Lecture Notes on Principal Component Analysis

Laurenz Wiskott*

Revision history (selection):
11 March 2004, first version
14 February 2006, major revision
27 October 2009, added the SVD section
21 February 2013, minor revision

Contents

*Institut für Neuroinformatik, Ruhr-Universität Bochum, http://www.ini.rub.de/PEOPLE/wiskott/.

1 Intuition

1.1 Problem statement

Experimental data to be analyzed is often represented as a number of vectors of fixed dimensionality. A single vector could for example be a set of temperature measurements across Germany. Taking such a vector of measurements at different times results in a number of vectors that altogether constitute the data. Each vector can also be interpreted as a point in a high dimensional space. Then the data are simply a cloud of points in this space (if one ignores the temporal order, otherwise it would be a trajectory).

When analyzing such data one often encounters the problem that the dimensionality of the data points is too high to be visualized or analyzed with some particular technique. Thus the problem arises to reduce the dimensionality of the data in some optimal way.

To keep things simple we insist that the dimensionality reduction is done linearly, i.e. we are looking for a low-dimensional linear subspace of the data space, onto which the data can be projected. As a criterion for what the optimal subspace might be it seems reasonable to require that it should be possible to reconstruct the original data points from the reduced ones as well as possible. Thus if one were to project the data back from the low-dimensional space into the original high-dimensional space, the reconstructed data points should lie as close as possible to the original ones, with the mean squared distance between original and

reconstructed data points being the reconstruction error. The question is, how can we find the linear subspace that minimizes this reconstruction error.

It is useful and common practice to remove the mean value from the data first before doing the dimensionality reduction as stated above. Thus, we assume zero mean data throughout. As a result variances and 2nd moments are the same. This justifies the slightly confusing common practice to speak of variances but write the equations for 2nd moments. Please keep that in mind.

1.2 Projection and reconstruction error

The task of principal component analysis (PCA) is to reduce the dimensionality of some high-dimensional data points by linearly projecting them onto a lower-dimensional space in such a way that the reconstruction error made by this projection is minimal. In order to develop an intuition for PCA we first take a closer look at what it means to project the data points and to reconstruct them. Figure 1 illustrates the process. (a) A few data

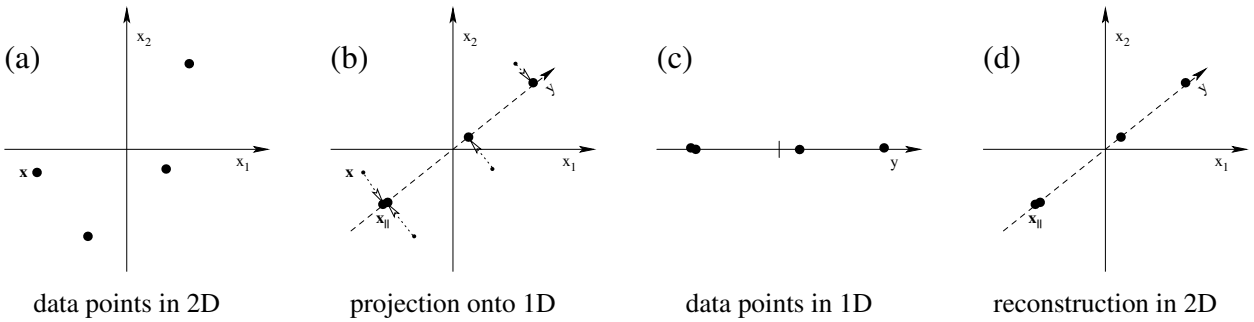


Figure 1: Projection of 2D data points onto a 1D subspace and their reconstruction.

points are given in a two-dimensional space and are represented by two-dimensional vectors $\mathbf{x} = (x_1, x_2)$. (b) In order to reduce the dimensionality down to one, we have to choose a one-dimensional subspace defined by a unit vector $\mathbf{v}$ and project the data points onto it, which can be done by

$$\mathbf{x}_\parallel := \mathbf{v}\mathbf{v}^T\mathbf{x}\,. \tag{1}$$

(c) The points can now be represented by just one number,

$$y := \mathbf{v}^T\mathbf{x}\,, \tag{2}$$

and we do not care that they originally came from a two-dimensional space. (d) If we want to reconstruct the original two-dimensional positions of the data points as well as possible, we have to embed the one-dimensional space in the original two-dimensional space in exactly the orientation used during the projection,

$$\mathbf{x}_\parallel \overset{(1,2)}{=} \mathbf{v}y\,. \tag{3}$$

However, we cannot recover the accurate 2D-position; the points remain on the one-dimensional subspace. The reconstruction error is therefore the average distance of the

original 2D-positions from the one-dimensional subspace (the length of the projection arrows in (b)). For mathematical convenience one actually takes the average squared distance

$$E := \langle \|\mathbf{x}^\mu - \mathbf{x}_{||}^\mu\|^2 \rangle_\mu \tag{4}$$

$$= \frac{1}{M} \sum_{\mu=1}^{M} \sum_{i=1}^{I} (x_i^\mu - x_{||\,i}^\mu)^2 \,, \tag{5}$$

where μ indicates the different data points, M the number of data points, and I the dimensionality of the data vectors.

1.3 Reconstruction error and variance

The question now is how we can find the direction of the one-dimensional subspace that minimizes the reconstruction error. For that it is interesting to inspect more closely what happens as we rotate the subspace. Figure 2 illustrates the projection onto two different

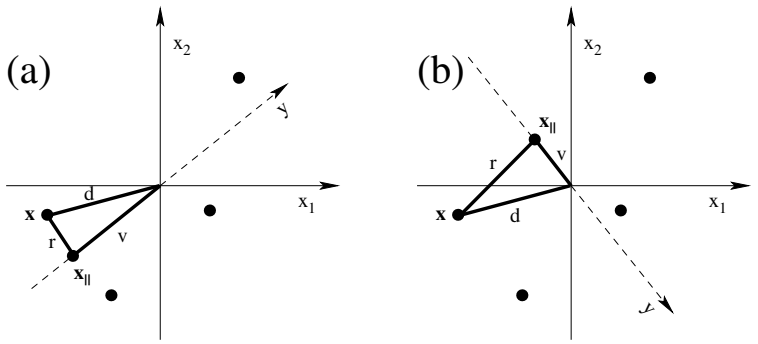


Figure 2: Variance of the projected data and reconstruction error as the linear subspace is rotated.

subspaces. Focus just on the one point $\mathbf{x}$ and its projection $\mathbf{x}_{||}$. d is the distance of $\mathbf{x}$ from the origin, r is the distance of $\mathbf{x}$ from $\mathbf{x}_{||}$ in the subspace, and v is the distance of $\mathbf{x}_{||}$ from the origin. r and v depend on the direction of the subspace while d does not. Interestingly, since the triangles between $\mathbf{x}$, $\mathbf{x}_{||}$, and the origin are right-angled, r and v are related by Pythagoras' theorem, i.e. $r^2 + v^2 = d^2$. We know that r^2 contributes to the reconstruction error. v^2 on the other hand contributes to the variance of the projected data within the subspace. Thus we see that the sum over the reconstruction error plus the variance of the projected data is constant and equals the variance of the original data. Therefore, minimizing the reconstruction error is equivalent to maximizing the variance of the projected data.

