67577 – Intro. to Machine Learning **Fall semester, 2003/4**

Lecture 9: Algebraic Representations I: PCA

Lecturer: Amnon Shashua *Scribe: Amnon Shashua* [1]

In the previous lecture we saw that the VC dimension d of the concept class C plays an important role in designing learning algorithms. We have seen that for a given training sample size m, the lower d is the better accuracy ϵ and confidence δ one could obtain by simply finding a concept $h \in C$ which minimizes the sample error. Conversely, for a fixed accuracy and confidence parameters, the smaller d is the smaller m needs to be:

$$m = O\left(\frac{1}{\epsilon}\log\frac{1}{\delta} + \frac{d}{\epsilon}\log\frac{1}{\epsilon}\right).$$

The large margin principle used by the support vector machine minimizes d while looking for a consistent hypothesis. Another possibility for reducing d is to reduce the dimension n of the input space X — as for example, the VC dimension of separating hyperplanes is $n + 1$. There are two possible ways to achieve the dimension reduction: (i) select a susbset of coordinates ("feature selection"), or (ii) compress the data into a lower dimensional representation ("feature extraction").

In this lecture we will focus on feature extraction from a very specific (and constrained) stanpoint. We would be looking for a mixing (linear combination) of the input coordinates such that we obtain a linear projection from R^n to R^q for some $q < n$. In doing so we wish to reduce the redundancy while preserving as much as possible the variance of the data. From a statistical standpoint this is achieved by transforming to a new set of variables, called principal components, which are uncorrelated so that the first few retain most of the variation present in all of the original coordinates. For example, in an image processing application the input images are highly redundant where neighboring pixel values are highly correlated. The purpose of feature extraction would be to transform the input image into a vector of output components with the least redundancy possible. Form a geometric standpoint, this is achieved by finding the "closest" (in least squares sense) linear q-dimensional susbspace to the m sample points S. The new subspace is a lower dimensional "best approximation" to the sample S. These two, equivalent, perspectives on data compression (dimensionality reduction) form the central idea of *principal component analysis* (PCA) which probably the oldest (going back to Pearson 1901) and best known of the techniques of multivariate analysis in statistics. The computation of PCA is very simple and the definition is straightforward, but has a wide variety of different applications, a number of different derivations, quite a number of different terminologies (especially outside the statistical literature) and is the basis for quite a number of variations on the basic technique.

We will also describe a non-linear extension of PCA known as Kernel-PCA, but the focus would be mostly on PCA itself and its analysis from a couple of vantage points: (i) PCA as an optimal reconstruction after a dimension reduction, i.e., data compression, and (ii) PCA for redundancy reduction (decorrelation) of the output components.

[1]

9.1 PCA: Statistical Perspective

Let $\mathbf{x}_1, ..., \mathbf{x}_m \in R^n$ be our sample data S of vectors in R^n, arranged as columns of a matrix A. It will be convenient to assume that the data is centered, i.e., $\sum \mathbf{x}_i = 0$. If the data is not centered we can always center it by computing the mean vector $\mu = (1/m)\sum_i \mathbf{x}_i$ and replace the original data sample with the new sample $\mathbf{x}_i - \mu$. In a statistical sense, the coordinates of the vector $\mathbf{x} \in R^n$ are considered as random variables, thus a row in the matrix A is the sample of values of a particular random variable, drawn from some unknown probability distribution, associated with the row position. We wish to find a new *basis* $\mathbf{u}_1, ..., \mathbf{u}_q$ (arranged as columns of a matrix U), where $q \leq min(n, m)$, such that the coordinates of the original input vectors (or the projection onto the subspace spanned by the $\mathbf{u}_i$) in the new basis, $\mathbf{y} = U^\top \mathbf{x}$, have certain desirable properties.

9.1.1 Maximizing the Variance of Output Coordinates

The property we would like to maximize is that the projection of the sample data on the new axes is as *spread* as possible. To start this analysis, assume $q = 1$, i.e., the n components of the input vector $\mathbf{x}$ are reduced to a single output component $y = \mathbf{u}^\top \mathbf{x}$. We are looking for a single vector $\mathbf{u} \in R^n$ whose direction *maximizes the variance* of the output component y.

