

Similarity search and clustering in low dimensions - notes

Principal Component Analysis - Cont'

Sariel Har-Peled

10/17/2000

The following is taken from [Jol86].

1 Preliminaries

1.1 Linear Algebra

The following can be found in any standard linear algebra book.

Definition 1.1 For a matrix $A \in \mathbb{R}^{m \times n}$, we define the $\text{trace}(A)$ to be the sum of elements in the diagonal of A . That is

$$\text{trace}(A) = \sum_{i=1}^{\min(m,n)} a_{ii}.$$

In particular, if A is a squared matrix, then its trace is just the sum of its eigenvalues, and furthermore, if X is an invertible matrix, then

$$\text{trace}(XAX^{-1}) = \text{trace}(A).$$

Corollary 1.2 Given a matrix $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$, then $M = AB$ can be written as

$$M = \sum_{k=1}^n A_k^{\downarrow} B_k^{-},$$

where A_k^{\downarrow} is the k -th column of A (written as a column vector), and B_k^{-} is the k -th row of B (in row format).

Proof: Let C_k be the matrix $A_k^{\downarrow} B_k^{-}$. Let $C_k[i, j] = a_{ik} b_{kj}$ denote the entry (i, j) of C_k . Let $D = \sum_{k=1}^n C_k$. Clearly,

$$D[i, j] = \sum_{k=1}^n C_k[i, j] = \sum_{k=1}^n a_{ik} b_{kj}.$$

On the other hand, by definition, we have:

$$M[i, j] = \sum_{k=1}^n a_{ik} b_{kj} = D[i, j].$$

■

2 PCA Properties

2.1 Algebraic Properties of PCA

Let $A = [\alpha_1, \dots, \alpha_p]$ be the matrix of orthonormal vectors computed in the PCA process (where α_i is a column vector). We can map the regular vector \mathbf{x} into the new variables, by $\mathbf{z} = A^T \mathbf{x}$ (i.e., we compute for each PC its value for the given sample \mathbf{x}). Let $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_p)$ be the principal values computed by the PCA of the covariance matrix S . Clearly, $S = A\Lambda A^T$ (since $SA = A\Lambda$).

Lemma 2.1 *For any integer $1 \leq q \leq p$, consider the orthonormal linear transformation*

$$y = B^T x,$$

where y is a q -element vector and B^T is a $q \times p$ matrix and let $S_y = B^T S B$ be the covariance matrix for y . Then the trace of S_y , denoted by $\text{trace}(S_y)$ is maximized by taking $B = A_q$, where A_q consists of the first q columns of A .

Proof: Let β_k be the k -th column of B . Now, $\beta_k = \sum_{j=1}^p c_{jk} \alpha_j$, where $c_{jk}, j = 1, \dots, p, k = 1, \dots, q$ are appropriate constants. Let $C = \{c_{jk}\} \in \mathbb{R}^{p \times q}$, and then $B = AC$. Then,

$$B^T S B = C^T A^T S A C = C^T \Lambda C = \sum_{j=1}^p \lambda_j c_j^T c_j,$$

by the mind boggling and totally amazing Corollary 1.2, where c_j is the j -th row of C . Therefore

$$\delta = \text{trace}(B^T S B) = \sum_{j=1}^p \lambda_j \text{trace}(c_j^T c_j) = \sum_{j=1}^p \lambda_j c_j c_j^T = \sum_{j=1}^p \lambda_j \sum_{k=1}^q c_{jk}^2.$$

Furthermore, $C = A^T A C = A^T B$ and $C^T C = B^T A A^T B = B^T B = I_q$ since B is orthogonal, and hence

$$q = \sum_{k=1}^q \sum_{j=1}^p c_{jk}^2 = \sum_{j=1}^p \sum_{k=1}^q c_{jk}^2.$$

Thus, C is also orthogonal, and let D be an extension of C to a $p \times p$ orthogonal matrix (by adding $p - q$ columns to the matrix). The rows of D are orthogonal (since $D^T = D^{-1}$, and $DD^T = DD^{-1} = I$). In particular, for $j = 1, \dots, p$, we have

$$\sum_{k=1}^q c_{jk}^2 \leq \sum_{k=1}^p d_{jk}^2 = 1,$$

since D is an extension of D and the rows of D are orthonormal.

Thus, since $\lambda_1 \geq \lambda_2 \geq \dots$ to maximize δ we need to set $\sum_{k=1}^q c_{jk} = 1$, for $k = 1, \dots, q$, and $c_{jk} = 0$ for $k = q+1, \dots, p$. Which is achieved if we set $B^T = A_q^T$. ■

Note, that in the above, the only fact that we used is that S , the correlation matrix, is symmetric. From S we extracted the normalized eigenvectors, and the eigenvalues.