1.4 Covariance matrix

How can we determine the direction of maximal variance? The first we can do is to determine the variances of the individual components. If the data points (or vectors) are written as $\mathbf{x} = (x_1, x_2)^T$ (T indicates transpose), then the variances of the first and second component can be written as $C_{11} := \langle x_1 x_1 \rangle$ and $C_{22} := \langle x_2 x_2 \rangle$, where angle brackets indicate averaging over all data points. Please remember that these are strictly speaking 2nd moments and not

variances, but since we assume zero mean data that does not make a difference. If C_{11} is large compared to C_{22}, then the direction of maximal variance is close to $(1,0)^T$, while if C_{11} is small, the direction of maximal variance is close to $(0,1)^T$. (Notice that variance doesn't have a polarity, so that one could use the inverse vector $(-1,0)^T$ instead of $(1,0)^T$ equally well for indicating the direction of maximal variance.)

But what if C_{11} is of similar value as C_{22}, like in the example of Figure 1? Then the covariance between the two components, $C_{12} := \langle x_1 x_2 \rangle$, can give us additional information (notice that $C_{21} := \langle x_2 x_1 \rangle$ is equal to C_{12}). A large positive value of C_{12} indicates a strong correlation between x_1 and x_2 and that the data cloud is extended along the $(1,1)^T$ direction. A negative value would indicate anti-correlation and an extension along the $(-1,1)^T$ direction. A small value of C_{12} would indicate no correlation and thus little structure of the data, i.e. no prominent direction of maximal variance. The variances and covariances are conveniently arranged in a matrix with components

$$\bullet \quad C_{ij} := \langle x_i x_j \rangle \,, \tag{6}$$

which **is called covariance matrix (remember, assuming zero mean data)**[1]. It can easily be shown that the components obey the relation

$$\bullet \quad C_{ij}^2 \leq C_{ii} C_{jj} \,. \tag{7}$$

It is also easy to see that scaling the data by a factor α scales the covariance matrix by a factor α^2. Figure 3 shows several data clouds and the corresponding covariance matrices.

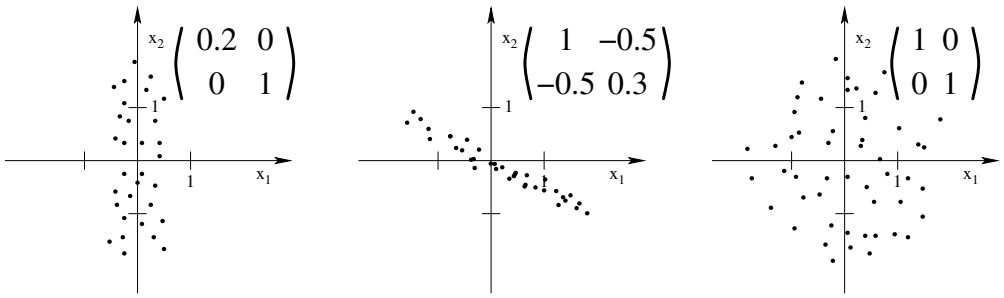


Figure 3: Several data distributions and their covariance matrices.

1.5 Covariance matrix and higher order structure

Notice that **the covariance matrix only gives you information about the general extent of the data (the second order moments). It does not give you any information about the higher-order structure of the data cloud**. Figure 4 shows different data distributions that all have the same covariance matrix. Thus as long as we consider only the covariance matrix, i.e. second order moments, we can always assume a Gaussian data distribution with an ellipsoid shape, because the covariance matrix does not represent any more structure in any case.

[1] Important text (but not inline formulas) is set bold face; • marks important formulas worth remembering; ◦ marks less important formulas, which I also discuss in the lecture.

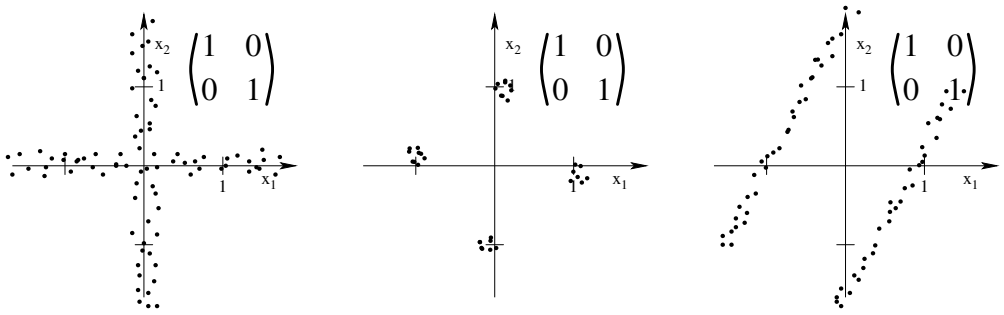


Figure 4: Different data distributions with identical covariance matrices.

1.6 PCA by diagonalizing the covariance matrix

Now that we have learned that the covariance matrix in principle contains the information about the direction of maximal variance the question arises how we can get at this information. From Figure 3 (a) and (b) we can see that there are two fundamentally different situations: in (a) the data cloud is aligned with the axes of the coordinate system and the covariance matrix is diagonal; in (b) the data cloud is oblique to the axes and the matrix is not diagonal. In the former case the direction of maximal variance is simply the axis belonging to the largest value on the diagonal of the covariance matrix. In the latter case, we cannot directly say what the direction of maximal variance might be. Thus, since the case of a diagonal covariance matrix is so much simpler, the strategy we are going to take is to make a non-diagonal covariance matrix digonal by rotating the coordinate system accordingly. This is illustrated in Figure 5. From linear algebra we know that diagonalizing a matrix can be done by solving the corresponding eigenvalue equation. It will turn out that the eigenvectors of the covariance matrix point into the directions of maximal (and minimal) variance and that the eigenvalues are equal to the variances along these directions. Projecting the data onto the eigenvectors with largest eigenvalues is therefore the optimal linear dimensionality reduction.

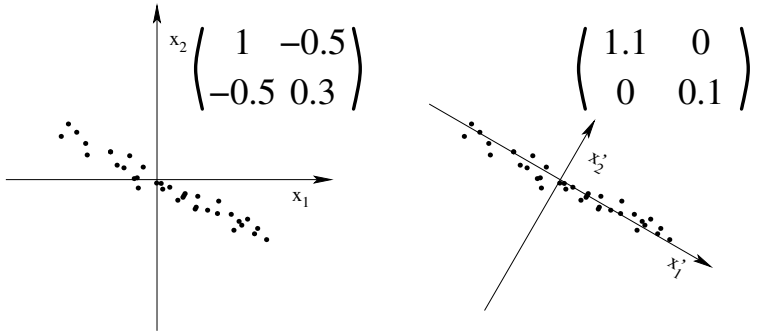


Figure 5: Diagonalizing the covariance matrix by rotating the coordinate system.