Formally, we are looking for a unit vector $\mathbf{u}$ which maximizes $\sum_i (\mathbf{u}^\top \mathbf{x}_i)^2$ (see Appendix A for basic statistical definitions). In other words, the projected points onto the axis represented by the vector $\mathbf{u}$ are as spread as possible (in a least squares sense). In vector notation, the optimization problem takes the following form:

$$\max_{\mathbf{u}} \frac{1}{2}\|\mathbf{u}^\top A\|^2 \qquad \textit{subject to} \qquad \mathbf{u}^\top \mathbf{u} = 1$$

The Lagrangian of the problem is:

$$L(\mathbf{u}, \lambda) = \frac{1}{2}\mathbf{u}^\top A A^\top \mathbf{u} - \lambda(\mathbf{u}^\top \mathbf{u} - 1)$$

By taking the partial derivative $\partial L / \partial \mathbf{u} = 0$ we obtain the following necessary condition (see Appendix B):

$$AA^\top \mathbf{u} = \lambda \mathbf{u},$$

which tells us that $\mathbf{u}$ is an eigenvector of the $n \times n$ (symmetric and positive definite) matrix $AA^\top$. There are n eigenvectors associated with $AA^\top$ and we can easily convince ourselves that we are looking for the one associated with the maximal eigenvalue: substitute $\lambda \mathbf{u}$ instead of $AA^\top \mathbf{u}$ in the criterion function $\mathbf{u}^\top AA^\top \mathbf{u}$ to obtain $\lambda(\mathbf{u}^\top \mathbf{u}) = \lambda$ and since the eigenvalues must be positive (since $AA^\top$ is positive definite), then the optimum is obtained for the maximal eigenvalue. The leading eigenvector $\mathbf{u}$ of $AA^\top$ is called the *first principal axis* of the data sample represented by the columns of the matrix A, and $y = \mathbf{u}^\top \mathbf{x}$ is called the first *principal component* of the data sample.

For convenience, we denote $\mathbf{u}_1 = \mathbf{u}$ and $\lambda_1 = \lambda$ as the leading eigenvector and eigenvalue of $AA^\top$. Next, we look for $y_2 = \mathbf{u}_2^\top \mathbf{x}$ which is *uncorrelated* with $y_1 = \mathbf{u}_1^\top \mathbf{x}$ and which has maximum variance (and so on for $\mathbf{u}_3, ..., \mathbf{u}_q$). Two random variables are uncorrelated if their covariance vanishes. By definition of covariance (see Appendix A) we obtain:

$$\begin{aligned} Cov(y_1 y_2) &= \sum_i (\mathbf{u}_1^\top \mathbf{x}_i)(\mathbf{u}_2^\top \mathbf{x}_i) = \mathbf{u}_1^\top (\sum_i \mathbf{x}_i \mathbf{x}_i^\top)\mathbf{u}_2 \\ &= \mathbf{u}_1^\top AA^\top \mathbf{u}_2 = \mathbf{u}_2^\top AA^\top \mathbf{u}_1 = \lambda_1 \mathbf{u}_1^\top \mathbf{u}_2 = 0 \end{aligned}$$

We can therefore use the condition $\mathbf{u}_1^\top \mathbf{u}_2 = 0$ to specify zero correlation between y_1, y_2. The functional to be optimized becomes:

$$\max_{\mathbf{u}_2} \frac{1}{2} \|\mathbf{u}_2^\top A\|^2 \qquad \textit{subject to} \qquad \mathbf{u}_2^\top \mathbf{u}_2 = 1, \quad \mathbf{u}_1^\top \mathbf{u}_2 = 0,$$

with the Lagrangian being:

$$L(\mathbf{u}_2, \lambda, \delta) = \frac{1}{2}\mathbf{u}_2^\top AA^\top \mathbf{u}_2 - \lambda(\mathbf{u}_2^\top \mathbf{u}_2 - 1) - \delta \mathbf{u}_1^\top \mathbf{u}_2.$$

