

# PCA, SVD, LSI, and Kernel PCA

## 1 Feature Vectors and Time Series

We now consider a sample  $x_1, \dots, x_N$  of “objects” (not necessarily vectors) and a feature map  $\Phi$  such that for any object  $x$  we have that  $\Phi(x) \in R^D$  is a vector called the feature vector of  $x$ . In practice much work goes into “feature engineering” — the design of the feature map.

For example,  $x$  might be an English document and  $\Phi(x)$  the word histogram feature vector defined as follows.

$$\Phi_i(x) = \frac{c_i(x)}{|x|} \quad (1)$$

$$c_i(x) = \text{the number of times word } i \text{ occurs in document } x \quad (2)$$

$$|x| = \text{the length of } x \text{ — the total number of word occurrences} \quad (3)$$

In general, given a sample of objects  $x_1, \dots, x_N$ , we define the following *data matrix* (sometimes called the *design matrix*).

$$\Phi_{t,i} = \Phi_i(x_t) \quad (4)$$

We now distinguish two vector spaces. A vector of the form  $\Phi(x) \in R^D$  will be called a *feature vector*. We will occasionally use  $w$  or  $\beta$  to denote an arbitrary feature vector. We will use  $i$  and  $j$  for feature indices. A vector  $\alpha \in R^N$  with components  $\alpha_t$  for  $1 \leq t \leq N$  will be called a *time series*. We will use  $t$  and  $s$  for time indices. The data matrix  $\Phi$  has one time index and one feature index.

Given a sample we can define a sample mean and covariance matrix as follows where  $\mathbf{1}$  is the constant time series with  $\mathbf{1}_t = 1$ .

$$\begin{aligned} \hat{\mu} &= \frac{1}{N} \sum_{t=1}^n \Phi(x_t) \\ &= \frac{1}{N} \mathbf{1}^T \Phi \end{aligned}$$

$$\begin{aligned} \hat{\Sigma} &= \frac{1}{N} \sum_{t=1}^N (\Phi(x_t) - \hat{\mu})(\Phi(x_t) - \hat{\mu})^T \\ &= \frac{1}{N} \sum_{t=1}^N \Phi(x_t)\Phi^T(x_t) - \hat{\mu}\Phi^T(x_t) - \Phi(x_t)\hat{\mu}^T + \hat{\mu}\hat{\mu}^T \end{aligned}$$

$$\begin{aligned}
&= \left( \frac{1}{N} \sum_{t=1}^N \Phi(x_t) \Phi^T(x_t) \right) - \hat{\mu} \left( \frac{1}{N} \sum_{t=1}^N \Phi^T(x_t) \right) - \left( \frac{1}{N} \sum_{t=1}^N \Phi(x_t) \right) \hat{\mu}^T + \hat{\mu} \hat{\mu}^T \\
&= \Phi^T \Phi - \hat{\mu} \hat{\mu}^T
\end{aligned}$$

## 2 Principle Component Analysis (PCA)

It is often convenient to be able to represent high dimensional data by a low dimensional vector. For example, we might want to represent an image of a face by various characteristics like the length of the nose, the placement of the eyes, and other “parameters” of a face. We can think of an image of a face as a high dimensional vector with one dimension for every pixel of the image. Let  $\Phi(x_t)$  be an image with a feature for every pixel of the image. We can then compute the eigenvectors of  $\hat{\Sigma}$ . In principle component analysis we compute the  $G$  eigenvectors with largest eigenvalue. We can find the eigenvector of largest eigenvalue by repeatedly replacing an initially random vector  $x$  by the normalization of  $\Sigma x$ . Once we have found this “most significant” eigenvector we can work in the subspace orthogonal to this eigenvector and find the second most significant eigenvector and so on. Let  $\beta^1, \dots, \beta^G$  be  $G$  orthonormal eigenvectors of  $\hat{\Sigma}$  with the  $G$  largest eigenvalues. We can define a “reduced” feature map  $\Psi$  with  $\Psi(x) \in R^G$  as follows.

$$\Psi_k(x) = (x - \hat{\mu}) \cdot \beta^k \quad (5)$$

The feature map  $\Psi$  is the PCA feature map.

PCA can be viewed as the solution of a least squares optimization problem. Suppose we wish to define a  $G$ -dimensional feature vector  $\Phi(x)$  by a  $k \times D$  matrix  $A$  and offset  $a$  as follows.

$$\Psi(x) = A\Phi(x) + a \quad (6)$$

Suppose we also want to be able to “decompress” a feature vector  $\beta$  into an input (image)  $x$  using a  $D \times G$  “inverse” matrix  $B$  and constant  $b$  with the decomposition given by  $x' = B\beta + b$ . We can define the “optimal”  $A$ ,  $a$ ,  $B$  and  $b$  as follows.

$$A^*, a^*, B^*, b^* = \operatorname{argmin}_{A, a, B, b} \sum_{t=1}^N \|\Phi(x_t) - (B\Psi(x_t) + b)\|^2 \quad (7)$$

It can be shown that PCA gives the solution to this problem as follows where  $\beta^1, \dots, \beta^G$  are orthonormal eigenvectors of  $\hat{\Sigma}$  with the  $G$  largest eigenvalues.

$$(A^*x)_k = x \cdot \beta^k \quad (8)$$

$$a^* = -A^*\hat{\mu} \quad (9)$$

$$B^*w = \sum_{k=1}^G w_k \beta^k \quad (10)$$

$$b^* = \hat{\mu} \quad (11)$$

### 3 Singular Value Decomposition (SVD)

SVD is a general method of factoring matrices and of finding low-rank approximation to matrices. We describe a use of SVD in dimensionality reduction similar to PCA.