2 Formalism

2.1 Definition of the PCA-optimization problem

The problem of principal component analysis (PCA) can be formally stated as follows.

Principal Component Analysis (PCA): Given a set $\{\mathbf{x}^\mu : \mu = 1, ..., M\}$ of I-dimensional data points $\mathbf{x}^\mu = (x_1^\mu, x_2^\mu, ..., x_I^\mu)^T$ with zero mean, $\langle \mathbf{x}^\mu \rangle_\mu = \mathbf{0}_I$, find an orthogonal matrix $\mathbf{U}$ with determinant $|\mathbf{U}| = +1$ generating the transformed data points $\mathbf{x}'^\mu := \mathbf{U}^T\mathbf{x}^\mu$ such that for any given dimensionality P the data projected onto the first P axes, $\mathbf{x}'^\mu_{||} := (x'^\mu_1, x'^\mu_2, ..., x'^\mu_P, 0, ..., 0)^T$, have the smallest

$$\textbf{reconstruction error} \quad E := \langle \|\mathbf{x}'^\mu - \mathbf{x}'^\mu_{||}\|^2 \rangle_\mu \tag{8}$$

among all possible projections onto a P-dimensional subspace. The row vectors of matrix $\mathbf{U}$ define the new axes and are called the *principal components*.

Some remarks: (i) $\langle \mathbf{x}^\mu \rangle_\mu$ indicates the mean over all M data points indexed with μ. To simplify the notation we will from now on drop the index μ and indicate averages over the data points by $\langle \cdot \rangle$. (ii) If one has non-zero-mean data, one typically removes the mean before applying PCA. Even though all the math is valid also for non-zero-mean data, the results would typically be undesired and nonintuitive. (iii) Since matrix $\mathbf{U}$ is orthogonal and has determinant value +1, it corresponds simply to a rotation of the data $\mathbf{x}$. Thus, the 'shape' of the data cloud remains the same, just the 'perspective' changes. $|\mathbf{U}| = -1$ would imply a mirror reflection of the data distribution and is often permitted, too. Note also that one can interpret the multiplication with matrix $\mathbf{U}^T$ either as a rotation of the data or as a rotation of the coordinate system. Either interpretation is valid. (iv) Projecting the data $\mathbf{x}'$ onto the P-dimensional linear subspace spanned by the first P axes is simply done by setting all components higher than P to zero. This can be done, because we still have an orthonormal coordinate system. If $\mathbf{U}$ and therefore the new coordinate system were not orthogonal then the projection became a mathematically more complex operation. (v) The reconstruction error has to be minimal for any P. This has the advantage that we do not need to decide on P before performing PCA. Often P is actually choosen based on information obtained during PCA and governed by a constraint, such as that the reconstruction error should be below a certain threshold.

2.2 Matrix $\mathbf{V}^T$: Mapping from high-dimensional old coordinate system to low-dimensional new coordinate system

Assume some data points x are given in an I-dimensional space and a linear subspace is spanned by P orthonormal vectors

$$\circ \quad \mathbf{v}_p \;:=\; (v_{1p}, v_{2p}, ..., v_{Ip})^T \tag{9}$$

$$\circ \quad \text{with} \quad \mathbf{v}_p^T \mathbf{v}_q \;=\; \delta_{pq} \;:=\; \begin{cases} 1 & \text{if } p = q \\ 0 & \text{otherwise} \end{cases} . \tag{10}$$

We will typically assume $P < I$ and speak of a high(I)-dimensional space and a low(P)-dimensional (sub)space. However, $P = I$ may be possible as a limiting case as well.

Arranging these vectors in a **matrix** yields

$$\bullet \quad \mathbf{V} \;:=\; (\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_P) \tag{11}$$

$$\overset{(9)}{=} \begin{pmatrix} v_{11} & v_{12} & ... & v_{1P} \\ v_{21} & v_{22} & ... & v_{2P} \\ \vdots & & \ddots & \vdots \\ v_{I1} & v_{I2} & ... & v_{IP} \end{pmatrix} . \tag{12}$$

This matrix **can be used to map the data points x into the subspace spanned by the vectors $\mathbf{v}_p$** yielding

$$\circ \quad \mathbf{y} := \mathbf{V}^T \mathbf{x} \,, \tag{13}$$

see also figure 6. If $P < I$ then the dimensionality is reduced and some information is lost;

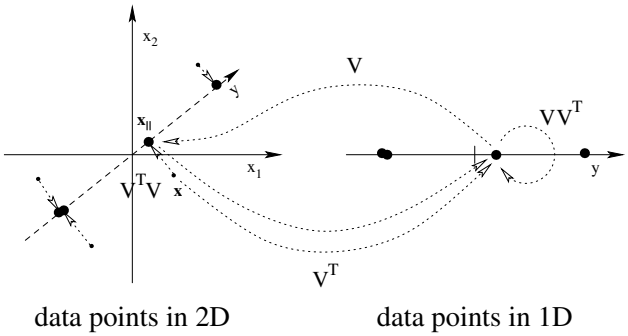


Figure 6: The effect of matrices $\mathbf{V}^T$ and $\mathbf{V}$ and combinations thereof for an example of a mapping from 2D to 1D.

if $P = I$ all information is preserved. In any case the mapped data are now represented in a new coordinate system the axes of which are given by the vectors $\mathbf{v}_p$. With $P = 2$ and $I = 3$, for example, we have

$$\mathbf{y} = \begin{pmatrix} \mathbf{v}_1^T \mathbf{x} \\ \mathbf{v}_2^T \mathbf{x} \end{pmatrix} = \begin{pmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \end{pmatrix} \mathbf{x} = \mathbf{V}^T \mathbf{x}$$

$$\text{or} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} v_{11} & v_{21} & v_{31} \\ v_{12} & v_{22} & v_{32} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \mathbf{V}^T \mathbf{x} \,.$$

Note that $\mathbf{y}$ is P-dimensional while $\mathbf{x}$ is I-dimensional.

It is important to realize that we have done two things here: firstly, **we have moved the points from the high-dimensional space onto the low-dimensional subspace** (the points that were already in the subspace have not been moved, of course) **and** secondly, **we have represented the moved points in a new coordinate system** that is particularly suitable for the low-dimensional subspace. Thus, we went from the high-dimensional space and the old coordinate system to the low-dimensional subspace and a new coordinate system. Note also that points in the high-dimensional space can generally not be represented accurately in the new coordinate system, because it does not have enough dimensions.