By taking the partial derivative with respect to $\mathbf{u}_2$ we obtain the necessary condition:

$$AA^\top \mathbf{u}_2 - \lambda \mathbf{u}_2 - \delta \mathbf{u}_1 = 0.$$

Multiply the equation by $\mathbf{u}_1$ from the left:

$$\mathbf{u}_1^\top AA^\top \mathbf{u}_2 - \lambda \mathbf{u}_1^\top \mathbf{u}_2 - \delta \mathbf{u}_1^\top \mathbf{u}_1 = 0,$$

and noting from above that $\mathbf{u}_1^\top AA^\top \mathbf{u}_2 = \mathbf{u}_1^\top \mathbf{u}_2 = 0$ we obtain $\delta = 0$. As a result we obtain:

$$AA^\top \mathbf{u}_2 = \lambda \mathbf{u}_2,$$

so once more we have that $\lambda, \mathbf{u}_2$ form an eigenvalue/eigenvector pair of $AA^\top$. As before, λ should be as large as possible. Assuming that $AA^\top$ does not have repeated eigenvalues (a complication which we will not consider here) λ should be the next highest eigenvalue after λ_1 and $\mathbf{u}_2$ the corresponding eigenvector (note that $\lambda \neq \lambda_1$ because otherwise it follows that $\mathbf{u}_1 = \mathbf{u}_2$ which contradicts the constraint $\mathbf{u}_1^\top \mathbf{u}_2 = 0$). By induction, it can be shown that the remaining principal vectors $\mathbf{u}_3, ..., \mathbf{u}_q$ are the decreasing order eigenvactors of $AA^\top$ and the variance of the i'th principal component $y_i = \mathbf{u}_i^\top \mathbf{x}$ is λ_i.

Taken together, the PCA is the solution of the following optimization problem:

$$\max_{\mathbf{u}_1,...,\mathbf{u}_q} \frac{1}{2} \sum_i \|\mathbf{u}_i^\top A\|^2 \qquad \textit{subject to} \qquad \mathbf{u}_i^\top \mathbf{u}_i = 1, \quad \mathbf{u}_i^\top \mathbf{u}_j = 0, \quad i \neq j = 1, ..., q.$$

It will be useful for later to write the optimization function in a more concise manner as follows. Let U be the $n \times q$ matrix whose columns are $\mathbf{u}_i$ and $D = diag(\lambda_1, ..., \lambda_q)$ is an $q \times q$ diagonal matrix and $\lambda_1 > \lambda_2 > ... > \lambda_q$. Then from above we have that $U^\top U = I$ and $AA^\top U = UD$. Using the fact that $trace(\mathbf{x}\mathbf{y}^\top) = \mathbf{x}^\top \mathbf{y}$, $trace(AB) = trace(BA)$ and $trace(A+B) = trace(A) + trace(B)$ we can convert $\sum_i \|\mathbf{u}_i^\top A\|^2$ to $trace(U^\top AA^\top U)$ as follows:

$$\begin{aligned} \sum_i \mathbf{u}_i^\top AA^\top \mathbf{u}_i &= \sum_i trace(A^\top \mathbf{u}_i \mathbf{u}_i^\top A) = trace(A^\top (\sum_i \mathbf{u}_i \mathbf{u}_i^\top) A) \\ &= trace(A^\top U U^\top A) = trace(U^\top AA^\top U) \end{aligned}$$

Thus, PCA becomes the solution of the following optimization function:

$$\max_{U \in R^{n \times q}} trace(U^\top AA^\top U) \qquad \textit{subject to} \qquad U^\top U = I. \tag{9.1}$$

The solution, as saw above, is that $U = [\mathbf{u}_1, ..., \mathbf{u}_q]$ consists of the decreasing order eigenvectors of $AA^\top$. At the optimum, $trace(U^\top AA^\top U)$ is equal to $trace(D)$ which is equal to the sum of eigenvalues $\lambda_1 + ... + \lambda_q$.

It is worthwhile noting that when $q = n$, $UU^\top = U^\top U = I$, and the PCA transform is a change of basis in R^n known as Karhunen-Loeve transform.

To conclude, the PCA transform looks for q orthogonal direction vectors (called the principal axes) such that the projection of input sample vectors onto the principal directions has the maximal spread, or equivalently that the variance of the output coordinates $\mathbf{y} = U^\top \mathbf{x}$ is maximal. The principal directions are the leading (with respect to descending eigenvalues) q eigenvectors of the matrix $AA^\top$. When $q = n$, the principal directions form a basis of R^n with the property of maximizing the variance of the coordinates of the sample input vectors.

9.1.2 Decorrelation: Diagonalization of the Covariance Matrix

In the previous section we saw that PCA generates a new coordinate system $\mathbf{y} = U^\top \mathbf{x}$ where the coordinates $y_1, ..., y_q$ of $\mathbf{x}$ in the new system are *uncorrelated*. This means that the covariance matrix over the principle components should be diagonal. In this section we will explore this perspective in more detail.