3 Fitting A Point-Set with Least Square Error

Let $\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_n$ be n measurements of p random variables; that is $\tilde{\mathbf{x}}_i = (\tilde{x}_{i1}, \dots, \tilde{x}_{ip})$. Let $M_j = (1/n) \sum_{i=1}^n \tilde{x}_{ij}$ denote the averages of the j -th variable. Note, that M_j is the sample approximation to the mean of the j -th variable X_j . Similarly, we can approximate the variance and covariance of the variables X_1, \dots, X_p . Indeed, let

$$\mathbf{x}_i = (x_{i1}, \dots, x_{ip}) = \tilde{\mathbf{x}}_i - (M_1, \dots, M_p) = (\tilde{x}_{i1} - M_1, \dots, \tilde{x}_{ip} - M_p)$$

be the sampled normalized, so that the origin is their average. In particular, we can not compute

$$\text{cov}(X_u, X_v) = E \left[(X_u - E(X_u)) (X_v - E(X_v)) \right]$$

but we can approximate it:

$$W_{uv} = \sum_{i=1}^n (\tilde{x}_{iu} - M_u) (\tilde{x}_{iv} - M_v) = \sum_{i=1}^n x_{iu} x_{iv}.$$

The question is of course, by what do we need to divide W_{uv} to get a good approximation to $\text{cov}(X_u, X_v)$?

Well,

$$E \left[W_{uv} \right] = (n-1) \text{cov}(X_u, X_v)$$

skipping a long and not so interesting sequence of arguments (see <http://www.math.uah.edu/stat/sample/sample9.html>). Thus, we set our estimation to the covariance of X_u, X_v to be $W_{uv}/(n-1)$. This is known as the *sample covariance*. In particular, let \mathcal{X} be the matrix having \mathbf{x}_i as its i -th line. Then,

$$\mathcal{S} = \frac{1}{n-1} \mathcal{X}^T \mathcal{X} = \sum_{i=1}^n x_{iu} x_{iv},$$

is the *sample covariance matrix*. As above, let A be the orthogonal matrix formed by the PCA algorithm applied to \mathcal{S} .

Let B an orthogonal $p \times q$ matrix, and let $\mathbf{y}_i = B^T \mathbf{x}_i$, for $i = 1, \dots, n$. The \mathbf{y}_i s are the projection of the x_i into a subspace spanned by the columns of B . Note, that the location of \mathbf{y}_i in the original space of B is

$$P_B(\mathbf{x}_i) = B(B^T \mathbf{x}_i).$$

Is it possible to find the best possible projection? A projection is intuitively good, if the distance between a point and its projection is small. In particular, let

$$LS_B(\mathcal{X}) = LS_B(\mathbf{x}_1, \dots, \mathbf{x}_n) = \sum_{i=1}^n \left\| \mathbf{x}_i - P_B(\mathbf{x}_i) \right\|^2$$

be the sum of the squared distances between the points and their projections.

Theorem 3.1 *The least-square distance LS_B is minimized when $B = A_q$; that is, the matrix formed by taking the first q columns of the matrix A .*

Proof: Let $\mathbf{m}_i = P_B(\mathbf{x}_i)$. Let $r_i = \mathbf{x}_i - \mathbf{m}_i$. Clearly,

$$LS_B(\mathcal{X}) = \sum_{i=1}^n r_i^T r_i.$$

Also, $r_i^T \mathbf{m}_i = 0$. Now,

$$\mathbf{x}_i^T \mathbf{x}_i = (\mathbf{m}_i + r_i)^T (\mathbf{m}_i + r_i) = \mathbf{m}_i^T \mathbf{m}_i + r_i^T r_i,$$

and $r_i^T r_i = \mathbf{x}_i^T \mathbf{x}_i - \mathbf{m}_i^T \mathbf{m}_i$. Thus,

$$LS_B(\mathcal{X}) = \sum_{i=1}^n r_i^T r_i = \sum_{i=1}^n (\mathbf{x}_i^T \mathbf{x}_i - \mathbf{m}_i^T \mathbf{m}_i).$$

However, $\sum_{i=1}^n \mathbf{x}_i^T \mathbf{x}_i$ is a given quantity. Thus minimizing $LS_B(\mathcal{X})$ is equivalent to maximizing $\sum_{i=1}^n \mathbf{m}_i^T \mathbf{m}_i$. However, B is orthogonal, which means that $\|\mathbf{m}_i\| = \|\mathbf{y}_i\|$, where $\mathbf{y} = B\mathbf{x}_i$. In particular,

$$\begin{aligned} \sum_{i=1}^n \mathbf{m}_i^T \mathbf{m}_i &= \sum_{i=1}^n y_i^T y_i = \sum_{i=1}^n \mathbf{x}_i^T B B^T \mathbf{x}_i = \sum_{i=1}^n \text{trace}((B^T \mathbf{x}_i) (\mathbf{x}_i^T B)) \\ &= \text{trace} \left(B^T \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T \right) B \right) = \text{trace}(B^T \mathcal{X}^T \mathcal{X} B) \\ &= (n-1) \text{trace}(B^T \mathcal{S} B). \end{aligned}$$

However, by Lemma 2.1. we know that the above trace is maximized when $B = A_q$. And we had proved the theorem. \blacksquare

The above theorem state that the best strategy for optimizing fitting a point-set with a q dimensional linear subspace, when using the least squared measure, is by doing PCA, and projecting the the space spanned by the first j -coordinates.

The PCA has the properties that minimizes the distortion of the embedding (in some sense).

References

[Jol86] I.T. Jolliffe. *Principal Component Analysis*. Springer series in statistics. Springer-Verlag, 1986. ISBN: 0387962697.