2.3 Matrix V: Mapping from low-dimensional new coordinate system to subspace in old coordinate system

Interestingly, since the vectors $\mathbf{v}_p$ are orthonormal, **matrix V can also be used to transform the points back from the new to the old coordinate system, although, the lost dimensions cannot be recovered**, of course. Thus the mapped points $\mathbf{y}$ in the new coordinate system become points $\mathbf{x}_{||}$ in the old coordinate system and are given by

$$\begin{aligned} \circ \quad \mathbf{x}_{||} &:= \mathbf{V}\mathbf{y} && (14) \\ \circ &\overset{(13)}{=} \mathbf{V}\mathbf{V}^T\mathbf{x}\,. && (15) \end{aligned}$$

$\mathbf{y}$ **and** $\mathbf{x}_{||}$ are equivalent representations, i.e. they **contain the same information, just in different coordinate systems.**

2.4 Matrix $(\mathbf{V}^T\mathbf{V})$: Identity mapping within new coordinate system

Before we look at the combined matrix $\mathbf{V}\mathbf{V}^T$ consider $\mathbf{V}^T\mathbf{V}$. **The latter is** obviously **a $P \times P$-matrix and performs a transformation from the new (low-dimensional) coordinate system to the old (high-dimensional) coordinate system (14) and back again (13).** The back-transformation implies a dimensionality reduction, but **since all points** in the old coordinate system come from the new coordinate system and therefore **lie within the low-dimensional subspace already, the** mapping onto the low-dimensional space does not discard any information. Thus, only the back and forth (or rather forth and back) transformation between the two coordinate systems remains and that in **combination is without any effect** either. This means that $\mathbf{V}^T\mathbf{V}$ **is the identity matrix,** which can be easily verified

$$\begin{aligned} \left(\mathbf{V}^T\mathbf{V}\right)_{pq} &= \mathbf{v}_p^T\mathbf{v}_q \overset{(10)}{=} \delta_{pq} && (16) \\ \Longleftrightarrow \quad \mathbf{V}^T\mathbf{V} &= \mathbf{1}_P && (17) \end{aligned}$$

with $\mathbf{1}_P$ indicating the identity matrix of dimensionality P. With $P = 2$, for example, we have

$$\mathbf{V}^T\mathbf{V} = \begin{pmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \end{pmatrix} (\mathbf{v}_1\mathbf{v}_2) = \begin{pmatrix} \mathbf{v}_1^T\mathbf{v}_1 & \mathbf{v}_1^T\mathbf{v}_2 \\ \mathbf{v}_2^T\mathbf{v}_1 & \mathbf{v}_2^T\mathbf{v}_2 \end{pmatrix} \overset{(10)}{=} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

2.5 Matrix ($\mathbf{V}\mathbf{V}^T$): Projection from high- to low-dimensional (sub)space within old coordinate system

As we have seen above (15) **the combined matrix $\mathbf{V}\mathbf{V}^T$ maps the points x onto the low-dimensional subspace** but in contrast to matrix $\mathbf{V}^T$ alone the mapped points are represented **within the old coordinate system** and not the new one. It turns out that **this is a projection operation with the characterizing property that** it does not make a difference whether you apply it once or twice, i.e. $\mathbf{P}\mathbf{P} = \mathbf{P}$. Let us therefore define the projection matrix

$$\bullet \quad \mathbf{P} := \mathbf{V}\mathbf{V}^T \tag{18}$$

and verify that

$$\circ \quad \mathbf{P}\mathbf{P} \overset{(18)}{=} \mathbf{V}\underbrace{\mathbf{V}^T\mathbf{V}}_{\mathbf{1}_P}\mathbf{V}^T \tag{19}$$

$$\circ \quad \overset{(17)}{=} \mathbf{V}\mathbf{V}^T \tag{20}$$

$$\circ \quad \overset{(18)}{=} \mathbf{P}\,. \tag{21}$$

A closer look at $\mathbf{P}$ shows that

$$\circ \quad \mathbf{P} := \mathbf{V}\mathbf{V}^T \tag{22}$$

$$\circ \quad = (\mathbf{v}_1, ..., \mathbf{v}_P) \begin{pmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_P^T \end{pmatrix} \tag{23}$$

$$\overset{(9)}{=} \begin{pmatrix} v_{11} & v_{12} & \cdots \\ v_{21} & \ddots & \\ v_{31} & \ddots & \\ \vdots & & v_{IP} \end{pmatrix} \begin{pmatrix} v_{11} & v_{21} & v_{31} & \cdots \\ v_{12} & \ddots & \ddots & \\ \vdots & & & v_{IP} \end{pmatrix} \tag{24}$$

$$= \begin{pmatrix} \sum_p v_{1p}v_{1p} & \sum_p v_{1p}v_{2p} & \cdots \\ \sum_p v_{2p}v_{1p} & \ddots & \\ \vdots & & \sum_p v_{Ip}v_{Ip} \end{pmatrix} \tag{25}$$

$$= \sum_{p=1}^{P} \begin{pmatrix} v_{1p}v_{1p} & v_{1p}v_{2p} & \cdots \\ v_{2p}v_{1p} & \ddots & \\ \vdots & & v_{Ip}v_{Ip} \end{pmatrix} \tag{26}$$

$$\circ \quad = \sum_{p=1}^{P} \mathbf{v}_p\mathbf{v}_p^T\,. \tag{27}$$

$\mathbf{P}$ is obviously an $I \times I$-matrix. **If $P = I$ then** projecting from the old to the new and back to the old coordinate system causes no information loss and $\mathbf{P} = \mathbf{1}_I$. **The smaller P the more information is lost** and the more does $\mathbf{P}$ differ from the identity matrix. Consider,

for example

$$\mathbf{v}_1 := \frac{1}{2}(\sqrt{2},-1,1)^T \;\Rightarrow\; \mathbf{v}_1\mathbf{v}_1^T = \frac{1}{4}\begin{pmatrix} 2 & -\sqrt{2} & \sqrt{2} \\ -\sqrt{2} & 1 & -1 \\ \sqrt{2} & -1 & 1 \end{pmatrix},$$

$$\mathbf{v}_2 := \frac{1}{2}(0,\sqrt{2},\sqrt{2})^T \;\Rightarrow\; \mathbf{v}_2\mathbf{v}_2^T = \frac{1}{4}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 2 & 2 \\ 0 & 2 & 2 \end{pmatrix}, \text{ and}$$

$$\mathbf{v}_3 := \frac{1}{2}(-\sqrt{2},-1,1)^T \;\Rightarrow\; \mathbf{v}_3\mathbf{v}_3^T = \frac{1}{4}\begin{pmatrix} 2 & \sqrt{2} & -\sqrt{2} \\ \sqrt{2} & 1 & -1 \\ -\sqrt{2} & -1 & 1 \end{pmatrix}$$

for which you can easily verify that $\mathbf{P}$ (27) successively becomes the identity matrix as you take more of the $\mathbf{v}_p\mathbf{v}_p^T$-terms.

2.6 Variance

The variance of a multi-dimensional data set is defined as the sum over the variances of its components. Since we assume zero-mean data, we have

$$\bullet \quad \mathrm{var}(\mathbf{x}) := \sum_{i=1}^{I} \langle x_i^2 \rangle \tag{28}$$

$$\circ \quad = \left\langle \sum_{i=1}^{I} x_i^2 \right\rangle \tag{29}$$

$$\bullet \quad = \langle \mathbf{x}^T\mathbf{x} \rangle \tag{30}$$

This also holds for the projected data, of course, $\mathrm{var}(\mathbf{y}) = \langle \mathbf{y}^T\mathbf{y} \rangle$.

