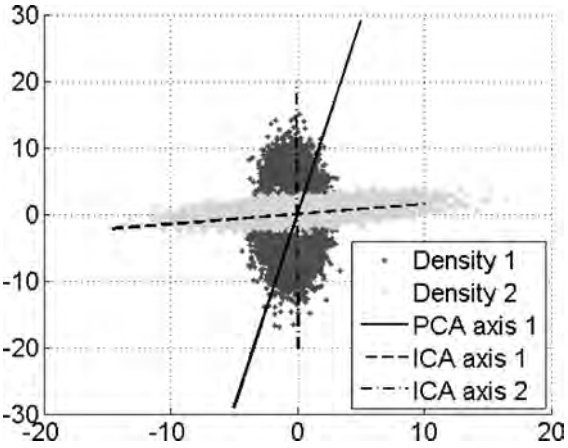

Figure 8.1 Principal and independent components for two bivariate normal densities. The second PCA axis is orthogonal to the one shown. The two ICA axes are aligned with the stretch directions.

resentation of the data. Whether such elimination is possible must be decided on the merits of a specific analysis. Techniques for choosing the optimal number of principal components are described in Section 8.3.4. Here, dimensionality reduction occurs in the transformed space. Reducing the number of original, nontransformed variables is discussed in Chapter 18.

The two rotations described in this chapter disregard class labels. They apply to the entire data, signal and background, and therefore are not used for supervised learning. Sometimes analysts apply principal component analysis and claim that it improves separation of signal and background by a consequent classification algorithm. Such an improvement can occur by accident but is not guaranteed; the effect of the rotation could be just the opposite. Class-conscious linear techniques are described in Chapter 11.

8.3 Principal Component Analysis (PCA)

Principal component analysis is one of the oldest statistical tools. It was proposed in Pearson (1901). Similar formalism was later developed in other fields. The terms "principal component analysis", "Hotelling transform" (Hotelling, 1933) and "Karhunen–Loeve transform" sometimes mean the same thing and sometimes mean slightly different things, depending on what is viewed as the "standard" PCA. We follow the approach by Hotelling using modern notation. PCA is described in many textbooks.

Nonlinear PCA, factor analysis (PCA with noise), and other extensions of PCA, although of potential interest to physics analysis, are not covered here.

In-depth reviews of PCA can be found in Jolliffe (2002) and Abdi and Williams (2010).

8.3.1 Theory

Suppose $\boldsymbol{X}$ is a random column vector with D elements. Without loss of generality, we can assume that the expectation of any element in $\boldsymbol{X}$ is 0: $E\boldsymbol{X} = \mathbf{0}_{D\times 1}$. If this is not the case, we can center $\boldsymbol{X}$ by subtracting its expectation.

We seek a linear transformation $\mathbf{W}$ such that the transformed variables in vector $\boldsymbol{Z}$,

$$\boldsymbol{Z} = \mathbf{W}\boldsymbol{X}\,, \tag{8.1}$$

are uncorrelated. Here, $\mathbf{W}$ is a $D \times D$ matrix, and $\boldsymbol{Z}$ has the same dimensionality as $\boldsymbol{X}$. If the expectation of any element in $\boldsymbol{X}$ is 0, so is the expectation of any element in $\boldsymbol{Z}$. Two scalar random variables with zero expected values are uncorrelated if their expected product is zero. In matrix notation, we require that the covariance matrix

$$\boldsymbol{\Sigma}_{ZZ} \equiv E(\boldsymbol{Z}\boldsymbol{Z}^{\mathsf{T}}) \tag{8.2}$$

be diagonal. Since $\boldsymbol{Z}$ is a column vector, $\boldsymbol{Z}\boldsymbol{Z}^{\mathsf{T}}$ is a $D \times D$ matrix. If we substitute $\boldsymbol{Z} = \mathbf{W}\boldsymbol{X}$ in the equation above, we obtain

$$\boldsymbol{\Sigma}_{ZZ} = \mathbf{W}\boldsymbol{\Sigma}_{XX}\mathbf{W}^{\mathsf{T}}\,, \tag{8.3}$$

or equivalently

$$\boldsymbol{\Sigma}_{XX} = \mathbf{V}\boldsymbol{\Sigma}_{ZZ}\mathbf{V}^{\mathsf{T}}\,, \tag{8.4}$$

where $\mathbf{V}$ is the inverse of $\mathbf{W}$. To avoid writing the sub-indices from now on, we define $\boldsymbol{\Sigma}_{XX} \equiv \boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_{ZZ} \equiv \boldsymbol{\Lambda}$ and finally obtain

$$\boldsymbol{\Sigma} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathsf{T}}\,. \tag{8.5}$$

