Principal components

Statistical Learning, 2011

Niels Richard Hansen
September 29, 2011

Let Σ be a $p \times p$ symmetric, positive semidefinite matrix, which is diagonalized as

$$\Sigma = V \Lambda V^T$$

where V is orthogonal and Λ is diagonal with the eigenvalues $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$ in the diagonal. We assume throughout that the eigenvalues are sorted in decreasing order and that the columns of V, $v_1, \ldots, v_p$, are in the corresponding order so that v_i is an eigenvector with eigenvalue λ_i. Define

$$H_q = \mathrm{span}\{v_1, v_2, \ldots, v_q\}$$

as the subspace spanned by the q first columns of V and let $H_q^{\perp}$ denote the orthogonal complement of H_q, which is then spanned by the remaining $p - q$ columns of V. Let also $H_0 = \{\emptyset\}$ with $H_0^{\perp} = \mathbf{R}^p$.

Result 1:

$$\lambda_q = v_q^T \Sigma v_q = \max_{\substack{\beta \in H_{q-1}^{\perp} \\ ||\beta|| = 1}} \beta^T \Sigma \beta. \tag{1}$$

Proof. For $q = 1$ and $\beta \in \mathbb{R}^p$ we have

$$\beta^T \Sigma \beta = \beta^T V \Lambda V^T \beta = \sum_{i=1}^{p} \beta^T v_i \lambda_i v_i^T \beta = \sum_{i=1}^{p} (\beta^T v_i)^2 \lambda_i.$$

Since $v_1, \ldots, v_p$ form an orthonormal basis $\sum_{i=1}^{p} (\beta^T v_i)^2 = ||\beta||^2$, hence if $||\beta|| = 1$ we see that $\beta^T \Sigma \beta$ is a convex combination of the eigenvalues $\lambda_1 \geq \ldots \geq \lambda_p$, which is thus maximized for $\beta^T v_1 = 1$. That is, for $\beta = v_1$. The maximum equals λ_1.

For general q we assume that $\beta \in H_{q-1}^{\perp}$ with $||\beta|| = 1$. Then the formula above reduces to

$$\beta^T \Sigma \beta = \sum_{i=q}^{p} (\beta^T v_i)^2 \lambda_i.$$

By the same argument as above we conclude that the sum is maximized for $\beta = v_q$ and the maximum equals λ_q. □

Consider then an $N \times p$ matrix $\mathbf{X}$, let $\Sigma = \mathbf{X}^T \mathbf{X}$, and let Λ and V be as above for this Σ. Let x_i denote the transposed of the i'th row in $\mathbf{X}$. We can think of the x_i's as

N p-dimensional observations. In the following let $K_q \subseteq \mathbf{R}^p$ denote a q-dimensional subspace of $\mathbf{R}^p$. The *rank-q-reconstruction error* of the data in $\mathbf{X}$ is defined as

$$\min_{K_q} \sum_{i=1}^{N} \min_{z \in K_q} ||x_i - z||^2,$$

where the outer minimum is over all q-dimensional subspace. The rank-q-reconstruction error is given in terms of a q-dimensional subspace that minimizes the total sums of squared distances from the observations to the subspace. The $z \in K_q$ that minimizes $||x_i - z||^2$ is the orthogonal projection onto K_q. It is not a priori clear that the outer minimum is attained, but we will show that this is indeed the case.

Result 2: *With $K_q = span\{w_1, \dots, w_q\}$ for any given q orthonormal vectors it holds that*

$$\sum_{i=1}^{N} \min_{z \in K_q} ||x_i - z||^2 = \sum_{i=1}^{N} ||x_i||^2 - \sum_{j=1}^{q} w_j^T \Sigma w_j.$$

Proof. With $W_q = [w_1 \dots w_q]$ we have that $P = W_q W_q^T$ is the orthogonal projection onto K_q and we find that

$$\begin{aligned} \sum_{i=1}^{N} \min_{z \in K_q} ||x_i - z||^2 &= \sum_{i=1}^{N} ||x_i - Px_i||^2 \\ &= \sum_{i=1}^{N} ||x_i||^2 - ||Px_i||^2 \end{aligned}$$

where we have used Pythagoras $||x_i||^2 = ||x_i - Px_i||^2 + ||Px_i||^2$. Then we have that

$$\begin{aligned} \sum_{i=1}^{N} ||Px_i||^2 &= \sum_{i=1}^{N} x_i^T P x_i \\ &= \mathrm{tr}(\mathbf{X} P \mathbf{X}^T) \\ &= \mathrm{tr}(W_q W_q^T \mathbf{X}^T \mathbf{X}) \\ &= \mathrm{tr}(W_q^T \Sigma W_q) \\ &= \sum_{j=1}^{q} w_j^T \Sigma w_j. \end{aligned}$$

□

Result 3: *The rank-q-reconstruction error is attained by $K_q = H_q$.*

Proof. We prove this by induction. For $q = 1$, Result 1 above shows that $\beta^T \Sigma \beta$ is maximized for $\beta = v_1$. Thus, by Result 2, the rank-1-reconstruction error is minimized by $K_1 = H_1$.