2.7 Reconstruction error

The reconstruction error E is defined as the mean square sum over the distances between the original data points $\mathbf{x}$ and the projected ones $\mathbf{x}_\parallel$. If we define the *orthogonal vectors* (D: Lotvektoren)

$$\circ \quad \mathbf{x}_\perp = \mathbf{x} - \mathbf{x}_\parallel \tag{31}$$

(in contrast to the projected vectors $\mathbf{x}_\parallel$) we can write **the reconstruction error** as the variance of the orthogonal vectors and find

$$
\begin{aligned}
\bullet|\circ \quad E &\overset{(8,31)}{=} \langle {\mathbf{x}_\perp}^T \mathbf{x}_\perp \rangle && (32)\\
&\overset{(31)}{=} \langle (\mathbf{x}-\mathbf{x}_\parallel)^T (\mathbf{x}-\mathbf{x}_\parallel) \rangle && (33)\\
\circ &\overset{(15)}{=} \langle (\mathbf{x}-\mathbf{V}\mathbf{V}^T\mathbf{x})^T (\mathbf{x}-\mathbf{V}\mathbf{V}^T\mathbf{x}) \rangle && (34)\\
\circ &= \langle \mathbf{x}^T\mathbf{x} - 2\mathbf{x}^T\mathbf{V}\mathbf{V}^T\mathbf{x} + \mathbf{x}^T\mathbf{V}\underbrace{(\mathbf{V}^T\mathbf{V})}_{=\mathbf{1}_P}\mathbf{V}^T\mathbf{x} \rangle && (35)\\
\circ &\overset{(17)}{=} \langle \mathbf{x}^T\mathbf{x} \rangle - \langle \mathbf{x}^T\mathbf{V}\underbrace{(\mathbf{V}^T\mathbf{V})}_{=\mathbf{1}_P}\mathbf{V}^T\mathbf{x} \rangle && (36)\\
\circ &\overset{(15)}{=} \langle \mathbf{x}^T\mathbf{x} \rangle - \langle {\mathbf{x}_\parallel}^T \mathbf{x}_\parallel \rangle && (37)\\
\bullet &\overset{(36,13)}{=} \langle \mathbf{x}^T\mathbf{x} \rangle - \langle \mathbf{y}^T\mathbf{y} \rangle \,. && (38)
\end{aligned}
$$

This means that the reconstruction error **equals the difference between the variance of the data minus the variance of the projected data**. Thus, this verifies our intuition that minimizing the reconstruction error is equivalent to maximizing the variance of the projected data.

2.8 Covariance matrix

We have already argued heuristically that the covariance matrix $\mathbf{C}_x$ with $C_{xij} := \langle x_i x_j \rangle$ plays an important role in performing PCA. It is convenient to write the **covariance matrix in vector notation**:

$$
\bullet \quad \mathbf{C}_x := \left\langle \mathbf{x}\mathbf{x}^T \right\rangle \;=\; \frac{1}{M}\sum_\mu \mathbf{x}^\mu \mathbf{x}^{\mu T} \,. \tag{39}
$$

It is an easy exercise to show that this definition is equivalent to the componentwise one given above. Since $(\mathbf{x}\mathbf{x}^T)^T = \mathbf{x}\mathbf{x}^T$ (remember $(\mathbf{A}\mathbf{B})^T = \mathbf{B}^T\mathbf{A}^T$), one can also see that $\mathbf{C}_x$ **is symmetric**, i.e. $\mathbf{C}_x^T = \mathbf{C}_x$.

2.9 Eigenvalue equation of the covariance matrix

Since the covariance matrix is symmetric, its eigenvalues are real and a set of orthogonal eigenvectors always exists. In mathematical terms, **for a given covariance matrix $\mathbf{C}_x$ we can always find a complete set of real eigenvalues λ_i and corresponding eigenvectors $\mathbf{u}_i$ such that**

$$
\begin{aligned}
\circ \quad \mathbf{C}_x\mathbf{u}_i &= \mathbf{u}_i\lambda_i \quad \text{(eigenvalue equation)}\,, && (40)\\
\bullet \quad \lambda_i &\geq \lambda_{i+1} \quad \text{(eigenvalues are ordered)}\,, && (41)\\
\circ \quad \mathbf{u}_i^T\mathbf{u}_j &= \delta_{ij} \quad \text{(eigenvectors are orthonormal)}\,. && (42)
\end{aligned}
$$

If we combine the eigenvectors into an orthogonal matrix $\mathbf{U}$ and the eigenvalues into a diagonal matrix $\mathbf{\Lambda}$,

$$\bullet \quad \mathbf{U} := (\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_I)\,, \tag{43}$$

$$\bullet \quad \mathbf{\Lambda} := \mathrm{diag}(\lambda_1, \lambda_2, ..., \lambda_I)\,, \tag{44}$$

then we can rewrite (42) and (40) as

$$\bullet \quad \mathbf{U}^T\mathbf{U} \overset{(42,43)}{=} \mathbf{1}_I \quad \text{(matrix } \mathbf{U} \text{ is orthogonal)}\,, \tag{45}$$

$$\circ \quad \Longleftrightarrow \quad \mathbf{U}\mathbf{U}^T = \mathbf{1}_I \quad \text{(since } \mathbf{U}^{-1} = \mathbf{U}^T \text{ and } \mathbf{U} \text{ is quadratic)}\,, \tag{46}$$

$$\bullet \quad \mathbf{C}_x\mathbf{U} \overset{(40,43,44)}{=} \mathbf{U}\mathbf{\Lambda} \quad \text{(eigenvalue equation)}\,, \tag{47}$$

$$\circ \quad \overset{(45)}{\Longleftrightarrow} \mathbf{U}^T\mathbf{C}_x\mathbf{U} = \mathbf{\Lambda} \tag{48}$$

$$\circ \quad \overset{(45,46)}{\Longleftrightarrow} \mathbf{C}_x = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T\,. \tag{49}$$

2.10 Total variance of the data x

Given the eigenvector matrix $\mathbf{U}$ and the eigenvalue matrix $\mathbf{\Lambda}$ it is easy to compute the total variance of the data

$$\bullet|\circ \quad \langle \mathbf{x}^T\mathbf{x} \rangle = \langle \mathrm{tr}(\mathbf{x}^T\mathbf{x}) \rangle \quad (\text{since } s = \mathrm{tr}(s) \text{ for any scalar } s) \tag{50}$$

$$\circ = \langle \mathrm{tr}(\mathbf{x}\mathbf{x}^T) \rangle \quad (\text{since } \mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA}) \text{ for any matrices } \mathbf{A}, \mathbf{B}) \tag{51}$$

$$\circ = \mathrm{tr}(\langle \mathbf{x}\mathbf{x}^T \rangle) \quad (\text{since } \mathrm{tr}(\cdot) \text{ and } \langle\cdot\rangle \text{ commute}) \tag{52}$$

$$\circ \overset{(39)}{=} \mathrm{tr}(\mathbf{C}_x) \tag{53}$$

$$\circ \overset{(46)}{=} \mathrm{tr}(\mathbf{U}\mathbf{U}^T\mathbf{C}_x) \tag{54}$$

$$\circ = \mathrm{tr}(\mathbf{U}^T\mathbf{C}_x\mathbf{U}) \tag{55}$$

$$\circ \overset{(48)}{=} \mathrm{tr}(\mathbf{\Lambda}) \tag{56}$$

$$\bullet \overset{(44)}{=} \sum_i \lambda_i\,. \tag{57}$$

Thus **the total variance of the data is simply the sum of the eigenvalues of its covariance matrix**.