The right-hand side gives an eigenvalue decomposition (EVD) of the covariance matrix $\boldsymbol{\Sigma}$. Matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Lambda}$ have identical eigenvalues and are said to be similar. Matrix $\mathbf{V}$ is orthogonal, $\mathbf{V}^{\mathsf{T}}\mathbf{V} = \mathbf{V}\mathbf{V}^{\mathsf{T}} = \mathbf{I}_{D\times D}$, where $\mathbf{I}$ is an identity matrix with 1 on the main diagonal and 0 elsewhere.

The covariance matrix $\boldsymbol{\Sigma}$ is symmetric and therefore can be always decomposed in this form. It is trivial to show that the covariance matrix is positive semidefinite, that is, $\boldsymbol{a}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{a} \geq 0$ for any vector $\boldsymbol{a}$. The eigenvalues of $\boldsymbol{\Sigma}$, or equivalently the diagonal elements of $\boldsymbol{\Lambda}$, are nonnegative. It is easy to see that the mth diagonal element of $\boldsymbol{\Lambda}$ gives the variance of $\boldsymbol{X}$ along the mth principal component. In the real world, EVD is subject to numerical errors. Sometimes your software application can find negative eigenvalues, especially if you work in high dimensions.

Base vectors in the $\boldsymbol{Z}$ space are called *principal components*. To get the mth principal component, set the mth element of $\boldsymbol{z}$ to 1 and the rest to 0. Directions of the principal components in the $\boldsymbol{X}$ space, known as *loadings* in the statistics literature, are given by the columns of matrix $\mathbf{V}$. Projections of a column vector $\boldsymbol{x}$ onto the principal axes, known as *scores* in the statistics literature, can be found by taking D dot products, $\boldsymbol{x}^{\mathrm{T}}\mathbf{V}$.

The eigenvalue decomposition is defined up to a permutation of diagonal elements in $\boldsymbol{\Lambda}$ with a corresponding permutation of columns in $\mathbf{V}$. If we sort the diagonal elements in $\boldsymbol{\Lambda}$ in descending order, $\lambda_{11} \geq \lambda_{22} \geq \ldots \geq \lambda_{DD}$, we commit to a unique ordering of eigenvalues. The decomposition is still not unique as we can flip the sign of any column in $\mathbf{V}$.

PCA can be applied to the covariance matrix $\boldsymbol{\Sigma}$ or to the correlation matrix $\boldsymbol{C} = \boldsymbol{\Omega}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{\Omega}^{-1/2}$, where $\boldsymbol{\Omega} = \mathbf{diag}(\boldsymbol{\Sigma})$ is a matrix with the main diagonal set to that of $\boldsymbol{\Sigma}$ and all off-diagonal elements set to zero. The choice between *covariance PCA* and *correlation PCA* should be based on the nature of analyzed variables. If the variables are measured in different units and have substantially different standard deviations, correlation PCA should be preferred. If the variables are measured in a similar fashion and their standard deviations can be compared to each other in a meaningful way, covariance PCA would be appropriate.

If $\boldsymbol{X}$ is drawn from a multivariate normal distribution, its principal components are aligned with the orthogonal normal axes. If two normal random variables are uncorrelated, they are independent. On occasion you can hear people say that PCA requires multivariate normality and finds independent variables. This merely describes one particular, although important, case of PCA. In general, PCA is not restricted to normal distributions. Independence is a stronger requirement than zero correlation. For normal random variables, these two requirements happen to be equivalent. For nonnormal variables, a PCA transformation does not necessarily produce new independent variables. A technique that attempts to obtain independent variables by a linear transformation is called Independent Component Analysis and described in Section 8.4.