The covariance matrix $\mathbf{\Sigma}_x$ of the sample data $\mathbf{x}_1, ..., \mathbf{x}_m$ with zero mean is

$$(1/m) \sum_i \mathbf{x}_i \mathbf{x}_i^\top = (1/m) AA^\top,$$

therefore the matrix $AA^\top$ we derived above is a scaled version of the covariance of the sample data (see Appendix A). The scale factor $1/m$ was unimportant in the process above because the eigenvectors are of unit norm, thus any scale of $AA^\top$ would produce the same set of eigenvectors.

The off-diagonal entries of the covariance matrix $\mathbf{\Sigma}_x$ represent the correlation (a measure of statistical dependence) between the i'th and j'th component vectors, i.e., the entries of the input vectors $\mathbf{x}$. The existence of correlations among the components (features) of the input signal is a sign of redundancy, therefore from the point of view of transforming the input representation into one which is *less* redundant, we would like to find a transformation $\mathbf{y} = U^\top \mathbf{x}$ with an output representation $\mathbf{y}$ which is associated with a diagonal covariance matrix $\mathbf{\Sigma}_y$, i.e., the components of $\mathbf{y}$ are uncorrelated.

Formally, $\mathbf{\Sigma}_y = (1/m) \sum_i \mathbf{y}_i \mathbf{y}_i^\top = (1/m) U^\top AA^\top U$, therefore we wish to find an $n \times q$ matrix for which $U^\top AA^\top U$ is diagonal. If in addition, we would require that the *variance* of the output coordinates is maximized, i.e., $trace(U^\top AA^\top U)$ is maximal (but then we need to constrain the length of the column vectors of U, i.e., set $\|\mathbf{u}_i\| = 1$) then we would get a unique solution for U where the columns are orthonormal and are defined as the first q eigenvectors of the covariance matrix $\mathbf{\Sigma}_x$. This is exactly the optimization problem defined by eqn. (7.1).

We see therefore that PCA "decorrelates" the input data. Decorrelation and statistical independence are not the same thing. If the coordinates are statistically independent then the covariance matrix is diagonal, but it does not follow that uncorrelated variables must be statistically independent — covariance is just one measure of dependence. In fact, the covariance is a measure of pairwise dependency only. However, it is a fact that uncorrelated variables are statistically independent if they have a multivariate normal distribution (a Gaussian). In other words, if the sample data $\mathbf{x}$ are drawn from a probability distribution $p(\mathbf{x})$ which has Gaussian form, the PCA transforms the sample data into a statistically independent set of variables $\mathbf{y} = U^\top \mathbf{x}$. The details are explained below.

Recall that a multivariate normal distribution of the random variables $\mathbf{x} = (x_1, ..., x_n)^\top$ is defined as $p(\mathbf{x}) \approx N(\mu, \mathbf{\Sigma})$:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} |\mathbf{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^\top \mathbf{\Sigma}^{-1} (\mathbf{x}-\mu)}.$$

Also recall that a linear combination of the variables produces also a normal distribution $N(U^\top \mu, U^\top \mathbf{\Sigma} U)$, therefore choose U such that $\mathbf{\Sigma}_y = U^\top \mathbf{\Sigma} U$ is a diagonal matrix $\mathbf{\Sigma}_y = diag(\sigma_1^2, ..., \sigma_n^2)$. We have in that case:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} \prod_i \sigma_i} e^{-\frac{1}{2}\sum_i \left(\frac{x_i - \mu_i}{\sigma_i}\right)^2}$$

which can be written as a product of univariate normal distributions $p_{x_i}(x_i)$:

$$p(\mathbf{x}) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{1/2}\sigma_i} e^{-\frac{1}{2}\left(\frac{x_i - \mu_i}{\sigma_i}\right)^2} = \prod_{i=1}^{n} p_{x_i}(x_i),$$

which proves the assertion that decorrelated normally distributed variables are statistically independent.