Notice that on the way of this proof we have shown some very general properties. From line (50) to (53) we have shown that the total variance of some multi-dimensional data equals the trace of its covariance matrix. From line (53) to (55) we have shown that the trace remains invariant under any orthogonal transformation of the coordinate system. This implies that the total variance of some multi-dimensional data is invariant under any orthogonal transformation such as a rotation. This is intuitively clear.

2.11 Diagonalizing the covariance matrix

We can now use matrix $\mathbf{U}$ to **transform the data such that the covariance matrix becomes diagonal.** Define $\mathbf{x}' := \mathbf{U}^T\mathbf{x}$ and denote the new covariance matrix by $\mathbf{C}'_x$. We

have

$$\bullet \quad \mathbf{x}' := \mathbf{U}^T\mathbf{x} \tag{58}$$

$$\bullet|\circ \quad \mathbf{C}'_x := \left\langle \mathbf{x}'\mathbf{x}'^T \right\rangle \tag{59}$$

$$\circ \overset{(58)}{=} \left\langle (\mathbf{U}^T\mathbf{x})(\mathbf{U}^T\mathbf{x})^T \right\rangle \tag{60}$$

$$\circ = \mathbf{U}^T \left\langle \mathbf{x}\mathbf{x}^T \right\rangle \mathbf{U} \tag{61}$$

$$\circ \overset{(39)}{=} \mathbf{U}^T\mathbf{C}_x\mathbf{U}\,, \tag{62}$$

$$\bullet \overset{(48)}{=} \mathbf{\Lambda} \tag{63}$$

and find that the transformed data $\mathbf{x}'$ have a diagonal covariance matrix. Working with $\mathbf{x}'$ instead of $\mathbf{x}$ will simplify further analysis without loss of generality.

2.12 Variance of y for a diagonalized covariance matrix

Now that we have the data represented in a coordinate system in which the covariance matrix is diagonal, we can try to answer the question, which is the P-dimensional subspace that minimizes the reconstruction error. Our intuition would predict that it is simply the space spanned by the first P eigenvectors. To show this analytically, we **take an arbitrary set of P orthonormal vectors $\mathbf{v}'_p$, and with $\mathbf{V}' := (\mathbf{v}'_1, \mathbf{v}'_2, ..., \mathbf{v}'_P)$** we **compute** the variance of $\mathbf{y}$.

$$\circ \quad \mathbf{y} := \mathbf{V}'^T\mathbf{x}' \tag{64}$$

$$\circ \implies \langle \mathbf{y}^T\mathbf{y} \rangle \overset{(64)}{=} \langle \mathbf{x}'^T\mathbf{V}'\mathbf{V}'^T\mathbf{x}' \rangle \tag{65}$$

$$= \langle \mathrm{tr}(\mathbf{x}'^T\mathbf{V}'\mathbf{V}'^T\mathbf{x}') \rangle \quad (\text{since } s = \mathrm{tr}(s) \text{ for any scalar } s) \tag{66}$$

$$= \langle \mathrm{tr}(\mathbf{V}'^T\mathbf{x}'\mathbf{x}'^T\mathbf{V}') \rangle \quad (\text{since } \mathrm{tr}(\mathbf{ABC}) = \mathrm{tr}(\mathbf{BCA}) \text{ if defined}) \tag{67}$$

$$\circ \overset{(59)}{=} \mathrm{tr}(\mathbf{V}'^T\mathbf{C}'_x\mathbf{V}') \quad (\text{since } \mathrm{tr}(\cdot) \text{ and } \langle\cdot\rangle \text{ commute}) \tag{68}$$

$$\circ \overset{(63)}{=} \mathrm{tr}(\mathbf{V}'^T\mathbf{\Lambda}\mathbf{V}') \tag{69}$$

$$\circ = \sum_i \lambda_i \sum_p (v'_{ip})^2\,. \quad (\text{as one can work out on a sheet of paper}) \tag{70}$$

2.13 Constraints of matrix $\mathbf{V}'$

Note that, **since the vectors $\mathbf{v}'_p$ are orthonormal, $\mathbf{V}'$** can always be completed to an orthogonal $I \times I$-matrix by adding $I - P$ additional orthonormal vectors. Since we know that an orthogonal matrix has normalized row as well as column vectors, we see that, by taking away the $I - P$ additional column vectors, we are left with the constraints

$$\circ \quad \sum_i (v'_{ip})^2 = 1 \quad (\text{column vectors of } \mathbf{V}' \text{ have norm one})\,, \tag{71}$$

$$\circ \implies \sum_{ip} (v'_{ip})^2 = P \quad (\text{square sum over all matrix elements equals } P)\,, \tag{72}$$

$$\circ \quad \sum_p (v'_{ip})^2 \leq 1 \quad (\text{row vectors of } \mathbf{V}' \text{ have norm less or equal one})\,. \tag{73}$$

Notice that Constraint (72) is a direct consequence of Constraint (71) and does not need to be verified separately in the following considerations.

2.14 Finding the optimal subspace

Since the variance (70) of y as well as the constraints (71, 72, 73) of Matrix $\mathbf{V}'$ are linear in $(v'_{ip})^2$, maximization of the variance $\langle \mathbf{y}^T\mathbf{y} \rangle$ is obviously achieved by putting as much 'weight' as possible on the large eigenvalues, which are the first ones. The simplest way of doing that is to set

$$\circ \quad v'_{ip} := \delta_{ip} := \begin{cases} 1 & \text{if } i = p \\ 0 & \text{otherwise} \end{cases} , \tag{74}$$

with the Kronecker symbol δ_{ip}.

Since $I \geq P$ we can verify the constraints

$$\sum_p (v'_{ip})^2 \overset{(74)}{=} \sum_p \delta_{ip}^2 = \left\{ \begin{array}{ll} 1 & \text{if } i \leq P \\ 0 & \text{otherwise} \end{array} \right\} \leq 1 , \tag{75}$$

$$\sum_i (v'_{ip})^2 \overset{(74)}{=} \sum_i \delta_{ip}^2 = \delta_{pp}^2 = 1 , \tag{76}$$

and see from (75) that there is actually as much 'weight' on the first, i.e. large, eigenvalues as Constraint (73) permits.