The derivation of PCA shown here is due to Hotelling. Originally, PCA was proposed by Pearson who approached from a different angle. Define a rotation $\boldsymbol{Z} = \mathbf{W}\boldsymbol{X}$. Transform $\boldsymbol{Z}$ back to the original space $\boldsymbol{X} = \mathbf{W}^{\mathrm{T}}\boldsymbol{Z}$ and take $E(||\boldsymbol{X} - \mathbf{W}^{\mathrm{T}}\mathbf{W}\boldsymbol{X}||^2)$ to be the reconstruction error induced by the transformation $\mathbf{W}$. Set the reconstruction error to zero by choosing $\mathbf{W} = \mathbf{V}^{\mathrm{T}}$ with $\mathbf{V}$ defined in (8.5).

8.3.2 Numerical Implementation

In practice, the covariance matrix $\boldsymbol{\Sigma}$ usually needs to be estimated from data. Let $\mathbf{X}$ be an $N \times D$ matrix with one row per observation and one column per variable. Think of $\mathbf{X}$ as a set of observed instances of the random column vector $\boldsymbol{X}$, transposed and concatenated vertically. First, center $\mathbf{X}$ by subtracting the mean of every column from all elements in this column. Since most usually the true mean is not known, use the observed mean instead. Then estimate the covariance matrix by

putting $\hat{\boldsymbol{\Sigma}} = \mathbf{X}^{\mathsf{T}}\mathbf{X}/(N-1)$. The $1/(N-1)$ factor is needed to get an unbiased estimate of the covariance matrix assuming a multivariate normal distribution. It is absorbed in the definition of $\boldsymbol{\Lambda}$ and has no effect on the principal components.

Numerically, PCA can be carried out in various ways. Modern software packages often use singular value decomposition (SVD). One advantage of SVD is not having to compute the covariance matrix $\mathbf{X}^{\mathsf{T}}\mathbf{X}$. Instead we decompose

$$\frac{\mathbf{X}}{\sqrt{N-1}} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathsf{T}} . \tag{8.6}$$

In the full SVD decomposition, $\mathbf{U}$ is of size $N \times N$, $\mathbf{S}$ is of size $N \times D$, and $\mathbf{V}$ is of size $D \times D$. Matrices $\mathbf{U}$ and $\mathbf{V}$ are orthogonal, and matrix $\mathbf{S}$ is diagonal. A nonquadratic diagonal matrix is defined by putting $s_{ij} = 0$ for any pair i and j except $i = j$.

In physics analysis, $\mathbf{X}$ is often very tall, $N \gg D$. Computing an $N \times N$ matrix $\mathbf{U}$ in this case can consume a lot of memory and time. Fortunately, this is not necessary; because $N - D$ bottom rows of $\mathbf{S}$ are filled with zeros, the last $N - D$ columns of $\mathbf{U}$ can be discarded. In the thin version of SVD, $\mathbf{U}$ is $N \times D$ and $\mathbf{S}$ is $D \times D$. Note that $\mathbf{U}$ is no longer orthogonal: $\mathbf{U}^{\mathsf{T}}\mathbf{U} = \mathbf{I}_{D\times D}$ holds, but $\mathbf{U}\mathbf{U}^{\mathsf{T}}$ is not an identity matrix.

Substituting (8.6) into (8.5), we obtain a simple relation between eigen and singular value decompositions,

$$\boldsymbol{\Lambda} = \mathbf{S}^2 , \tag{8.7}$$

for a square matrix $\mathbf{S}$. The elements of $\mathbf{S}$ are standard deviations along principal components. The matrix $\mathbf{V}$ is the same in both decompositions.

When we square a diagonal matrix, we also square its condition number, defined as the ratio of the largest and smallest eigenvalues. Generally, matrices with large condition numbers pose problems for numerical analysis. If the condition number is too large, the matrix is said to be ill-conditioned. For this reason, PCA implementations often prefer SVD of $\mathbf{X}$ over EVD of $\mathbf{X}^{\mathsf{T}}\mathbf{X}$.

PCA is included in many data analysis software suites. One example is function `pca` available from the Statistics Toolbox in MATLAB (or function `princomp` in older MATLAB releases).

Numerical issues in matrix operations are described in many books. We recommend Press *et al.* (2002) and Moler (2008).