9.2 PCA: Optimal Reconstruction

A different, yet equivalent, perspective on the PCA transformation is as an optimal reconstruction (in a least squares sense) after a dimension reduction. We are given a sample data as before $\mathbf{x}_1, ..., \mathbf{x}_m$ and we are looking for a *small* number of orthonormal principal vectors $\mathbf{u}_1, ..., \mathbf{u}_q$ where $q < min(n, k)$ which define a q-dimensional linear subspace of R^n which *best* approximate the original input vectors in a least squares sense. In other words, the projection $\hat{\mathbf{x}}_i$ of the sample points $\mathbf{x}_i$ onto the q-dimensional subspace should minimize $\sum_i \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2$ over all possible q-dimensional subspaces of R^n.

Let $\mathcal{U}$ be the subspace spanned by the principal vectors (columns of U) and let P be the $n \times n$ projection matrix mapping a point $\mathbf{x} \in R^n$ onto its projection $\hat{\mathbf{x}} \in \mathcal{U}$. From the definition of projection, the vector $\mathbf{x} - \hat{\mathbf{x}}$ must be orthogonal to the subspace $\mathcal{U}$. Let $\mathbf{y} = (y_1, ..., y_q)$ be the coordinates of $\hat{\mathbf{x}}$ with respect to the principal vectors, i.e., $U\mathbf{y} = \hat{\mathbf{x}}$. Then, from orthogonality we have that $(\mathbf{x} - U\mathbf{y})^\top U\mathbf{w} = 0$ for all vectors $\mathbf{w} \in R^n$. Since this is true for all $\mathbf{w}$ then $U^\top U\mathbf{y} - U^\top \mathbf{x} = 0$. Therefore, $\mathbf{y} = (U^\top U)^{-1} U^\top \mathbf{x}$ and as a result the projection matrix P becomes:

$$P = U(U^\top U)^{-1} U^\top,$$

satisfying $P\mathbf{x} = \hat{\mathbf{x}}$. In the case the columns of U are orthonormal, $U^\top U = I$, we have $P = UU^\top$. We are ready now to describe the optimization problem on U: we wish to find an orthonormal set of principal vectors, $U^\top U = I$, such that $\sum_i \|\mathbf{x}_i - UU^\top \mathbf{x}_i\|^2$ is minimized.

Note that $\sum_i \|\mathbf{x}_i - UU^\top \mathbf{x}_i\|^2 = \|A - UU^\top A\|_F^2$ where $\|B\|_F^2 = \sum_{i,j} b_{ij}^2$ is the square *Frobenious* norm of a matrix. The optimal reconstruction problem therefore becomes:

$$\min_U \|A - UU^\top A\|_F^2 \qquad \textit{subject to} \qquad U^\top U = I.$$

We will show now that:

$$\text{argmin}_U \|A - UU^\top A\|_F^2 = \text{argmax } trace(U^\top A A^\top U),$$

which shows that the optimal reconstruction problem is solved by PCA (recall Eqn. 7.1).

From the identity $\|B\|_F^2 = trace(BB^\top)$, we have:

$$\|A - UU^\top A\|_F^2 = trace((A - UU^\top A)(A - UU^\top A)^\top).$$

Expanding the right hand side gives us:

$$\begin{aligned} trace((A-UU^\top A)(A-UU^\top A)^\top) &= trace(AA^\top) - trace(AA^\top UU^\top) \\ &- trace(UU^\top AA^\top) + trace(UU^\top AA^\top UU^\top) \end{aligned}$$

The second and third term are equal (commutativity of trace) and is also equal to the 4th term due to commutativity of the trace and $U^\top U = I$. Taken together:

$$\|A - UU^\top A\|_F^2 = trace(AA^\top) - trace(U^\top AA^\top U).$$

To conclude, we have proven that by taking the first q eigenvectors of $AA^\top$ we obtain a linear subspace which is *as close as possible* (in a least squares sense) to the original sample data. Hence, PCA can be viewed as a vehicle for optimal reconstruction after dimension reduction.

9.3 The Case $n >> m$

Consider the situation where n, the dimension of the input vectors, is relatively large compared to the number of sample vectors m. For example, consider input vectors representing 50 × 50 sized images of faces, i.e., $n = 2500$, where $m = 100$. In other words, we are looking for a small number of “face templates” (known as “eigenfaces”) which approximate well the original set of 100 face images. In this case, $AA^\top$ is very large, 2500×2500, yet the number of non-vanishing eigenvalues cannot be higher than 100. Given that the eigendecomposition process is $O(2500^3)$, the computational burden would be very high. However, it is possible to perform an eigendecomposition on $A^\top A$ (a 100×100 matrix) instead, as shown next.