2.15 Interpretation of the result

What does it mean to set $v'_{ip} := \delta_{ip}$? **It means that $\mathbf{V}'$ projects the data $\mathbf{x}'$ onto the first P axes, which in fact is a projection onto the first P eigenvectors of the covariance matrix $\mathbf{C}_x$.** Thus, **if we define**

$$\bullet|\circ \quad \mathbf{V} := \mathbf{U}\mathbf{V}' \tag{77}$$

$$\bullet \overset{(43,\,74)}{=} (\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_P) \tag{78}$$

we can go back to the original coordinate system and **find**

$$\bullet|\circ \quad \mathbf{y} \overset{(64)}{=} \mathbf{V}'^T\mathbf{x}' \tag{79}$$

$$\circ \overset{(58)}{=} \mathbf{V}'^T\mathbf{U}^T\mathbf{x} \tag{80}$$

$$\bullet \overset{(77)}{=} \mathbf{V}^T\mathbf{x} , \tag{81}$$

which we know **has maximal variance.** Thus, if we start from the original data $\mathbf{x}$ we would set $\mathbf{v}_p := \mathbf{u}_p$.

The variance of y is

$$\bullet|\circ \quad \langle \mathbf{y}^T \mathbf{y} \rangle \overset{(70)}{=} \sum_{i=1}^{I} \lambda_i \sum_{p=1}^{P} (v'_{ip})^2 \tag{82}$$

$$\circ \overset{(74)}{=} \sum_{i=1}^{I} \lambda_i \sum_{p=1}^{P} \delta_{ip}^2 \tag{83}$$

$$\bullet = \sum_{i=1}^{P} \lambda_i \,, \tag{84}$$

which is the sum over the first P largest eigenvalues of the covariance matrix. Likewise one can determine **the reconstruction error** as

$$\bullet|\circ \quad E \overset{(38)}{=} \langle \mathbf{x}^T \mathbf{x} \rangle - \langle \mathbf{y}^T \mathbf{y} \rangle \tag{85}$$

$$\circ \overset{(57,\ 84)}{=} \sum_{i=1}^{I} \lambda_i - \sum_{j=1}^{P} \lambda_j \tag{86}$$

$$\bullet = \sum_{i=P+1}^{I} \lambda_i \,. \tag{87}$$

Notice that this is just one optimal set of weights. We have seen above that **the projected data**, like any multi-dimensional data, **can be rotated arbitrarily without changing its variance** and therefore without changing its reconstruction error. This is equivalent to a rotation of the projection vectors $\mathbf{v}_p$ within the space spanned by the first eigenvectors.

2.16 PCA Algorithm

2.17 Intuition of the Results

Eigenvalue spectrum.

Projection onto a low-dimensional eigenspace.

Visualization of eigenvectors.

2.18 Whitening or sphering

Sometimes it is desirable **to transform a data set such that it has variance one in all directions**. Such a normalization operation **is called *whitening* or *sphering*.** The latter term is quite intuitive, because a spherical data distribution has the same variance in all directions. Intuitively speaking sphering requires to stretch and compress the data distribution along the axes of the principal components such they have variance one. Technically speaking one first rotates the data into a coordinate system where the covariance matrix is diagonal, then performs the stretching along the axes, and then rotates the data back into the original coordinate system. Principal component analysis obviously gives all the required information. The eigenvectors of the covariance matrix provide the axes of the

new coordinate system and the eigenvalues λ_i indicate the variances and therefore how much one has to stretch the data. If the original variance is λ_i then one obviously has to stretch by a factor of $1/\sqrt{\lambda_i}$ to get variance one. Thus, **sphering is achieved by multiplying the data with a sphering matrix**

$$\bullet \quad \mathbf{W} := \mathbf{U}\,\mathrm{diag}\left(\frac{1}{\sqrt{\lambda_1}}, \frac{1}{\sqrt{\lambda_2}}, ..., \frac{1}{\sqrt{\lambda_I}}\right)\mathbf{U}^T \tag{88}$$

$$\bullet \quad \hat{\mathbf{x}} := \mathbf{W}\mathbf{x}\,. \tag{89}$$

If the final orientation of the data does not matter, the sphering matrix is often defined without the first $\mathbf{U}$. It is easy to verify that **the sphering matrix is symmetrical, the sphered data $\hat{\mathbf{x}}$ have a unit covariance matrix**,

$$\begin{aligned}
\mathbf{C}_{\hat{\mathbf{x}}} &:= \langle \hat{\mathbf{x}}\hat{\mathbf{x}}^T \rangle && (90)\\
&\overset{(89)}{=} \mathbf{W}\langle \mathbf{x}\mathbf{x}^T \rangle \mathbf{W}^T && (91)\\
&\overset{(39,88)}{=} \mathbf{U}\,\mathrm{diag}\left(\frac{1}{\sqrt{\lambda_i}}\right)\mathbf{U}^T\,\mathbf{C}_{\mathbf{x}}\,\mathbf{U}\,\mathrm{diag}\left(\frac{1}{\sqrt{\lambda_i}}\right)\mathbf{U}^T && (92)\\
&\overset{(48)}{=} \mathbf{U}\,\mathrm{diag}\left(\frac{1}{\sqrt{\lambda_i}}\right)\mathbf{\Lambda}\,\mathrm{diag}\left(\frac{1}{\sqrt{\lambda_i}}\right)\mathbf{U}^T && (93)\\
&\overset{(44)}{=} \mathbf{U}\,\mathbf{1}\,\mathbf{U}^T && (94)\\
&\overset{(46)}{=} \mathbf{1}\,, && (95)
\end{aligned}$$

and they have variance one in all directions, since for any projection vector $\mathbf{n}$ of norm one the variance $\langle (\mathbf{n}^T\hat{\mathbf{x}})^2 \rangle$ of the projected data is

$$\begin{aligned}
\langle (\mathbf{n}^T\hat{\mathbf{x}})^2 \rangle &= \mathbf{n}^T \langle \hat{\mathbf{x}}\hat{\mathbf{x}}^T \rangle \mathbf{n} && (96)\\
&\overset{(95)}{=} \mathbf{n}^T\mathbf{n} && (97)\\
&= 1 && (98)
\end{aligned}$$

Similarly one can show that **the sphered data projected onto two orthogonal vectors are uncorrelated**.

2.19 Singular value decomposition +

Sometimes one has fewer data points than dimensions. For instance one might have 100 images with 10000 pixels each. Then doing direct PCA is very inefficient and the following method, known as *singular value decomposition* (SVD), is helpful.

Let $\mathbf{x}^\mu, \mu = 1, ..., M$ be the I-dimensional data with $M < I$. For convenience we write the data in one $I \times M$-matrix

$$\mathbf{X} := (\mathbf{x}^1, ..., \mathbf{x}^M)\,. \tag{99}$$

The second-moment matrix can then be written as

$$\mathbf{C}_1 := \mathbf{X}\mathbf{X}^T/M\,, \tag{100}$$

and its eigenvalue equation and decomposition read

$$\begin{aligned} \mathbf{C}_1\mathbf{U}_1 &= \mathbf{U}_1\mathbf{\Lambda}_1 && (101)\\ \iff \quad \mathbf{C}_1 &= \mathbf{U}_1\mathbf{\Lambda}_1\mathbf{U}_1^T\,. && (102) \end{aligned}$$

The data represented in the coordinate system of the eigenvectors is

$$\mathbf{Y}_1 \;:=\; \mathbf{U}_1^T\mathbf{X}\,, \tag{103}$$

which is still high-dimensional.