8.3.3 Weighted Data

In physics analysis, observations are often weighted. Let $\boldsymbol{w}$ be a vector with weights for observations (rows) in matrix $\mathbf{X}$. Here, vector $\boldsymbol{w}$ has nothing to do with the transformation matrix $\mathbf{W}$; unfortunately, both entities are most usually denoted by the same letter. Suppose the weights have been normalized to sum to 1. The observed weighted mean for variable d; $d = 1, \ldots, D$; is then

$$\hat{\mu}_d = \sum_{n=1}^{N} w_n x_{nd} . \tag{8.8}$$

The observed weighted covariance for variables i and j is given by

$$\hat{\sigma}_{ij} = \sum_{n=1}^{N} w_n \left(x_{ni} - \hat{\mu}_i\right)\left(x_{nj} - \hat{\mu}_j\right) . \tag{8.9}$$

This estimate of the covariance is biased. For unweighted data drawn from a multivariate normal distribution, the unbiased estimate of the covariance matrix is given by $\mathbf{X}^T\mathbf{X}/(N-1)$. But if we set all weights in (8.9) to $1/N$, we would get $\mathbf{X}^T\mathbf{X}/N$. A small corrective factor fixes this discrepancy:

$$\hat{\sigma}_{ij} = \frac{\sum_{n=1}^{N} w_n \left(x_{ni} - \hat{\mu}_i\right)\left(x_{nj} - \hat{\mu}_j\right)}{1 - \sum_{n=1}^{N} w_n^2} . \tag{8.10}$$

Multiplying all elements of the covariance matrix by the same factor does not change the principal components.

An equivalent way of estimating the covariance matrix would be to use $\mathbf{WX}$, where $\mathbf{W}$ is a diagonal $N \times N$ matrix. The nth element on its main diagonal is set to

$$\sqrt{\frac{w_n}{1 - \sum_{n=1}^{N} w_n^2}} ,$$

and off-diagonal elements are set to zero. PCA of weighted data can be then carried out just like the ordinary PCA, either by EVD of $\hat{\boldsymbol{\Sigma}} = \mathbf{X}^T\mathbf{W}^2\mathbf{X}$ or by SVD of $\mathbf{WX}$.

8.3.4
How Many Principal Components Are Enough?

For N observations and D variables, we can find, after centering, $M_{\max} = \max(N-1, D)$ principal components at most. The number of principal components is limited by the rank of matrix $\mathbf{X}$ and can be smaller than $M_{\max}$ if some variables (or observations) are linear combinations of other variables (or observations). Yet for sufficiently large N and D the number of principal components can be impractically high.

Often it is possible to keep just a few largest principal components and discard the rest. This strategy is justified if the condition number of the estimated covariance matrix $\mathbf{X}^T\mathbf{X}$ for covariance PCA or estimated correlation matrix for correlation PCA is large. Under multivariate normality, a formal procedure described in Bartlett (1950) can be used to test the equality of all eigenvalues for covariance PCA. In practice, a number of heuristic techniques, not requiring the normality assumption, can be deployed for deciding how many principal components deserve to be kept.

One simple approach is to select as many components as needed to keep the fraction of the total variance explained by the first M components,

$$\delta_M = \frac{\sum_{m=1}^{M} \lambda_m}{\sum_{m=1}^{M_{\max}} \lambda_m} , \tag{8.11}$$

above a specified threshold. Here, λ_m is the mth diagonal element of matrix $\boldsymbol{\Lambda}$. Jolliffe (2002) recommends setting this threshold to a value between 0.7 and 0.9.

Another simple approach is to plot eigenvalues λ_m versus m and find where the slope of the plot goes from steep to flat, implying that the extra components add little information. This point is called an *elbow*, and this plot is called an *elbow* or *scree plot*. In a slightly modified version of this technique, plot the difference between the adjacent eigenvalues $\lambda_m - \lambda_{m+1}$ to detect the point where the difference approaches zero.

These two simple techniques are subjective and can produce inconclusive results. We apply these techniques to the ionosphere data available from the UCI repository (Frank and Asuncion, 2010) and show the results in Figure 8.2. There is no well-defined elbow on the scree plot. Based on the two plots, we could decide to retain at least seven components.