#### 3.1 SVD Machine for Dimensionality Reduction

SVD can be computed more efficiently than naive PCA when  $D \gg N$ , i.e., there are many more features than objects in the sample. In this case the time series space has lower dimension ( $N$ ) than does the feature vector space (dimension  $D$ ). For the word histogram feature map on documents it is common to have  $D \gg N$ . Rather than work with the covariance matrix, we work with the Gram matrix defined as follows.

$$K_{s,t} = \Phi(x_s) \cdot \Phi(x_t) \quad (12)$$

$$K = \Phi\Phi^T \quad (13)$$

$K$  is also symmetric and positive semi-definite. The eigenvectors of  $K$  are time series. Let  $\alpha^1, \dots, \alpha^G$  be orthonormal eigenvectors of  $K$  with eigenvalues  $\lambda_1, \dots, \lambda_G$  respectively. We can define the reduced dimensionality feature map  $\Psi$  as follows.

$$\Psi_k(x) = \frac{1}{\sqrt{\lambda_k}} \alpha^k \cdot \Phi\Phi(x) \quad (14)$$

$$(\Phi\Phi(x))_t = \Phi(x_t) \cdot \Phi(x) \quad (15)$$

Note that equations (12), (14) and (15) imply that  $\Psi(x)$  can be computed provided that we can compute  $K(x,y) = \Phi(x) \cdot \Phi(y)$  even if we cannot compute the feature vector  $\Phi(x)$ . In some cases we can have  $D = \infty$  but can still compute  $K(x,y)$ . The reduced feature vector  $\Psi(x)$  can still be computed in this case using (12), (14) and (15).

### 3.2 The Similarity with PCA

SVD is closely related to PCA. SVD solves the following optimization problem for a given reduced dimension  $G$  where the matrix  $A$  must be  $G \times D$  and the matrix  $B$  must be  $N \times G$ .

$$A^*, B^* = \operatorname{argmin}_{A, B} \sum_{t=1}^N \|\Phi(x_t) - BA\Phi(x_t)\|^2 \quad (16)$$

$$= \operatorname{argmin}_{A, B} \|\Phi - BA\|^2 \quad (17)$$

Equation (17) suggests that there is a symmetry between  $\Phi$  and  $\Phi^T$ . It is possible to solve (17) either by computing eigenvectors of  $\Phi^T\Phi$  or by computing eigenvectors  $\Phi\Phi^T$ . We define the correlation matrix  $\hat{\Gamma}$  as follows.

$$\hat{\Gamma} = \frac{1}{N} \sum_{t=1}^N \Phi(x) \Phi^T(x) \quad (18)$$

$$= \frac{1}{N} \Phi^T \Phi \quad (19)$$

Like the covariance matrix, the correlation matrix  $\hat{\Gamma}$  is symmetric and positive semi-definite and hence has orthogonal eigenvectors. Let  $\beta^1, \dots, \beta^G$  be orthonormal eigenvectors of  $\hat{\Gamma}$  with the  $G$  largest eigenvalues. We will now show that the following definition of  $\Psi(x)$  is equivalent to that given above.

$$\Psi_k(x) = \Phi(x) \cdot \beta^k \quad (20)$$

Equation (20) is the same as PCA except that it uses eigenvectors of the correlation matrix  $\hat{\Gamma}$  rather than eigenvectors of the covariance matrix  $\hat{\Sigma}$ . To show the equivalence of the two definitions of  $\Psi$  we first note that if  $\alpha$  is an eigenvector of  $K = \Phi\Phi^T$  with eigenvalue  $\lambda$  then we have the following.

$$\Phi^T \Phi (\Phi^T \alpha) = \Phi^T (\Phi \Phi^T) \alpha \quad (21)$$

$$= \Phi^T \lambda \alpha \quad (22)$$

$$= \lambda (\Phi^T \alpha) \quad (23)$$

Hence, if  $\alpha$  is an eigenvector of  $K = \Phi\Phi^T$  then  $\Phi^T \alpha$  is an eigenvector of  $\hat{\Gamma} = (1/N)\Phi^T \Phi$  with the same eigenvalue. Equation (14) is equivalent to  $\Psi_k(x) = (1/\sqrt{\lambda_k})(\Phi^T \alpha^k) \cdot \Phi(x)$ . This implies that  $\Psi_k(x)$  as defined by (20) is proportional to  $\Psi_k(x)$  as defined by (14). The comments in the next subsection imply that if  $\alpha$  is a unit norm eigenvector of  $K$  then  $\Phi^T \alpha$  has norm  $\sqrt{\lambda_i}$  which implies that the two definitions are the same. Note that a unit norm eigenvector of  $K = (1/N)\Phi\Phi^T$  is the same as a unit norm eigenvector of  $\Phi\Phi^T$  — the factor of  $1/N$  is removed when we normalize the eigenvector.

