-th principal component of $X$. The variability of $X$ in the direction of any k-th principal component is the square of the k-th singular value: $\sigma_k^2$.*

*Proof.* The first principal component is:

$$w_1^{(X)} = \arg\max_{|w|=1} \sum_{i=1}^{n} (x_i \cdot w)^2 = \arg\max_{|w|=1} \; w^T X^T X w$$

Let $w_k$ be the $k$-th column of $W$. This is known as a *right singular vector* of $X$ corresponding to the singular value $\sigma_k$. Then one easily sees, using the singular value decomposition, that $w_k^T X^T X w_k = \sigma_k^2$. Note that the columns of $W$ form an orthonormal basis of $\mathbb{R}^p$, so any unit vector $w \in \mathbb{R}^p$ can be written as a linear combination $w = c^k w_k$ of the columns of $W$ with $\sum_k c_k^2 = 1$. With this notation, we seek to maximize $w^T X^T X w = \sum_k c_k^2 \sigma_k^2$ as a function of the $c_k$, with the constraint $\sum_k c_k^2 = 1$. As simple computation with Lagrange multipliers shows that this quantity is maximized when $c_1 = 1$ and $c_k = 0$ for $k \neq 1$. In other words, the first column $w_1$ of $W$ is a first principal component.

For $k > 1$, first note that the $p \times p$ matrix $W^T w_k w_k^T$ has all rows zeros except for the $k$-th row, which is $w_k^T$. Using this fact, one uses the singular value decomposition of $X$ to show that the first $p - k + 1$ singular values of the matrix $X - X \sum_{j=1}^{k-1} w_j w_j^T$ are $\sigma_k, \sigma_{k+1}, \ldots, \sigma_p$, with corresponding right singular vectors $w_k, w_{k+1}, \ldots, w_p$ (the remaining singular values are zero). The result follows easily; we omit the remaining details. □

## 4. Alternative Definition

**Definition 4.1.** A first principal component for $X$ is any vector that satisfies the following equivalent conditions:

(1) $w = \text{argmax}_{w \in \mathbb{R}^p} \ \frac{w^T X^T X w}{w^T w}$
(2) $w = \text{argmax}_{w \in \mathbb{R}^p} \ \sum_{i=1}^{n} \left( \frac{x_i \cdot w}{w \cdot w} \right)^2$
(3) $w$ is a right singular vector of $X$ with maximal singular value.
(4) $w$ is an eigenvector of $X^T X$ with maximal norm eigenvalue.
(5) $w$ is the first column of $Q$, where $X = P \Sigma Q^T$ is a singular value decomposition of $X$, ordered so that the diagonal entries of $\Sigma$ are non-decreasing.

Note that the singular values of $X$ are the squares of the eigenvalues of the real symmetric matrix $X^T X$.

## 5. Appendix: The Spectral Theorem

Here is a special case of the Spectral Theorem:

**Proposition 5.1.** *Let $A \in \mathbb{R}^{n \times n}$ be a real symmetric matrix, so $A^T = A$. Then there exists an orthogonal matrix $Q \in O(n)$ and real numbers $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ such that:*

$$A = Q \cdot \text{Diag}\,(\lambda_1, \ldots, \lambda_n) \cdot Q^T$$

In other words, $A$ is diagonalizable with real eigenvalues, and the columns of $Q$ form an orthonormal eigenbasis.

*Sketch of proof.* We proceed by induction on $n$. The base case of $n = 1$ is immediate. For $n > 1$, let $\lambda \in \mathbb{C}$ be an eigenvalue of $A$. We argue that $\lambda \in \mathbb{R}$. Indeed, let $v \in \mathbb{C}^n$ be any unit eigenvector for $A$ with eigenvalue $\lambda$. Then $\overline{v}^T v = 1$ and $Av = \lambda v$. We observe that:

$$\lambda = \overline{v}^T A v = \overline{v}^T \overline{A}^T v = \overline{\overline{v}^T A v}^T = \overline{\lambda}$$

It follows that $\lambda$ is real. Now, the square matrix $A - \lambda I_n$ is real and non-invertible. Hence, there exists a unit vector $w \in \mathbb{R}^n$ in the kernel of this matrix. The vector $w$ is a real eigenvector for $A$ with eigenvalue $\lambda$. Since we took $w$ to be a unit vector, we may extend $w$ to an orthogonal matrix:

$$Q_1 = \begin{bmatrix} w & P_1 \end{bmatrix}$$

where $P_1 \in \mathbb{R}^{n \times (n-1)}$ has orthonormal columns. One computes:

$$Q_1^T A Q_1 = \begin{bmatrix} \lambda & 0 \\ 0 & P_1^T A P_1 \end{bmatrix}$$

where the columns and rows are blocked as $1 + (n-1)$. Since $P_1^T A P_1$ is a real symmetric matrix, the result now follows from the induction hypothesis. $\square$