The two simple techniques described above could be applied to either covariance or correlation PCA. Peres-Neto *et al.* (2005) evaluate a number of less subjective approaches for correlation PCA aimed at discovering nontrivial principal components. A component is *nontrivial* if it has sizable contributions from two or more variables. The case of trivial components corresponds to an identity correlation matrix when all variables are normally distributed and independent. The presence of nontrivial components would be seen in a departure of some eigenvalues from one. The algorithms reviewed in Peres-Neto *et al.* (2005) search for nontrivial components with large eigenvalues.

Even if data were drawn from a perfectly spherical pdf, unequal eigenvalues would be observed due to random fluctuations. The largest eigenvalue would be always above one. To estimate the significance of the mth component, we could compare the observed mth eigenvalue with the distribution for the mth eigenvalue found in data with trivial components only. One of the more accurate algorithms for identifying nontrivial components in data with trivial and nontrivial components mixed combines this approach and a permutation test. This algorithm works as fol-

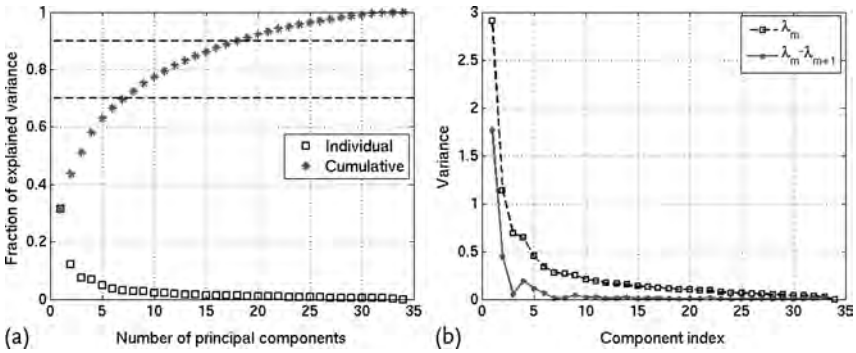

Figure 8.2 (a) Variance of an individual component normalized to the total variance versus component index (squares) and explained fraction of total variance versus the number of components (stars). (b) Two versions of the scree plot. Both plots are for the ionosphere data.

lows. Run PCA on the input data to obtain the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{M_{\max}}$. Generate R replicas of the input data for sufficiently large R. In each replica, shuffle values for each variable at random. Apply PCA to each replica r and record the eigenvalues, $\lambda_1^{(r)} \geq \lambda_2^{(r)} \geq \cdots \geq \lambda_{M_{\max}}^{(r)}$. Compute p-values for $M_{\max}$ null hypotheses "the mth component is trivial". To compute the p-value for the mth hypothesis, count the number of eigenvalues found in the shuffled data $\{\lambda_m^{(r)}\}_{r=1}^R$ above the observed eigenvalue λ_m. A low p-value indicates that λ_m is statistically large and likely produced by a nontrivial component. For large values of $M_{\max}$, we should account for effects of multiple testing, to be discussed in Sections 10.4 and 18.3.2.

We apply this algorithm to the ionosphere data using $R = 1000$ replicas. The p-values for the five largest components are exactly zero, and so is the p-value for the last (34th) component. The p-values for all other components are exactly one. We conclude that the five largest components are not trivial. The 34th component does not represent a significant effect. The second variable in the ionosphere data has zero variance, and the rank of the input matrix is therefore at most 33. The respective eigenvalue differs from zero by a value of the order of 10^{-31} due to floating-point error. The 34th eigenvalue in every shuffled replica is equally meaningless and should be ignored.

Note that trivial components can be essential for explaining the data. For instance, one variable with large variance and small correlation with the rest of the variables could be responsible for most observed variance. The described algorithm, as well as the other algorithms in Peres-Neto *et al.* (2005), by design would fail to detect its significance.