Let the columns of Q be the first $q < m$ eigenvectors of $A^\top A$, i.e., $A^\top AQ = QD$ where D is diagonal containing the corresponding eigenvalues. After pre-multiplying both sides by A we obtain:

$$AA^\top(AQ) = (AQ)D,$$

from which we conclude that AQ contains the first q eigenvectors (but un-normalized) of $AA^\top$. We have therefore that $U = AQD^{-\frac{1}{2}}$ because:

$$U^\top U = D^{-\frac{1}{2}}Q^\top A^\top AQD^{-\frac{1}{2}} = D^{-\frac{1}{2}}DD^{-\frac{1}{2}} = I,$$

where we used the fact that $Q^\top A^\top AQ = D$. Note that eigenvalues of $A^\top A$ and $AA^\top$ are the same (because $AA^\top(AQD^{-\frac{1}{2}}) = (AQD^{-\frac{1}{2}})D$).

9.4 Kernel PCA

We can take the case $n >> m$ described in the previous section one step further and consider such large values of n which are practically uncomputable — a situation which results when mapping the original input vectors to a high dimensional space: $\phi(\mathbf{x})$ where $\phi : R^n \rightarrow \mathcal{F}$ for which $dim(\mathcal{F}) >> n$. For example, $\phi(\mathbf{x})$ representing the d'th order monomials of the coordinates of $\mathbf{x}$, i.e., $dim(\mathcal{F}) = \binom{n+d-1}{d}$ which is exponential in d. The mappings of interest are those which are paired with a non-linear kernel function: $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^\top \phi(\mathbf{x}')$ (see Lecture 5).

Performing PCA on $A = [\phi(\mathbf{x}_1), ..., \phi(\mathbf{x}_m)]$ is equivalent to finding the non-linear surface in R^n (the nature of the non-linearity depends on the choice of $\phi()$) which best approximates the original sample data $\mathbf{x}_1, ..., \mathbf{x}_k$. The problem is that $AA^\top$ is not computable — however $A^\top A$ is computable because $(A^\top A)_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$.

From the previous section, $U = AQD^{-\frac{1}{2}} = AV$ contains the first q eigenvectors of $AA^\top$(where Q and D are computable). Since A itself is not computable we cannot represent U explicitly, but we can project a new vector $\phi(\mathbf{x})$ onto the principal directions $\mathbf{u}_1, ..., \mathbf{u}_q$ and obtain the principal components, i.e., the output vector $\mathbf{y} = U^\top \phi(\mathbf{x})$, as follows. First, note that

$$\mathbf{u}_i = A\mathbf{v}_i = \sum_{j=1}^{q} v_{ij}\phi(\mathbf{x}_j),$$

where $V = [\mathbf{v}_1, ..., \mathbf{v}_q]$ and v_{ij} is the j'th coordinate of $\mathbf{v}_i$. Therefore,

$$y_i = \phi(\mathbf{x})^\top \mathbf{u}_i = \sum_{j=1}^{q} v_{ij} k(\mathbf{x}, \mathbf{x}_j).$$

Given the principal components (entries of $\mathbf{y} = U^\top \phi(\mathbf{x})$ of $\phi(\mathbf{x})$) we can measure, for example, the *distance* between $\phi(\mathbf{x})$ and the projection $\hat{\phi(\mathbf{x})} = UU^\top \phi(\mathbf{x}) = U\mathbf{y}$ onto the linear subspace spanned by $\mathbf{u}_1, ..., \mathbf{u}_q$ (without the need to explicitly compute the principal axes $\mathbf{u}_i$), as follows.

$$\begin{aligned}
\|\phi(\mathbf{x}) - \hat{\phi(\mathbf{x})}\|^2 &= \phi(\mathbf{x})^\top \phi(\mathbf{x}) + \hat{\phi(\mathbf{x})}^\top \hat{\phi(\mathbf{x})} - 2\phi(\mathbf{x})^\top \hat{\phi(\mathbf{x})} \\
&= k(\mathbf{x}, \mathbf{x}) + \mathbf{y}^\top U^\top U\mathbf{y} - 2\phi(\mathbf{x})^\top (UU^\top \phi(\mathbf{x})) \\
&= k(\mathbf{x}, \mathbf{x}) - \mathbf{y}^\top \mathbf{y} - 2\mathbf{y}^\top \mathbf{y} \\
&= k(\mathbf{x}, \mathbf{x}) - \|\mathbf{y}\|^2
\end{aligned}$$

A Variance, Covariance, etc.