### 3.3 SVD as General Matrix Factorization

In general, singular value decomposition is a statement about an arbitrary matrix  $\Phi$  with arbitrary dimensions  $N \times D$ . Let  $G$  be the rank of  $\Phi$  which is no larger than the minimum of  $N$  and  $D$ . In general we can write  $\Phi$  as  $B\Lambda A$  where  $B$  is  $N \times G$ ,  $\Lambda$  is  $G \times G$ ,  $A$  is  $G \times D$  and where the vectors  $B_{\cdot,k}$  are orthonormal eigenvectors of  $\Phi\Phi^T$ , the vectors  $A_{k,\cdot}$  are orthonormal nonsingular eigenvectors  $\Phi^T\Phi$ , and  $\Lambda$  is diagonal with  $\Lambda_{k,k} = \sqrt{\lambda_k}$  where  $\lambda_k$  is the eigenvalue of  $B_{\cdot,k}$  which is also the eigenvalue of  $A_{k,\cdot}$ . The optimization problem (17) is solved by dropping the eigenvectors of smallest eigenvalue from the  $G$ -dimensional intermediate representation.

## 4 Latent Semantic Indexing (LSI)

SVD applied to English documents, with feature vectors defined by word histograms or related word count based feature vectors (such as tf-idf), is called latent semantic indexing (LSI). In this case the feature value  $\Psi_k(x_t)$  defined by (12), (14) and (15) gives the component of document  $x$  along the “topic” or “concept” dimension  $k$ .

## 5 Kernel PCA

Like SVD, Kernel PCA is appropriate for  $D \gg N$  as is common with word-based feature vectors for documents. Kernel PCA is SVD applied to the centered data matrix defined as follows.

$$\tilde{\Phi}_{t,i} = \Phi_i(x_t) - \hat{\mu}_i \tag{24}$$

For the centered data matrix we have the following.

$$\hat{\Sigma} = \frac{1}{N} \tilde{\Phi}^T \tilde{\Phi} \tag{25}$$

$$\tilde{K} = \tilde{\Phi} \tilde{\Phi}^T \tag{26}$$

So SVD on  $\tilde{\Phi}$  yields a representation of eigenvectors of the covariance matrix  $\hat{\Sigma}$  in terms of the eigenvectors of  $\tilde{K}$ . The trick is to compute  $\tilde{K}$  using only  $K(x, y)$ , i.e., without computing explicit feature vectors  $\Phi(x)$ . Note that the (uncentered) Gram matrix  $K$  can be computed from  $K(x, y)$  alone. One can show that the centered Gram matrix  $\tilde{K}$  can be written in terms of  $K$  as follows

where  $\mathbf{1}$  denotes the  $N \times N$  matrix in which every entry has the value  $1/N$ .

$$\tilde{K}_{s,t} = (\Phi(x_s) - \hat{\mu}) \cdot (\Phi(x_t) - \hat{\mu}) \quad (27)$$

$$= K(x_s, x_t) - \hat{\mu} \cdot \Phi(x_t) - \hat{\mu} \cdot \Phi(x_s) + \|\hat{\mu}\|^2 \quad (28)$$

$$\hat{\mu} = \frac{1}{N} \sum_{t=1}^N \Phi(x_t) \quad (29)$$

$$\hat{\mu} \cdot \Phi(x) = \frac{1}{N} \sum_{t=1}^N \Phi(x_t) \cdot \Phi(x) \quad (30)$$

$$= \frac{1}{N} \sum_{t=1}^N K(x, x_t) \quad (31)$$

$$\|\hat{\mu}\|^2 = \hat{\mu} \cdot \hat{\mu} \quad (32)$$

$$= \frac{1}{N^2} \sum_{t=1}^N \sum_{s=1}^N K(x_t, x_s) \quad (33)$$

Equations (28), (31) and (33) allow  $\tilde{K}$  to be computed entirely from inner products. Now let  $\alpha^1, \dots, \alpha^G$  be  $G$  orthonormal eigenvectors of  $\tilde{K}$  with the  $G$  largest eigenvalues  $\sigma_1^2, \dots, \sigma_G^2$ . By the arguments given for SVD dimension reduction, the following reduced feature map implements inner products with the principle eigenvectors of  $\hat{\Sigma} = (1/N)\tilde{\Phi}^T\tilde{\Phi}$ .

$$\Psi_k(x) = \frac{1}{\sigma_k} \alpha^k \cdot \tilde{\Phi}\Phi(x) \quad (34)$$

$$(\tilde{\Phi}\Phi(x))_t = (\Phi(x_t) - \hat{\mu}) \cdot \Phi(x) \quad (35)$$

$$= K(x_t, x) - \frac{1}{N} \sum_{s=1}^N K(x, x_s) \quad (36)$$

Equations (34) and (36) allow the PCA feature vector  $\Psi(x)$  to be computed entirely with inner products.