An alternative approach is to keep as many components as needed to satisfy bounds on the reconstruction error. These bounds could be set by the physics or by engineering tolerance constraints on the measured data. Here is one way to define the reconstruction error. Project data $\mathbf{X}$ onto the principal components, $\mathbf{Z} = \mathbf{XV}$. Then project the obtained scores $\mathbf{Z}$ back onto the original variables, $\hat{\mathbf{X}} = \mathbf{XVV}^{\mathsf{T}}$. The reconstructed matrix $\hat{\mathbf{X}}$ equals $\mathbf{X}$ because $\mathbf{V}$ is orthogonal. Let $\mathbf{V}_M$ be a matrix of PCA loadings for the first M components. To obtain this matrix, delete columns with indices $M + 1$ and higher in the full loading matrix $\mathbf{V}$. The reconstructed matrix $\hat{\mathbf{X}}_M = \mathbf{XV}_M\mathbf{V}_M^{\mathsf{T}}$ may not equal $\mathbf{X}$ because $\mathbf{V}_M$ may not be orthogonal. Let $\mathbf{R}^{(M)} = \hat{\mathbf{X}}_M - \mathbf{X}$ be a matrix of residuals. The Frobenius norm,

$$\|\mathbf{R}^{(M)}\|_F = \sqrt{\sum_{n=1}^{N}\sum_{d=1}^{D}\left(r_{nd}^{(M)}\right)^2}, \tag{8.12}$$

divided by the square root of the total number of elements in $\mathbf{X}$ can be used as the average reconstruction error. The maximal reconstruction error is given by the element of $\mathbf{R}^{(M)}$ with the maximal magnitude.

If we used the same data $\mathbf{X}$ to estimate $\mathbf{V}$ and to compute the reconstruction error, the error estimate would be biased low. We will discuss this phenomenon again in Chapter 9 in the context of supervised learning. If a large amount of data is available, we can find the loadings $\mathbf{V}$ using one dataset and estimate $\mathbf{R}^{(M)}$ using another

set. If there is not enough data, we can use cross-validation. Split $\mathbf{X}$ into $\mathcal{K}$ disjoint subsets with $N/\mathcal{K}$ observations (rows) per subset, on average. Take the first subset out of $\mathbf{X}$. Run PCA on the remaining $\mathcal{K}-1$ subsets to estimate $\mathbf{V}^{(1)}$. Apply the found loadings $\mathbf{V}^{(1)}$ to the held-out subset to obtain $\hat{\mathbf{X}}_M^{(1)}$; $M = 1, \ldots, M_{\max}$. Repeat for the remaining $\mathcal{K}-1$ subsets. Form the reconstructed matrix $\hat{\mathbf{X}}_M$ by concatenating the reconstructed subset matrices $\hat{\mathbf{X}}_M^{(k)}$; $k = 1, \ldots, \mathcal{K}$. This concatenation is unambiguous because every observation (row) in $\mathbf{X}$ can be found in one subset matrix only. The residual matrix $\mathbf{R}^{(M)}$ is then defined in the usual way.

We compute the average and maximal reconstruction error for the ionosphere data by 10-fold cross-validation. The results are shown in Figure 8.3. The average error steadily decreases as the number of principal components grows. The maximal error shows no improvement over the value obtained using just the first component until the number of components exceeds 20.

A more sophisticated technique for optimizing the number of components, based on the reconstruction error, is described in Krzanowski (1987).

8.3.5
Example: Apply PCA and Choose the Optimal Number of Components

Contents

- Load data
- Center the data
- Perform PCA
- Plot the explained variance
- Make a scree plot
- Find nontrivial eigenvalues
- Partition the data in 10 folds for cross-validation
- Estimate reconstruction error by cross-validation.

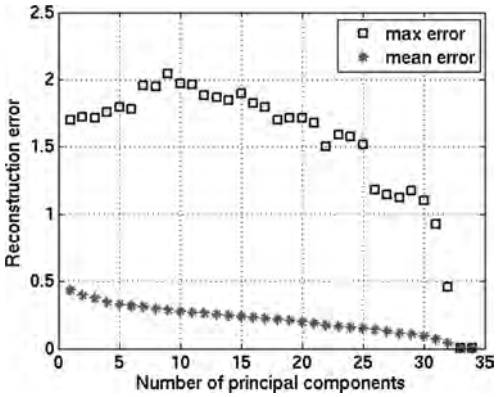

Figure 8.3 Average and maximal reconstruction error versus the number of principal components for the ionosphere data.