Principal Components Analysis

Kenneth Tay [*]

July 23, 2019

1 What is a principal component?

Let $X \in \mathbb{R}^{n \times p}$ be our data matrix, and assume that the columns of X have been centered. The sample covariance matrix is $S = X^T X/n$, and since it is a real symmetric matrix, it admits an eigendecomposition:

$$X^T X = V \Lambda V^T,$$

where V and Λ are $p \times p$ matrices, with V being orthogonal and Λ being diagonal. The eigenvectors $v_j \in \mathbb{R}^p$ (i.e. the columns of V) are called the **principal component directions** of X, while the derived variables $z_j = X v_j \in \mathbb{R}^n$ are called the **principal components** of X. The z_j are also sometimes called **principal component scores**.

1.1 Connection to singular value decomposition (SVD)

Any matrix X admits a singular value decomposition $X = UDV^T$, where $U \in \mathbb{R}^{n \times n}$ and $V \in \mathbb{R}^{p \times p}$ are orthogonal, and $D \in \mathbb{R}^{n \times p}$ is diagonal, in the sense that non-zero entries only occur on the diagonal. The columns of U span the column space of X, while the columns of V span the row space of X.

It turns out that this V is the same as the V obtained from the eigendecomposition of $X^T X$, and $D^2 = \Lambda$ from before. Thus, we can use a matrix's SVD to obtain the PC directions V, as well as the PCs themselves $XV = UD$.

[*] Minor additions by Stephen Bates, 2019

2 Interpretations of PCA

2.1 PC directions as variance maximizers

The first principal component direction v_1 has the property that $z_1 = Xv_1$ has the largest sample variance of all normalized linear combinations of the X_j. The proof is fairly straightforward: if X is centered, then maximizing the variance of normalized linear combinations of the X_j is given by

$$\text{maximize}_{v_1} \quad v_1^T X^T X v_1 \quad \text{subject to} \quad \|v_1\|_2^2 = 1,$$

which has solution equal to the first eigenvector of $X^T X$.

Subsequent principal components z_j have maximum variance subject to being orthogonal to the earlier ones.

It is this property that makes principal components a popular dimension reduction technique. When we collect data X, we hope that there is variation in X. (Imagine if the value for a feature X_j is always 1: that is not very informative for either supervised or unsupervised learning!) Principal components allows us to summarize the p features into $k \ll p$ derived variables in a way that preserve variance in the data optimally.

2.2 PC directions as best approximating linear manifold

Let's think of our n observations $x_1, \dots, x_n$ as living in $\mathbb{R}^p$. (Assume still that the features have been centered.) We want to find the rank-q linear manifold that "best" approximates them. One measure of "best" is the linear manifold which minimizes the sum of squared distances from the points to the manifold. If we parametrize the manifold by $f(\lambda) = \mu + V_q\lambda$, where $\mu \in \mathbb{R}^p$ and $V_q \in \mathbb{R}^{p\times q}$ and V_q orthogonal, then we have to solve

$$\text{minimize}_{\mu,\lambda_i,V_q} \quad \sum_{i=1}^{N} \|x_i - \mu - V_q\lambda_i\|^2.$$

We can partially optimize to obtain $\hat{\mu} = \bar{x} = 0$, $\hat{\lambda}_i = V_q^T(x_i - \bar{x}) = V_q^T x_i$, and so to find V_q it remains to solve

$$\begin{aligned}\text{argmin}_{V_q} \quad \sum_{i=1}^{N} \|x_i - V_q V_q^T x_i\|^2 &= \text{argmin}_{V_q} \quad \sum_{i=1}^{N} C - 2x_i^T V_q V_q^T x_i + x_i^T V_q V_q^T V_q V_q^T x_i \\ &= \text{argmax}_{V_q} \quad \sum_{i=1}^{N} x_i^T V_q V_q^T x_i \\ &= \text{argmax}_{V_q} \text{tr}(V_q^T (X^T X) V_q).\end{aligned}$$

The solution of this problem is the q largest eigenvectors of $X^T X$, i.e. the PC directions for the top q principal components.

3 Connection to ridge regression

Recall that in ridge regression with regularization parameter λ, the coefficient estimates are given by $\hat{\beta}_\lambda = (X^T X + \lambda I)^{-1} X^T y$. Plugging in the SVD of X into this formula, we get

$$\begin{aligned}
\hat{\beta}_\lambda &= (VD^2V^T + \lambda I)^{-1} VDU^T y \\
&= [V(D^2 + \lambda I)V^T]^{-1} VDU^T y \\
&= V\text{diag}\left(\frac{1}{d_j^2 + \lambda}\right) DU^T y \\
&= V\text{diag}\left(\frac{d_j}{d_j^2 + \lambda}\right) U^T y,
\end{aligned}$$

so the fitted values from ridge regression are

$$X\hat{\beta}_\lambda = \sum_{j=1}^{p} u_j \frac{d_j^2}{d_j^2 + \lambda} u_j^T y.$$

When there is no regularization (i.e. OLS), we have $X\hat{\beta} = UU^T y$. Note that we can interpret $U^T y$ as the coordinate of y w.r.t. the orthonormal basis U. Thus, what ridge regression is doing is shrinking the coordinates of y w.r.t. U, and it is shrinking less when $\frac{d_j}{d_j^2+\lambda}$ is large, i.e. d_j is large, which corresponds to the top principal component directions.

4 Generalizations of PCA

4.1 Kernel PCA

The principal components can also be computed from the inner-product (gram) matrix $K = XX^T$. If we look at the doubly-centered version of the gram matrix $\widetilde{K} = (I - 11^T/N)K(I - 11^T/N)$, this is equal to $\widetilde{K} = UD^2U^T$, and we can compute our PCs $Z = UD$.

In regular PCA, $K_{ij} = \langle x_i, x_j \rangle$ is the Euclidean inner product of features related to the ith and jth observation. In kernel PCA, we transform the features to some other space (typically of much higher dimension), $x_i \mapsto h(x_i)$, and take $K_{ij} = \langle h(x_i), h(x_j) \rangle$ as the Euclidean inner product in that new space. The neat thing is that while h is of much higher dimension, we choose "kernels" so that we never have to deal with h explicitly. For example, with the Gaussian kernel, we compute $K_{ij} = K(x_i, x_j) = \exp(-\lambda \|x_i - x_j\|^2)$.

4.2 Sparse versions of PCA

Note that the first PC $z_1 = Xv_1$ is a linear combination of the X_j's. In general v_1 is not sparse, meaning that we need all p features in order to compute the first PC. Sometimes, it is desirable for v_1 to be sparse so that our principal components depend on only a handful of features.

There have been a few attempts to define sparse principal components in different ways. The first attempt was due to Jolliffe et al. 2003, which uses the "maximal variance" property of the first PC but penalizes the vector v_1:

$$\text{maximize}_{v_1} \quad v_1^T X^T X v_1 \quad \text{subject to} \quad \|v_1\|_2^2 = 1, \|v_1\|_1 \leq c.$$

This optimization problem is difficult to solve. The most popular definition for sparse PCA is probably due to Zou et al. 2006, where v_1 is the solution to

$$\text{minimize}_{v_1, \alpha} \quad \left\| X - X v_1 \alpha^T \right\|_F^2 + \lambda \|v_1\|_2^2 + \mu \|v_1\|_1 \quad \text{subject to} \quad \|\alpha\|_2^2 = 1.$$

This can be solved via an iterative method (fix v_1 and minimize w.r.t. α, fix α and minimize w.r.t. v_1).