Now interpret the data matrix $\mathbf{X}$ transposed, i.e. swap the data point index for the dimension index. In our example this would correspond to having 10000 data points in a 100-dimensional space, which is, of course, much easier to deal with. We get the same equations as above just with $\mathbf{X}$ transposed.

$$\begin{aligned} \mathbf{C}_2 &:= \mathbf{X}^T\mathbf{X}/I\,, && (104)\\ \mathbf{C}_2\mathbf{U}_2 &= \mathbf{U}_2\mathbf{\Lambda}_2 && (105)\\ \iff \quad \mathbf{C}_2 &= \mathbf{U}_2\mathbf{\Lambda}_2\mathbf{U}_2^T\,, && (106)\\ \mathbf{Y}_2 &:= \mathbf{U}_2^T\mathbf{X}^T\,. && (107) \end{aligned}$$

The interesting property of matrix $\mathbf{Y}_2$ now is that its rows are eigenvectors of matrix $\mathbf{C}_1$, as can be shown easily.

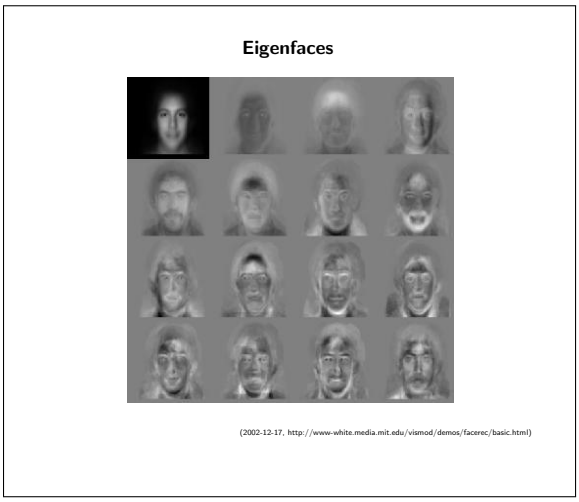
$$\begin{aligned} \mathbf{C}_1\mathbf{Y}_2^T &\overset{(100,107)}{=} (\mathbf{X}\mathbf{X}^T/M)(\mathbf{X}\mathbf{U}_2) && (108)\\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X}/I)\mathbf{U}_2 I/M && (109)\\ &\overset{(104)}{=} \mathbf{X}\mathbf{C}_2\mathbf{U}_2 I/M && (110)\\ &\overset{(105)}{=} \mathbf{X}\mathbf{U}_2\mathbf{\Lambda}_2 I/M && (111)\\ &\overset{(107)}{=} \mathbf{Y}_2^T\mathbf{\Lambda}_2 I/M\,. && (112) \end{aligned}$$

The corresponding eigenvalues are eigenvalues of $\mathbf{C}_2$ scaled by I/M.

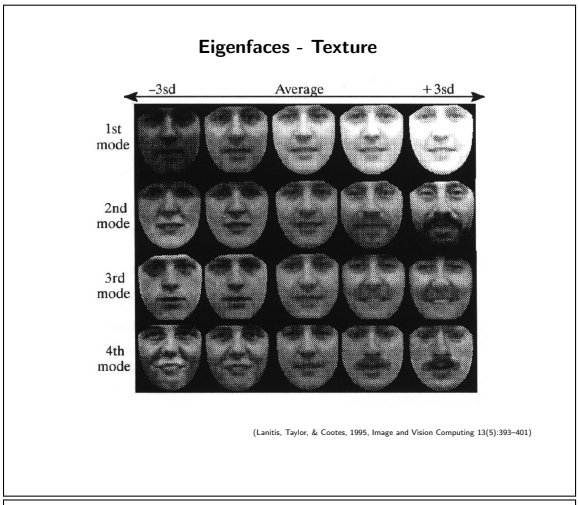
However, $\mathbf{Y}_2$ yields only M eigenvectors and eigenvalues. The other eigenvalues are all zero, because M data points can only produce M non-zero variance dimensions or, in other words, M data points together with the origin can only span an M-dimensional subspace. The missing $(I-M)$ eigenvectors must all be orthogonal to the first M ones and orthogonal to each other but can otherwise be quite arbitrary, since their eigenvalues are all equal. A Gram-Schmidt orthogonalization procedure can be used to generate them.

3 Application

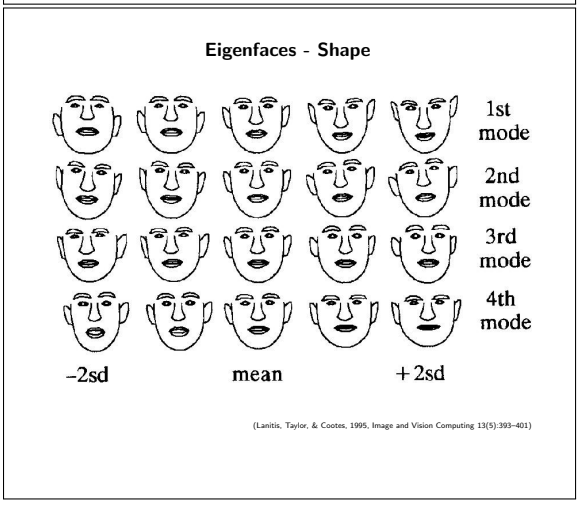
3.1 Face processing



If one carefully shifts and scales many face images such that the eyes are in register, i.e. at identical positions, and performs PCA (or SVD) on them, meaningful eigenvectors can be calculated and are called eigenfaces. These are the principal grey value variations that distinguish faces from each other. The figure shows first the mean and then the eigenfaces ordered in rows. The first eigenface obviously accounts for the mean grey value, the second one for the difference in color between hair and face, the third one for illumination from the side, the fourth and seventh one at least partially for a beard. The higher components become increasingly difficult to interpret. The projection of face images onto the first eigenfaces is a suitable representation for face recognition, (cf. Turk and Pentland, 1991).



In this example (Lanitis et al., 1995) many faces were warped to a standard geometry and then PCA was applied to calculate the eigenvectors, which are called eigenfaces. The Figure visualizes the first four eigenfaces by varying the average face (middle) along the direction of the eigenvectors by up to ± 3 standard deviations. The first eigenface (mode) accounts for overall illumination, the other three for some combination of beard, gender, and mimic expression.



In this example (Lanitis et al., 1995) PCA was applied to the geometry of faces. A graph with some standard structure was mapped onto many faces and the concatenated vector of xy-positions of the nodes (not shown in the graphs) of the graphs serve as the data for PCA. Visualization is done relative to the average geometry (middle) by adding the eigenvectors up to ± 2 standard deviations. The first three components mainly account for the orientation of the head along the three rotational axes. The fourth component accounts for some variation in width and mimic expression.

4 Acknowledgment

I thank Agnieszka Grabska-Barwinska for working out the proof of singular value decomposition.

References

Lanitis, A., Taylor, C. J., and Cootes, T. F. (1995). An automatic face identification system using flexible appearance models. *Image and Vision Computing*, 13(5):393–401.

Turk, M. and Pentland, A. (1991). Eigenfaces for recognition. *J. of Cognitive Neuroscience*, 3(1):71–86.