For the induction step we assume that the rank-$(q-1)$-reconstruction error for $q \geq 2$ is attained for $K_{q-1} = H_{q-1}$. Let K_q denote any q-dimensional subspace. Take a unit vector $w_q \in H_{q-1}^{\perp} \cap K_q$ – the latter being a non-empty subspace by a dimensions consideration. By Result 1 we know that $w_q^T \Sigma w_q \leq v_q \Sigma v_q = \lambda_q$. Moreover, the $(q-1)$-dimensional subspace, $\tilde{K}_{q-1}$, of K_q orthogonal to w_q, has by the induction hypothesis reconstruction error larger than H_{q-1}. Letting $w_1, \ldots, w_{q-1}$ denote an orthonormal basis for $\tilde{K}_{q-1}$ it follows from Result 2 that

$$\begin{aligned}
\sum_{i=1}^{N} \min_{z \in K_q} ||x_i - z||^2 &= \sum_{i=1}^{N} ||x_i||^2 - \sum_{j=1}^{q} w_j^T \Sigma w_j \\
&= \sum_{i=1}^{N} ||x_i||^2 - \sum_{j=1}^{q-1} w_j^T \Sigma w_j - w_q^T \Sigma w_q \\
&= \sum_{i=1}^{N} \min_{z \in \tilde{K}_{q-1}} ||x_i - z||^2 - w_q^T \Sigma w_q \\
&\geq \sum_{i=1}^{N} \min_{z \in H_{q-1}} ||x_i - z||^2 - v_q^T \Sigma v_q \\
&= \sum_{i=1}^{N} \min_{z \in H_q} ||x_i - z||^2.
\end{aligned}$$

□

It is a conclusion from the result above that the sequence of subspaces that minimize rank-q-reconstruction errors for $q = 1, \ldots, p$ is a nested sequence. Although this may seem reasonable and intuitive, it is not a priori obvious from the definition of the rank-q-reconstruction error. Thus to have a complete proof it is paramount **not** to assume nestedness of the subspaces in the argument above. With an a priori assumption of nestedness the argument becomes trivial in the light of Results 1 and 2.

For the curious, there is an alternative proof, which does the optimization directly instead of by induction, and which is essentially a generalization of the argument of Result 1. We give it below.

Proof. (of Result 3, alternative version) If $w_1, \ldots, w_q$ is any orthonormal set of q vectors it follows from Result 2 that we need to maximize the quantity

$$\sum_{j=1}^{q} w_j^T \Sigma w_j = \sum_{j=1}^{q} \sum_{i=1}^{p} \lambda_i (w_j^T v_i)^2 = \sum_{i=1}^{p} \lambda_i \sum_{j=1}^{q} (w_j^T v_i)^2$$

over the set of w_i's. For $q = 1$ the solution is given by Result 1. The difficulty, in general, is to show that, under the constraint that the w_j's are orthogonal and that

$\sum_{i=1}^{p}(w_j^T v_i)^2 = 1$ for $j = 1, \ldots, q$ (the w_i's have unit norm), the above quantity is maximized by, for instance, $w_j = v_j$. To give a correct argument we first argue that

$$\sum_{j=1}^{q}(w_j^T v_i)^2 \leq 1$$

for $i = 1, \ldots, p$. This follows from the fact that $w_1, \ldots, w_q$ can be enlarged to an orthonormal basis $w_1, \ldots, w_p$ and then $\sum_{j=1}^{p}(w_j^T v_i)^2 = ||v_i||^2 = 1$. Moreover,

$$\sum_{i=1}^{p}\sum_{j=1}^{q}(w_j^T v_i)^2 = q.$$

Thus the quantity we seek to maximize can be written as

$$\sum_{i=1}^{p}\lambda_i a_i$$

with $a_i \in [0,1]$ and $\sum_{i=1}^{p} a_i = q$. In this form it is clear that we maximize the quantity by taking $a_1 = \ldots = a_q = 1$ and $a_{q+1} = \ldots = a_p = 0$. As a final remark this optimum is attained by taking $w_j = v_j$ for $j = 1, \ldots, q$. □

We may note that the rank-q-reconstrution error equals $\lambda_{q+1} + \ldots + \lambda_p$.

If $\mathbf{X} = UDV^T$ denotes the singular value decomposition we have $\lambda_i = d_i^2$. The q first columns of V form an orthonormal basis of H_q and the coefficients for the projection of the the x_i's onto this subspace are in the first q columns of UD. The terminology is usually that these coefficients are called the *principal components*, that is, the first column of UD is the first principal component etc. The columns of V are called the *principal component vectors.*

In practice, it is common to center the x_i's before the computation of the principal components, and usually the centering is done using the column means of $\mathbf{X}$. In that case, Σ equals the empirical covariance matrix up to a factor $N-1$. This gives a second interpretation of the projections. The projection onto the first principal component vector is the one-dimensional projection that maximizes the (empirical) variance, the projection onto the second principal component maximizes the (empirical) variance subject to being (empirically) uncorrelated with the first projection etc. Sometimes, the columns of $\mathbf{X}$ are, in addition, scaled by the empirical standard deviation. Then Σ becomes proportional to the empirical correlation matrix instead.