Let X, Y be two random variables and let $f(x, y)$ be some function on $x \in X, y \in Y$, and let $p(x, y)$ be the probability of the event x and y occurring together. The expectation $E[f(x, y)]$ is defined:

$$E[f(x, y)] = \sum_{x\in X}\sum_{y\in Y} f(x, y)p(x, y)$$

. The mean, variance and covariance are defined:

$$\begin{aligned}
\mu_x &= E[X] = \sum_x\sum_y xp(x, y) \\
\mu_y &= E[Y] = \sum_x\sum_y yp(x, y) \\
\sigma_x^2 &= Var[X] = E[(x - \mu_x)^2] = \sum_x\sum_y (x - \mu_x)^2 p(x, y) \\
\sigma_y^2 &= Var[Y] = E[(y - \mu_y)^2] = \sum_x\sum_y (y - \mu_y)^2 p(x, y) \\
\sigma_{xy} &= Cov(XY) = E[(x - \mu_x)(y - \mu_y)] = \sum_x\sum_y (x - \mu_x)(y - \mu_y)p(x, y)
\end{aligned}$$

In vector-matrix notation, let $\mathbf{x}$ represent the n random variables of $X_1, ..., X_n$, i.e., $\mathbf{x} = (x_1, ..., x_n)^\top$ is an instance vector and $p(\mathbf{x})$ is the probability of the instance occurrence. Then the mean is a vector μ and the covariance matrix E are defined:

$$\begin{aligned}
\mu &= \sum_{\mathbf{x}\in\{X_1,...,X_n\}} \mathbf{x}p(\mathbf{x}) \\
E &= \sum_{\mathbf{x}} (\mathbf{x} - \mu)(\mathbf{x} - \mu)^\top p(\mathbf{x})
\end{aligned}$$

Note that the covariance matrix E is the linear superposition of rank-1 matrices $(\mathbf{x}-\mu)(\mathbf{x}-\mu)^\top$ with coefficients $p(\mathbf{x})$. The diagonal of E containes the variances of the variables $x_1, ..., x_n$. For a uniform distribution and a sample data S consisting of m points, let $A = [\mathbf{x}_1 - \mu, ..., \mathbf{x}_m - \mu]$ be the matrix whose columns consist of the points centered around the mean: $\mu = \frac{1}{m}\sum_i \mathbf{x}_i$. The (sample) covariance matrix is $E = \frac{1}{m}AA^\top$.

B Derivatives of Matrix Operations: Scalar Functions of a Vector

The two most important examples of a scalar function of a vector $\mathbf{x}$ are the linear form $\mathbf{a}^\top\mathbf{x}$ and the quadratic form $\mathbf{x}^\top A\mathbf{x}$ for some square matrix A.

$$\begin{aligned} d(\mathbf{a}^\top\mathbf{x}) &= \mathbf{a}^\top d\mathbf{x} \\ d(\mathbf{x}^\top A\mathbf{x}) &= (d\mathbf{x})^\top A\mathbf{x} + \mathbf{x}^\top A(d\mathbf{x}) \\ &= \left((d\mathbf{x})^\top A\mathbf{x}\right)^\top + \mathbf{x}^\top A(d\mathbf{x}) \\ &= \mathbf{x}^\top(A + A^\top)d\mathbf{x} \end{aligned}$$

where the derivative $d(\mathbf{x}^\top A\mathbf{x})$ using the rule of products $d(f \cdot g) = (df) \cdot g + f \cdot (dg)$ where $g = A\mathbf{x}$ and $f = \mathbf{x}^\top$ and noting that $d(A\mathbf{x}) = Ad\mathbf{x}$. Thus, $\frac{d}{d\mathbf{x}}(\mathbf{a}^\top\mathbf{x}) = \mathbf{a}^\top$ and $\frac{d}{d\mathbf{x}}(\mathbf{x}^\top A\mathbf{x})) = \mathbf{x}^\top(A + A^\top)$. If A is symmetric then $\frac{d}{d\mathbf{x}}(\mathbf{x}^\top A\mathbf{x})) = (2A\mathbf{x})^\top$.